Exact results for the stochastic SIS epidemic model in densely populated environments

Tingting Chen[1], Zhen Jin[1,2], Guirong Liu[1,2,*], Chen Jia[3,*]

[1] School of Mathematics and Statistics, Shanxi University, Taiyuan, Shanxi 030006, China.

[2] Key Laboratory of Complex Systems and Data Science of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China.

[3] Applied and Computational Mathematics Division, Beijing Computational Science Research Center, Beijing 100193, China.

[*] Correspondence: lgr5791@sxu.edu.cn (G. Liu); chenjia@csrc.ac.cn (C. Jia)

Abstract

In this study, we investigate the stochastic dynamics of an extended SIS epidemic model in densely populated environments within a Markov jump process framework. We solve the master equation in closed form and obtain exact solutions of the time-dependent distribution of the number of infected individuals, the quasi-stationary distribution, the extinction time distribution of the epidemic, and the distribution of the first-passage time at which the number of infections reaches a certain threshold. The approximated quasi-stationary distribution and mean extinction time are also derived using the large deviation theory. Interestingly, we find that the first nonzero eigenvalue of the generator matrix of the Markovian model characterizes the extinction rate of the epidemic, while the second nonzero eigenvalue characterizes its outbreak rate. We also examine the stochastic bifurcation for our model based on the time evolution of the probability distribution and the bifurcation threshold of the basic reproduction number for the stochastic SIS model is shown to be large than that for its deterministic counterpart. Finally, we demonstrate that analyzing the first-passage time distribution can offer early warning for interventions and optimize the allocation of emergency beds.

1 Introduction

The spread of infectious diseases poses a significant threat to human populations, making it a critical concern for public health [1, 2]. Epidemic models have become essential tools for understanding the transmission mechanisms of infectious diseases, providing a theoretical framework to analyze past outbreaks and predict the future trajectory of epidemics. Common epidemic models include the SI (susceptible-infected), SIS (susceptible-infected-susceptible), and SIR (susceptible-infected-removed) compartment models, which typically classify individuals into different compartments based on their disease status [3–5]. These models are chosen based on the pathological characteristics of specific epidemics.

Over the past three decades, significant progress has been made in the mathematical modeling and theoretical analysis of various epidemic models, extensively studied from both deterministic and stochastic perspectives. Classical epidemic models assume homogeneous mixing within the population and disease transmission occurs through effective contact between susceptible and infected compartments. When the population size is very large, random fluctuations in disease transmission can be ignored, and the evolution of the number of individuals in each compartment is typically modeled by a set of ordinary differential equations (ODEs). Research within this framework has focused on various aspects such as threshold dynamics [6], bifurcation analysis [7], parameter inference [8], and optimal control [9]. The classical framework has also been enriched through the integration of network topologies [10], spatial diffusion [11, 12], time delays [13], and intervention measures [14], enabling deeper insights into the spread of epidemics.

However, when the population size is relatively small, random fluctuations in disease transmission can no longer be ignored, and the evolution of the system should be modeled within a stochastic framework. There are typically two types of stochastic epidemic models: stochastic differential equation (SDE) models and continuous-time Markov

chain (CTMC) models [5, 15, 16]. In SDE models, the population size in each compartment is treated as a real-valued continuous variable, whereas in CTMC models, the population size is treated as an integer-valued discrete variable, and hence provides a more accurate description. The probability distribution of the population size in each compartment is governed by a Fokker-Planck equation for SDE models and by a master equation for CTMC models. At the center of the stochastic dynamic theory of infectious diseases is a limit theorem proved by Kurtz in the 1970s [17, 18], which states that, in the limit of large population size, the CTMC model converges to the deterministic ODE model with probability one over any finite time interval. Roughly speaking, in Kurtz's theory, ODE models can be viewed as first-order approximations of CTMC models when the population size is large, while SDE models serve as second-order approximations (also referred to as diffusion approximations) of CTMC models [18].

Due to the complexity of CTMC models, various approximation techniques have been developed, including diffusion approximations [18, 19], van Kampen's system-size expansion [20], Wentzel-Kramers-Brillouin (WKB) approximations [21, 22], and branching process approximations [23, 24]. In general, the first three approximation methods provide accurate predictions for relatively large population sizes, whereas branching process approximations are mainly valid during the initial stages of an epidemic. These techniques have been extensively used to study key topics such as threshold dynamics [15, 16], stochastic bifurcation analysis [25, 26], moment closure strategies [27], estimation of the time to extinction [28, 29], and quasi-stationary distribution prior to disease extinction [30, 31].

While various approximation techniques have been extensively studied, exact results for stochastic epidemic models remain quite limited due to the nonlinear interactions between compartments. In this work, we present a systematic study of an extended stochastic SIS model in crowded environments within a CTMC framework. We derive closed-form solutions for several important quantities, including the time-dependent distribution of the number of infected individuals, the quasi-stationary distribution, the extinction time distribution of the epidemic, and the first-passage time distribution. The classical SIS model describes the dynamics of infectious diseases where recovered individuals do not develop immunity and can be re-infected after recovery. However, in densely populated areas, such as public transportation, schools, and hospitals, an infected individual may simultaneously infect multiple susceptible individuals. Such non-pairwise interaction mechanisms in disease transmission are sometimes modeled and analyzed using simplicial contagion models [32–34]. Our work extends the classical SIS model by incorporating simultaneous transmission to multiple individuals, making it more suitable for infectious diseases that spread rapidly in densely populated and crowded environments. This generalization significantly increases the mathematical complexity, since the spectra of the classical SIS model are all real, whereas those of the extended SIS model may be complex.

The paper is organized as follows. In Sec. 2, we propose an extended stochastic SIS model in densely populated environments and compare it with its deterministic counterpart. In Sec. 3, we analytically solve the master equation of the stochastic model and derive the exact time-dependent distribution of the number of infected individuals expressed in terms of the eigenvalues of the generator matrix. We find that the time evolution of the probability distribution may exhibit two distinct patterns, and the stochastic bifurcation between these patterns is examined in detail. In Sec. 4, we use the analytical time-dependent distribution to obtain the exact quasi-stationary distribution, which is then analyzed and compared with the WKB approximation. In Sec. 5, we investigate the extinction rate and extinction time of the epidemic. We find that the first nonzero eigenvalue of the generator matrix characterizes the extinction rate of the epidemic, while the second nonzero eigenvalue characterizes its outbreak rate. Additionally, we also derive the exact distribution and expected value of the extinction time and compare them with the WKB approximation. In Sec. 6, we derive the exact distribution of the first-passage time at which the number of infections reaches a given threshold, and we demonstrate how analyzing this distribution can guide infectious disease control. We conclude in Sec. 7.

2 An extended SIS model

2.1 Classical SIS model

In some infectious diseases, such as gonorrhea, chlamydia, influenza, the common cold, and respiratory syncytial virus, infection does not confer long-lasting immunity [35–37]. Individuals who recover from the disease return to the susceptible state immediately and can be infected again. Such diseases are typically modeled within the classical SIS

framework (Fig. 1(a)). Specifically, we consider a closed population of size N, where each individual is classified as either susceptible or infected. Let $N_S(t)$ and $N_I(t)$ denote the numbers of susceptible and infected individuals at time t, respectively, such that $N_S(t) + N_I(t) = N$. The densities of the two subpopulations (compartments) are then given by

$$x(t) = \frac{N_S(t)}{N}, \quad y(t) = \frac{N_I(t)}{N}.$$

To proceed, let α be the disease transmission rate, i.e. the average number of contacts per person per unit time that result in disease transmission between an infected and a susceptible individual. Since an infected individual contacts a susceptible individual with probability N_S/N, the average number of new infections per unit time is $\alpha N_I N_S/N$. In addition, let γ denote the recovery rate of an infected individual. Then γN_I represents the average number of infected individuals recovering from the disease per unit time. It was recently pointed out [38] that the above epidemic model is essentially equivalent to a chemical reaction system governed by mass-action kinetics, where random collisions between molecules result in reactions, and the reaction rate is proportional to the concentrations of the reactants. Specifically, the classical SIS model corresponds to the following chemical reaction system [39]:

$$S + I \xrightarrow{\alpha} 2I, \quad I \xrightarrow{\gamma} S,$$

where S denotes a susceptible individual and I denotes an infected individual. The first reaction represents the infection of a susceptible individual, while the second reaction represents the recovery of an infected individual.

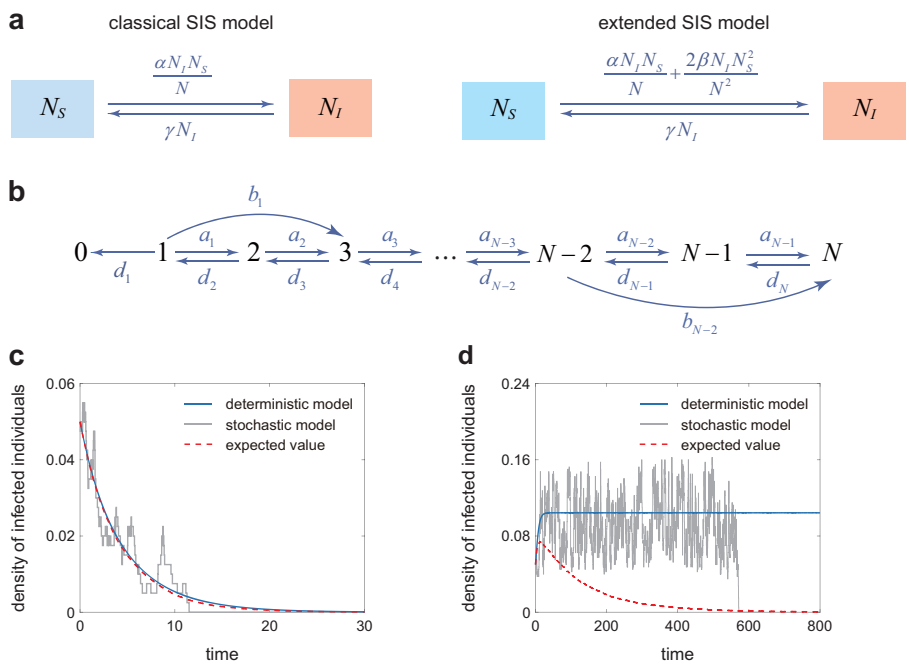

Figure 1: **Deterministic and stochastic SIS epidemic models.** **(a)** Illustration of the classical and extended deterministic SIS models. **(b)** Transition diagram of the Markovian dynamics for the stochastic (extended) SIS model. **(c)** Comparison of the trajectory of $x(t)$ for the deterministic SIS model (blue curve), the trajectory of $N_I(t)/N$ for the stochastic SIS model (grey curve), and the expected value of $N_I(t)/N$ for the stochastic SIS model (red curve) when $R_0 < 1$. The model parameters are chosen as $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 1$, $R_0 = 0.8$, and $N = 400$. **(d)** Same as in (c) but for $R_0 > 1$. The model parameters are chosen as $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 1$, $R_0 = 1.2$, and $N = 400$. In (c),(d), the deterministic trajectory is the solution to Eq. (2), the stochastic trajectory is generated using Gillespie's stochastic simulation algorithm, the expected value is computed as the sample mean of 10^4 stochastic trajectories, and the initial condition is chosen as $N_I(0) = 0.05N$.

When the total population size N is very large, the evolution of the numbers of susceptible and infected individuals is described by the deterministic SIS model

$$\begin{cases} \dfrac{\mathrm{d}N_S}{\mathrm{d}t} = -\alpha \dfrac{N_I N_S}{N} + \gamma N_I, \\ \dfrac{\mathrm{d}N_I}{\mathrm{d}t} = \alpha \dfrac{N_I N_S}{N} - \gamma N_I. \end{cases}$$

Since the total population size $N_s + N_I = N$ remains constant, the above system can be reduced to the following single-variable ODE:

$$\frac{\mathrm{d}N_I}{\mathrm{d}t} = \alpha \frac{N_I(N - N_I)}{N} - \gamma N_I,$$

which describes the evolution of the number of infected individuals. Sometimes, it is more convenient to consider the density of infected individuals, $x(t) = N_I(t)/N$, whose evolution is governed by the ODE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha x(1 - x) - \gamma x.$$

2.2 Extended SIS model

The classical SIS model assumes that each infected individual can transmit the disease to only one susceptible individual per contact. However, in densely populated settings, such as public transportation, schools, and hospitals, the number of individuals involved in each contact can increase significantly. This may result in multiple individuals being infected simultaneously during a single contact event. To capture this phenomenon, we propose an extended SIS model in which each infected individual can simultaneously contact two susceptible individuals and transmit the disease to both (Fig. 1(a)). Specifically, let β denote the rate of simultaneous infection, i.e. the average number of contacts per person per unit time that result in simultaneous transmission of the disease to two susceptible individuals. It is clear that an infected individual comes into contact with two susceptible individuals with probability

$$\frac{\binom{N_S}{2}}{\binom{N}{2}} = \frac{N_S(N_S - 1)}{N(N - 1)}.$$

Hence the average number of new infections resulting from simultaneous transmission per unit time is given by

$$\frac{2\beta N_I N_S(N_S - 1)}{N(N - 1)}.$$

Similarly, the extended SIS model is equivalent to the chemical reaction system

$$S + I \xrightarrow{\alpha} 2I, \quad 2S + I \xrightarrow{\beta} 3I, \quad I \xrightarrow{\gamma} S, \tag{1}$$

where the first reaction represents disease transmission to a single susceptible individual, the second reaction describes simultaneous transmission to two susceptible individuals in crowded environments, and the third corresponds to the recovery of an infected individual. Note that the extended SIS model reduces to the classical one when $\beta = 0$. Similar reaction system has also be used to model the regulation of stem cell dynamics [40].

When the total population size N is very large, the evolution of the numbers of susceptible and infected individuals is described by the following system of deterministic ODEs:

$$\begin{cases} \dfrac{\mathrm{d}N_S}{\mathrm{d}t} = -\alpha \dfrac{N_I N_S}{N} - 2\beta \dfrac{N_I N_S^2}{N^2} + \gamma N_I, \\ \dfrac{\mathrm{d}N_I}{\mathrm{d}t} = \alpha \dfrac{N_I N_S}{N} + 2\beta \dfrac{N_I N_S^2}{N^2} - \gamma N_I, \end{cases}$$

where we have used the approximations of $N_s(N_s - 1) \approx N_s^2$ and $N(N - 1) \approx N^2$ when $N \gg 1$. Since the total population size $N_s + N_I = N$ is conserved, the evolution of the density of infected individuals, $x(t) = N_I(t)/N$, is governed by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha x(1 - x) + 2\beta x(1 - x)^2 - \gamma x. \tag{2}$$

Clearly, the interval $[0, 1]$ is positively invariant with respect to this system, and the system must have a disease-free equilibrium $x = 0$. Using the next generation matrix approach [41], it is easy to see that the basic reproduction number of the extended SIS model is given by

$$R_0 = \frac{\alpha + 2\beta}{\gamma}.$$

When $R_0 \leq 1$, the disease-free equilibrium globally asymptotically stable within $[0, 1]$; when $R_0 > 1$, the disease-free equilibrium is unstable and the system also has a positive endemic equilibrium

$$x^* = \begin{cases} 1 - \dfrac{\sqrt{\alpha^2 + 8\beta\gamma} - \alpha}{4\beta}, & \beta \neq 0, \\ 1 - \dfrac{\gamma}{\alpha}, & \beta = 0, \end{cases}$$

which is globally asymptotically stable within $(0, 1]$. In what follows, unless otherwise stated, we always focus on the extended SIS model since it provides a more accurate description of how infectious diseases spread in densely populated environments than the classical SIS model.

2.3 Stochastic SIS model

When the total population size N is relatively small, stochasticity cannot be ignored and it is necessary to consider a stochastic epidemic model where the numbers of susceptible and infected individuals take integer values rather than real values as in the deterministic model. When stochastically modeled, the evolution of the extended SIS system is governed by a Markov jump process $(N_S(t), N_I(t))_{t\geq 0}$ with state space $\{(m, n) \in \mathbb{N}^2 | m + n = N\}$, where m and n denote the numbers of susceptible and infected individuals, respectively. Clearly, the Markovian model can only make the following three types of transitions:

$$(m, n) \xrightarrow{\alpha nm/N} (m - 1, n + 1)$$

due to disease transmission to a single susceptible individual,

$$(m, n) \xrightarrow{\beta nm(m-1)/(N(N-1))} (m - 2, n + 2)$$

due to simultaneous transmission to two susceptible individuals, and

$$(m, n) \xrightarrow{\gamma n} (m + 1, n - 1)$$

due to recovery of an infected individual. Note that here the transition rates for the three events are exactly the same as those in the deterministic model (Fig. 1(a)). Since $N_S(t) + N_I(t) = N$, we only need to consider the dynamics of the single-variable process $N_I(t)$, which is a Markov jump process with state space $S = \{0, 1, \ldots, N\}$. Similarly, the process $N_I(t)$ can make the following three types of transitions:

$$n \xrightarrow{a_n} n + 1, \quad n \xrightarrow{b_n} n + 2, \quad n \xrightarrow{d_n} n - 1,$$

with transition rates

$$a_n = \frac{\alpha n(N - n)}{N}, \quad b_n = \frac{\beta n(N - n)(N - n - 1)}{N(N - 1)}, \quad d_n = \gamma n.$$

The transition diagram of the process $N_I(t)$ is shown in Fig. 1(b). Clearly, we have $a_0 = b_0 = d_0 = 0$ and thus $0 \in S$ is an absorbing state. The other states $1, \ldots, N$ are all non-absorbing and form a communicating class [42]. Let $P_n(t)$ denote the probability of having n infected individuals at time t. Then the evolution of the Markovian model is governed by the Kolmogorov forward equation (also known as the master equation)

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = a_{n-1}P_{n-1}(t) + b_{n-2}P_{n-2}(t) + d_{n+1}P_{n+1}(t) - q_nP_n(t), \tag{3}$$

where $q_n = a_n + b_n + d_n$ denotes the total transition rate from state n to all other states and $P_n(t) \equiv 0$ by default for any $n \notin S$. Let $P(t) = (P_0(t), \ldots, P_N(t))$ denote the probability distribution of the number of infected individuals at time t. Then the master equation can be written in matrix form as

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = P(t)Q, \tag{4}$$

where

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ d_1 & -q_1 & a_1 & b_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & -q_2 & a_2 & b_2 & \cdots & 0 & 0 \\ 0 & 0 & d_3 & -q_3 & a_3 & \cdots & 0 & 0 \\ 0 & 0 & 0 & d_4 & -q_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -q_{N-1} & a_{N-1} \\ 0 & 0 & 0 & 0 & 0 & \cdots & d_N & -q_N \end{pmatrix} \tag{5}$$

is the generator matrix of the Markovian model. Since $0 \in S$ is an absorbing state, the first row of Q is the zero vector.

The relationship between the deterministic ODE model and the stochastic CTMC model was rigorously proved by Kurtz [17, 18]. Let $N_I(t)$ denote the number of infected individuals for the stochastic model and $x(t)$ denote the density of infected individuals for the deterministic model. Kurtz's theorem states that for any $T > 0$,

$$\lim_{N\to\infty} \sup_{t\le T} \left| \frac{N_I(t)}{N} - x(t) \right| = 0, \quad \text{a.s.}$$

The theorem states that, in the limit of large population size, the density of the stochastic model converges to that of the deterministic model over any finite time interval with probability one. We emphasize that although the deterministic and stochastic models exhibit similar dynamic behavior over finite time periods, their long-term behavior can be fundamentally different. Since the Markov jump process $N_I(t)$ has a finite state space with $0 \in S$ being the unique absorbing state, the ergodic theorem implies that the disease will eventually go extinct with probability one for the stochastic SIS model. This is in sharp contrast to the predictions of the deterministic model, in which the density of infected individuals converges to a positive endemic equilibrium x^* when $R_0 > 1$.

Fig. 1(c),(d) compare the time evolution of the density of infected individuals $x(t)$ for the deterministic SIS model, the time evolution of the density of infected individuals $y(t) = N_I(t)/N$ for the stochastic SIS model, and the time evolution of the expected value of $y(t)$ for the stochastic SIS model. Clearly, the two models have similar dynamic behaviors when $R_0 < 1$. However, the long-term behaviors for the two models are totally different when $R_0 > 1$ — the deterministic trajectory tends to the positive endemic equilibrium x^*, while the stochastic trajectory first fluctuates around the deterministic trajectory over a finite period of time, and eventually becomes zero due to extinction of the disease. Moreover, the expected value $\mathbb{E}[y(t)]$ of the density of infected individuals tends to zero for the stochastic model, which is inconsistent with the predictions of the deterministic model (Fig. 1(d)).

3 Time-dependent distribution

3.1 Exact solution

We next focus on the time evolution of the stochastic SIS model by computing the probability distribution of the number of infected individuals at each time. Understanding the time evolution of the distribution provides insight into how the disease spreads. Following the method proposed in [43], we next solve the master equation in time and thus obtain the time-dependent distribution $P_n(t)$ of infected individuals. The solution to Eq. (4) can be represented using matrix exponential as

$$P(t) = P(0)e^{Qt},$$

where $P(0) = (P_0(0), \ldots, P_N(0))$ denotes the initial distribution of infected individuals. In general, it is difficult to compute the matrix exponential exactly since we need to calculate the eigenvalues and eigenvectors of the generator matrix Q. However, the special structure of Q allows us to bypass the eigenvector calculation. By Cauchy's integral formula for matrices [44], for any continuous function f, we have

$$f(Q) = \frac{1}{2\pi i} \oint_C f(z)(zI - Q)^{-1} \mathrm{d}z,$$

where C is an arbitrary simple closed curve in the complex plane that contains all the eigenvalues of Q in its interior. If we take $f(z) = P(0)e^{zt}$, then we obtain

$$P(t) = \frac{1}{2\pi i} \oint_C P(0)(zI - Q)^{-1} e^{zt} \mathrm{d}z. \tag{6}$$

Suppose that the initial number of infected individuals is $n_0 \in S$; then the initial distribution is given by $P_n(0) = \delta_{n_0}(n)$, where $\delta_{n_0}(n)$ a Kronecker delta that takes the value of 1 when $n = n_0$ and the value of 0 otherwise. We will explain later how to extend our results to more general initial distributions. Under the point initial distribution, Eq. (6) can be simplified as

$$P_n(t) = \frac{1}{2\pi i} \oint_C (zI - Q)^{-1}_{n_0+1,n+1} e^{zt} \mathrm{d}z, \tag{7}$$

where $(zI - Q)^{-1}_{i,j}$ denotes the element of the matrix $(zI - Q)^{-1}$ in the ith row and the jth column. By Cramer's rule of calculating the inverse matrix, we can prove that

$$(zI - Q)^{-1}_{n_0+1,n+1} = \frac{1}{u_{N+1}(z)} \begin{cases} h_{n_0,n}(z), & n_0 < n, \\ u_n(z)v_n(z), & n_0 = n, \\ d_{n+1} \dots d_{n_0} u_n(z) v_{n_0}(z), & n_0 > n, \end{cases} \tag{8}$$

where $u_n(z)$, $0 \le n \le N+1$, are polynomials defined recursively by

$$\begin{gathered} u_n(z) = (z + q_{n-1})u_{n-1}(z) - a_{n-2}d_{n-1}u_{n-2}(z) - b_{n-3}d_{n-1}d_{n-2}u_{n-3}(z), \\ u_0(z) = 1, \quad u_1(z) = z, \quad u_2(z) = (z + q_1)u_1(z), \end{gathered} \tag{9}$$

with $u_{N+1} = \det(zI - Q)$ being the characteristic polynomial of Q, $v_n(z)$, $0 \le n \le N$, are polynomials defined recursively by

$$\begin{gathered} v_n(z) = (z + q_{n+1})v_{n+1}(z) - a_{n+1}d_{n+2}v_{n+2}(z) - b_{n+1}d_{n+3}d_{n+2}v_{n+3}(z), \\ v_N(z) = 1, \quad v_{N-1}(z) = z + q_N, \quad v_{N-2}(z) = (z + q_{N-1})v_{N-1}(z) - a_{N-1}d_N, \end{gathered}$$

and $h_{n_0,n}(z)$ are functions defined as

$$\begin{aligned} h_{n_0,n}(z) = {} & \delta_1(n - n_0)[b_{n-2}d_{n-1}v_n(z)u_{n_0-1}(z)] \\ & + (-1)^{n-n_0+1}\big\{[a_{n-1}v_n(z) + d_{n+1}b_{n-1}v_{n+1}(z)]w_{n_0,n-n_0}(z) \\ & - b_{n-2}(z + q_{n-1})v_n(z)w_{n_0,n-n_0-1}(z) + b_{n-2}b_{n-3}d_{n-1}v_n(z)w_{n_0,n-n_0-2}(z)\big\}. \end{aligned}$$

Here $w_{n_0,i}(z)$, $1 \le i \le n - n_0$, are another sequence of polynomials defined recursively by

$$\begin{gathered} w_{n_0,i}(z) = -a_{n_0+i-2}w_{n_0,i-1}(z) + b_{n_0+i-3}(z + q_{n_0+i-2})w_{n_0,i-2}(z) - b_{n_0+i-3}b_{n_0+i-4}d_{n_0+i-2}w_{n_0,i-3}(z), \\ w_{n_0,1}(z) = u_{n_0}(z), \quad w_{n_0,2}(z) = -a_{n_0}w_{n_0,1}(z) - b_{n_0-1}d_{n_0}u_{n_0-1}(z), \\ w_{n_0,3}(z) = -a_{n_0+1}w_{n_0,2}(z) + b_{n_0}(z + q_{n_0+1})w_{n_0,1}(z). \end{gathered}$$

Combining Eqs. (7) and (8), we obtain

$$P_n(t) = \frac{1}{2\pi i} \oint_C \frac{g_{n_0,n}(z)}{u_{N+1}(z)} e^{zt} \mathrm{d}z,$$

where $g_{n_0,n}(z)$ are polynomials defined as

$$g_{n_0,n}(z) = \begin{cases} h_{n_0,n}(z), & n_0 < n, \\ u_n(z)v_n(z), & n_0 = n, \\ d_{n+1}d_{n+2} \dots d_{n_0} u_n(z) v_{n_0}(z), & n_0 > n. \end{cases}$$

Since $u_{N+1}(z)$ is the characteristic polynomial of Q, we have

$$u_{N+1}(z) = (z-\lambda_0)^{r_0} \dots (z-\lambda_l)^{r_l},$$

where $\lambda_0, \dots, \lambda_l$ are all pairwise distinct eigenvalues of Q with $r_0, \dots, r_l$ being their multiplicities, respectively. We can then apply Cauchy's residue theorem

$$\oint_C f(z)\mathrm{d}z = 2\pi i \sum_k \mathrm{Res}(f;\varepsilon_k),$$

where ε_k are all isolated singularities of f inside the simple closed curve C. In our current case, all the isolated singularities are the eigenvalues λ_k and thus

$$\oint_C \frac{g_{n_0,n}(z)}{u_{N+1}(z)} e^{zt}\mathrm{d}z = 2\pi i \sum_{k=0}^{l} \frac{1}{(r_k-1)!} \frac{\mathrm{d}^{r_k-1}}{\mathrm{d}z^{r_k-1}} \left. \frac{g_{n_0 n}(z)e^{zt}}{\prod_{j\neq k}(z-\lambda_j)^{r_j}} \right|_{z=\lambda_k}.$$

Therefore, once we have known all the eigenvalues of Q, the time-dependent solution of Eq. (3), i.e. the time-dependent distribution of infected individuals, is given by

$$P_n(t) = \sum_{k=0}^{l} \frac{1}{(r_k-1)!} \frac{\mathrm{d}^{r_k-1}}{\mathrm{d}z^{r_k-1}} \left. \frac{g_{n_0,n}(z)e^{zt}}{\prod_{j\neq k}(z-\lambda_j)^{r_j}} \right|_{z=\lambda_k}.$$

In most cases, the eigenvalues of Q are mutually different (in fact, any matrix can be approximated by matrices with non-degenerate eigenvalues to any degree of accuracy). In this case, we have $l = N$ and $r_0 = \dots = r_l = 1$, and thus the time-dependent distribution can be simplified as

$$P_n(t) = \sum_{k=0}^{N} \frac{g_{n_0,n}(\lambda_k)e^{\lambda_k t}}{\prod_{j\neq k}(\lambda_k-\lambda_j)}, \tag{10}$$

where $\lambda_0, \dots, \lambda_N$ are all the eigenvalues of Q, i.e. all the zeros of the characteristic polynomial $\det(zI-Q)$.

Thus far, we have obtained the exact time-dependent solution of the master equation for the stochastic SIS model under a point initial distribution. In the general case that the system starts from a general initial distribution $p_n(0) = \pi_n$, time-dependent distribution is given by

$$P_n(t) = \sum_{n_0=0}^{N} \pi_{n_0} P_n(t|n_0),$$

where $P_n(t|n_0)$ is the time-dependent distribution when the system starts from the fixed state n_0.

As a check of our exact time-dependent solution, we compare it with the numerical one obtained by solving the master equation Eq. (4) directly using the MATLAB function "ODE45" (Fig. 2). As expected, the two solutions agree perfectly at all times. Interestingly, we observe that the stochastic SIS model can produce two different patterns of time evolution. When R_0 is small, the distribution of infected individuals is unimodal at all times (Fig. 2(a)); when R_0 is large, the distribution is unimodal at small and large times, and becomes bimodal with both a zero and a nonzero peak at intermediate times (Fig. 2(b)). The two modes of the bimodal distribution correspond to the disease-free and endemic equilibria, respectively.

In practice, the distribution of the number of infected individuals can be measured through parallel observations across multiple regions. Specifically, suppose that the transmission range of the disease can be divided into several sub-regions of similar size and with comparable epidemiological characteristics, such as different cities or communities implementing identical containment strategies. The numbers of infected individuals in these sub-regions can be approximately regarded as statistically independent and identically distributed samples, and thus empirical distributions of infected individuals at multiple discrete time points can be statistically obtained. The exact time-dependent distribution derived in this paper can then be used to construct a likelihood function, allowing all model parameters to be inferred via the maximum likelihood method [16, 45]. In this way, our analytical results enable the integration of observed data with the theoretical model to achieve accurate parameter inference.

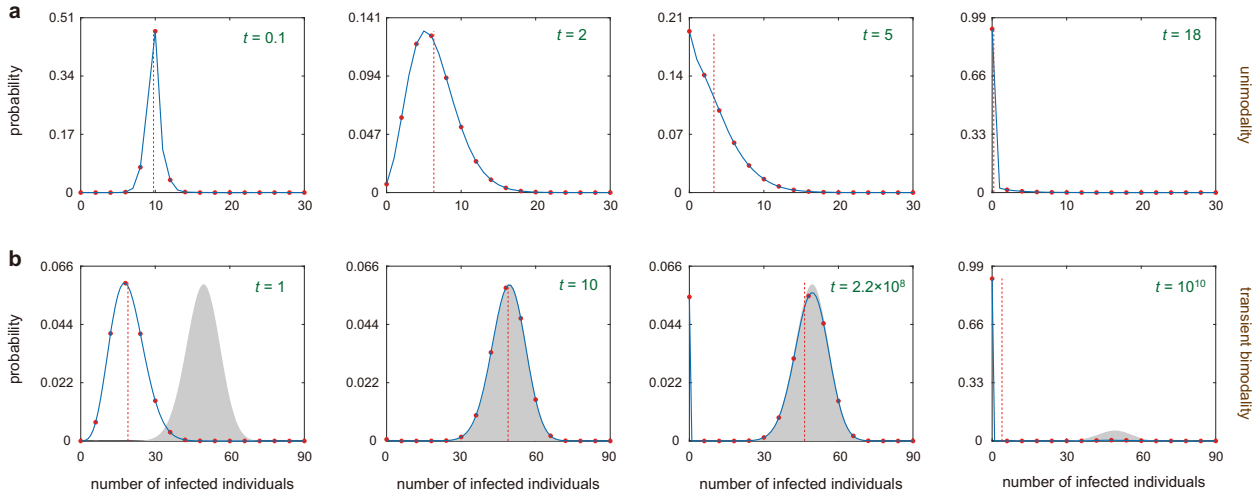

Figure 2: **Two distinct patterns of time evolution for the stochastic SIS model. (a)** When R_0 is small, the distribution of infected individuals is unimodal at small times. The blue curve shows the exact time-dependent distribution given by Eq. (10), the red circles show the numerical one obtained from stochastic simulations, and the red dashed vertical line shows the mean number of infected individuals. The model parameters are chosen as $\alpha = 0.3$, $\beta = 0.06$, $\gamma = 0.6$, $R_0 = 0.7$, and $N = 100$. **(b)** When R_0 is large, the distribution of infected individuals becomes bimodal with both a zero and a non-zero peak at intermediate times and then reverts back to being unimodal at large times. Prior to disease extinction, the distribution can be approximately stationary. The grey region shows the quasi-stationary distribution of infected individuals computed using Eq. (19). The model parameters are chosen as $\alpha = 0.8$, $\beta = 0.4$, $\gamma = 0.6$, $R_0 = 2.67$, and $N = 100$. In (a),(b), the initial condition is chosen as $N_I(0) = 0.1N$.

3.2 Convergence to the extinction state

We have seen that the disease must eventually go extinct for the stochastic SIS model since the state space is finite with $0 \in S$ being the unique absorbing state. Here we will validate this fact using the exact time-dependent solution. By the Perron-Frobenius theorem [46], when the system is ergodic, the generator matrix Q must have zero as its eigenvalue with multiplicity one and all other eigenvalues must have negative real parts, i.e.

$$0 = \lambda_0 > \mathrm{Re}(\lambda_1) \geq \mathrm{Re}(\lambda_2) \geq \cdots \geq \mathrm{Re}(\lambda_l).$$

In fact, we can prove a stronger result. Let $\tilde{Q}$ be the matrix obtained from Q by removing the first row and the first column. Since the first row of Q is the zero vector, all eigenvalues of $\tilde{Q}$ are exactly all nonzero eigenvalues of Q. It has been proved in [47] that if all non-absorbing states form a communicating class (this is automatically satisfied for the stochastic SIS model), then the eigenvalue of $\tilde{Q}$ with the maximal real part, i.e. the eigenvalue λ_1, must be real and has multiplicity one. This indicates that all the eigenvalues of Q can be arranged so that

$$0 = \lambda_0 > \lambda_1 > \mathrm{Re}(\lambda_2) \geq \cdots \geq \mathrm{Re}(\lambda_l).$$

Therefore, the only term in Eq. (10) independent of time t is the first term and all other terms tend to zero exponentially fast as $t \to \infty$. Since the system is ergodic, the stationary solution of the master equation is independent of the choice of the initial state. As a result, we can take $n_0 = N$, allowing us to focus on to the case of $n_0 \geq n$, which is now

$$\lim_{t\to\infty} P_n(t) = \frac{d_{n+1} \cdots d_N u_n(0)}{\prod_{j\neq 0}(-\lambda_j)},$$

where we have used the fact that $v_N(0) = 1$. Since the term $1/\prod_{j\neq 0}(-\lambda_j)$ contributes the same to each term, we can treat it as a normalizing factor. We can also do the same with $d_{n+1} \cdots d_N$, which contributes the same as $(d_1 \cdots d_n)^{-1}$ up to normalization. Hence the stationary solution has the following simplified expression:

$$\lim_{t\to\infty} P_n(t) = K \frac{u_n(0)}{d_1 \cdots d_n} = K \frac{u_n(0)}{\gamma^n n!},$$

where K is a normalization constant. By induction, it is easy to see that $u_n(0) = \delta_0(n)$, where $\delta_0(n)$ is a Kronecker delta which takes the value of 1 when $n = 0$ and takes the value of 0 otherwise. Since $u_n(0) = 0$ for any $n \geq 1$, we finally obtain

$$\lim_{t\to\infty} P_n(t) = \delta_0(n) = \begin{cases} 1, & n = 0, \\ 0, & n = 1, \cdots, N. \end{cases}$$

This demonstrates that the stationary distribution of the stochastic SIS model is the point distribution at $n = 0$, corresponding to the extinction state, and the time-dependent distribution converges to this point distribution in the long-term limit.

We emphasize that there is an remarkable difference between the classical and extended SIS models. For the classical SIS model ($\beta = 0$), the Markovian model shown in Fig. 1(b) reduces to a birth-death process and the generator matrix Q given in Eq. (5) reduces to a tridiagonal matrix. In this case, it is well known that the eigenvalues of Q are all real numbers [48]. However, for the extended SIS model ($\beta \neq 0$), the first two eigenvalues, λ_0 and λ_1, are always real, while all other eigenvalues can be complex numbers. To validate this, we compute the first eight eigenvalues of Q under different choices of β (Table 1). Clearly, all the eigenvalues are real when $\beta = 0$, as expected. The number of complex eigenvalues increases with the increase of β, while the first two eigenvalues are always real.

eigenvalues	R_0	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
$\beta = 0$	0.1	0	-0.92	-1.90	-2.93	-4.02	-5.15	-6.33	-7.56
$\beta = 0.5$	1	0	-0.52	-1.89	-3.45	-5.17	-5.83	-6.80+1.78**i**	-6.80-1.78**i**
$\beta = 12$	24	0	-0.002	-6.61	-9.07	-11.36+3.42**i**	-11.36-3.42**i**	-15.20+6.53**i**	-15.20-6.53**i**
$\beta = 24$	48	0	-0.0002	-12.32+1.20**i**	-12.32-1.20**i**	-16.60+2.28**i**	-16.60-2.28**i**	-25.48+6.64**i**	-25.48-6.64**i**

Table 1: **The first eight eigenvalues of the generator matrix Q for the stochastic SIS model as β increases.** The model parameters are chosen as $\alpha = 0.1$, $\gamma = 1$, and $N = 7$.

3.3 Dynamical phase diagram and stochastic bifurcation

Recall that the deterministic SIS model has only one equilibrium (disease-free equilibrium) when $R_0 \leq 1$ and has two equilibria (disease-free and endemic equilibria) when $R_0 > 1$. A bifurcation occurs at $R_0 = 1$. Actually, the bifurcation for stochastic epidemic models has also been widely studied using the technique of branching process approximation [23, 24]. It is well known that during the early stage of the disease, the stochastic epidemic model can be approximated by a branching process. This approximation also reveals a bifurcation threshold at $R_0 = 1$ — the disease undergoes a minor outbreak when $R_0 \leq 1$, whereas a major outbreak becomes possible when $R_0 > 1$.

We emphasize that the branching process approximation is only valid during the early stage of the disease. Here we revisit the stochastic bifurcation for the stochastic SIS model by analyzing the time evolution of its probability distribution without relying on any approximations (similar stochastic bifurcation analysis has been performed in stochastic gene networks [49] as well as in stochastic enzyme kinetics [50]). In general, an equilibrium of the deterministic model corresponds to a mode of the probability distribution of the stochastic model. Intuitively, if the deterministic model has only one equilibrium, then the stochastic model should have unimodal distributions at all time points; if the deterministic has two equilibria, then the stochastic model may have bimodal distributions at certain time points — the two modes of the bimodal distribution correspond to the two equilibria of the deterministic model.

We next investigate the shape of the time-dependent distribution for the stochastic SIS model. Following previous studies [23, 24], we assume that initially there is only one infected individual, i.e. $N_I(0) = 1$. According to simulations, we find that the stochastic SIS model can only produce a unimodal distribution with only one peak or a bimodal distribution with both a zero and a nonzero peak. A distribution with more than two peaks is not observed. We stress that the time-dependent distribution of the number of infected individuals, $P_n(t)$, must be unimodal when t is very small or when t is very large. The distribution is unimodal at small times because the initial distribution of the system is the point distribution at $n = 1$, which is unimodal. The distribution is unimodal at large times because the stationary distribution is the point distribution at $n = 0$ (the extinction state), which is also unimodal.

To further understand the shape of the time-dependent distribution, we classify the dynamic behavior of the stochastic SIS model into two different phases: (i) the distribution remains unimodal at all times (Fig. 2(a)); (ii) the distribution is unimodal at small and large times and is bimodal at intermediate times (Fig. 2(b)). Roughly speaking, the two modes of the bimodal distribution correspond to the disease-free and endemic equilibria of the deterministic model, respectively. To distinguish between them, we refer to (i) as unimodality and (ii) as transient bimodality. Note that unimodality corresponds to a minor outbreak of the disease and bimodality corresponds to a major outbreak.

To determine the regions of the two phases in the parameter space, we illustrate the $\beta/\alpha - R_0$ phase diagram of the stochastic SIS model in Fig. 3 under three different choices of the total population size N. Here β/α is the ratio of the simultaneous infection rate to the normal infection rate and R_0 is the basic reproduction number. Intriguingly, we find that the system undergoes a stochastic bifurcation from unimodality to transient bimodality as R_0 increases; the bifurcation threshold of R_0 always exceeds 1 and increases as β/α increases. Note that this is different from the predictions of the branching process approximation, where $R_0 = 1$ is the bifurcation point. From Fig. 3(a)-(c), we can see that the bifurcation threshold of R_0 decreases and approaches 1 with the increase of N. The indicates that the stochastic bifurcation reduces to the deterministic bifurcation when the population size is very large. Our results also suggest that deterministic models of epidemic dynamics may underestimate the threshold for disease outbreak when the population size is relatively small.

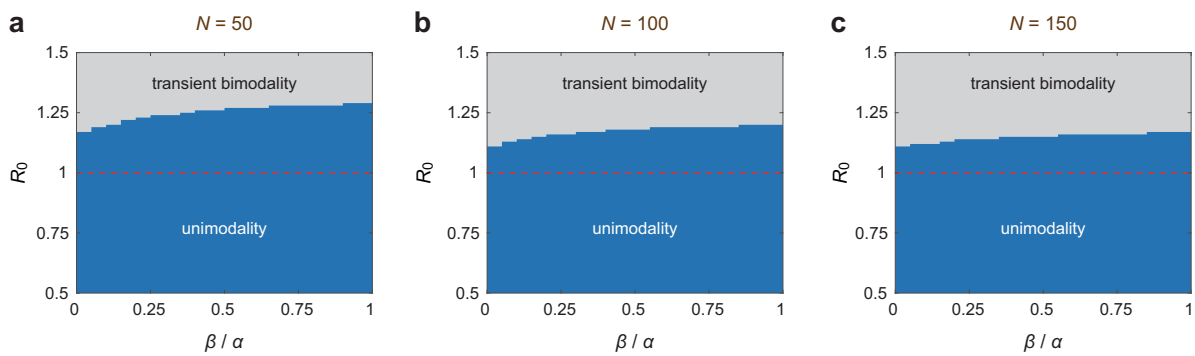

Figure 3: **Dynamical phase diagrams for the stochastic SIS model in the $\beta/\alpha - R_0$ plane as the total population size size N increases.** The system can produce two different patterns of time evolution: (i) unimodality (blue region) where the distribution is unimodal at small times, and (ii) transient bimodality (grey region) where the distribution becomes bimodal at intermediate times and then reverts back to being unimodal at long times. **(a)** Case of $N = 50$. **(b)** Case of $N = 100$. **(c)** Case of $N = 150$. In (a)-(c), the model parameters are chosen as $\gamma = 0.6$ and the initial condition is chosen as $N_I(0) = 1$.

4 Quasi-stationary distribution

4.1 Exact solution

Recall that when $R_0 > 1$, the deterministic SIS model has an unstable disease-free equilibrium and a positive endemic equilibrium, which is globally asymptotically stable. According to Kurtz's theorem, when the total population size N is large, the trajectory of the stochastic SIS model fluctuates around the endemic equilibrium x^* for an extended period before eventually leading to extinction over a much longer timescale (Fig. 1(d)). Prior to disease extinction, the probability distribution of the stochastic SIS model can be approximately stationary. This approximately stationary distribution is widely known as the quasi-stationary (or metastable) distribution [31, 47].

Mathematically, the quasi-stationary distribution is defined as the stationary distribution of the stochastic SIS model conditioned on non-extinction. Note that, given that the disease has not gone extinct at time t, the probability distribution of the number of infected individuals at time t is given by

$$\tilde{P}_n(t) = \frac{P_n(t)}{1 - P_0(t)}, \quad n = 1, \cdots, N. \tag{11}$$

Taking derivatives on both sides yields

$$\frac{\mathrm{d}\tilde{P}_n(t)}{\mathrm{d}t} = \frac{\mathrm{d}P_n(t)}{\mathrm{d}t}\frac{1}{1-P_0(t)} + \frac{\tilde{P}_n(t)}{1-P_0(t)}\frac{\mathrm{d}P_0(t)}{\mathrm{d}t}. \tag{12}$$

For convenience, we write the generator matrix as $Q = (q_{mn})_{0\le m,n\le N}$, where q_{mn}, $m \neq n$ is the transition rate from state m to state n and $q_{mm} = -\sum_{m\neq n} q_{mn}$. It then follows from Eq. (4) that

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \sum_{m=0}^{N} P_m(t)q_{mn} = \sum_{m=1}^{N} P_m(t)q_{mn}, \tag{13}$$

where we have used the fact that $q_{0n} = 0$ since the first row of Q is the zero vector. Moreover, applying Eq. (4) yields

$$\frac{\mathrm{d}P_0(t)}{\mathrm{d}t} = \gamma P_1(t). \tag{14}$$

Inserting Eqs. (13) and (14) into Eq. (12), we obtain

$$\frac{\mathrm{d}\tilde{P}_n(t)}{\mathrm{d}t} = \sum_{m=1}^{N} \tilde{P}_m(t)q_{mn} + \gamma\tilde{P}_1(t)\tilde{P}_n(t).$$

Therefore, conditioned on non-extinction, the evolution of the distribution $\tilde{P}(t) = (\tilde{P}_1(t), \cdots, \tilde{P}_N(t))$ of infected individuals is governed by

$$\frac{\mathrm{d}\tilde{P}(t)}{\mathrm{d}t} = \tilde{P}(t)\tilde{Q} + \gamma\tilde{P}_1(t)\tilde{P}(t),$$

where $\tilde{Q}$ is the matrix obtained from Q by deleting the first row and the first column. It has been proved [47] that this system has a unique stationary solution

$$\tilde{P}^{ss} = (\tilde{P}_1^{ss}, \ldots, \tilde{P}_N^{ss}) = \lim_{t\to\infty} \tilde{P}(t), \tag{15}$$

which is referred to as the quasi-stationary distribution of the stochastic SIS model. Clearly, the quasi-stationary distribution $\tilde{P}^{ss}$ must satisfy

$$\tilde{P}^{ss}\tilde{Q} = -\gamma\tilde{P}_1^{ss}\tilde{P}^{ss}. \tag{16}$$

Note that Eq. (16) is difficult to solve since it is nonlinear. We now use the exact time-dependent distribution obtained previously to compute the quasi-stationary distribution. Note that the convergence in Eq. (15) is independent of the choice of the initial state. Hence we can take $n_0 = N$, allowing us to focus on to the case of $n_0 \ge n$. It then follows from Eq. (10) that

$$P_n(t) = \sum_{k=0}^{N} \frac{d_{n+1}\ldots d_N u_n(\lambda_k)e^{\lambda_k t}}{\prod_{j\neq k}(\lambda_k - \lambda_j)}. \tag{17}$$

In particular, taking $n = 0$ in the above equation yields

$$P_0(t) = \sum_{k=0}^{N} \frac{d_1\ldots d_N e^{\lambda_k t}}{\prod_{j\neq k}(\lambda_k - \lambda_j)}. \tag{18}$$

Inserting Eqs. (17) and (18) into Eq. (11), we obtain the exact time-dependent distribution of infected individuals conditioned on non-extinction:

$$\tilde{P}_n(t) = \frac{P_n(t)}{1-P_0(t)} = \frac{\sum_{k=0}^{N} \frac{d_{n+1}\ldots d_N u_n(\lambda_k)e^{\lambda_k t}}{\prod_{j\neq k}(\lambda_k-\lambda_j)}}{1-\sum_{k=0}^{N} \frac{d_1\ldots d_N e^{\lambda_k t}}{\prod_{j\neq k}(\lambda_k-\lambda_j)}}.$$

It then follows from Eq. (15) that

$$\begin{aligned}\tilde{P}_n^{ss} &= \lim_{t\to\infty} \tilde{P}_n(t) \\ &= \lim_{t\to\infty} \frac{\lambda_1 \frac{d_{n+1}\dots d_N u_n(\lambda_1)}{\prod_{j\neq 1}(\lambda_1-\lambda_j)} + \sum_{k=2}^{N} \lambda_k \frac{d_{n+1}\dots d_N u_n(\lambda_k) e^{(\lambda_k-\lambda_1)t}}{\prod_{j\neq k}(\lambda_k-\lambda_j)}}{-\lambda_1 \frac{d_1\dots d_N}{\prod_{j\neq 1}(\lambda_1-\lambda_j)} - \sum_{k=2}^{N} \lambda_k \frac{d_1\dots d_N e^{(\lambda_k-\lambda_1)t}}{\prod_{j\neq k}(\lambda_k-\lambda_j)}} = -\frac{u_n(\lambda_1)}{d_1\dots d_n},\end{aligned} \tag{19}$$

where we have used the fact that $\lambda_1 > \mathrm{Re}(\lambda_2)\dots \geq \mathrm{Re}(\lambda_N)$. This gives the explicit expression of the quasi-stationary distribution of infected individuals. In particular, it follows from Eq. (19) that

$$\tilde{P}_1^{ss} = -\frac{u_1(\lambda_1)}{d_1} = -\frac{\lambda_1}{\gamma}.$$

This equation, together with Eq. (16), yields

$$\tilde{P}^{ss}\tilde{Q} = \lambda_1 \tilde{P}^{ss}. \tag{20}$$

This clearly demonstrates that the quasi-stationary distribution is exactly the left eigenvector of $\tilde{Q}$ associated with the eigenvalue λ_1. From Fig. 2(b), we can see that when R_0 is large, the time-dependent distribution of infected individuals initially approaches the quasi-stationary distribution at a relatively short timescale, then becomes bimodal, and eventually leads to extinction at a much longer timescale.

4.2 Stochastic bifurcation

Fig. 4(a) shows the quasi-stationary distribution for the stochastic SIS model computed using Eq. (19). Interestingly, we find that the quasi-stationary distribution may have two different shapes. When β is small, the quasi-stationary distribution is always unimodal (red solid curve). However, when β is large, the distribution may be bimodal with a peak at $n = 1$ and the other peak at $n > 1$ (blue solid curve).

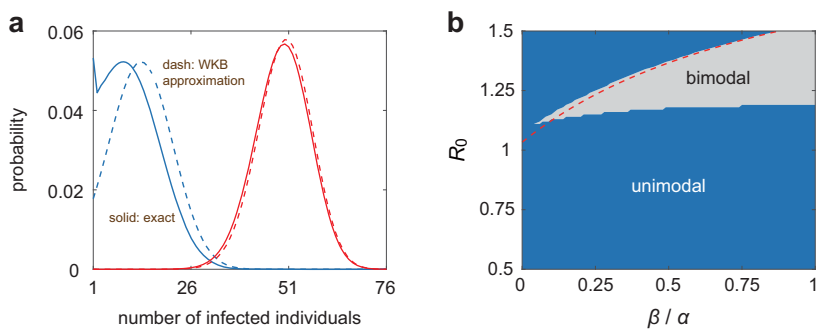

Figure 4: **Quasi-stationary distribution of infected individuals for the stochastic SIS model. (a)** Comparison between the exact quasi-stationary distribution (blue and red solid curves) given by Eq. (19) and the WKB approximation (blue and red dashed curves) computed using Eq. (21). The model parameters are chosen as $\alpha = 0.12$, $\beta = 0.33$, $\gamma = 0.6$, $R_0 = 1.3$, and $N = 100$ for the blue curves, and are chosen as $\alpha = 1.1$, $\beta = 0.11$, $\gamma = 0.6$, $R_0 = 2.2$, and $N = 100$ for the red curves. **(b)** Phase diagrams for the quasi-stationary distribution in the $\beta/\alpha - R_0$ plane. The system can produce two different shapes of quasi-stationary distributions: unimodal (blue region) and bimodal (grey region). The red dashed line shows the graph of the curve $R_0 = (1 + 2\beta/\alpha)/(1 + \beta/\alpha)$. Theoretically, it can be proved that the bimodal region is (approximately) below this red dashed line. The model parameters are chosen as $\gamma = 0.6$ and $N = 100$.

To determine the regions for the two distinct shapes of quasi-stationary distributions in the parameter space, we illustrate the $\beta/\alpha - R_0$ phase diagram in Fig. 4(b). It is clear that the quasi-stationary distribution is always unimodal for the classical SIS model ($\beta = 0$). When β is relatively large, we find that the extended SIS model may display a bimodal distribution when R_0 is neither too large nor too small. Recall that when R_0 is very large, the system produces transient bimodality (Fig. 3(b)) and hence the distribution of infected individuals must be bimodal at at least one time

point. However, the quasi-stationary distribution is always unimodal when R_0 is very large. This is a crucial difference between the quasi-stationary distribution and the time-dependent distribution.

To explain why the quasi-stationary distribution is always unimodal when R_0 is very large, note that if the quasi-stationary distribution is bimodal, then it must has a peak at $n = 1$, implying that $\tilde{P}_1^{ss} > \tilde{P}_2^{ss}$. From Eq. (19), the condition of $\tilde{P}_1^{ss} > \tilde{P}_2^{ss}$ is equivalent to

$$\frac{u_1(\lambda_1)}{d_1} < \frac{u_2(\lambda_1)}{d_1 d_2}.$$

From Eq. (9), this can be rewritten as

$$\lambda_1 + \alpha\frac{N-1}{N} + \beta\frac{(N-1)(N-2)}{N(N-1)} < \gamma.$$

When R_0 is very large, the first nonzero eigenvalue λ_1 can be ignored, i.e. $\lambda_1 \approx 0$ (Table 1), and hence the above condition reduces to

$$R_0 = \frac{\alpha + 2\beta}{\gamma} < \frac{1 + 2\frac{\beta}{\alpha}}{\frac{N-1}{N} + \frac{(N-2)}{N}\frac{\beta}{\alpha}} \approx \frac{1 + 2\frac{\beta}{\alpha}}{1 + \frac{\beta}{\alpha}}.$$

This indicates that a bimodal quasi-stationary only takes place when $R_0 < (1 + 2\beta/\alpha)/(1 + \beta/\alpha)$ (shown by the region under the red dashed curve in Fig. 4(b)) and hence fails to be observed when R_0 is very large. In summary, we find that the quasi-stationary distribution tends to be bimodal when β is relatively large and when R_0 is neither too large nor too small. In addition, comparing Figs. 3(b) and 4(b), we find that the bimodal region of quasi-stationary distribution is much smaller than the region of transient bimodality. This is intuitive because if the quasi-stationary distribution is bimodal, then the distribution of infected individuals must be bimodal at at least one time point and hence the system exhibits transient bimodality.

4.3 WKB approximation

Thus far, we have computed the quasi-stationary distribution of the stochastic SIS model in closed form. However, the explicit expression given in Eq. (19) is complicated. Next we apply the WKB theory to derive an approximate analytical solution in the limit of large population size. For simplicity, here we only focus on the case of $R_0 > 1$. In this case, the deterministic model has a disease-free equilibrium which is unstable and a positive endemic equilibrium which is globally asymptotic stable. For convenience, let $x = n/N$ denote the density of infected individuals and we employ the WKB ansatz

$$\tilde{P}_n^{ss} \asymp e^{-NS(x)}, \tag{21}$$

where $S(x)$ is called the quasi-potential of the stochastic SIS model. The rationality of this ansatz can be rigorously established by using Freidlin-Wentzell-type large deviation theory [51, 52]. In Supplementary Sec. 1, we prove that the quasi-potential is given by

$$S(x) = \int_{x^*}^{x} L(u)\mathrm{d}u, \tag{22}$$

where $x^* > 0$ is the endemic equilibrium of the deterministic system and

$$L(x) = \ln\left[\frac{\sqrt{[\alpha + \beta(1-x)]^2 + 4\beta\gamma} - [\alpha + \beta(1-x)]}{2\beta(1-x)}\right]. \tag{23}$$

Note that Eqs. (21) and (22) give an approximate analytical expression for the quasi-stationary distribution.

From Eq. (23), it is easy to check that $L'(x) > 0$ for any $x \geq 0$ and $L(x^*) = 0$. This shows that $L(x)$ is strictly increasing, $L(x) \leq 0$ for any $x \leq x^*$, and $L(x) > 0$ for any $x > x^*$. It then follows from Eq. (21) that

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{-NS(x)} = -Ne^{S(x)}L(x).$$

Hence $e^{-NS(x)}$ is monotonically increasing when $x \leq x^*$ and is monotonically decreasing when $x \geq x^*$. This clearly shows that the WKB approximation of the quasi-stationary distribution must be unimodal.

To evaluate the performance of the WKB approximation, we compare it with the exact quasi-stationary distribution obtained from Eq. (19) in Fig. 4(a). We find that the WKB approximation is in excellent agreement with the exact solution when the quasi-stationary distribution is unimodal (shown by the red solid and dashed lines). However, it may deviate significantly from the exact solution when the quasi-stationary distribution is bimodal (shown by the blue solid and dashed lines). We have proved that the WKB approximation is always unimodal and hence fails to capture the bimodal characteristics of the quasi-stationary distribution.

5 Extinction time distribution

5.1 Relaxation rate for the deterministic model

We have seen that the disease must eventually go extinct for the stochastic SIS model. Next we investigate the extinction rate and extinction time for the stochastic dynamics. Before doing this, we first focus on the relaxation rate for the deterministic model.

Recall that the deterministic SIS model given in Eq. (2) has an unique global asymptotic stable equilibrium x_e, which is zero when $R_0 \leq 1$ and is positive when $R_0 > 1$. In the long-term limit, the deterministic system converges to this equilibrium at exponential speed: there exists a constant $C > 0$ such that

$$|x(t) - x_e| \leq Ce^{-vt}, \quad t \geq 0.$$

Here the optimal (maximal) parameter v characterizes the exponential converge rate to the equilibrium x_e, which is also called the relaxation rate in the physics literature [53]. We next compute the relaxation rate v analytically. To this end, we make the ansatz

$$x(t) \approx x_e + Ce^{-vt},$$

which is a good approximation when t is sufficiently large. Inserting it into Eq. (2) yields

$$v \approx -(\alpha + 2\beta - \gamma) + 2(\alpha + 4\beta)x_e - 6\beta x_e^2 - [6\beta x_e - (\alpha + 4\beta)]Ce^{-vt} - 2\beta C^2 e^{-2vt}.$$

Taking $t \to \infty$ in the above equation, we obtain

$$v = -(\alpha + 2\beta - \gamma) + 2(\alpha + 4\beta)x_e - 6\beta x_e^2.$$

Note that $x_e = 0$ is the disease-free equilibrium when $R_0 \leq 1$ and $x_e = x^*$ is the endemic equilibrium when $R_0 > 1$; hence we obtain the following explicit expression of the relation rate:

$$v = \begin{cases} \gamma - \alpha - 2\beta, & R_0 \leq 1, \\ \frac{\alpha+4\beta}{4\beta}\sqrt{\alpha^2 + 8\beta\gamma} - \frac{\alpha^2}{4\beta} - 2\gamma, & R_0 > 1. \end{cases}$$

We next investigate the connection between the relaxation rate v for the deterministic model and the eigenvalues of the generator matrix Q for the stochastic model. Recall that Q must have the zero eigenvalue $\lambda_0 = 0$. Fig. 5(a) illustrates the relaxation rate v, the absolute value of the first nonzero eigenvalue of Q (we have seen that the first nonzero eigenvalue must be real and negative), i.e. $|\lambda_1|$, and the absolute value of the real part of the second nonzero eigenvalue of Q, i.e. $|\mathrm{Re}(\lambda_2)|$, as a function of the basic reproduction number R_0. It is clear that the relaxation rate v for the deterministic model decreases with R_0 when $R_0 \leq 1$ and increases with R_0 when $R_0 > 1$. Interestingly, we find that the relaxation rate v is very close to $|\lambda_1|$ when R_0 is relatively small, whereas it is very closed to $|\mathrm{Re}(\lambda_2)|$ when R_0 is relatively large. This provides deep insights into the first and second nonzero eigenvalues of the generator matrix for the stochastic SIS model. Roughly speaking, for the stochastic dynamics, the first nonzero eigenvalue λ_1 characterizes the rate of disease extinction (since $|\lambda_1| \approx v = \gamma(1 - R_0)$ when $R_0 < 1$), i.e. the convergence rate to the extinction state, while the second nonzero eigenvalue λ_2 characterizes the rate of disease outbreak (since $|\mathrm{Re}(\lambda_2)| \approx v$ when $R_0 > 1$), i.e. the convergence rate to the quasi-stationary state.

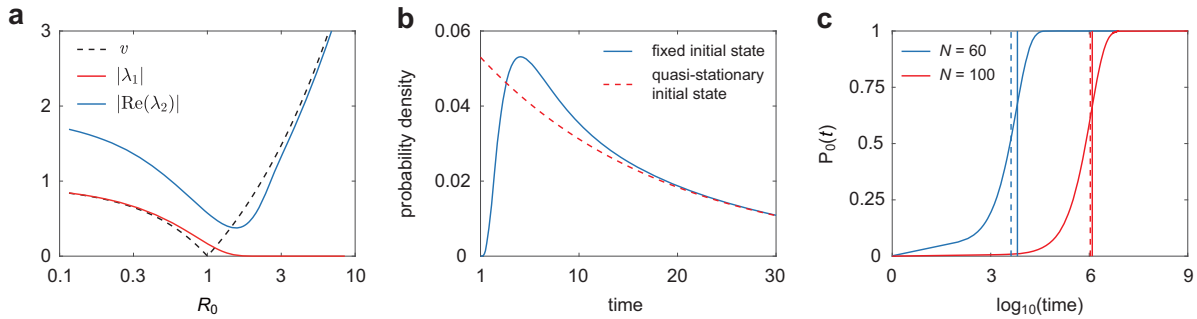

Figure 5: **Extinction rate and extinction time for the stochastic SIS model. (a)** Comparison of the relaxation rate v to the global asymptotically stable equilibrium for the deterministic model, the absolute value of the first nonzero eigenvalue $|\lambda_1|$ for the stochastic model, and the absolute value of the real part of the second nonzero eigenvalue $|\mathrm{Re}(\lambda_2)|$ for the stochastic model as R_0 varies. The model parameters are chosen as $\alpha = 0.09$, $\gamma = 0.95$, and $N = 100$. **(b)** Comparison of the extinction time distributions when the system starts from a fixed initial state (blue curve) and when the system starts from the quasi-stationary distribution (red dashed curve). The model parameters are chosen as $\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.6$, $R_0 = 1.17$, and $N = 60$. **(c)** Comparison the exact and approximate mean extinction times. The blue and red curves show the time evolution of $P_0(t)$ under two different choices of N. The solid vertical lines show the exact mean extinction time computed using Eq. (26) and the dashed vertical lines show the WKB approximation computed using Eq. (27). The model parameters are chosen as $\alpha = 0.8$, $\beta = 0.2$, $\gamma = 0.6$, and $R_0 = 2$. In (b),(c), the initial condition is chosen to be $N_I(0) = 0.1N$.

As a summary, we emphasize that the stationary distribution (extinction state) of the stochastic SIS model corresponds to the eigenvector associated with the zero eigenvalue λ_0 and the convergence rate to the extinction state is characterized by the first nonzero eigenvalue λ_1. Similarly, the quasi-stationary distribution of the stochastic SIS model corresponds to the eigenvector associated with the first nonzero eigenvalue λ_1 and the convergence rate to the quasi-stationary state is characterized by the second nonzero eigenvalue λ_2.

5.2 Exact extinction time distribution

We next exmaine the extinction dynamics of the disease. Let $T = \inf\{t \geq 0 : N_I(t) = 0\}$ denote the time to extinction for the stochastic SIS model. We next compute the distribution of the extinction time. Since $n = 0$ is the absorbing state, it is easy to see that the events $T \leq t$ and $N_I(t) = 0$ are equivalent. The probability of $N_I(t) = 0$ has been obtained analytically in Eq. (18). Hence it is easy to compute the distribution of the extinction time T. Suppose that the initial number of infected individuals is $n_0 \geq 1$. It follows from Eq. (10) that

$$\begin{aligned} \mathrm{Prob}(T \leq t) &= \mathrm{Prob}(N_I(t) = 0) = P_0(t|n_0) \\ &= \sum_{k=0}^{N} \frac{d_1 \cdots d_{n_0} v_{n_0}(\lambda_k) e^{\lambda_k t}}{\prod_{j \neq k}(\lambda_k - \lambda_j)}. \end{aligned} \tag{24}$$

Differentiating this equation with respect to t, we obtain the explicit expression of the probability density $f(t)$ of the extinction time T:

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Prob}(T \leq t) = \sum_{k=1}^{N} \frac{d_1 \cdots d_{n_0} \lambda_k v_{n_0}(\lambda_k) e^{\lambda_k t}}{\prod_{j \neq k}(\lambda_k - \lambda_j)}. \tag{25}$$

The gives the exact extinction time distribution when the system starts from a fixed initial state.

In some cases, it is more important to study the extinction time distribution starting from the quasi-stationary state rather than from a fixed initial state. We next compute the extinction time distribution when the system starts from the quasi-stationary distribution $\tilde{P}^{ss}$. Since $\tilde{P}^{ss}$ is stationary solution of Eq. (12), we have

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} + \tilde{P}_n^{ss} \frac{\mathrm{d}P_0(t)}{\mathrm{d}t} = 0, \quad n = 1, \cdots, N,$$

where we have used the fact that $\tilde{P}_n(t) = \tilde{P}_n^{ss}$ for any $t \geq 0$ and $n \geq 1$ when the system starts from the quasi-stationary distribution. Moreover, it follows from Eq. (14) that $P_0'(t) = \gamma P_1(t)$. This indicates that

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} + \gamma \tilde{P}_n^{ss} P_1(t) = 0, \quad n = 1, \cdots, N,$$

Taking $n = 1$ in the above equation, we obtain

$$\frac{\mathrm{d}P_1(t)}{\mathrm{d}t} + \gamma \tilde{P}_1^{ss} P_1(t) = 0.$$

Solving this equation yields $P_1(t) = \tilde{P}_1^{ss} e^{-\gamma \tilde{P}_1^{ss} t}$. Since $P_0'(t) = \gamma P_1(t)$, we finally obtain the probability density $f(t)$ of the extinction time T when the system starts from the quasi-stationary distribution:

$$f(t) = \frac{\mathrm{d}P_0(t)}{\mathrm{d}t} = \gamma \tilde{P}_1^{ss} e^{-\gamma \tilde{P}_1^{ss} t} = -\lambda_1 e^{\lambda_1 t},$$

where we have used the fact that $\lambda_1 = -\gamma \tilde{P}_1^{ss}$. This demonstrates that the extinction time follows an exponential distribution with rate $-\lambda_1$ when the system starts from the quasi-stationary distribution. This is consistent with the results obtained in [47].

Fig. 5(b) compares the extinction time distributions when the system starts from the quasi-stationary distribution and when it starts from a fixed initial state. In the former case, the extinction time distribution is exponential and thus monotonically decreasing (shown by the red dashed curve). In the latter case, the extinction time distribution may be non-exponential and bell-shaped, increasing initially and then decreasing (shown by the blue curve).

5.3 Exact mean extinction time and the WKB approximation

We next compute the mean extinction time for the stochastic SIS model. We first consider the case where the initial number of infected individuals is $n_0 \geq 1$. From Eq. (25), the mean extinction time is given by

$$\begin{aligned} \mathbb{E}(T) &= \int_0^\infty t f(t) \mathrm{d}t \\ &= \sum_{k=1}^{N} \frac{d_1 \cdots d_{n_0} \lambda_k v_{n_0}(\lambda_k) \int_0^\infty t e^{\lambda_k t} \mathrm{d}t}{\prod_{j \neq k} (\lambda_k - \lambda_j)} \\ &= \sum_{k=1}^{N} \frac{d_1 \cdots d_{n_0} v_{n_0}(\lambda_k)}{\lambda_k \prod_{j \neq k} (\lambda_k - \lambda_j)}. \end{aligned} \tag{26}$$

This gives the explicit expression of the mean extinction time when the system starts from a fixed initial state. When the system starts from the quasi-stationary distribution, we have proved that the extinction time follows an exponential distribution with parameter $-\lambda_1$. Hence the mean extinction time is $\mathbb{E}(T) = -1/\lambda_1$. This again shows that the first nonzero eigenvalue of Q characterizes the convergence speed to the extinction state.

Thus far, we have obtained the mean extinction time for the stochastic SIS model. However, the explicit expression given in Eq. (26) is complicated and it is not clear how the population size affects the mean extinction time. Next we use the WKB theory to derive an approximate analytical solution of the mean extinction time when the total population size is large. For simplicity, here we only focus on the case of $R_0 > 1$. According to the Freidlin-Wentzell-type large deviation theory [51, 52], the leading-order asymptotic behaviour of the mean extinction time is given by

$$\mathbb{E}_{\tilde{P}^{ss}}(T) \asymp e^{NS(0)} = e^{-N \int_0^{x^*} L(u) \mathrm{d}u}, \tag{27}$$

where $S(0) = S(0) - S(x^*)$ represents the energy barrier between the disease-free equilibrium, i.e. the extinction state, and the endemic equilibrium x^*, and the function $L(u)$ is given by Eq. (23). Clearly, the mean extinction time increases exponentially with the total population size N.

To validate the analytical results, we focus on the evolution of $P_0(t)$ when the system starts from a fixed initial state. Fig. 5(c) illustrates the time evolution of $P_0(t)$ obtained from Eq. (24), the exact mean extinction time, and the approximate mean extinction time predicted using the WKB theory under different population sizes. It can be seen that when N is large, $P_0(t)$ (shown by the colored curve) is close to zero at small times and is close one at large times. In particular, there is a sharp critical transition of $P_0(t)$ from zero to one at intermediate times. We emphasize that the increase of $P_0(t)$ is due to disease extinction, and hence the critical transition must occur around the mean extinction time. We find that the approximate mean extinction time predicted by the WKB theory (shown by the dashed vertical line) is very closed to the exact mean extinction time (shown by the solid vertical line). This suggests that the WKB approximation indeed provides an excellent prediction of the mean extinction time.

6 First-passage time distribution

In addition to the extinction time, public health authorities are particularly concerned with the time at which the number of infected individuals first reaches or exceeds a specified threshold — an indicator known as the first-passage time. This indicator provides a quantitative basis for constructing early warning systems and supports the design of graded response strategies. Moreover, it plays a critical role in guiding both the timing and intensity of intervention measures, thereby helping to delay the epidemic peak and optimize the allocation of resources [54, 55].

Specifically, let $T_M = \inf\{t : N_I(t) \geq M\}$ denote the time at which the number of infected individuals first reaches or exceeds a given threshold $M \geq 1$. We next compute the distribution of the first-passage time T_M for the stochastic SIS model. Suppose that the initial number of infected individuals is $n_0 \in [1, M)$ and let

$$F(n_0, M, t) = \text{Prob}(T_M \leq t | N_I(0) = n_0)$$

be the cumulative distribution function of T_M. Note that within the small time interval $[0, \Delta t]$, the system can only make three types of transitions (Fig. 1(b)): a susceptible individual is infected with probability $a_{n_0}\Delta t + o(\Delta t)$, two susceptible individuals are simultaneously infected with probability $b_{n_0}\Delta t + o(\Delta t)$, and an infected individual recovers with probability $d_{n_0}\Delta t + o(\Delta t)$. On the other hand, it is clear that the probability of having more than one transitions during $[0, \Delta t]$ is $o(\Delta t)$, and the probability of having no transitions is $1 - (a_{n_0} + b_{n_0} + d_{n_0})\Delta t + o(\Delta t)$. Hence it follows that

$$\begin{aligned} F(n_0, M, t) = {} & a_{n_0}\Delta t F(n_0 + 1, M, t - \Delta t) + b_{n_0}\Delta t F(n_0 + 2, M, t - \Delta t) + d_{n_0}\Delta t F(n_0 - 1, M, t - \Delta t) \\ & + [1 - (a_{n_0} + b_{n_0} + d_{n_0})\Delta t]\, F(n_0, M, t - \Delta t) + o(\Delta t). \end{aligned}$$

Subtracting $F(n_0, M, t - \Delta t)$ on both sides of the above equation and letting $\Delta t \to 0$, we find that $F(n_0, M, t)$ is governed by the following Kolmogorov backward equation:

$$\frac{\mathrm{d}F(n_0, M, t)}{\mathrm{d}t} = a_{n_0}F(n_0 + 1, M, t) + b_{n_0}F(n_0 + 2, M, t) + d_{n_0}F(n_0 - 1, M, t) - q_{n_0}F(n_0, M, t), \tag{28}$$

where the boundary conditions are given by $F(0, M, t) = 0$ and $F(n_0, M, t) = 1$ for any $n_0 \geq M$. To proceed, let

$$F_M(t) = (F(1, M, t), \cdots, F(M - 1, M, t))^T$$

be a column vector whose components are $F(n_0, M, t)$ for all $n_0 \in [1, M)$. Then Eq. (28) can be written in matrix form as

$$\frac{\mathrm{d}F_M(t)}{\mathrm{d}t} = \tilde{Q}_{M-1}F_M(t) + b, \tag{29}$$

where $b = (0, 0, \ldots, b_{M-2}, a_{M-1} + b_{M-1})^T$ and

$$\tilde{Q}_{M-1} = \begin{pmatrix} -q_1 & a_1 & b_1 & 0 & \cdots & 0 & 0 \\ d_2 & -q_2 & a_2 & b_2 & \cdots & 0 & 0 \\ 0 & d_3 & -q_3 & a_3 & \cdots & 0 & 0 \\ 0 & 0 & d_4 & -q_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -q_{M-2} & a_{M-2} \\ 0 & 0 & 0 & 0 & \cdots & d_{M-1} & -q_{M-1} \end{pmatrix}$$

is the matrix obtained from $\tilde{Q}$ by retaining its first $M-1$ rows and $M-1$ columns. Since $F_M(0) = 0$ is the zero vector, it follows from Eq. (29) that

$$F_M(t) = e^{\tilde{Q}_{M-1}t} F_M(0) + (e^{\tilde{Q}_{M-1}t} - I)\tilde{Q}_{M-1}^{-1} b = (e^{\tilde{Q}_{M-1}t} - I)\tilde{Q}_{M-1}^{-1} b. \tag{30}$$

Similarly to the derivation of the exact time-dependent distribution, we compute $e^{\tilde{Q}_{M-1}t}$ and $\tilde{Q}_{M-1}^{-1}$ analytically using matrix-valued Cauchy's integral formula and Cauchy's residue theorem. Then $F_M(t)$ can be computed in closed form as (see Supplementary Sec. 2 for the proof):

$$\begin{aligned} F(n_0, M, t) = \frac{1}{\tilde{u}_M(0)} \Bigg\{ & \sum_{l=1}^{M-1} [-b_{M-2}\tilde{g}_{l,M-2}(0) - (a_{M-1} + b_{M-1})\tilde{g}_{l,M-1}(0)] \sum_{m=1}^{M-1} \frac{\tilde{g}_{n_0,l}(\tilde{\lambda}_m) e^{\tilde{\lambda}_m t}}{\prod_{n \neq m}(\tilde{\lambda}_m - \tilde{\lambda}_n)} \\ & + b_{M-2}\tilde{g}_{n_0,M-2}(0) + (a_{M-1} + b_{M-1})\tilde{g}_{n_0,M-1,}(0) \Bigg\}, \end{aligned} \tag{31}$$

where $\tilde{g}_{k,j}(z)$ are polynomials defined as

$$\tilde{g}_{k,j}(z) = \begin{cases} \tilde{h}_{k,j}(z), & k < j, \\ \tilde{u}_j(z)\tilde{v}_j(z), & k = j, \\ d_{j+1}d_{j+2}\ldots d_k \tilde{u}_j(z)\tilde{v}_k(z), & k > j, \end{cases}$$

and $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{M-1}$ are all the eigenvalues of $\tilde{Q}_{M-1}$ (for simplicity, we assume that these eigenvalues are non-degenerate and hence $\tilde{Q}_{M-1}$ has $M-1$ distinct eigenvalues). Here $\tilde{u}_j(z)$, $1 \leq j \leq M$, are polynomials defined recursively by

$$\begin{gathered} \tilde{u}_j(z) = (z + q_{j-1})\tilde{u}_{j-1}(z) - a_{j-2}d_{j-1}\tilde{u}_{j-2}(z) - b_{j-3}d_{j-1}d_{j-2}\tilde{u}_{j-3}(z), \\ \tilde{u}_1(z) = 1, \quad \tilde{u}_2(z) = (z + q_1), \quad \tilde{u}_3(z) = (z + q_2)\tilde{u}_2(z) - a_1 d_2, \end{gathered}$$

with $\tilde{u}_M(z) = \det(zI - \tilde{Q}_{M-1})$ being the characteristic polynomial of $\tilde{Q}_{M-1}$, $\tilde{v}_k(z)$, $1 \leq k \leq M-1$, are polynomials defined recursively by

$$\begin{gathered} \tilde{v}_k(z) = (z + q_{k+1})\tilde{v}_{k+1}(z) - a_{k+1}d_{k+2}\tilde{v}_{k+2}(z) - b_{k+1}d_{k+3}d_{k+2}\tilde{v}_{k+3}(z), \\ \tilde{v}_{M-1}(z) = 1, \quad \tilde{v}_{M-2}(z) = z + q_{M-1}, \quad \tilde{v}_{M-3}(z) = (z + q_{M-2})\tilde{v}_{M-2}(z) - a_{M-2}d_{M-1}, \end{gathered}$$

and $\tilde{h}_{k,j}(z)$ are functions defined as

$$\begin{aligned} \tilde{h}_{k,j}(z) = \delta_1(j-k)[b_{j-2}d_{j-1}\tilde{v}_j(z)\tilde{u}_{k-1}(z)] + (-1)^{j-k+1}\big\{ & [a_{j-1}\tilde{v}_j(z) + d_{j+1}b_{j-1}\tilde{v}_{j+1}(z)]\tilde{w}_{k,j-k}(z) \\ & - b_{j-2}(z + q_{j-1})\tilde{v}_j(z)\tilde{w}_{k,j-k-1}(z) + b_{j-2}b_{j-3}d_{j-1}\tilde{v}_j(z)\tilde{w}_{k,j-k-2}(z)\big\}, \end{aligned}$$

where $\tilde{w}_{k,i}(z)$, $1 \leq i \leq j-k$, are another sequence of polynomials defined recursively by

$$\begin{gathered} \tilde{w}_{k,i}(z) = -a_{k+i-2}\tilde{w}_{k,i-1}(z) + b_{k+i-3}(z + q_{k+i-2})\tilde{w}_{k,i-2}(z) - b_{k+i-3}b_{k+i-4}d_{k+i-2}\tilde{w}_{k,i-3}(z), \\ \tilde{w}_{k,1}(z) = \tilde{u}_k(z), \quad \tilde{w}_{k,2}(z) = -a_k\tilde{w}_{k,1}(z) - b_{k-1}d_k\tilde{u}_{k-1}(z), \\ \tilde{w}_{k,3}(z) = -a_{k+1}\tilde{w}_{k,2}(z) + b_k(z + q_{k+1})\tilde{w}_{k,1}(z). \end{gathered}$$

Differentiating Eq. (31) with respect to t, we finally obtain the probability density of the first-passage time T_M:

$$f(n_0, M, t) = \frac{1}{\tilde{u}_M(0)} \sum_{l=1}^{M-1} [-b_{M-2}\tilde{g}_{l,M-2}(0) - (a_{M-1} + b_{M-1})\tilde{g}_{l,M-1}(0)] \sum_{m=1}^{M-1} \frac{\tilde{\lambda}_m \tilde{g}_{n_0,l}(\tilde{\lambda}_m) e^{\tilde{\lambda}_m t}}{\prod_{n \neq m}(\tilde{\lambda}_m - \tilde{\lambda}_n)}. \tag{32}$$

As expected, the analytical solution coincides perfectly with the numerical one obtained using the stochastic simulation algorithm (Fig. 6(a)). Moreover, we emphasize that for any given threshold $M > n_0$, there is a positive probability that the disease will have gone extinct before the number of infected individuals ever reaches M. Consequently, the first-passage time T_M can take an infinite value with positive probability, implying that its expected value must diverge to infinity, i.e. $\mathbb{E}[T_M] = \infty$.

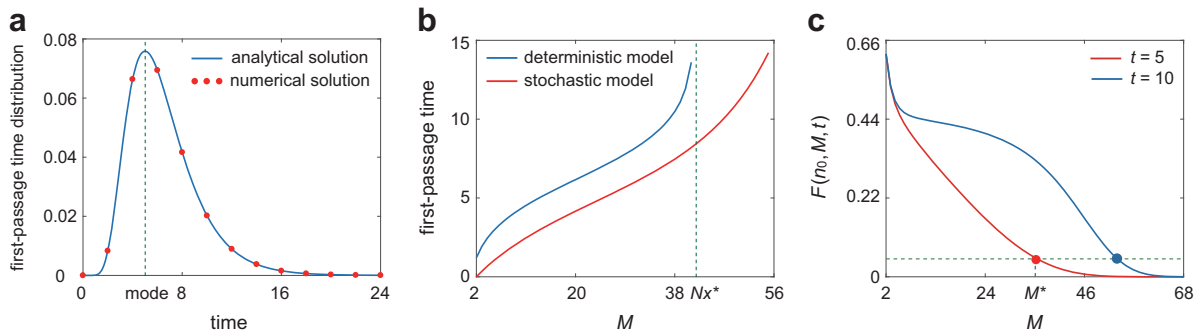

Figure 6: **First-passage time distribution for the stochastic SIS model.** **(a)** Comparison of the analytical first-passage time distribution given in Eq. (32) (blue curve) and the numerical one (red dots) obtained from stochastic simulations. The green dashed vertical line shows the mode (peak value) of the distribution. The model parameters are chosen as $N = 100$, $\alpha = 0.8$, $\beta = 0.2$, $\gamma = 0.6$, and $R_0 = 2$. The initial condition is chosen as $N_I(0) = 1$ and the critical threshold is chosen as $M = 30$. **(b)** Comparison of the first-passage time for the deterministic model (blue curve) and the mode of the first-passage time distribution for the stochastic model (red curve) as the threshold M varies. The green dashed vertical line shows the endemic equilibrium Nx^* for the deterministic model. **(c)** The probability that the number of infections has already reached the threshold M before time t, i.e. $F(n_0, M, t)$, as a function of M. Given the significance level of $\alpha = 0.05$, the critical threshold M^* satisfying $F(n_0, M^*, t) = \alpha$ gives an estimate of the emergency bed capacity at time t (with 95% confidence). In (b),(c), the model parameters and initial condition are chosen to be the same as in (a).

Fig. 6(b) compares the first-passage time for the deterministic model with the mode of the first-passage time distribution for the stochastic model (we use the mode here rather than the mean since the expected value of T_M diverges to infinity), both as functions of the threshold M. Clearly, the deterministic model predicts a longer first-passage time than the stochastic model. In particular, the discrepancy becomes pronounced when the threshold M approaches the endemic equilibrium Nx^*. This occurs because the deterministic system can never actually reach the endemic equilibrium, leading to an infinite first-passage time when $M \geq Nx^*$, whereas the stochastic system still yields a finite value. In the context of infectious disease control, the first-passage time at which the number of infections reaches a critical threshold is often regarded as a warning signal for intervention measures. As shown in Fig. 6(b), the deterministic model may severely overestimate the first-passage time, potentially delaying critical response actions and thereby increasing the scale and risk of disease outbreaks. By contrast, the stochastic model provides a more timely and reliable warning signal for interventions.

Next we demonstrate how analyzing the first-passage time distribution can inform the allocation of emergency beds. Fig. 6(c) illustrates $F(n_0, M, t)$ as a function of M for two different values of t, where $F(n_0, M, t)$ denotes the probability that the number of infections has already reached the threshold M by time t. For any given significance level $\alpha > 0$, one can determine the critical threshold M^* satisfying $F(n_0, M^*, t) = \alpha$ (Fig. 6(c)). For example, if we choose $\alpha = 0.05$, then with 95% probability the number of infections will not have reached M^* by time t. This provides practical guidance for short-term healthcare resource planning: by setting the emergency bed capacity to M^*, hospitals can meet disease control demand at time t with 95% confidence. Thus analysis of the first-passage time for the stochastic SIS model offers a quantitative basis for early warning and control.

7 Conclusions and discussion

In this work, we provide a systematic study of the stochastic dynamics of an extended SIS epidemic model in densely populated environments using a Markov jump process framework. Unlike the classical SIS model, which only includes the infection of a susceptible individual and the recovery of an infected individual, the extended SIS model allows for simultaneous transmission of the disease to two susceptible individuals due to crowded conditions. It is well known that the deterministic SIS model has a positive endemic equilibrium when the basic reproduction number is greater than one. However, the stochastic SIS model will eventually go extinct with probability one, since the state space is finite and the extinction state is an absorbing state of the system.

Using complex analysis techniques including matrix-valued Cauchy's integral formula and Cauchy's residue theorem, we analytically solve the master equation for the stochastic SIS model and derive closed-form solutions for the time-dependent distribution of the number of infected individuals as well as the distribution of the first-passage time at which the number of infected individuals reaches a certain threshold. Prior to disease extinction, the stochastic SIS model approaches an approximately stationary state corresponding to the endemic equilibrium of the deterministic model. This analytical time-dependent distribution is then used to compute the quasi-stationary distribution of infected individuals and the extinction time distribution of the disease. Moreover, the quasi-potential of the stochastic SIS model is derived using the large deviation theory, which is then used to construct WKB approximations of the quasi-stationary distribution and the mean extinction time in the limit of large population size.

Within the Markov jump process framework, the generator matrix for the classical SIS model has only real eigenvalues. In contrast, the extended SIS model has a zero eigenvalue and a real first nonzero eigenvalue, but all other eigenvalues may be complex due to the simultaneous infection of two susceptible individuals. The extinction state is shown to correspond to the eigenvector associated with the zero eigenvalue, while the quasi-stationary state corresponds to the eigenvector associated with the first nonzero eigenvalue. We also investigate the extinction and outbreak rates of the infectious disease. Interestingly, we find that the first nonzero eigenvalue of the generator matrix characterizes the extinction rate of the epidemic, while the second nonzero eigenvalue characterizes the outbreak rate.

We also study the stochastic bifurcation for the stochastic SIS model by analyzing the time evolution of the distribution of infected individuals, and find that the system may exhibit two distinct dynamic patterns: (i) unimodality, where the distribution remains unimodal at all times; and (ii) transient bimodality, where the distribution is unimodal at early and late times but becomes bimodal at intermediate times. The bifurcation threshold of the basic reproduction number for the stochastic model is shown to be larger than that for its deterministic counterpart. Similarly, we examine the shape of the quasi-stationary distribution and find that it may be unimodal or bimodal. The parameter region for bimodality in the quasi-stationary distribution is shown to be much narrower than that for transient bimodality. Finally, we demonstrate that analyzing the first-passage time distribution provides a quantitative basis for early warning of intervention measures and rational planning of medical resources, such as hospital beds and drug inventories, thereby improving the efficiency of resource allocation and emergency response capabilities.

Our study also has some limitations. First, we only consider the case where at most two susceptible individuals can be infected simultaneously. However, in densely populated settings or during highly contagious phases of an outbreak, more than two individuals may become infected simultaneously. Second, to ensure analytical tractability, we have not considered the birth of susceptible individuals and the death of susceptible and infected ones [56]. Such extensions would lead to a more complex Markov process, for which analytical treatment may be much more challenging. We aim to explore these more complex scenarios in future work to gain deeper insights into infectious disease dynamics and improve epidemic control strategies. We also anticipate that our analytical results can be used to perform accurate parameter inference for stochastic epidemic dynamics based on empirical probability distributions obtained from multi-region parallel observations at multiple discrete time points.

Acknowledgements

G. Liu acknowledges support from NSFC grants Nos.12371494, 12231012, and 11971279, as well as Shanxi Provincial Key Research and Development Project No.202202020101010. C. Jia acknowledges support from NSFC

grants No. 12271020.

References

[1] Anderson, R. M. & May, R. M. *Infectious diseases of humans: dynamics and control* (Oxford university press, Oxford, 1991).

[2] Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).

[3] Diekmann, O. & Heesterbeek, J. A. P. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation* (John Wiley & Sons, New York, 2000).

[4] Brauer, F., Van den Driessche, P. & Wu, J. *Mathematical epidemiology* (Springer, Berlin, 2008).

[5] Martcheva, M. *An introduction to mathematical epidemiology* (Springer, New York, 2015).

[6] Van den Driessche, P. & Watmough, J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**, 29–48 (2002).

[7] Hadeler, K. P. & Van den Driessche, P. Backward bifurcation in epidemic control. *Math. Biosci.* **146**, 15–35 (1997).

[8] Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A. & Jewell, C. P. Novel coronavirus 2019-nCoV (COVID-19): early estimation of epidemiological parameters and epidemic size estimates. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **376**, 20200265 (2021).

[9] Hansen, E. & Day, T. Optimal control of epidemics with limited resources. *J. Math. Biol.* **62**, 423–451 (2011).

[10] Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925–979 (2015).

[11] Allen, L. J., Bolker, B., Lou, Y. & Nevai, A. Asymptotic profiles of the steady states for an SIS epidemic reaction-diffusion model. *Discrete Contin. Dyn. Syst. Ser. A* **21**, 1–20 (2008).

[12] Magal, P., Webb, G. F. & Wu, Y. On the basic reproduction number of reaction-diffusion epidemic models. *SIAM J. Appl. Math.* **79**, 284–304 (2019).

[13] Ruan, S. Delay differential equations in single species dynamics. In *Delay differential equations and applications*, 477–517 (Springer, 2006).

[14] Ferguson, N. M. *et al.* Strategies for mitigating an influenza pandemic. *Nature* **442**, 448–452 (2006).

[15] Allen, L. J. *An introduction to stochastic processes with applications to biology* (CRC press, New York, 2010).

[16] Andersson, H. & Britton, T. *Stochastic epidemic models and their statistical analysis* (Springer Science & Business Media, New York, 2012).

[17] Kurtz, T. G. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8**, 344–356 (1971).

[18] Kurtz, T. G. Strong approximation theorems for density dependent Markov chains. *Stoch. Proc. Appl.* **6**, 223–240 (1978).

[19] Guy, R., Larédo, C. & Vergu, E. Approximation of epidemic models by diffusion processes and their statistical inference. *J. Math. Biol.* **70**, 621–646 (2015).

[20] Van Kampen, N. G. *Stochastic processes in physics and chemistry* (Elsevier, 1992).

[21] Dykman, M. I., Schwartz, I. B. & Landsman, A. S. Disease extinction in the presence of random vaccination. *Phys. Rev. Lett.* **101**, 078101 (2008).

[22] Assaf, M. & Meerson, B. WKB theory of large deviations in stochastic populations. *J. Phys. A: Math. Theor.* **50**, 263001 (2017).

[23] Ball, F. & Donnelly, P. Strong approximations for epidemic models. *Stoch. Proc. Appl.* **55**, 1–21 (1995).

[24] Britton, T. Stochastic epidemic models: a survey. *Math. Biosci.* **225**, 24–35 (2010).

[25] Kuske, R., Gordillo, L. F. & Greenwood, P. Sustained oscillations via coherence resonance in SIR. *J. Theor. Biol.* **245**, 459–469 (2007).

[26] Forgoston, E., Bianco, S., Shaw, L. B. & Schwartz, I. B. Maximal sensitive dependence and the optimal path to epidemic extinction. *Bull. Math. Biol.* **73**, 495–514 (2011).

[27] Krishnarajah, I., Cook, A., Marion, G. & Gibson, G. Novel moment closure approximations in stochastic epidemics. *Bull. Math. Biol.* **67**, 855–873 (2005).

[28] Nåsell, I. Extinction and quasi-stationarity in the Verhulst logistic model. *J. Theor. Biol.* **211**, 11–27 (2001).

[29] Doering, C. R., Sargsyan, K. V. & Sander, L. M. Extinction times for birth-death processes: Exact results, continuum asymptotics, and the failure of the Fokker–Planck approximation. *Multiscale Model. Sim.* **3**, 283–299 (2005).

[30] Nåsell, I. On the quasi-stationary distribution of the stochastic logistic epidemic. *Math. Biosci.* **156**, 21–40 (1999).

[31] Assaf, M. & Meerson, B. Spectral theory of metastability and extinction in birth-death systems. *Phys. Rev. Lett.* **97**, 200602 (2006).

[32] Iacopini, I., Petri, G., Barrat, A. & Latora, V. Simplicial models of social contagion. *Nat. Commun.* **10**, 2485 (2019).

[33] Jhun, B., Jo, M. & Kahng, B. Simplicial SIS model in scale-free uniform hypergraph. *J. Stat. Mech.: Theory Exp.* **2019**, 123207 (2019).

[34] Fan, J., Yin, Q., Xia, C. & Perc, M. Epidemics on multilayer simplicial complexes. *Proc. R. Soc. Lond. A.* **478**, 20220059 (2022).

[35] Castillo-Chavez, C., Huang, W. & Li, J. Competitive exclusion in gonorrhea models and other sexually transmitted diseases. *SIAM J. Appl. Math.* **56**, 494–508 (1996).

[36] Heikkinen, T. & Järvinen, A. The common cold. *The Lancet* **361**, 51–59 (2003).

[37] Hall, C. B., Walsh, E. E., Long, C. E. & Schnabel, K. C. Immunity to and frequency of reinfection with respiratory syncytial virus. *J. Infect. Dis.* **163**, 693–698 (1991).

[38] Simon, C. M. The SIR dynamic model of infectious disease transmission and its analogy with chemical kinetics. *PeerJ Phys. Chem.* **2**, e14 (2020).

[39] Anderson, D. F. & Kurtz, T. G. *Stochastic analysis of biochemical systems*, vol. 674 (Springer, 2015).

[40] García-Tejera, R., Schumacher, L. & Grima, R. Regulation of stem cell dynamics through volume exclusion. *Proc. R. Soc. Lond. A.* **478**, 20220376 (2022).

[41] van den Driessche, P. & Watmough, J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**, 29–48 (2002).

[42] Norris, J. R. *Markov Chains* (Cambridge University Press, Cambridge, 1998).

[43] Smith, S. & Shahrezaei, V. General transient solution of the one-step master equation in one dimension. *Phys. Rev. E* **91**, 062119 (2015).

[44] Higham, N. J. *Functions of matrices: theory and computation* (SIAM, Philadelphia, 2008).

[45] Wu, B. & Jia, C. Parameter inference and nonequilibrium identification for Markov networks based on coarse-grained observations. *Phys. Rev. Lett.* **134**, 087103 (2025).

[46] Berman, A. & Plemmons, R. J. *Nonnegative Matrices in the Mathematical Sciences* (Academic Press, New York, 1979).

[47] Darroch, J. N. & Seneta, E. On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *J. Appl. Probab.* **4**, 192–196 (1967).

[48] Machihara, F. On the property of eigenvalues of some infinitesimal generator. *Oper. Res. Lett.* **2**, 123–126 (1983).

[49] Jia, C. & Grima, R. Dynamical phase diagram of an auto-regulating gene in fast switching conditions. *J. Chem. Phys.* **152** (2020).

[50] Holehouse, J., Sukys, A. & Grima, R. Stochastic time-dependent enzyme kinetics: Closed-form solution and transient bimodality. *J. Chem. Phys.* **153** (2020).

[51] Agazzi, A., Dembo, A. & Eckmann, J.-P. Large deviations theory for Markov jump models of chemical reaction networks. *Ann. Appl. Probab.* **28**, 1821–1855 (2018).

[52] Freidlin, M. I. & Wentzell, A. D. *Random perturbations of dynamical systems* (Springer, Heidelberg, 1998).

[53] Jia, C., Qian, H., Chen, M. & Zhang, M. Q. Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks. *J. Chem. Phys.* **148** (2018).

[54] Hatchett, R. J., Mecher, C. E. & Lipsitch, M. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc. Natl. Acad. Sci. USA* **104**, 7582–7587 (2007).

[55] Proverbio, D., Kemp, F., Magni, S. & Gonçalves, J. Performance of early warning signals for disease re-emergence: A case study on COVID-19 data. *PLoS Comput. Biol.* **18**, e1009958 (2022).

[56] Zhu, C. Critical result on the threshold of a stochastic SIS model with saturated incidence rate. *Physica A* **523**, 426–437 (2019).