

# A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons

Yi-Cheng Chen, Ping-En Lu, *Graduate Student Member, IEEE*, Cheng-Shang Chang, *Fellow, IEEE*, and Tzu-Hsuan Liu

**Abstract**—In this paper, we conduct mathematical and numerical analyses for COVID-19. To predict the trend of COVID-19, we propose a time-dependent SIR model that tracks the transmission and recovering rate at time  $t$ . Using the data provided by China authority, we show our one-day prediction errors are almost less than 3%. The turning point and the total number of confirmed cases in China are predicted under our model. To analyze the impact of the undetectable infections on the spread of disease, we extend our model by considering two types of infected persons: detectable and undetectable infected persons. Whether there is an outbreak is characterized by the spectral radius of a  $2 \times 2$  matrix. If  $R_0 > 1$ , then the spectral radius of that matrix is greater than 1, and there is an outbreak. We plot the phase transition diagram of an outbreak and show that there are several countries on the verge of COVID-19 outbreaks on Mar. 2, 2020. To illustrate the effectiveness of social distancing, we analyze the independent cascade model for disease propagation in a configuration random network. We show two approaches of social distancing that can lead to a reduction of the effective reproduction number  $R_e$ .

**Index Terms**—COVID-19, SARS-CoV-2, Coronavirus, time-dependent SIR model, undetectable infection, herd immunity, superspreader, independent cascade, social distancing.

## 1 INTRODUCTION

AT the beginning of December 2019, the first COVID-19 victim was diagnosed with the coronavirus in Wuhan, China. In the following weeks, the disease spread widely in China mainland and other countries, which causes global panic. The virus has been named “SARS-CoV-2,” and the disease it causes has been named “coronavirus disease 2019 (abbreviated “COVID-19”). There have been 80,151 people infected by the disease and 2,943 deaths until Mar. 2, 2020 according to the official statement by the Chinese government. To block the spread of the virus, there are some strategies such as city-wide lockdown, traffic halt, community management, social distancing, and propaganda of health education knowledge that have been adopted by the governments of China and other countries in the world.

Unlike the Severe Acute Respiratory Syndrome (SARS) and other infectious diseases, one problematic characteristic of COVID-19 is that there are asymptomatic infections (who have very mild symptoms). Those asymptomatic infections are unaware of their contagious ability, and thus get more people infected before they are detected [1]. The transmission rate can increase dramatically in this circumstance. According to the recent report from WHO [2], only 87.9% of COVID-19 patients have a fever, and 67.7% of them have a dry cough. If we use body temperature as a means to detect COVID-19 infected cases, then more than 10% of infected persons cannot be detected.

Due to the recent development of the epidemic, we are inter-

ested in addressing the following important questions for COVID-19:

- (Q1) Is it possible to contain COVID-19? Are the commonly used measures, such as city-wide lockdown, traffic halt, community management, and propaganda of health education knowledge, effective in containing COVID-19?
- (Q2) If COVID-19 can be contained, when will be the peak of the epidemic, and when will it end?
- (Q3) How do the undetectable infections affect the spread of disease?
- (Q4) If COVID-19 cannot be contained, what is the ratio of the population that needs to be infected in order to achieve herd immunity?
- (Q5) How effective are the social distancing approaches, such as reduction of interpersonal contacts and canceling mass gatherings in controlling COVID-19?

For (Q1), we analyze the cases in China and aim to predict how the virus spreads in this paper. Specifically, we propose using a time-dependent susceptible-infected-recovered (SIR) model to analyze and predict the number of infected persons and the number of recovered persons (including deaths). In the traditional SIR model, it has two time-invariant variables: the transmission rate  $\beta$  and the recovering rate  $\gamma$ . The transmission rate  $\beta$  means that each individual has on average  $\beta$  contacts with randomly chosen others per unit time. On the other hand, the recovering rate  $\gamma$  indicates that individuals in the infected state get recovered or die at a fixed average rate  $\gamma$ . The traditional SIR model neglects the time-varying property of  $\beta$  and  $\gamma$ , and it is too simple to precisely and effectively predict the trend of the disease. Therefore, we propose using a time-dependent SIR model, where both the transmission rate  $\beta$  and the recovering rate  $\gamma$  are functions of time  $t$ . Our idea is to use machine learning methods to track the transmission rate

• The authors are with the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan R.O.C.  
E-mail: yichengchen@gapp.nthu.edu.tw, j94223@gmail.com, cschang@ee.nthu.edu.tw, and carinaliu@gapp.nthu.edu.tw.

Manuscript received April 08, 2020; revised June 18, 2020; accepted September 12, 2020. This work was supported by the Ministry of Science and Technology (MOST) of Taiwan (R.O.C.) under Project MOST108-2221-E007-016-MY3. (Corresponding author: Ping-En Lu.)

$\beta(t)$  and the recovering rate  $\gamma(t)$ , and then use them to predict the number of the infected persons and the number of recovered persons at a certain time  $t$  in the future. Our time-dependent SIR model can dynamically adjust the crucial parameters, such as  $\beta(t)$  and  $\gamma(t)$ , to adapt accordingly to the change of control policies, which differs from the existing SIR and SEIR models in the literature, e.g., [3], [4], [5], [6], [7], and [8]. For example, we observe that city-wide lockdown can lower the transmission rate substantially from our model. Most data-driven and curve-fitting methods for the prediction of COVID-19, e.g., [9], [10], and [11] seem to track data perfectly; however, they are lack of physical insights of the spread of the disease. Moreover, they are very sensitive to the sudden change in the definition of confirmed cases on Feb. 12, 2020 in the Hubei province. On the other hand, our time-dependent SIR model can examine the epidemic control policy of the Chinese government and provide reasonable explanations. Using the data provided by the National Health Commission of the People's Republic of China (NHC) [12], we show that the one-day prediction errors for the numbers of confirmed cases are almost less than 3% except for Feb. 12, 2020, which is unpredictable due to the change of the definition of confirmed cases.

For (Q2), the basic reproduction number  $R_0$ , defined as the expected number of additional infections (secondary cases) by one typical infected person before it recovers in a wholly susceptible population during the infection period [13], [14], is one of the commonly used metrics to check whether the disease will become an outbreak, and what proportion of the population needs to be vaccinated to eradicate the disease. In fact, at any given time, different proportions of the population are immune to any given disease. For this, the effective reproduction number  $R_e(t)$  is used to quantify the instantaneous spread of disease in the partially susceptible population [14]. Knowing  $R_0$  and  $R_e(t)$  in advance can help governments make more accurate epidemic prevention policies. In the classical SIR model,  $R_0$  is simply  $\beta/\gamma$  as an infected person takes (on average)  $1/\gamma$  days to recover, and during that period time, it will be in contact with (on average)  $\beta$  persons. In our time-dependent SIR model, the effective reproduction number at time  $t$ , denoted by  $R_e(t)$ , is defined as  $\beta(t)/\gamma(t)$ . If  $R_e(t) > 1$ , the disease will spread exponentially and infects a certain fraction of the total population. On the contrary, the disease will eventually be contained. Therefore, by observing the change of  $R_e(t)$  with respect to time or even predict  $R_e(t)$  in the future, we can check whether certain epidemic control policies are effective or not. Using the data provided by the National Health Commission of the People's Republic of China (NHC) [12], we show that the turning point (peak), defined as the day that the effective reproduction number is less than 1, is predicted to be Feb. 17, 2020. Moreover, the disease in China will end in about 6 weeks after its peak in our (deterministic) model if the current contagious disease control policies are maintained in China. In that case, the total number of confirmed cases is predicted to be around 80,000 cases in China under our (deterministic) model.

For (Q3), we extend our SIR model to include two types of infected persons: detectable infected persons (type I) and undetectable infected persons (type II). With probability  $w_1$  (resp.  $w_2$ ), an infected person is of type I (resp. II), where  $w_1 + w_2 = 1$ . Type I (resp. II) infected persons have the transmission rate  $\beta_1$  (resp.  $\beta_2$ ) and the recovering rate  $\gamma_1$  (resp.  $\gamma_2$ ). The basic reproduction

number in this model is

$$R_0 = w_1 \frac{\beta_1}{\gamma_1} + w_2 \frac{\beta_2}{\gamma_2}. \quad (1)$$

In practice, type I infected persons have a lower transmission rate than that of type II infected persons (as type I infected persons can be isolated). For such a model, whether the disease is controllable is characterized by the spectral radius of a  $2 \times 2$  matrix. If the spectral radius of that matrix is larger than 1, then there is an outbreak. On the other hand, if it is smaller than 1, then there is no outbreak. One interesting result is that the spectral radius of that matrix is larger (resp. smaller) than 1 if the basic reproduction number  $R_0$  in (1) is larger (resp. smaller) than 1. The curve that has the spectral radius equal to 1 is known as the percolation threshold curve in a phase transition diagram [15]. Using the historical data from Jan. 22, 2020 to Mar. 2, 2020 from the GitHub of Johns Hopkins University [16], we extend our study to some other countries, including Japan, Singapore, South Korea, Italy, and Iran. Our numerical results show that there are several countries, including South Korea, Italy, and Iran, that are above the percolation threshold curve, and they are on the verge of COVID-19 outbreaks on Mar. 2, 2020. In comparison with other time-dependent epidemic models in the literature (see, e.g., the well-known tsIR model [17], [18] and the recent Science article [19]), our model further considers the effect of undetectable infected persons (see also the comments in [19] on the preprint version of our paper [20]). A preprint version of our time-dependent SIR model [20] has been cited by Oliver Wyman (one of the well-known global management consulting firms) as one of the two key references in their white paper of COVID-19 Pandemic Navigator Core Model [21]. There they further include government response actions and Google COVID-19 Community Mobility reports for analyzing the effect of undetectable infected persons. This shows the potential to further extend our model to yield more accurate predictions.

The British prime minister, Boris Johnson, once suggested having a sufficiently high fraction of individuals infected by COVID-19 and recovered from the disease to achieve herd immunity. To address the question in (Q4), we argue that herd immunity corresponds to the reduction of the number of susceptible persons in the SIR model, and herd immunity can be achieved after at least  $1 - \frac{1}{R_0}$  fraction of individuals being infected and recovered from the COVID-19.

For (Q5), we consider two commonly used approaches for social distancing: (i) allowing every person to keep its interpersonal contacts up to a fraction of its normal contacts, and (ii) canceling mass gatherings. For the analysis of social distancing, we have to take the social network (and its network structure) into account. For this, we consider the independent cascade (IC) model for disease propagation in a random network specified by a degree distribution  $p_k, k = 0, 1, 2, \dots$ . The IC model has been widely used for the study of the influence maximization problem in viral marketing (see, e.g., [22]). In the IC model, an infected node can transmit the disease to a neighboring susceptible node (through an edge) with a certain propagation probability. Repeatedly continuing the propagation, we have a subgraph that contains the set of infected nodes in the long run. By relating the propagation probabilities in the IC model to the transmission rates and recovering rates in the SIR model, we show two results for social distancing: (i) for the social distancing approach that allows every person to keep its interpersonal contacts up to (on average) a

fraction  $a$  of its normal contacts, the effective reproduction number is reduced by a factor of  $a^2$ , and (ii) for the social distancing approach that cancels mass gatherings by removing nodes with the number of edges larger than or equal to  $k_0$ , the effective reproduction number is reduced by a factor of  $\frac{\sum_{k=0}^{k_0-2} kq_k}{\sum_{k=0}^{\infty} kq_k}$ , where  $q_k$  is the excess degree distribution of  $p_k$ .

The rest of the paper is organized as follows: In Section 2, we propose the time-dependent SIR model. We then extend the model to the SIR model with undetectable infected persons in Section 3. In Section 4, we consider the independent cascade model for disease propagation in a random network specified by a degree distribution. In Section 5, we conduct several numerical experiments to illustrate the effectiveness of our models. In Section 6, we put forward some discussions and suggestions to control COVID-19. The paper is concluded in Section 7.

## 2 THE TIME-DEPENDENT SIR MODEL

### 2.1 Susceptible-infected-recovered (SIR) Model

In the typical mathematical model of infectious disease, one often simplify the virus-host interaction and the evolution of an epidemic into a few basic disease states. One of the simplest epidemic model, known as the susceptible-infected-recovered (SIR) model [15], includes three states: the susceptible state, the infected state, and the recovered state. An individual in the *susceptible state* is one who does not have the disease at time  $t$  yet, but may be infected if one is in contact with a person infected with the disease. The *infected state* refers to an individual who has a disease at time  $t$  and may infect a susceptible individual potentially (if they come into contact with each other). The *recovered state* refers to an individual who is either recovered or dead from the disease and is no longer contagious at time  $t$ . Also, a recovered individual will not be back to the susceptible state anymore. The reason for the number of deaths is counted in the recovered state is that, from an epidemiological point of view, this is basically the same thing, regardless of whether recovery or death does not have much impact on the spread of the disease. As such, they can be effectively eliminated from the potential host of the disease [23]. Denote by  $S(t)$ ,  $X(t)$  and  $R(t)$  the numbers of susceptible persons, infected persons, and recovered persons at time  $t$ . Summing up the above SIR model, we believe it is very similar to the COVID-19 outbreak, and we will adopt the SIR model as our basic model in this paper.

In the traditional SIR model, it has two time-invariant variables: the transmission rate  $\beta$  and the recovering rate  $\gamma$ . The transmission rate  $\beta$  means that each individual has on average  $\beta$  contacts with randomly chosen others per unit time. On the other hand, the recovering rate  $\gamma$  indicates that individuals in the infected state get recovered or die at a fixed average rate  $\gamma$ . The traditional SIR model neglects the time-varying property of  $\beta$  and  $\gamma$ . This assumption is too simple to precisely and effectively predict the trend of the disease. Therefore, we propose the time-dependent SIR model, where both the transmission rate  $\beta$  and the recovering rate  $\gamma$  are functions of time  $t$ . Such a time-dependent SIR model is much better to track the disease spread, control, and predict the future trend.

### 2.2 Differential Equations for the Time-dependent SIR Model

For the traditional SIR model, the three variables  $S(t)$ ,  $X(t)$  and  $R(t)$  are governed by the following differential equations (see, e.g., the book [15]):

$$\begin{aligned}\frac{dS(t)}{dt} &= \frac{-\beta S(t)X(t)}{n}, \\ \frac{dX(t)}{dt} &= \frac{\beta S(t)X(t)}{n} - \gamma X(t), \\ \frac{dR(t)}{dt} &= \gamma X(t).\end{aligned}$$

We note that

$$S(t) + X(t) + R(t) = n, \quad (2)$$

where  $n$  is the total population. Let  $\beta(t)$  and  $\gamma(t)$  be transmission rate and recovering rate at time  $t$ . Replacing  $\beta$  and  $\gamma$  by  $\beta(t)$  and  $\gamma(t)$  in the differential equations above yields

$$\frac{dS(t)}{dt} = \frac{-\beta(t)S(t)X(t)}{n}, \quad (3)$$

$$\frac{dX(t)}{dt} = \frac{\beta(t)S(t)X(t)}{n} - \gamma(t)X(t), \quad (4)$$

$$\frac{dR(t)}{dt} = \gamma(t)X(t). \quad (5)$$

The three variables  $S(t)$ ,  $X(t)$  and  $R(t)$  still satisfy (2).

Now we briefly explain the intuition of these three equations. Equation (3) describes the difference of the number of susceptible persons  $S(t)$  at time  $t$ . If we assume the total population is  $n$ , then the probability that a randomly chosen person is in the susceptible state is  $S(t)/n$ . Hence, an individual in the infected state will contact (on average)  $\beta(t)S(t)/n$  people in the susceptible state per unit time, which implies the number of newly infected persons is  $\beta(t)S(t)X(t)/n$  (as there are  $X(t)$  people in the infected state at time  $t$ ). On the contrary, the number of people in the susceptible state will decrease by  $\beta(t)S(t)X(t)/n$ . Additionally, as every individual in the infected state will recover with rate  $\gamma(t)$ , there are (on average)  $\gamma(t)X(t)$  people recovered at time  $t$ . This is shown in (5) that illustrates the difference of  $R(t)$  at time  $t$ . Since three variables  $S(t)$ ,  $X(t)$  and  $R(t)$  still satisfy (2), we have

$$\frac{dX(t)}{dt} = -\left(\frac{dS(t)}{dt} + \frac{dR(t)}{dt}\right),$$

which is the number of people changing from the susceptible state to the infected state minus the number of people changing from the infected state to the recovered state (see (4)).

### 2.3 Discrete Time Time-dependent SIR Model

Due to the COVID-19 data is updated in days [12], we revise the differential equations in (3), (4), and (5) into discrete time difference equations:

$$S(t+1) - S(t) = \frac{-\beta(t)S(t)X(t)}{n}, \quad (6)$$

$$X(t+1) - X(t) = \frac{\beta(t)S(t)X(t)}{n} - \gamma(t)X(t), \quad (7)$$

$$R(t+1) - R(t) = \gamma(t)X(t). \quad (8)$$

Again, the three variables  $S(t)$ ,  $X(t)$  and  $R(t)$  still satisfy (2).

In the beginning of the disease spread, the number of confirmed cases is very low, and most of the population are in the susceptible state. Hence, for our analysis of the initial stage of

COVID-19, we assume  $\{S(t) \approx n, t \geq 0\}$ , and further simplify (7) as follows:

$$X(t+1) - X(t) = \beta(t)X(t) - \gamma(t)X(t). \quad (9)$$

From the difference equations above, one can easily derive  $\beta(t)$  and  $\gamma(t)$  of each day. From (8), we have

$$\gamma(t) = \frac{R(t+1) - R(t)}{X(t)}. \quad (10)$$

Using (8) in (9) yields

$$\beta(t) = \frac{[X(t+1) - X(t)] + [R(t+1) - R(t)]}{X(t)}. \quad (11)$$

Given the historical data from a certain period  $\{X(t), R(t), 0 \leq t \leq T-1\}$ , we can measure the corresponding  $\{\beta(t), \gamma(t), 0 \leq t \leq T-2\}$  by using (10) and (11). With the above information, we can use machine learning methods to predict the time varying transmission rates and recovering rates.

## 2.4 Tracking Transmission Rate $\beta(t)$ and Recovering Rate $\gamma(t)$ by Ridge Regression

In this subsection, we track and predict  $\beta(t)$  and  $\gamma(t)$  by the commonly used Finite Impulse Response (FIR) filters in linear systems. Denote by  $\hat{\beta}(t)$  and  $\hat{\gamma}(t)$  the *predicted* transmission rate and recovering rate. From the FIR filters, they are predicted as follows:

$$\begin{aligned} \hat{\beta}(t) &= a_1\beta(t-1) + a_2\beta(t-2) + \dots + a_J\beta(t-J) + a_0 \\ &= \sum_{j=1}^J a_j\beta(t-j) + a_0, \end{aligned} \quad (12)$$

$$\begin{aligned} \hat{\gamma}(t) &= b_1\gamma(t-1) + b_2\gamma(t-2) + \dots + b_K\gamma(t-K) + b_0 \\ &= \sum_{k=1}^K b_k\gamma(t-k) + b_0, \end{aligned} \quad (13)$$

where  $J$  and  $K$  are the orders of the two FIR filters ( $0 < J, K < T-2$ ),  $a_j, j = 0, 1, \dots, J$ , and  $b_k, k = 0, 1, \dots, K$  are the coefficients of the impulse responses of these two FIR filters.

There are several widely used machine learning methods for the estimation of the coefficients of the impulse response of an FIR filter, e.g., ordinary least squares (OLS), regularized least squares (i.e., ridge regression), and partial least squares (PLS) [24]. In this paper, we choose the ridge regression as our estimation method that solves the following optimization problem:

$$\min_{a_j} \sum_{t=J}^{T-2} (\beta(t) - \hat{\beta}(t))^2 + \alpha_1 \sum_{j=0}^J a_j^2, \quad (14)$$

$$\min_{b_k} \sum_{t=K}^{T-2} (\gamma(t) - \hat{\gamma}(t))^2 + \alpha_2 \sum_{k=0}^K b_k^2, \quad (15)$$

where  $\alpha_1$  and  $\alpha_2$  are the regularization parameters.

## 2.5 Tracking the Number of Infected Persons $\hat{X}(t)$ and the Number of Recovered Persons $\hat{R}(t)$ of the Time-dependent SIR Model

In this subsection, we show how we use the two FIR filters to track and predict the number of infected persons and the number of recovered persons in the time-dependent SIR model. Given a period of historical data  $\{X(t), R(t), 0 \leq t \leq T-1\}$ , we first

measure  $\{\beta(t), \gamma(t), 0 \leq t \leq T-2\}$  by (10) and (11). Then we solve the ridge regression (with the objective functions in (14) and (15) and the constraints in (12) and (13)) to learn the coefficients of the FIR filters, i.e.,  $a_j, j = 0, 1, \dots, J$  and  $b_k, k = 0, 1, \dots, K$ . Once we learn these coefficients, we can predict  $\hat{\beta}(t)$  and  $\hat{\gamma}(t)$  at time  $t = T-1$  by the trained ridge regression in (12) and (13).

Denote by  $\hat{X}(t)$  (resp.  $\hat{R}(t)$ ) the predicted number of infected (resp. recovered) persons at time  $t$ . To predict  $\hat{X}(t)$  and  $\hat{R}(t)$  at time  $t = T$ , we simply replace  $\beta(t)$  and  $\gamma(t)$  by  $\hat{\beta}(t)$  and  $\hat{\gamma}(t)$  in (8) and (9). This leads to

$$\hat{X}(T) = (1 + \hat{\beta}(T-1) - \hat{\gamma}(T-1))X(T-1), \quad (16)$$

$$\hat{R}(T) = R(T-1) + \hat{\gamma}(T-1)X(T-1). \quad (17)$$

To predict  $\hat{X}(t)$  and  $\hat{R}(t)$  for  $t > T$ , we estimate  $\hat{\beta}(t)$  and  $\hat{\gamma}(t)$  by using (12) and (13). Similar to those in (16) and (17), we predict  $\hat{X}(t)$  and  $\hat{R}(t)$  as follows:

$$\hat{X}(t+1) = (1 + \hat{\beta}(t) - \hat{\gamma}(t))\hat{X}(t), \quad t \geq T, \quad (18)$$

$$\hat{R}(t+1) = \hat{R}(t) + \hat{\gamma}(t)\hat{X}(t), \quad t \geq T. \quad (19)$$

The detailed steps of our tracking/predicting method are outlined in Algorithm 1.

---

### ALGORITHM 1: Tracking Discrete Time Time-dependent SIR Model

---

**Input:**  $\{X(t), R(t), 0 \leq t \leq T-1\}$ , Regularization parameters  $\alpha_1$  and  $\alpha_2$ , Order of FIR filters  $J$  and  $K$ , Prediction window  $W$ .

**Output:**  $\{\beta(t), \gamma(t), 0 \leq t \leq T-2\}$ ,  $\{\hat{\beta}(t), \hat{\gamma}(t), t \geq T-1\}$ , and  $\{\hat{X}(t), \hat{R}(t), t \geq T\}$ .

- 1: Measure  $\{\beta(t), \gamma(t), 0 \leq t \leq T-2\}$  using (11) and (10) respectively.
  - 2: Train the ridge regression using (14) and (15).
  - 3: Estimate  $\hat{\beta}(T-1)$  and  $\hat{\gamma}(T-1)$  by (12) and (13) respectively.
  - 4: Estimate the number of infected persons  $\hat{X}(T)$  and recovered persons  $\hat{R}(T)$  on the next day  $T$  using (16) and (17) respectively.
  - 5: **while**  $T \leq t \leq T+W$  **do**
  - 6:   Estimate  $\hat{\beta}(t)$  and  $\hat{\gamma}(t)$  in (12) and (13) respectively.
  - 7:   Predict  $\hat{X}(t+1)$  and  $\hat{R}(t+1)$  using (18) and (19) respectively.
  - 8: **end while**
- 

We note that this *deterministic* epidemic model is based on the mean-field approximation for  $X(t)$  and  $R(t)$ . Such an approximation is a result of the law of large numbers. Therefore, when  $X(t)$  and  $R(t)$  are relatively small, the mean-field approximation may not be as accurate as expected. In those cases, one might have to resort to stochastic epidemic models, such as Markov chains.

## 3 THE SIR MODEL WITH UNDETECTABLE INFECTED PERSONS

According to the recent report from WHO [2], only 87.9% of COVID-19 patients have a fever, and 67.7% of them have a dry cough. This means there exist asymptomatic infections. Recent studies in [7] and [25] also pointed out the existence of the asymptomatic carriers of COVID-19. Those people are unaware



of their contagious ability, and thus get more people infected. The transmission rate can increase dramatically in this circumstance. If there is no method like mass testing to detect those asymptomatic infections, those people will be undetectable.

An undetectable infected person in this paper is defined as an infected person that is not yet detectable or detected under the current epidemic prevention policies. To take the undetectable infected persons into account, we propose the SIR model with undetectable infected persons in this section. We assume that there are two types of infected persons. The individuals who are detectable (with obvious symptoms) are categorized as type I infected persons, and the asymptomatic individuals who are undetectable are categorized as type II infected persons. For an infected individual, it has probability  $w_1$  to be type I and probability  $w_2$  to be type II, where  $w_1 + w_2 = 1$ . Besides, those two types of infected persons have different transmission rates and recovering rates, depending on whether they are under treatment or isolation or not. We denote  $\beta_1(t)$  and  $\gamma_1(t)$  as the transmission rate and the recovering rate of type I at time  $t$ . Similarly,  $\beta_2(t)$  and  $\gamma_2(t)$  are the transmission rate and the recovering rate for type II at time  $t$ .

### 3.1 The Governing Equations for the SIR Model with Undetectable Infected Persons

Now we derive the governing equations for the SIR model with two types of infected persons. Let  $X_1(t)$  (resp.  $X_2(t)$ ) be the number of type I (resp. type II) infected persons at time  $t$ . Similar to the derivation of (7), (8) in Subsection 2.3, we assume that  $\{S(t) \approx n, t \geq 0\}$  in the initial stage of the epidemic and split  $X(t)$  into two types of infected persons. We have the following difference equations:

$$X_1(t+1) - X_1(t) = \beta_1 X_1(t) w_1 + \beta_2 X_2(t) w_1 - \gamma_1 X_1(t), \quad (20)$$

$$X_2(t+1) - X_2(t) = \beta_1 X_1(t) w_2 + \beta_2 X_2(t) w_2 - \gamma_2 X_2(t), \quad (21)$$

$$R(t+1) - R(t) = \gamma_1 X_1(t) + \gamma_2 X_2(t), \quad (22)$$

where  $\beta_1, \beta_2, \gamma_1$ , and  $\gamma_2$  are constants. It is noteworthy that those constants can also be time-dependent as we have in Section 2. However, in this section, we set them as constants to focus on the effect of undetectable infected persons. Rewriting (20) and (21) in the matrix form yields the following matrix equation:

$$\begin{bmatrix} X_1(t+1) \\ X_2(t+1) \end{bmatrix} = \begin{bmatrix} 1 + \beta_1 w_1 - \gamma_1 & \beta_2 w_1 \\ \beta_1 w_2 & 1 + \beta_2 w_2 - \gamma_2 \end{bmatrix} \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix},$$

where  $w_2 = 1 - w_1$ . Let  $\mathbf{A}$  be the transition matrix of the above system equations, i.e.,

$$\mathbf{A} = \begin{bmatrix} 1 + \beta_1 w_1 - \gamma_1 & \beta_2 w_1 \\ \beta_1 w_2 & 1 + \beta_2 w_2 - \gamma_2 \end{bmatrix}. \quad (23)$$

It is well-known (from linear algebra) such a system is stable if the spectral radius (the largest absolute value of the eigenvalue) of  $\mathbf{A}$  is less than 1. In other words,  $X_1(t+1)$  and  $X_2(t+1)$  will converge gradually to finite constants when  $t$  goes to infinity. In that case, there will not be an outbreak. On the contrary, if the spectral radius is greater than 1, there will be an outbreak, and the number of infected persons will grow exponentially with respect to time  $t$  (at the rate of the spectral radius).

In addition,  $w_1$  and  $w_2$  can be further written in a time-dependent form. If we assume  $\gamma_1 = \gamma_2 = \gamma$ , according to (20) and (21), we can derive the following expression,

$$X(t+1) = (1 + \beta_1 w_1(t) + \beta_2(1 - w_1(t)) - \gamma)X(t). \quad (24)$$

According to (24), if  $\beta_1, \beta_2$ , and  $\gamma$  are assumed to be constants over a period of time, we can measure the change of  $w_1$  over time, which is

$$w_1(t) = \frac{\frac{X(t+1)}{X(t)} - \beta_2 + \gamma - 1}{\beta_1 - \beta_2}. \quad (25)$$

The ratio between  $w_1$  and  $w_2$  will change with the government's epidemic prevention policy due to the progress of the epidemic. Usually, after a new epidemic prevention policy is promulgated, the policy will have a significant impact on the ratio of  $w_1$  and  $w_2$  over a period of time, and thus  $w_1$  usually appears in the form of a step function.

### 3.2 The Basic Reproduction Number

To further examine the stability condition of such a system, we let

$$R_0 = w_1 \frac{\beta_1}{\gamma_1} + w_2 \frac{\beta_2}{\gamma_2}. \quad (26)$$

Note that  $R_0$  is simply the basic reproduction number of a newly infected person as an infected person can further infect on average  $\beta_1/\gamma_1$  (resp.  $\beta_2/\gamma_2$ ) persons if it is of type I (resp. type II) and that happens with probability  $w_1$  (resp.  $w_2$ ). In the following theorem, we show that there is no outbreak if  $R_0 < 1$  and there is an outbreak if  $R_0 > 1$ . Thus,  $R_0$  in (26) is known as the percolation threshold for an outbreak in such a model [15].

**Theorem 1.** *If  $R_0 < 1$ , then the spectral radius of  $\mathbf{A}$  in (23) is less than 1 and there is no outbreak of the epidemic. On the other hand, if  $R_0 > 1$ , then the spectral radius of  $\mathbf{A}$  in (23) is larger than 1 and there is an outbreak of the epidemic.*

**Proof.** (Theorem 1)

First, we note that  $\gamma_1$  and  $\gamma_2$  are recovering rates and they cannot be larger than 1 in the discrete-time setting, i.e., it takes at least one day for an infected person to recover. Thus, the matrix  $\mathbf{A}$  is a positive matrix (with all its elements being positive). It then follows from the Perron-Frobenius theorem that the spectral radius of the matrix is the larger eigenvalue of the  $2 \times 2$  matrix.

Now we find the larger eigenvalue of the matrix  $\mathbf{A}$ . Let  $\mathbf{I}$  be the  $2 \times 2$  identity matrix and

$$\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{I}. \quad (27)$$

Then

$$\tilde{\mathbf{A}} = \begin{bmatrix} \beta_1 w_1 - \gamma_1 & \beta_2 w_1 \\ \beta_1 w_2 & \beta_2 w_2 - \gamma_2 \end{bmatrix}. \quad (28)$$

Let

$$z_1 = \beta_1 w_1 - \gamma_1 + \beta_2 w_2 - \gamma_2, \quad (29)$$

and

$$z_2 = \beta_1 w_1 \gamma_2 + \beta_2 w_2 \gamma_1 - \gamma_1 \gamma_2. \quad (30)$$

It is straightforward to show that the two eigenvalues of  $\tilde{\mathbf{A}}$  are

$$\lambda_1 = \frac{1}{2}(z_1 + \sqrt{z_1^2 + 4z_2}), \quad (31)$$

and

$$\lambda_2 = \frac{1}{2}(z_1 - \sqrt{z_1^2 + 4z_2}). \quad (32)$$

Note that  $\lambda_1 \geq \lambda_2$ . In view of (27), the larger eigenvalue of the transition matrix  $\mathbf{A}$  is  $1 + \lambda_1$ .

If  $R_0 < 1$ , we know that  $z_2 < 0$ ,  $w_1 \frac{\beta_1}{\gamma_1} < 1$ , and  $w_2 \frac{\beta_2}{\gamma_2} < 1$ . Thus, we have from (29) that  $z_1 < 0$ . In view of (31), we conclude that

$$\lambda_1 < \frac{1}{2}(z_1 + |z_1|) = 0.$$

This shows that  $1 + \lambda_1 < 1$  and the spectral radius of  $\mathbf{A}$  is less than 1.

On the other hand, if  $R_0 > 1$ , then  $z_2 > 0$  and we have from (31) that

$$\lambda_1 > \frac{1}{2}(z_1 + |z_1|) \geq 0.$$

This shows that  $1 + \lambda_1 > 1$  and the spectral radius of  $\mathbf{A}$  is larger than 1. ■

The relation between the system parameters and the phase transition will be shown in Subsection 5.6.

### 3.3 Herd Immunity

Herd immunity is one way to resist the spread of a contagious disease if a sufficiently high fraction of individuals are immune to the disease, especially through vaccination. One interesting strategy, once suggested by Boris Johnson, the British prime minister, is to have a sufficiently high fraction of individuals infected by COVID-19 and recovered from the disease to achieve herd immunity. The question is, what will be the fraction of individuals that need to be infected to achieve herd immunity for COVID-19.

To address such a question, we note that herd immunity corresponds to the reduction of the number of susceptible persons in the SIR model. In our previous analysis, we all assume that every person is susceptible to COVID-19 at the early stage and thus  $S(t)/n \approx 1$ . For the analysis of herd immunity, we assume that there is a probability  $h$  that a randomly chosen person is susceptible at time  $t$ . This is equivalent to that  $1 - h$  fraction of individuals are immune to the disease. Under such an assumption, we then have

$$\frac{S(t)}{n} \approx h. \quad (33)$$

In view of the difference equation for  $X(t)$  in (7), we can rewrite (20)-(22) to derive the governing equations for herd immunity as follows:

$$X_1(t+1) - X_1(t) = \beta_1 X_1(t) h w_1 + \beta_2 X_2(t) h w_1 - \gamma_1 X_1(t), \quad (34)$$

$$X_2(t+1) - X_2(t) = \beta_1 X_1(t) h w_2 + \beta_2 X_2(t) h w_2 - \gamma_2 X_2(t), \quad (35)$$

$$R(t+1) - R(t) = \gamma_1 X_1(t) + \gamma_2 X_2(t). \quad (36)$$

In comparison with the original governing equations in (20)-(22), the only difference is the change of the transmission rate of type I (resp. type II) from  $\beta_1$  to  $\beta_1 h$  (resp. from  $\beta_2$  to  $\beta_2 h$ ). Thus, herd immunity effectively reduces the transmission rates by a factor of  $h$ . As a direct consequence of Theorem 1, we have the following corollary.

**Corollary 2.** *For a contagious disease modeled by our SIR model with two types of infected persons that has  $R_0$  in (26) greater than 1, herd immunity can be achieved after at least  $1 - h^*$  fraction*

*of individuals being infected and recovered from the contagious disease, where*

$$h^* = \frac{1}{R_0}.$$

For more discussions on the effect of a possible limited immunity and its impact on herd immunity for COVID-19, we refer to the recent paper [26]. There the authors discussed possible consequences of reaching the COVID-19 herd immunity threshold in the absence of a vaccine. In particular, it is stated in [26] that depletion in healthcare resources will lead not only to elevated COVID-19 mortality but also to increased all-cause mortality.

## 4 THE INDEPENDENT CASCADE (IC) MODEL FOR DISEASE PROPAGATION IN NETWORKS

Our analysis in the previous section does not consider how the structure of a social network affects the propagation of a disease. There are other widely used policies, such as social distancing, that could not be modeled by our SIR model with undetectable infected persons in Section 3. To take the network structure into account, in this section, we consider the independent cascade (IC) model for disease propagation. The IC model was previously studied by Kempe, Kleinberg, and Tardos in [22] for the influence maximization problem in viral marketing. In the IC model, there is a social network modeled by a graph  $G = (V, E)$ , where  $V$  is the set of nodes, and  $E$  is the set of edges. An *infected* node can transmit the disease to a neighboring *susceptible* node (through an edge) with a certain propagation probability. As there are two types of infected persons in our model, we denote by  $\phi_1$  (resp.  $\phi_2$ ) the propagation probability that a type I (resp. type II) infected node transmits the disease to an (immediate) neighbor of the infected node. Once a neighboring node is infected, it becomes a type I (type II) infected node with probability  $w_1$  (resp.  $w_2$ ) and it can continue the propagation of the disease to its neighbors. Continuing the propagation, we thus form a subgraph of  $G$  that contains the set of infected nodes in the long run. Call such a subgraph the *infected subgraph*. One interesting question is how one controls the spread of the disease so that the total number of nodes in the infected subgraph remains small even when the total number of nodes is very large.

### 4.1 The Infected Tree in the Configuration Model

The exact network structure, i.e., the adjacency matrix of the network  $G$ , is in general very difficult to obtain for a large population. However, it might be possible to learn some characteristics of the network, in particular, the degree distribution of the nodes. The configuration model (see, e.g., the book [15]) is one family of random networks that are specified by degree distributions of nodes. A randomly selected node in such a random network has degree  $k$  with probability  $p_k$ . The edges of a node are randomly connected to the edges of the other nodes. As the edge connections are random, the infected subgraph appears to be a *tree* (with high probability) if one follows an edge of an infected node to propagate the disease to the other nodes in such a network. The tree assumption is one of the most important properties of the configuration model. Another crucial property of the configuration model is the *excess degree distribution*. The probability that one finds a node with degree  $k + 1$  along an edge connected to that node is

$$q_k = \frac{(k+1)p_{k+1}}{\sum_{\ell=0}^{\infty} (\ell+1)p_{\ell+1}}. \quad (37)$$

Thus, excluding the edge coming to the node, there are still  $k$  edges that can propagate the disease. This is also the reason why  $q_k$  is called the excess degree distribution. Note that the excess distribution  $q_k$  is in general different from the degree distribution  $p_k$ . They are the same when  $p_k$  is the Poisson degree distribution. In that case, the configuration model reduces to the famous Erdős-Rényi random graph.

As the infected subgraph is a tree in the configuration model, we are interested to know whether the size of the infected tree is finite. We say that there is no outbreak if the size of the infected tree of an infected node is finite with probability 1. Let  $u_1$  (resp.  $u_2$ ) be the probability that the size of the infected tree of a type I (resp. type II) node is finite via a specific one of its neighbors. Then,

$$u_1 = 1 - \phi_1 + \phi_1 \sum_{k=0}^{\infty} (w_1 q_k u_1^k + w_2 q_k u_2^k), \quad (38)$$

$$u_2 = 1 - \phi_2 + \phi_2 \sum_{k=0}^{\infty} (w_1 q_k u_1^k + w_2 q_k u_2^k). \quad (39)$$

To see the intuition of (38), we note that either the neighbor is infected or not infected. It is not infected with probability  $1 - \phi_1$ . On the other hand, it is infected with probability  $\phi_1$ . Then with probability  $w_1$  (resp.  $w_2$ ), the infected neighbor is of type I (resp. type II). Also, with probability  $q_k$ , the neighbor has additional  $k$  edges to transmit the disease. From the tree assumption, the probability that these  $k$  edges all have finite infected trees is  $u_1^k$  (resp.  $u_2^k$ ) if the infected neighbor is of type I (resp. type II). The equation in (39) follows from a similar argument.

Let

$$g_1(z) = \sum_{k=0}^{\infty} q_k z^k \quad (40)$$

be the moment generating function of the excess degree distribution. Then we can simplify (38) and (39) as follows:

$$u_1 = 1 - \phi_1 + \phi_1 w_1 g_1(u_1) + \phi_1 w_2 g_1(u_2), \quad (41)$$

$$u_2 = 1 - \phi_2 + \phi_2 w_1 g_1(u_1) + \phi_2 w_2 g_1(u_2). \quad (42)$$

From (41) and (42), we can solve  $u_1$  and  $u_2$  by starting from  $u_1^{(0)} = u_2^{(0)} = 0$  and

$$u_1^{(m+1)} = 1 - \phi_1 + \phi_1 w_1 g_1(u_1^{(m)}) + \phi_1 w_2 g_1(u_2^{(m)}), \quad (43)$$

$$u_2^{(m+1)} = 1 - \phi_2 + \phi_2 w_1 g_1(u_1^{(m)}) + \phi_2 w_2 g_1(u_2^{(m)}). \quad (44)$$

It is easy to show (by induction) that  $u_1^{(m+1)} \geq u_1^{(m)}$  and  $u_2^{(m+1)} \geq u_2^{(m)}$ . Thus, they converge to some fixed point solution  $u_1^*$  and  $u_2^*$  of (41) and (42).

## 4.2 Connections to the Previous SIR Model

Now we show the connections to the SIR model in Section 3 by specifying the propagation probabilities  $\phi_1$  and  $\phi_2$ .

Suppose that one end of a randomly selected edge is a type I node. Then this type I node will infect  $\beta_1/\gamma_1$  persons on average from the SIR model in Section 3. Since the average excess degree is  $\sum_{k=0}^{\infty} k q_k = g_1'(1)$ , the average number of neighbors infected by this type I node is  $\phi_1 g_1'(1)$ . In order for a type I node to infect the same average number of nodes in the SIR model in Section 3, we have

$$\phi_1 = \frac{\beta_1/\gamma_1}{g_1'(1)}. \quad (45)$$

Similarly for  $\phi_2$ , we have

$$\phi_2 = \frac{\beta_2/\gamma_2}{g_1'(1)}. \quad (46)$$

With the propagation probabilities  $\phi_1$  and  $\phi_2$  specified in (45) and (46), we have the following stability result.

**Theorem 3.** *For the IC model (for disease propagation) in a random network constructed by the configuration model, suppose that the propagation probabilities  $\phi_1$  and  $\phi_2$  are specified in (45) and (46). Then the size of the infected tree is finite with probability 1 if*

$$R_0 = w_1 \frac{\beta_1}{\gamma_1} + w_2 \frac{\beta_2}{\gamma_2} < 1.$$

*Under such a condition, there is no outbreak.*

**Proof.** (Theorem 3)

Let  $\mathbf{u} = (u_1, u_2)^T$  and  $\mathbf{e} = (1, 1)^T$ . It suffices to show that  $\mathbf{u} = \mathbf{e}$  is the unique solution for the system of equations in (41) and (42) if  $R_0 < 1$ . We prove this by contradiction. Suppose that there is a solution of (41) and (42) that either  $u_1 < 1$  or  $u_2 < 1$  when  $R_0 < 1$ .

Since the moment generating function in (40) is a convex function, we have from the first order Taylor's expansion for  $g_1(u_1)$  and  $g_1(u_2)$  that

$$g_1(u_1) \geq g_1(1) + (u_1 - 1)g_1'(1), \quad (47)$$

$$g_1(u_2) \geq g_1(1) + (u_2 - 1)g_1'(1). \quad (48)$$

Note that  $g_1(1) = \sum_{k=0}^{\infty} q_k = 1$ . Replacing (47) and (48) into (41) and (42), we have

$$u_1 \geq 1 + \phi_1 g_1'(1)(w_1 u_1 + w_2 u_2 - 1), \quad (49)$$

$$u_2 \geq 1 + \phi_2 g_1'(1)(w_1 u_1 + w_2 u_2 - 1). \quad (50)$$

Writing these two equations in the matrix form yields

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \geq \begin{bmatrix} \phi_1 g_1'(1)w_1 & \phi_1 g_1'(1)w_2 \\ \phi_2 g_1'(1)w_1 & \phi_2 g_1'(1)w_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 1 - \phi_1 g_1'(1) \\ 1 - \phi_2 g_1'(1) \end{bmatrix}.$$

This can be further simplified by using (45) and (46). Thus, we have

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \geq \begin{bmatrix} w_1 \beta_1/\gamma_1 & w_2 \beta_1/\gamma_1 \\ w_1 \beta_2/\gamma_2 & w_2 \beta_2/\gamma_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 1 - \beta_1/\gamma_1 \\ 1 - \beta_2/\gamma_2 \end{bmatrix}. \quad (51)$$

Let

$$\mathbf{B} = \begin{bmatrix} w_1 \beta_1/\gamma_1 & w_2 \beta_1/\gamma_1 \\ w_1 \beta_2/\gamma_2 & w_2 \beta_2/\gamma_2 \end{bmatrix}, \quad (52)$$

and

$$\mathbf{z} = \begin{bmatrix} 1 - \beta_1/\gamma_1 \\ 1 - \beta_2/\gamma_2 \end{bmatrix}. \quad (53)$$

We now rewrite (51) in the following matrix form:

$$\mathbf{u} \geq \mathbf{B}\mathbf{u} + \mathbf{z}. \quad (54)$$

It is straightforward to see that the two eigenvalues of  $\mathbf{B}$  are

$$\tilde{\lambda}_1 = 0, \text{ and } \tilde{\lambda}_2 = w_1 \frac{\beta_1}{\gamma_1} + w_2 \frac{\beta_2}{\gamma_2} = R_0.$$

Moreover, the eigenvector corresponding to the eigenvalue  $R_0$  is

$$\mathbf{v} = \left( \frac{\beta_1}{\gamma_1}, \frac{\beta_2}{\gamma_2} \right)^T.$$

Recursively expanding (54) for  $m$  times yields

$$\begin{aligned} \mathbf{u} &\geq \mathbf{B}^{m+1}\mathbf{u} + (\mathbf{I} + \mathbf{B} + \dots + \mathbf{B}^m)\mathbf{z} \\ &= \mathbf{B}^{m+1}\mathbf{u} + \mathbf{e} + R_0^m\mathbf{v}. \end{aligned} \quad (55)$$

Since  $R_0 < 1$ , both  $\mathbf{B}^{m+1}\mathbf{u}$  and  $R_0^m\mathbf{v}$  converge to the zero vectors as  $m \rightarrow \infty$ . Letting  $m \rightarrow \infty$  in (55) yields  $\mathbf{u} \geq \mathbf{e}$ . This contradicts to the assumption that either  $u_1 < 1$  or  $u_2 < 1$ . ■

### 4.3 Social Distancing

Social distancing is an effective way to slow down the spread of a contagious disease. One common approach of social distancing is to allow every person to keep its interpersonal contacts up to (on average) a fraction  $a$  of its normal contacts (see, e.g., [27], [28]). In our IC model, this corresponds to that every node randomly disconnects one of its edges with probability  $1 - a$ .

As in the previous subsection, we let  $u_1$  (resp.  $u_2$ ) be the probability that the size of the infected tree of a type I (resp. type II) node is finite via a specific one of its neighbors. Then,

$$u_1 = 1 - a^2\phi_1 + a^2\phi_1 \sum_{k=0}^{\infty} (w_1 q_k u_1^k + w_2 q_k u_2^k), \quad (56)$$

$$u_2 = 1 - a^2\phi_2 + a^2\phi_2 \sum_{k=0}^{\infty} (w_1 q_k u_1^k + w_2 q_k u_2^k). \quad (57)$$

To see (56), note that a neighboring node of an infected node can be infected only if (i) the edge connecting these two nodes is not removed (with probability  $a^2$ ), and (ii) the disease propagates through the edge (with the propagation probability  $\phi_1$ ). This happens with probability  $a^2\phi_1$ . Then with probability  $w_1$  (resp.  $w_2$ ), the infected neighbor is of type I (resp. type II). Also, with probability  $q_k$ , the neighbor has additional  $k$  edges to transmit the disease. From the tree assumption, the probability that these  $k$  edges all have finite infected trees is  $u_1^k$  (resp.  $u_2^k$ ) if the infected neighbor is of type I (resp. type II). The equation in (57) follows from a similar argument.

In comparison with the two equations in (38) and (39), we conclude that this approach of social distancing reduces the propagation probabilities  $\phi_1$  and  $\phi_2$  to  $a^2\phi_1$  and  $a^2\phi_2$ , respectively. As a direct consequence of Theorem 3, we have the following corollary.

**Corollary 4.** *Suppose that a social distancing approach allows every person to keep its interpersonal contacts up to (on average) a fraction  $a$  of its normal contacts. For the IC model (for disease propagation) in a random network constructed by the configuration model, the size of the infected tree is finite with probability 1 if*

$$a^2 R_0 < 1. \quad (58)$$

*Under such a condition, there is no outbreak.*

Another commonly used approach of social distancing is canceling mass gatherings. Such an approach aims to eliminate the effect of “superspreaders” who have lots of interpersonal contacts. For this, we consider a disease control parameter  $k_0$  and remove

nodes with the number of edges larger than or equal to  $k_0$  in our IC model. Analogous to the derivation of (38) and (39), we have

$$\begin{aligned} u_1 &= 1 - \phi_1 + \phi_1 \sum_{k=k_0-1}^{\infty} q_k \\ &\quad + \phi_1 \sum_{k=0}^{k_0-2} q_k (w_1 u_1^k + w_2 u_2^k), \end{aligned} \quad (59)$$

$$\begin{aligned} u_2 &= 1 - \phi_2 + \phi_2 \sum_{k=k_0-1}^{\infty} q_k \\ &\quad + \phi_2 \sum_{k=0}^{k_0-2} q_k (w_1 u_1^k + w_2 u_2^k). \end{aligned} \quad (60)$$

To see (59), we note that a type I infected person only infects a finite number of persons along an edge if (i) the disease does not propagate through the edge (with probability  $1 - \phi_1$ ), (ii) the disease propagates through the edge and the neighboring node is removed (with probability  $\phi_1 \sum_{k=k_0-1}^{\infty} q_k$ ), and (iii) the disease propagates through the edge and the neighboring node only infects a finite number of persons (with probability  $\phi_1 \sum_{k=0}^{k_0-2} q_k (w_1 u_1^k + w_2 u_2^k)$ ). The argument for (60) is similar.

Analogous to the stability result of Theorem 3, we have the following stability result for a social distancing approach that cancels mass gatherings.

**Theorem 5.** *Consider a social distancing approach that cancels mass gatherings by removing nodes with the number of edges larger than or equal to  $k_0$ . For the IC model (for disease propagation) in a random network constructed by the configuration model, suppose that the propagation probabilities  $\phi_1$  and  $\phi_2$  are specified in (45) and (46). Then the size of the infected tree is finite with probability 1 if*

$$\left( \frac{\sum_{k=0}^{k_0-2} k q_k}{\sum_{k=0}^{\infty} k q_k} \right) R_0 < 1. \quad (61)$$

*Under such a condition, there is no outbreak.*

**Proof.** (Theorem 5)

As in the proof of Theorem 3, we let  $\mathbf{u} = (u_1, u_2)^T$  and  $\mathbf{e} = (1, 1)^T$ . It suffices to show that  $\mathbf{u} = \mathbf{e}$  is the unique solution for the system of equations in (59) and (60) if the inequality in (61) is satisfied. We prove this by contradiction. Suppose that there is a solution of (59) and (60) that either  $u_1 < 1$  or  $u_2 < 1$  when the inequality in (61) is satisfied.

Since  $f(u) = u^k$  is a convex function for  $u \geq 0$  and  $k \geq 0$ , we have  $u^k \geq 1 + (u - 1)k$ . It then follows from (59) and (60) that

$$\begin{aligned} u_1 &\geq 1 - \phi_1 + \phi_1 \sum_{k=k_0-1}^{\infty} q_k + \phi_1 \sum_{k=0}^{k_0-2} q_k \\ &\quad + \phi_1 \sum_{k=0}^{k_0-2} k q_k (w_1 u_1 + w_2 u_2 - 1), \end{aligned} \quad (62)$$

$$\begin{aligned} u_2 &\geq 1 - \phi_2 + \phi_2 \sum_{k=k_0-1}^{\infty} q_k + \phi_2 \sum_{k=0}^{k_0-2} q_k \\ &\quad + \phi_2 \sum_{k=0}^{k_0-2} k q_k (w_1 u_1 + w_2 u_2 - 1). \end{aligned} \quad (63)$$



Writing these two equations in the matrix form and using (45) and (46) yields

$$\begin{aligned} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} &\geq \left( \sum_{k=0}^{k_0-2} kq_k \right) \begin{bmatrix} \phi_1 w_1 & \phi_1 w_2 \\ \phi_2 w_1 & \phi_2 w_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &+ \begin{bmatrix} 1 - \phi_1 (\sum_{k=0}^{k_0-2} kq_k) \\ 1 - \phi_2 (\sum_{k=0}^{k_0-2} kq_k) \end{bmatrix} \\ &= \frac{(\sum_{k=0}^{k_0-2} kq_k)}{g'_1(1)} \mathbf{B} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 1 - \frac{(\sum_{k=0}^{k_0-2} kq_k)}{g'_1(1)} \frac{\beta_1}{\gamma_1} \\ 1 - \frac{(\sum_{k=0}^{k_0-2} kq_k)}{g'_1(1)} \frac{\beta_2}{\gamma_2} \end{bmatrix}, \quad (64) \end{aligned}$$

where  $\mathbf{B}$  is the matrix in (52). Note that  $g'_1(1) = \sum_{k=0}^{\infty} kq_k$  is simply the expected excess degree. Following the same argument as that in Theorem 3, one can easily show that  $u_1 \geq 1$  and  $u_2 \geq 1$  when the inequality in (61) is satisfied. This contradicts to the assumption that either  $u_1 < 1$  or  $u_2 < 1$ . ■

Unfortunately, it is difficult to obtain an explicit expression for  $k_0$  to prevent an outbreak in (61). For this, we will resort to numerical computations in the next section.

## 5 NUMERICAL RESULTS

### 5.1 Dataset

In this section, we analyze and predict the trend of COVID-19 by using our time-dependent SIR model in Section 2 and the SIR model with undetectable infected persons in Section 3. For our analysis and prediction of COVID-19, we collect our dataset from the National Health Commission of the People's Republic of China (NHC) daily Outbreak Notification [12]. NHC announces the data as of 24:00 the day before. We collect the number of confirmed cases, the number of recovered persons, and the number of deaths from Jan. 15, 2020 to Mar. 2, 2020 as our dataset. The confirmed case is defined as the individual with positive real-time reverse transcription polymerase chain reaction (rRT-PCR) result. It is worth noting that in the Hubei province, the definition of the confirmed case has been relaxed to the clinical features since Feb. 12, 2020, while the other provinces use the same definition as before.

### 5.2 Parameter Setup

For our time-dependent SIR model, we set the orders of the FIR filters for predicting  $\beta(t)$  and  $\gamma(t)$  as 3, i.e.,  $J = K = 3$ . The stopping criteria of the model is set to  $X(t) \leq 0$ . Since the numbers of infected persons before Jan. 27, 2020 are too small to exhibit a clear trend (which may contain noises), we only use the data after Jan. 27, 2020 as our training data for predicting  $\beta(t)$  and  $\gamma(t)$ .

We use the scikit-learn library [29] (a third-party library of Python 3) to compute the ridge regression. The regularization parameters of predicting  $\beta(t)$  and  $\gamma(t)$  are set to 0.03 and  $10^{-6}$  respectively. Since the transmission rate  $\beta(t)$  is nonnegative, we set it to 0 if it is less than 0. Then, we use Algorithm 1 to predict the trend of COVID-19.

We note there are limitations for our prediction model to yield good results:

1. Our model is a deterministic epidemic model. It is based on the mean-field approximation for  $X(t)$  and  $R(t)$ . Such an approximation is a result of the law of large numbers. Therefore, when  $X(t)$  and  $R(t)$  are relatively small, the mean-field approximation may not be as accurate as expected. In those cases, one might have to resort to stochastic epidemic models, such as Markov chains. We will leave it as our future work.
2. The data can be extremely noisy at the beginning. As such, choosing a good starting day to train the ridge regression for prediction is quite important.
3. The recovering rates of several countries might not yet show a clear trend on Mar. 2, 2020. As such, it is difficult to predict the recovering rates for these countries. One remedy for this is to use a constant recovering rate of 1/30 days, which is the median recovery/death estimate by medical professionals. Another approach is to wait for a clear trend for the recovering rate. The reasons for the low recovering rate and the unclear trend of recovering rate are the lack of specific drugs and the shortage of medical resources. Please refer to Section 6 for more discussions on this issue.

### 5.3 Time Evolution of the Time-dependent SIR Model

In Figure 1, we show the time evolution of the number of infected persons and the number of recovered persons. The *circle-marked solid curves* are the real historical data by Mar. 2, 2020, and the *star-marked dashed curves* are our prediction results for the future. The prediction results imply that the disease will end in 6 weeks, and the number of the total confirmed cases would be roughly 80,000 if the Chinese government remains their control policy, such as city-wide lockdown and suspension of works and classes.

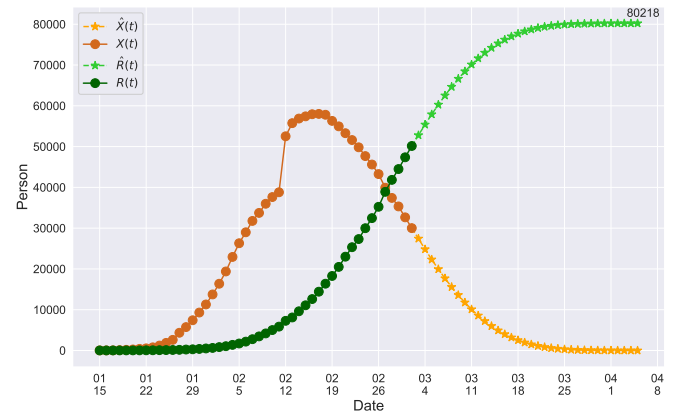


Figure 1: Time evolution of the time-dependent SIR model of the COVID-19. The circle-marked solid curve with dark orange (resp. green) color is the real number of infected persons  $X(t)$  (resp. recovered persons  $R(t)$ ), the star-marked dashed curve with light orange (resp. green) color is the predicted number of infected persons  $\hat{X}(t)$  (resp. recovered  $\hat{R}(t)$  persons).

In Figure 2, we show the measured  $\beta(t)$  and  $\gamma(t)$  from the real historical data. We can see that  $\beta(t)$  decreases dramatically, and  $\gamma(t)$  increases slightly. This is a direct result of the Chinese government that tries to suppress the transmission rate  $\beta(t)$  by city-wide lockdown and traffic halt. On the other hand, due to the

lack of effective drugs and vaccines for COVID-19, the recovering rate  $\gamma(t)$  grows relatively slowly. Additionally, there is a definition change of the confirmed case on Feb. 12, 2020 that makes the data related to Feb. 11, 2020 have no reference value. We mark these data points for  $\beta(t)$  and  $\gamma(t)$  with the gray dashed curve.

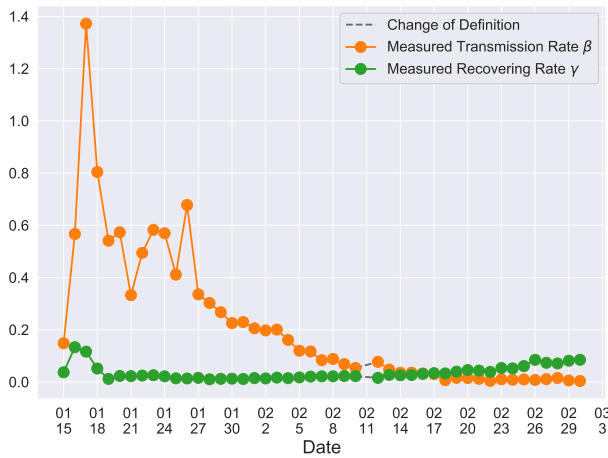


Figure 2: Measured transmission rate  $\beta(t)$  and recovering rate  $\gamma(t)$  of the COVID-19 from Jan. 15, 2020 to Feb. 19, 2020. The two curves are measured according to (11) and (10) respectively.

In an epidemic model, one crucial question is whether the disease can be contained and the epidemic will end, or whether there will be a pandemic that infects a certain fraction of the total population  $n$ . To answer this, one commonly used metric is the basic reproduction number  $R_0$  that is defined as the average number of additional infections by an infected person before it recovers in a wholly susceptible population [13]. In the classical SIR model,  $R_0$  is simply  $\beta/\gamma$  as an infected person takes (on average)  $1/\gamma$  days to recover, and during that period time, it will be in contact with (on average)  $\beta$  persons. In our time-dependent SIR model, the effective reproduction number at time  $t$ , denoted by  $R_e(t)$ , is defined as  $\beta(t)/\gamma(t)$ . If  $R_e(t) > 1$ , the disease will spread exponentially and infects a certain fraction of the total population  $n$ . On the contrary, the disease will eventually be contained. Therefore, by observing the change of  $R_e(t) > 1$  with respect to time or even predicting  $R_e(t) > 1$  in the future, we can check whether certain epidemic control policies are effective or not.

In Figure 3, we show the measured effective reproduction number  $R_e(t)$ , and the predicted effective reproduction number  $\hat{R}_e(t)$ . The blue circle-marked solid curve is the measured  $R_e(t)$  and the purple star-marked dashed curve is the predicted  $\hat{R}_e(t)$  (from Feb. 15, 2020). It is clear that  $R_e(t)$  has decreased dramatically since Jan. 28, 2020, and it implies that the control policies work in China. More importantly, it shows that the turning point is Feb. 17, 2020 when  $\hat{R}_e(t) < 1$ . In the following days after Feb. 17, 2020,  $X(t)$  will decrease exponentially, and that will lead to the end of the epidemic in China. Our model predicts precisely that  $R_e(t)$  will go less than 1 on Feb. 17, 2020 by 3 days in advance (Feb. 14, 2020). The results show that our model is very effective in tracking the characteristics of  $\beta(t)$  and  $\gamma(t)$ .

## 5.4 One-day Prediction

To show the precision of our model, we demonstrate the prediction results for the next day (one-day prediction) in Figure 4. It contains

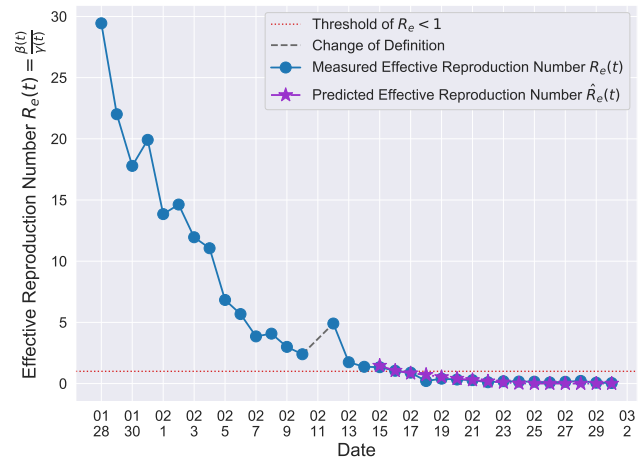


Figure 3: Effective reproduction number  $R_e(t)$  of the time-dependent SIR model of the COVID-19 in China. The circle-marked solid curve with blue color is the  $R_e(t)$  based on the given data from Jan. 27, 2020 to Feb. 20, 2020, the star-marked dashed curve with purple color is the predicted  $\hat{R}_e(t)$  based on the data from Jan. 27, 2020 to Feb. 15, 2020, and the dashed line with red color is the percolation threshold 1 for the effective reproduction number.

the predicted number of infected persons  $\hat{X}(t)$  (orange star-marked dashed curve), the predicted number of recovered persons  $\hat{R}(t)$  (green star-marked dashed curve), and the real number of infected and recovered persons (dark orange and dark green circle-marked solid curves) every day. The unpredictable days due to the change of the definition of the confirmed case on Feb. 12, 2020, are marked as gray. The predicted curves are extremely close to the measured curves (obtained from the real historical data). In this figure, we also annotate the predicted number of infected persons  $\hat{X}(t) = 27,433$  and the predicted number of recovered persons  $\hat{R}(t) = 52,785$  on Mar. 3, 2020.

We further examine our prediction accuracy in Figure 5. The error rates are all within  $\pm 3\%$  except for the predicted number of recovered persons  $\hat{R}(t)$  on Feb. 1, Feb. 3, and Feb. 5, 2020. The gray dashed line stands for the unpredictable points due to the change of definition of the confirmed case. However, from the prediction results after Feb. 16, 2020, we find that our model can still keep tracking  $\beta(t)$  and  $\gamma(t)$  accurately and overcome the impact of the change of the definition.

## 5.5 Basic Reproduction Numbers of Several Other Countries

In addition to the dataset for China, we also measure the effective reproduction number  $R_e(t)$  on Mar. 31, 2020 for several countries from the datasets in [16]. This is shown in the last column of Table 1. As the data for the cumulative numbers of recovered persons for these countries are noisy, we also show the estimated  $R_e(t)$  under various assumptions of the average time to recover  $1/\gamma$ . The  $R_e(t)$  values for the five countries, including United States of America, the United Kingdom, France, Iran, and Spain are very high. On the other hand, it seems that Italy is gaining control of the spread of the disease after the Italian government announces the lockdown and forbids the gatherings of people on Mar. 10,

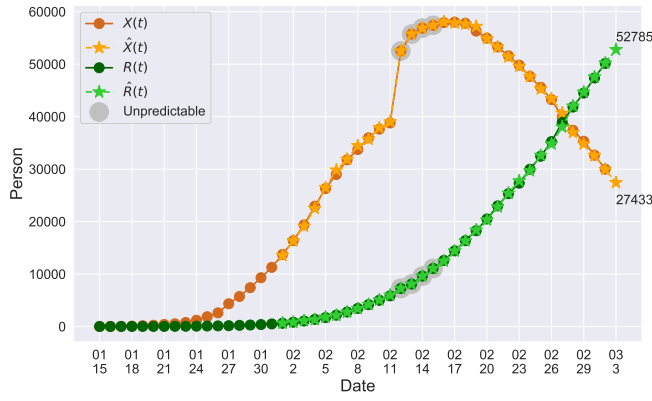


Figure 4: One-day prediction for the number of infected and recovered persons. The unpredictable points due to the change of definition of the confirmed case are marked as gray. The circle-marked solid curve with dark orange (resp. green) color is the real number of infected persons  $X(t)$  (resp. recovered persons  $R(t)$ ), the star-marked dashed curve with light orange (resp. green) color is the predicted number of infected persons  $\hat{X}(t)$  (resp. recovered persons  $\hat{R}(t)$ ).

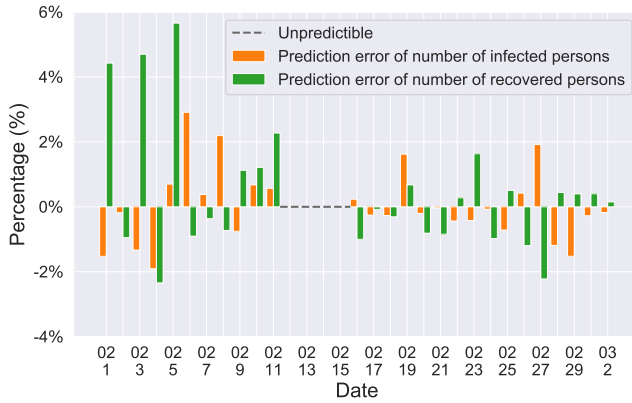


Figure 5: Errors of the one-day prediction of the number of infected and recovered persons. The unpredictable points due to the change of definition of the confirmed case on Feb. 12, 2020 are marked as the gray dash line.

2020. Also, both Germany and Republic of Korea are capable of controlling the spread of the disease.

## 5.6 The Effects of Type II Infected Persons

In this subsection, we show how undetectable (type II) infected persons affect the epidemic. In particular, we are interested in addressing the question of whether the existence of undetectable infected persons (type II) can cause an outbreak.

To carry out our numerical study, we need to fix some variables in the system of difference equations in (20)-(22). For the transmission rate (resp. recovering rate) of type I infected persons, i.e.,  $\beta_1$  (resp.  $\gamma_1$ ), we set it to be the measured  $\beta(t) = 0.00383$  (resp.  $\gamma(t) = 0.08493$ ) on Mar. 1, 2020 in China. The rationale behind this is that type I infected persons were detected and they were under treatment and isolation after Mar. 1, 2020 in China. Also, as there is no medicine for COVID-19, we may assume that

these two types of infected persons have the same recovering rate, i.e.,  $\gamma_2 = \gamma_1$ . In view of the system equation in (20)-(22), there are still two free variables  $w_2$  and  $\beta_2$ , where  $w_2$  is the probability that an infected person is of type II and  $\beta_2$  is the transmission rate of type II.

In Figure 6, we illustrate how  $w_2$  and  $\beta_2$  affect the outbreak of the COVID-19. Such a figure is known as the phase transition diagram in [15]. The black curve in Figure 6 is the curve when the spectral radius of the transition matrix  $\mathbf{A}$  in (23) equals to 1. This curve represents the percolation threshold of COVID-19. If  $w_2$  and  $\beta_2$  fall above the black curve (in the orange zone), then there will be an outbreak. On the contrary, if  $w_2$  and  $\beta_2$  fall below the black curve (in the yellow zone), then there will not be an outbreak. As shown in Figure 6, we would like to point out the importance of detecting an infected person. As long as more than 90% of those infected persons can be actually detected and properly isolated and treated, it is possible to contain the spread of the disease even if the transmission rate of type II infected persons, i.e.,  $\beta_2$ , is as high as 0.7. On the other hand, suppressing the transmission rate of type II infected persons can also be effective in controlling the disease while the detection rate is not that high. For example, wearing masks and washing hands can be an effective epidemic prevention mechanism to reduce  $\beta_2$ .

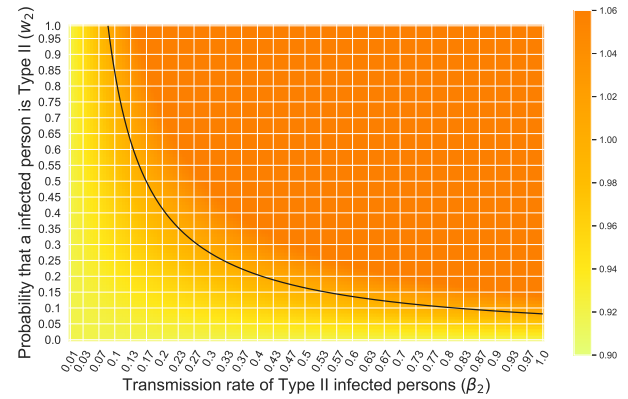


Figure 6: Phase transition diagram of an outbreak with respect to  $\beta_2$  and  $w_2$ . The black curve is the percolation threshold. The orange area means the disease will be an outbreak while the yellow area means the disease is under control.

In the following experiments, we extend our study to other countries, including Japan, Singapore, South Korea, Italy, and Iran. We collect the historical data from Jan. 22, 2020 to Mar. 2, 2020 from the GitHub of Johns Hopkins University [16]. For these countries, the transmission rates  $\beta(t)$  and the recovering rates  $\gamma(t)$  (measured from the time-dependent SIR model in Section 2) during the initial period with a rapid increase of the number of confirmed cases can be viewed as  $\beta_2$  and  $\gamma_2$ . This is because during that period of time, there is no epidemic prevention intervention, and all the infected persons are basically not detected. It is interesting to note that different countries might have different  $\beta_2$  and  $\gamma_2$ . On the other hand, only 87.9% of COVID-19 cases have a fever from the report of WHO [2]. If we use body temperature as a means to detect an infected person, then only 87.9% of COVID-19 cases can be detected. For this, we set  $w_1 = 87.9\%$ .

With  $\beta_1$  and  $\gamma_1$  specified in the previous study for China, we plot the phase transition diagram in Figure 7 in terms of the two

Country	Estimated $R_e(t)$ when the average time to recover $1/\gamma$ is					$R_e(t)$ on Mar. 31, 2020
	14 Days	21 Days	28 Days	35 Days	42 Days	
United States of America	2.13	3.20	4.26	5.33	6.39	12.59
The United Kingdom	1.89	2.83	3.77	4.72	5.66	8.90
France	2.58	3.86	5.15	6.44	7.73	4.76
Iran	1.25	1.88	2.51	3.14	3.76	4.51
Spain	1.63	2.45	3.27	4.08	4.90	3.47
Italy	0.79	1.19	1.59	1.98	2.38	3.08
Germany	1.49	2.24	2.98	3.73	4.48	2.80
Republic of Korea	0.44	0.66	0.88	1.11	1.33	1.68

Table 1: The estimated  $R_e(t)$  under various assumptions of the average time to recover ( $1/\gamma$ ) from COVID-19, and the measured  $R_e(t)$  on Mar. 31, 2020.

variables  $\beta_2$  and  $\gamma_2$ . Again, the black curve is the curve when the spectral radius of the transition matrix  $\mathbf{A}$  in (23) equals to 1. Such a curve represents the percolation threshold of a COVID-19 outbreak. If  $\beta_2$  and  $\gamma_2$  fall above the black curve (in the orange zone), then there will be an outbreak. On the contrary, if  $\beta_2$  and  $\gamma_2$  fall below the black curve (in the yellow zone), then there will not be an outbreak. The countries with large confirmed cases, including Japan, Singapore, South Korea, Italy, and Iran, are marked in Figure 7. From Figure 7, we observe that both Singapore and Japan are below the percolation threshold. But Japan is much closer to the percolation threshold. On the other hand, both South Korea and Italy are above the percolation threshold, and they are on the verge of a potential outbreak on Mar. 2, 2020. These two countries must implement epidemic prevention policies urgently. Not surprisingly, on Mar. 10, 2020, the Italian government announces the lockdown and forbids the gatherings of people. It is worth mentioning that there are two marks for Iran in the Figure. The one above the percolation threshold is the one without adding the number of deaths into the number of recovered persons. The other one below the percolation threshold is the one that adds the number of deaths into the number of recovered persons. For some unknown reason, the death rate in Iran is higher than the other countries. The high death rate seems to prevent an outbreak in Iran.

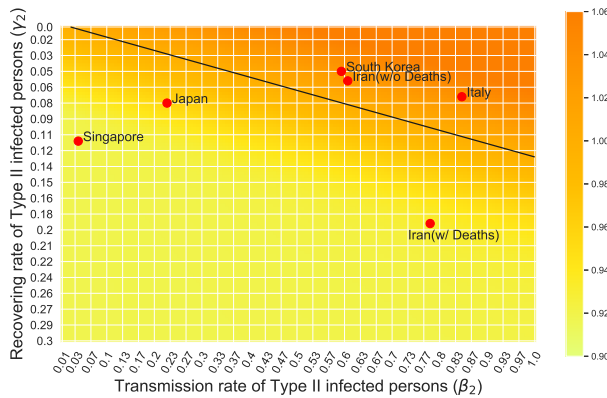


Figure 7: Phase transition diagram of an outbreak with respect to  $\beta_2$  and  $\gamma_2$ . The black curve is the percolation threshold. The orange area means the disease will be an outbreak, while the yellow area means the disease is under control.

In order to show that  $w_1(t)$  can change significantly over time to reflect the change of the government's epidemic prevention policy, we plot in Figure 8 the estimated  $w_1(t)$  from Feb. 23, 2020 to Mar. 18, 2020 in South Korea. Using the same experimental settings in this section, we set  $\beta_1$ ,  $\gamma_1$ , and  $\gamma_2$  to the measured values in the previous study for China, and set  $\beta_2$  to the measured value in South Korea. We set 5 days as a time interval and leverage the recursive formula in (25) to estimate the value of  $w_1(t)$  over time. In each time interval,  $w_1$  is obtained by minimizing the sum of the squared residuals under (25) [29]. As shown in Figure 8, we can see that  $w_1$  in the first three time intervals, starting from Feb. 23, 2020, increase significantly. If we compare this result with the dataset provided by Our World in Data [30], we can find that South Korea began to conduct a large number of COVID-19 testing on the public on Feb. 24, 2020, and increased the number of testing every day onward. Therefore, we can infer that epidemic prevention policy like mass testing can effectively increase the probability of  $w_1$  and greatly reduce the proportion of undetectable persons. From the above discussions, we know that  $w_1(t)$  can reflect the impact on a new epidemic prevention policy. Also, mass testing has a significant effect on reducing the proportion of undetectable persons.

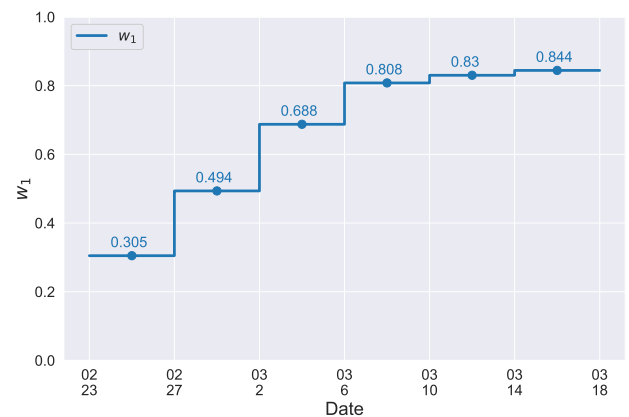


Figure 8: The estimated  $w_1$  as a function of 5-day intervals in South Korea.

## 5.7 The Effects of Social Distancing

In this subsection, we show the numerical results of the social distancing approach that cancels mass gatherings by removing



nodes with the number of edges larger than or equal to  $k_0$ . As shown in Theorem 5, the effective reproduction number is reduced by a factor of  $\frac{\sum_{k=0}^{k_0-2} kq_k}{\sum_{k=0}^{\infty} kq_k}$ , where  $q_k$  is the excess degree distribution of  $p_k$ . For this experiment, we use the dataset collected by [31] from Facebook. This dataset represents the verified Facebook page (with blue checkmark) networks of the artist category. The blue checkmark means Facebook has confirmed that an account is the authentic presence of the public figure, celebrity, or global brand it represents. Each node in the network represents the page, and edges between two nodes are mutual likes among them. This dataset is composed of 50,515 nodes and 819,306 edges. Some other properties are listed as follow: mean degree 32.4, max degree 1,469, diameter 11, and clustering coefficient 0.053. In Figure 9, we show the log-log plots of the degree distribution and the excess degree distribution of this dataset. The degree distribution appears to be a (truncated) Pareto distribution with the exponent 1.69 (the slope in the figure).

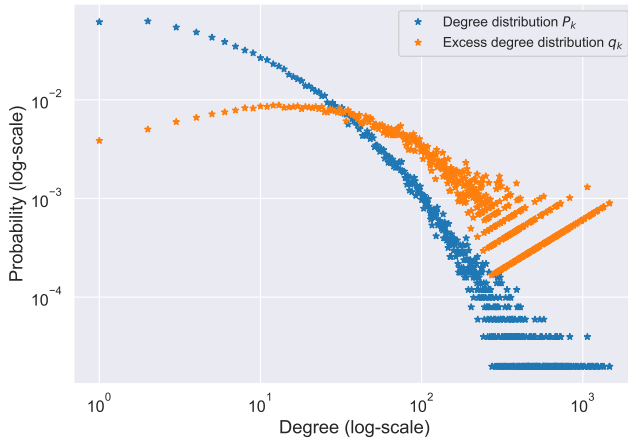


Figure 9: The degree distribution and the excess degree distribution of the Facebook dataset.

In Figure 10, we plot the reduction ratio  $\frac{\sum_{k=0}^{k_0-2} kq_k}{\sum_{k=0}^{\infty} kq_k}$  as a function of  $k_0$ . The ratio is between 0 and 1, and it is monotonically increasing in  $k_0$ . The reduction ratio represents how large a gathering is forbidden to reduce the corresponding ratio of  $R_e(t)$ . Using the  $R_e(t)$  values (on Mar. 31, 2020) in the last column of Table 1, we also show that the minimum  $k_0$ s to prevent an outbreak in Italy, U.S., and South Korea are 63, 195, and 435, respectively; moreover, the affected fraction of tail distributions are 13.1%, 2.2%, and 0.4%, respectively. In particular, if canceling mass gathering is the only measure used for controlling COVID-19 in the U.S. with the  $R_e(t)$  value of 12.59 on Mar. 31, 2020, then one can prevent an outbreak by “removing” all the nodes with degrees larger than or equal to 63 (in the Facebook dataset), and the removal might affect 13.1% of the nodes in the Facebook dataset.

## 6 DISCUSSIONS AND SUGGESTIONS

As of Mar. 2, 2020, it seems that COVID-19 has been gradually controlled in China since the prevention policies, such as city-wide lockdown was issued in China in Jan. 2020. Although the policies such as traffic halt, small community management, and city-wide lockdown can effectively reduce the transmission rates  $\beta_1$  and

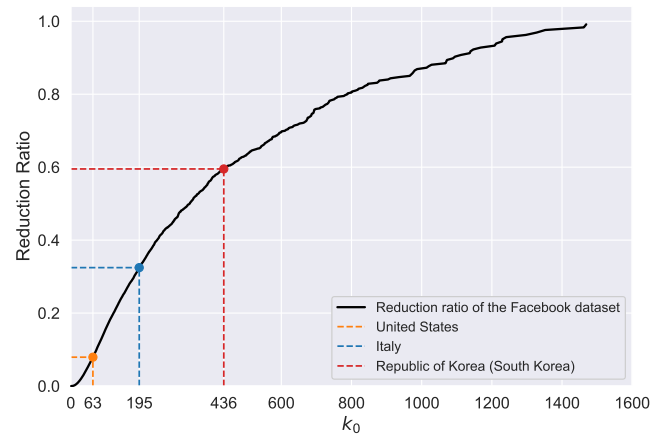


Figure 10: The reduction ratio  $\frac{\sum_{k=0}^{k_0-2} kq_k}{\sum_{k=0}^{\infty} kq_k}$  as a function of  $k_0$ . The minimum  $k_0$ s to prevent an outbreak in Italy, U.S., and South Korea are 63, 195, and 435, respectively.

$\beta_2$ ; however, these relatively extreme policies not only restrict the right of personal freedom but also affect the normal operation of society. These extreme policies forced several companies and factories to halt production, which impacts all sectors of the economy. Therefore, to strike a balance between the prevention of disease and ensuring the normal operation of society is crucial, and it is important to suggest the so-called “optimal” control policies.

For this, we would like to put forward some discussions and suggestions for controlling the spread of COVID-19 based on the observation made from the results of our system models. From our results in Theorem 1, Corollary 2, Theorem 3, Corollary 4, and Theorem 5, we know that there is no outbreak for a disease if the effective reproduction number

$$R_e = h \cdot s \cdot (w_1 \frac{\beta_1}{\gamma_1} + w_2 \frac{\beta_2}{\gamma_2}) < 1, \quad (65)$$

where  $h$  is the herd immunity reduction factor in Corollary 2 when  $1 - h$  fraction of individuals are immune to the disease,  $s$  is the social distancing reduction factor in Corollary 4 or Theorem 5,  $w_1$  (resp.  $w_2$ ) is the probability that an infected person can (resp. cannot) be detected, and  $\beta_1$  and  $\gamma_1$  (resp.  $\beta_2$  and  $\gamma_2$ ) are the transmission rate and the recovering rate of an infected person who can (resp. cannot) be detected. For COVID-19,  $\beta_1 \leq \beta_2$  and  $\gamma_1 \geq \gamma_2$  as an infected person, once detected, can be treated (to shorten the recovery time) and isolated (to reduce the transmission rate). As  $w_1 + w_2 = 1$ , it is thus preferable to have a much larger  $w_1$  than  $w_2$ .

To prevent an outbreak, one should minimize the value on the left-hand side of (65). Here we discuss several approaches for that.

1. Increasing the recovering rate  $\gamma_1$ : the most effective way to increase  $\gamma_1$  is to find anti-virus drugs; however, it takes time. Hence, we should focus on the other approaches to control the spread of the disease in this stage.
2. Reducing the herd immunity reduction factor  $h$ : the most effective way for this is to find vaccines to reduce the fraction of susceptible persons. Once again, it takes time, and we should focus on the other approaches to control the spread of the disease in this stage.

3. Decreasing the transmission rate  $\beta_1$  [32]: once an infected person is detected, it should be isolated to avoid extra infection on society and lower the transmission rate  $\beta_1$ . Quarantine of the persons who are suspected to be in contact with infected persons could also lower the transmission rates  $\beta_1$  and  $\beta_2$ .
4. Increasing the detection probability  $w_1$  (and thus reducing  $w_2$ ): mass testing can certainly increase  $w_1$ . In fact, South Korea did an outstanding job of drive-thru testing. If mass testing is not possible due to the limitation of medical resources, then measuring body temperature can also be an effective alternative, as 87.9% of the confirmed cases of COVID-19 have a fever. In addition to this, one can also track the travel history, occupation, contact, and cluster (TOCC) of the confirmed cases to narrow the range of the possible sources. These sources might contain undetectable infected persons, and testing the close contacts of these possible sources can thus increase  $w_1$  by detecting undetectable infected persons.
5. Decreasing the transmission rate  $\beta_2$ : propaganda of health education knowledge can reduce the transmission rate  $\beta_2$  substantially. For example, wearing masks in public and enclosed space, washing hands, avoiding touching your mouth, eyes, and nose are good ways to not only protect ourselves from being infected by the undetectable infected persons but also avoid infecting others.
6. Reducing the social distancing reduction factor  $s$ : as shown in Corollary 4 and Theorem 5, there are two practical approaches that can reduce the social distancing reduction factor  $s$ : (i) allowing every person to keep its interpersonal contacts up to a fraction of its normal contacts, and (ii) canceling mass gatherings.

## 7 CONCLUSION AND FUTURE WORK

In this paper, we conducted mathematical and numerical analyses for COVID-19. Our time-dependent SIR model is not only more adaptive than traditional static SIR models, but also more robust than direct estimation methods. Our numerical results show that one-day prediction errors for the number of infected persons  $X(t)$  and the number of recovered persons  $R(t)$  are within (almost) 3% for the dataset collected from the National Health Commission of the People's Republic of China (NHC) [12]. Moreover, we are capable of tracking the characteristics of the transmission rate  $\beta(t)$  and the recovering rate  $\gamma(t)$  with respect to time  $t$ , and precisely predict the future trend of the COVID-19 outbreak in China.

To address the impact of undetectable infections in COVID-19, we extended our SIR model by considering two types of infected persons: detectable infected persons and undetectable infected persons. Whether there is an outbreak in such a model is characterized by the spectral radius of a  $2 \times 2$  matrix that is closely related to the basic reproduction number  $R_0$ . In addition to our numerical analysis for China, we further extended our study to other countries, including Japan, Singapore, South Korea, Italy, and Iran.

To understand the effects of social distancing approaches, including the reduction of interpersonal contacts and canceling mass gatherings, we analyzed the IC model for disease propagation in the configuration model. By relating the propagation probabilities in the IC model to the transmission rates and recovering rates in

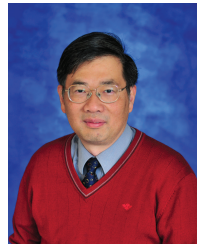
the SIR model, we showed these social distancing approaches can lead to a reduction of the effective reproduction number  $R_e$ .

Last but not least, based on the experimental results, some discussions and suggestions on epidemic prevention are proposed from the perspective of our models. One possible future work is to extend our deterministic SIR model to stochastic models so that super-spreading events (that show the high stochasticity under low prevalence [33]) can be included in the analysis. By doing so, one can further improve the precision of the prediction results.

## REFERENCES

- [1] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-m. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov *et al.*, "Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19)," *International journal of infectious diseases*, vol. 94, p. 154, 2020.
- [2] "Coronavirus disease (covid-19) outbreak," Jan 2020. [Online]. Available: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>
- [3] I. Nesteruk, "Statistics based predictions of coronavirus 2019-ncov spreading in mainland china," *MedRxiv*, 2020.
- [4] Y. Chen, J. Cheng, Y. Jiang, and K. Liu, "A time delay dynamical model for outbreak of 2019-ncov and the parameter identification," *arXiv preprint arXiv:2002.00418*, 2020.
- [5] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of covid-19 in china by dynamical modeling," *arXiv preprint arXiv:2002.06563*, 2020.
- [6] 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," *Journal of Evidence-Based Medicine*, 2020.
- [7] B. F. Maier and D. Brockmann, "Effective containment explains sub-exponential growth in confirmed cases of recent covid-19 outbreak in mainland china," *arXiv preprint arXiv:2002.07572*, 2020.
- [8] Q. Lin, S. Zhao, D. Gao, Y. Lou, S. Yang, S. S. Musa, M. H. Wang, Y. Cai, W. Wang, L. Yang *et al.*, "A conceptual model for the outbreak of coronavirus disease 2019 (covid-19) in wuhan, china with individual reaction and governmental action," *International journal of infectious diseases*, 2020.
- [9] S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He *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 the outbreak," *International Journal of Infectious Diseases*, 2020.
- [10] T. Zeng, Y. Zhang, Z. Li, X. Liu, and B. Qiu, "Predictions of 2019-ncov transmission ending via comprehensive methods," *arXiv preprint arXiv:2002.04945*, 2020.
- [11] Z. Hu, Q. Ge, L. Jin, and M. Xiong, "Artificial intelligence forecasting of covid-19 in china," *arXiv preprint arXiv:2002.07112*, 2020.
- [12] "Outbreak notification," Jan 2020. [Online]. Available: [http://www.nhc.gov.cn/xcs/yqtb/list\\_gzbd.shtml](http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml)
- [13] R. M. Anderson and R. M. May, "Vaccination and herd immunity to infectious diseases," *Nature*, vol. 318, no. 6044, pp. 323–329, 1985.
- [14] S. Zhao, S. S. Musa, J. T. Hebert, P. Cao, J. Ran, J. Meng, D. He, and J. Qin, "Modelling the effective reproduction number of vector-borne diseases: the yellow fever outbreak in luanda, angola 2015–2016 as an example," *PeerJ*, vol. 8, p. e8601, 2020.
- [15] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.
- [16] CSSEGISandData and J. H. University, "Covid-19," Feb 2020. [Online]. Available: <https://github.com/CSSEGISandData/COVID-19>
- [17] B. F. Finkenstädt and B. T. Grenfell, "Time series modelling of childhood diseases: a dynamical systems approach," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 49, no. 2, pp. 187–205, 2000.
- [18] A. D. Becker and B. T. Grenfell, "tsir: An r package for time-series susceptible-infected-recovered models of epidemics," *PLOS ONE*, vol. 12, no. 9, pp. 1–10, 09 2017. [Online]. Available: <https://doi.org/10.1371/journal.pone.0185528>
- [19] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann, "Inferring change points in the spread of covid-19 reveals the effectiveness of interventions," *Science*, 2020. [Online]. Available: <https://science.sciencemag.org/content/early/2020/05/14/science.abb9789>
- [20] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, "A time-dependent sir model for covid-19 with undetectable infected persons," *arXiv preprint arXiv:2003.00122*, 2020.

- [21] O. Wyman, "Covid-19 pandemic navigator core model: Overview of methodology and use cases," Apr 2020. [Online]. Available: <https://hlc-hlc21901.s3-us-west-2.amazonaws.com/public/White%20Paper%20OW%20Pandemic%20Navigator%20Core%20Model.pdf>
- [22] D. Kenpe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.
- [23] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Stable algorithms for link analysis," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 258–266.
- [24] B. S. Dayal and J. F. MacGregor, "Identification of finite impulse response models: methods and robustness issues," *Industrial & engineering chemistry research*, vol. 35, no. 11, pp. 4078–4090, 1996.
- [25] T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens, "Estimating the generation interval for covid-19 based on symptom onset data," *medRxiv*, 2020.
- [26] H. E. Randolph and L. B. Barreiro, "Herd immunity: Understanding covid-19," *Immunity*, vol. 52, no. 5, pp. 737–741, 2020.
- [27] C.-J. Chen, *Epidemiology: Principles and Methods*. Linking Publishing Company, 1999.
- [28] N. G. Becker, *Modeling to inform infectious disease control*. CRC Press, 2015, vol. 74.
- [29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [30] E. O.-O. Max Roser, Hannah Ritchie and J. Hasell, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020, <https://ourworldindata.org/coronavirus>.
- [31] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "Gemsec: Graph embedding with self clustering," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019*. ACM, 2019, pp. 65–72. [Online]. Available: <https://snap.stanford.edu/data/gemsec-Facebook.html>
- [32] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong *et al.*, "Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia," *New England Journal of Medicine*, 2020.
- [33] B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu, "Stochasticity and heterogeneity in the transmission dynamics of sars-cov-2," *arXiv preprint arXiv:2005.13689*, 2020.



**Cheng-Shang Chang** (S'85-M'86-M'89-SM'93-F'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1983, and the M.S. and Ph.D. degrees from Columbia University, New York, NY, USA, in 1986 and 1989, respectively, all in electrical engineering.

From 1989 to 1993, he was employed as a Research Staff Member with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Since 1993, he has been with the Department of Electrical Engineering, National Tsing Hua University, Taiwan, where he is a Tsing Hua Distinguished Chair Professor. He is the author of the book *Performance Guarantees in Communication Networks* (Springer, 2000) and the coauthor of the book *Principles, Architectures and Mathematical Theory of High Performance Packet Switches* (Ministry of Education, R.O.C., 2006). His current research interests are concerned with network science, big data analytics, mathematical modeling of the Internet, and high-speed switching.

Dr. Chang served as an Editor for *Operations Research* from 1992 to 1999, an Editor for the *IEEE/ACM TRANSACTIONS ON NETWORKING* from 2007 to 2009, and an Editor for the *IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING* from 2014 to 2017. He is currently serving as an Editor-at-Large for the *IEEE/ACM TRANSACTIONS ON NETWORKING*. He is a member of IFIP Working Group 7.3. He received an IBM Outstanding Innovation Award in 1992, an IBM Faculty Partnership Award in 2001, and Outstanding Research Awards from the National Science Council, Taiwan, in 1998, 2000, and 2002, respectively. He also received Outstanding Teaching Awards from both the College of EECS and the university itself in 2003. He was appointed as the first Y. Z. Hsu Scientific Chair Professor in 2002. He received the Merit NSC Research Fellow Award from the National Science Council, R.O.C. in 2011. He also received the Academic Award in 2011 and the National Chair Professorship in 2017 from the Ministry of Education, R.O.C. He is the recipient of the 2017 IEEE INFOCOM Achievement Award.



**Yi-Cheng Chen** received his B.S. degree in electrical engineering from National Taiwan University of Science and Technology, Taipei, Taiwan (R.O.C.), in 2018. He is currently pursuing the M.S. degree in the Institute of Communications engineering, National Tsing-Hua University, Hsinchu, Taiwan (R.O.C.).



**Tzu-Hsuan Liu** received the B.S. degree in communication engineering from National Central University, Taoyuan, Taiwan (R.O.C.), in 2018. She is currently pursuing the M.S. degree in the Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan (R.O.C.). Her research interest is in 5G wireless communication.



**Ping-En Lu** (GS'17) received his B.S. degree in communication engineering from the Yuan Ze University, Taoyuan, Taiwan (R.O.C.), in 2015. He is currently pursuing the Ph.D. degree in the Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan (R.O.C.). He won the ACM Multimedia 2017 Social Media Prediction (SMP) Challenge with his team in 2017. His research interest is in network science, efficient clustering algorithms, network embedding, and deep learning algorithms. He is

an IEEE Graduate Student Member.