Monte-Carlo simulation and analytic approximation of epidemic processes on large networks

Noémi Nagy, Péter L. Simon*

June 27, 2012

Institute of Mathematics, Eötvös Loránd University, Budapest, Hungary

Abstract

Low dimensional ODE approximations that capture the main characteristics of SIS-type epidemic propagation along a cycle graph are derived. Three different methods are shown that can accurately predict the expected number of infected nodes in the graph. The first method is based on the derivation of a master equation for the number of infected nodes. This uses the average number of SI edges for a given number of the infected nodes. The second approach is based on the observation that the epidemic spreads along the cycle graph as a front. We introduce a continuous time Markov chain describing the evolution of the front. The third method we apply is the subsystem approximation using the edges as subsystems. Finally, we compare the steady state value of the number of infected nodes obtained in the different ways.

Keywords: SIS epidemic, ODE approximation, network process

* corresponding author
email: simonp@cs.elte.hu

1 Introduction

Epidemic processes running on large networks attracted considerable interest in the last decade [1, 4, 6, 7]. Assuming a simple dynamics at the node level such as the SIS-model, when nodes can be susceptible or infected, leads to a stochastic process where the structure of the network will have an impact on how the infection spreads over the network. The mathematical model describing the process is a continuous time Markov chain with extremely large state space (2^N, where N is the number of nodes in the network), leading to the master equations that form a system of linear ordinary differential equations (consisting of 2^N equations) [10]. Several dynamical processes running on networks lead to the same kind of mathematical model, for example other epidemic dynamics, such as SIR (susceptible-infected-recovered) dynamics, spread of opinions through a population, propagation of neuronal activity on a neural network. Many models are studied in the monograph [1], the binary-state dynamics is investigated in [5]. The common feature of these models is that a graph with N nodes is given, and the nodes can be in one of M states. The transition between these states is described by independent Poisson processes, the transition rates of which are determined by the state of the neighbouring nodes. Hence the mathematical model is a continuous time Markov chain leading to the master equations that form a linear system of M^N equations. Solving these even numerically is impossible for the typical values of N simply due to the large number of equations. Two different approaches are used to overcome this difficulty. On one hand, the Monte-Carlo simulation of the stochastic process is carried out to get the average number of nodes in a given state as a function of time. On the other hand, several low-dimensional ODE systems have been derived as approximate models that can capture the exact dynamics in terms of the expected values of some well-defined quantities, such as the average number of nodes or edges in a given state. One of the most important mathematical questions is to establish a relation between the network structure and the dynamical behaviour of the ODE system describing the process. For random graphs [2] that can model complex networks, mean-field and pair-wise approximation models have been derived as low-dimensional approximations, see e.g. [5, 6]. In these approximating models some structural properties of the network, such as the average degree of the nodes, the degree correlation and the clustering of the network can be reflected by certain parameters. However, if the network has a special structure, i.e. it cannot be described as a random graph, then the use of the above parameters is not sufficient. Moreover, the performance of the mean-field and pair-wise approximations is strongly based on the randomness of the network, see Figure 1 below. This Figure shows that even in the simplest case, when the average degree is $n = 2$, the graph is a cycle graph and the SIS dynamics is considered, these approximations fail to predict the average number of infected nodes. In this paper our aim is to introduce and compare different low-dimensional ODE approximations when the network is given by a cycle graph and the process is the SIS dynamics.

The structure of the paper is as follows. In Section 2 we introduce the exact mathematical model, present the algorithm of the Monte-Carlo simulation and the well known low dimensional ODE approximations. In Section 3 we derive an approximating master equation, for which the state space consists of $N+1$ elements compared to the full state space with 2^N states. In Section 4 a new approach is applied that is based on the observation that the epidemic spreads along the cycle graph as a front. We introduce a continuous time Markov chain describing the evolution

of the length of the front and use this to get the number of infected nodes. In Section 5 we apply the subsystem approximation [9] using the edges as subsystems. Finally, in the concluding section the performance of the different approximations is compared to simulation.

2 SIS epidemic on networks: master equation, simulation and known approximations

In this Section we introduce the mathematical model that will be dealt with in the paper, then the algorithm of the Monte-Carlo simulation and the widely used ODE approximations will be presented briefly. The SIS type dynamics [3] is considered on a network with N nodes. The network is given by a graph with N nodes and bi-directional edges and without self-loops. The dynamics of the process has two key stages: transmission of the disease and recovery of infectious individuals. Infection is transmitted at rate τ across every (S, I) edge. Infectious individuals recover at rate γ. Upon recovery, infectious individuals become susceptible again. Both infection and recovery are modelled as independent Poisson processes. This means that in a short time interval δt, a susceptible individual with k infectious neighbours becomes infected with probability $1 - \exp(-k\tau\delta t)$, and an infectious individual recovers with probability $1 - \exp(-\gamma\delta t)$, independently of the state of its neighbours. The state space is the set $\{0,1\}^N$ and the process can be given by a continuous time Markov chain on this state space. The master equation, forming a system of 2^N linear ODEs, is formulated in the next subsection based on our results in [10]. Solving this system of ODEs is impossible due to the large number of equations. Therefore other approaches are applied to study the process. The first one is to apply Monte-Carlo simulation that will be presented in Subsection 2.2. Another possibility is to derive nonlinear ODE approximations of the master equation that will be shown in Subsection 2.3.

2.1 Master equation

The 2^N elements of the state space can be grouped into $N+1$ subsets as follows: $\mathcal{S}^k$ is the set of $\binom{N}{k}$ states with k number of Is. The elements of the class $\mathcal{S}^k$ are denoted by $\mathcal{S}^k_1, \mathcal{S}^k_2, \ldots, \mathcal{S}^k_{N_k}$, where $N_k = \binom{N}{k}$. The l-th element of state $\mathcal{S}^k_j$ will be denoted by $\mathcal{S}^k_j(l)$, thus $\mathcal{S}^k_j(l) = S$ if in the state $\mathcal{S}^k_j$ the l-th node is in state S, and $\mathcal{S}^k_j(l) = I$ if in the state $\mathcal{S}^k_j$ the l-th node is in state I.

Let us denote the probability of the system being in state $\mathcal{S}^k_i$ at time t by $X_{\mathcal{S}^k_i}$. Let

$$X_k := (X_{\mathcal{S}^k_1}, X_{\mathcal{S}^k_2}, \ldots, X_{\mathcal{S}^k_{N_k}})^T$$

be an N_k-dimensional vector for $k = 0, 1, \ldots, N$. The master equation takes the form

$$\dot{X}_k = A_{k-1}X_{k-1} + B_kX_k + C_{k+1}X_{k+1}, \quad k = 0, 1, \ldots, N,$$

where A_{-1} and C_{N+1} are zero matrices, see [10]. The entries in A_k are responsible for the infection process, while those in C_k determine recovery. Their exact formulation can be found in [10]. The matrix B_k is diagonal, with main diagonal $-(e_{k-1}C_k + e_{k+1}A_k)$ where $e_k = (1, \ldots, 1)$ is an N_k dimensional vector, every coordinate of which is 1. The master equation can be written in matrix form as

$$\dot{X} = PX, \tag{1}$$

where

$$P = \begin{pmatrix} B_0 & C_1 & 0 & 0 & 0 & 0 \\ A_0 & B_1 & C_2 & 0 & 0 & 0 \\ 0 & A_1 & B_2 & C_3 & 0 & 0 \\ 0 & 0 & A_2 & B_3 & C_4 & 0 \\ \vdots & \vdots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & A_{N-1} & B_N \end{pmatrix}.$$

We emphasize that even to formulate this matrix for a given graph is far from obvious from the algorithmic point of view since for a graph with N nodes this is a 2^N-by-2^N matrix.

2.2 Monte-Carlo simulation

Here we briefly present the exact algorithm of the simulation, since our approximating results will be compared to simulation ones. Let us denote the adjacency matrix of the graph by A and the states of the nodes at time t by $x(t) \in \{0,1\}^N$. The k-th coordinate of the state vector $x_k(t)$ is equal to 1 if the k-th node is infected, and it is 0 when the node is susceptible. Introducing the notation $x * y = (x_1y_1, x_2y_2, \dots, x_Ny_N)$ for two N-dimensional vectors x and y, the coordinates of $Ax(t) * (e - x(t))$ yield the number of infected neighbours of S nodes and they are zero for I nodes. In each step of the Monte-Carlo simulation a susceptible node can become infected or an infected one can become susceptible. In order to decide at which nodes happen transition a vector $r \in [0,1]^N$ containing random numbers is generated. The condition for the change in the k-th node in the time interval $t + \Delta t$ can be given in the form

$$r_k < \left(Ax(t) * (e - x(t))\tau + x(t)\gamma\right)_k \Delta t,$$

where Δt is chosen to be sufficiently small. Let us introduce the vector $v \in \{0,1\}^N$, the k-th coordinate of which is 1, if the above condition holds, that is

$$v = \frac{1}{2}\left(sign[Ax(t) * (e - x(t))\tau\Delta t + x(t)\gamma\Delta t - r] + e\right),$$

where the *sign* function is used coordinate-wise. Clearly, at the node k a change takes place in the time interval $[t, t + \Delta t]$ if and only if $v_k = 1$. Thus, after the small time Δt the state vector becomes

$$x(t + \Delta t) = x(t) + v * (e - 2x(t)), \tag{2}$$

since the coordinates of $e - 2x(t)$ are $+1$ or -1 when the corresponding nodes are susceptible and infected, respectively.

Then starting from an initial state vector $x(0)$ the state vectors at the times $\Delta t, 2\Delta t, \dots M\Delta t$ can be determined by (2). Once this algorithm is carried out then it is repeated several times and the average of the state vectors is taken as an approximation of the process.

2.3 Mean-field and pair-wise approximations

Concerning the SIS epidemic on a network the quantity of main interest is the expected value of infected nodes at a given time t, that is denoted by $[I](t)$. It was generally accepted that this

function satisfies the differential equation

$$[\dot{I}] = \tau[SI] - \gamma[I], \tag{3}$$

where $[SI]$ denotes the expected value of the SI edges. It was proved rigorously in [10] that this differential equation holds for an arbitrary graph. Obviously, this is not a proper differential equation for $[I]$, because $[SI]$ is also an unknown quantity. To close the differential equation we need to express $[SI]$ in terms of $[I]$. There is no exact relation between these two quantities, even for simple graphs, therefore approximating closure relations are used. For a regular random graph, for which every node has the same degree, denoted by n, a simple heuristic combinatorial argument yields the following approximate closure relation

$$[SI] \approx \frac{n}{N-1}(N - [I](t))[I](t).$$

Using this relation we get the mean-field equation

$$\dot{\tilde{I}} = \tau\frac{n}{N-1}\tilde{I}(N - \tilde{I}) - \gamma\tilde{I} \tag{4}$$

as an approximating differential equation, with the new approximate function $\tilde{I}$. To see the accuracy of this approximation, it is usually compared to the Monte-Carlo simulation in the literature. In the case of a completely connected graph the above relation performs well, while for a cycle graph its accuracy is poor. For regular random graphs it gives a much better approximation as it is shown in Figure 1.

The approximation in (4) is based on a relation between $[SI]$ and $[I]$. Another approach to close the differential equation (3) is to derive an equation for $[SI]$. This leads to the well known pair approximations, for a review see [7]. It has been proved in [11] and, by using a different approach, in [9] that for an arbitrary graph the following differential equations hold

$$\begin{aligned}
[\dot{I}] &= \tau[SI] - \gamma[I], & (5)\\
[\dot{SI}] &= \gamma([II] - [SI]) + \tau([SSI] - [ISI] - [SI]), & (6)\\
[\dot{II}] &= -2\gamma[II] + 2\tau([ISI] + [SI]), & (7)\\
[\dot{SS}] &= 2\gamma[SI] - 2\tau[SSI], & (8)
\end{aligned}$$

where $[II]$ and $[SS]$ denote the expected values of II and SS pairs, $[SSI]$ and $[ISI]$ denote the expected values of these types of triples. This system is still not closed, but using again a heuristic combinatorial consideration an approximate relation can be derived for the expected value of this triples, in terms of the expected value of the pairs. These formulas are called moment closure approximations and take the form [7]

$$[SSI] \simeq \frac{n-1}{n}\frac{[SS][SI]}{[S]}, \quad [ISI] \simeq \frac{n-1}{n}\frac{[SI]^2}{[S]}$$

using that $[IS] = [SI]$. Substituting these formulas into system (5) - (8) we get a closed system, that is called the pair approximation model. The comparison of the results obtained by this approximation and by simulation is shown in Figure 1 for a regular random graph and for the cycle graph.

3 Master equation for the number of infected nodes

In this Section we show a new approach to reduce the size of system (1) from 2^N to $N+1$. This can be achieved by introducing a new Markov chain with state space $\{0,1,\dots,N\}$. Let $x_k(t)$ denote the probability that there are k infected nodes at time t (with a given initial state that is not specified at the moment). Assuming that starting from state k the system can move to either state $k-1$ by recovery, or to state $k+1$ by infection, the master equations of the Markov chain take the form

$$\dot{x}_k = a_{k-1}x_{k-1} - (a_k + c_k)x_k + c_{k+1}x_{k+1}, \quad k = 0,\dots,N. \tag{9}$$

It is known [10] that for a completely connected graph with N nodes system (1) reduces to (9) with coefficients

$$a_k = \tau k(N-k), \quad c_k = \gamma k \quad \text{for} \quad k = 0,\dots,N \quad \text{and} \quad a_{-1} = c_{N+1} = 0, \tag{10}$$

by defining $x_k = e_k X_k$, i.e. x_k is the sum of the coordinates in X_k.

For homogeneous random graphs, where every node has $n(< N)$ links to other nodes in the network, the state space is much larger, because the above reduction, based on the symmetry of the complete graph, cannot be carried out. However, (9) can be used as an approximating system with the transition rates

$$a_k = \tau n k\frac{N-k}{N-1}, \quad c_k = \gamma k \quad \text{for} \quad k = 0,\dots,N \quad \text{with} \quad a_{-1} = c_{N+1} = 0. \tag{11}$$

These coefficients cannot be derived by using the lumping technique introduced in [10], hence (9) is not an exact system for homogeneous random graphs.

In the case of a cycle graph, which is a regular graph with nodes of degree 2, the above choice of the coefficients fails to work, the infection rate a_k is overestimated by the above formula. In this section we show two new approximating formulas for a_k that perform significantly better numerically in the cases we consider.

In order to derive the reduced system (9) from the full system (1) we lump the states in class $\mathcal{S}^k$ together for $k = 0,1,\dots,N$. Let us consider the master equation

$$\dot{X}_k = A_{k-1}X_{k-1} + B_k X_k + C_{k+1}X_{k+1}$$

belonging to the states in class $\mathcal{S}^k$ and take the sum of the equations in this system. Then for the new probabilities

$$x_k = e_k X_k = X_{\mathcal{S}_1^k} + X_{\mathcal{S}_2^k} + \ldots + X_{\mathcal{S}_{N_k}^k}$$

we get

$$\dot{x}_k = e_k A_{k-1}X_{k-1} + e_k B_k X_k + e_k C_{k+1}X_{k+1}.$$

Since in every coloumn of C_{k+1} there are $k+1$ entries that are equal to γ and the other entries are zeros (see [10]) we have $e_k C_{k+1} = (k+1)\gamma e_{k+1}$, hence

$$e_k C_{k+1}X_{k+1} = (k+1)\gamma e_{k+1}X_{k+1} = (k+1)\gamma x_{k+1}.$$

Moreover, we also know from [10] that $e_k A_{k-1} = \tau \cdot N_{SI}(\mathcal{S}^{k-1})$, where

$$N_{SI}(\mathcal{S}^{k-1}) = (N_{SI}(\mathcal{S}_1^{k-1}), N_{SI}(\mathcal{S}_2^{k-1}), \ldots, N_{SI}(\mathcal{S}_{N_k}^{k-1}))$$

and $N_{SI}(\mathcal{S}_j^{k-1})$ is the number of SI edges in the state $\mathcal{S}_j^{k-1}$. Hence the above differential equation for x_k takes the form

$$\dot{x}_k = \tau \cdot N_{SI}(\mathcal{S}^{k-1}) X_{k-1} + e_k B_k X_k + (k+1)\gamma x_{k+1}.$$

Since our goal is to derive the reduced system (9) we have to find a coefficient a_k for which

$$\tau \cdot N_{SI}(\mathcal{S}^k) X_k = a_k x_k$$

holds in exact or at least in approximating sense. Let us consider first the simplest case of a complete graph. Then in state $\mathcal{S}_j^k$ the number of SI edges is $k(N-k)$, for any $j = 1, 2, \ldots, N_k$, implying

$$N_{SI}(\mathcal{S}^k) = k(N-k)e_k. \tag{12}$$

Hence using that $x_k = e_k X_k$ we obtain

$$N_{SI}(\mathcal{S}^k) X_k = k(N-k) e_k X_k = k(N-k) x_k,$$

yielding $a_k = \tau k(N-k)$ as it was given in (10). For an arbitrary graph the number of SI edges in the states $\mathcal{S}_j^k$ are different for different values of j. Hence instead of (12) we can have only an approximating equation

$$N_{SI}(\mathcal{S}^k) \approx e_{SI}(k) e_k, \tag{13}$$

with a suitably chosen artificial parameter $e_{SI}(k)$ that will be referred to as average number of SI edges for the states with k infected nodes. Then repeating the above derivation we arrive to (9) with

$$a_k = \tau e_{SI}(k). \tag{14}$$

In the next two subsections we show two different methods to derive theoretical formulas for the average number of SI edges $e_{SI}(k)$.

3.1 Approximation of the expected value of the SI edges by state weighting

The main idea here is to express $e_{SI}(k)$ as a weighted average for the states in class $\mathcal{S}^k$ as

$$e_{SI}(k) = \sum_{j=1}^{N_k} N_{SI}(\mathcal{S}_j^k) \cdot w_{kj}, \quad k = 0, \ldots, N, \tag{15}$$

where w_k is an N_k dimensional vector of weights, and the sum of the weights is 1, i.e. $\sum_{j=1}^{N_k} w_{kj} = 1$.

The simplest choice for w_k is using equal weights, that is $w_{kj} = 1/N_k$ for all $j = 1, 2, \ldots, N_k$. If the graph is regular, i.e. all nodes have the same degree n, then simple combinatorial arguments show that this equal weighting leads to the coefficients given in (11). Thus we have the following proposition.

Proposition 1 *Let us assume that the graph is regular, i.e. all nodes have the same degree n. If all the weights are equal, that is $w_{kj} = 1/N_k$, then*

$$e_{SI}(k) = \frac{1}{N_k} \cdot \sum_{j=1}^{N_k} N_{SI}(\mathcal{S}_j^k) = \frac{n}{N-1} \cdot k \cdot (N-k). \tag{16}$$

According to our numerical investigation, for random regular graphs the results obtained from the master equation (9) with coefficients (11) are in good agreement with simulation. However, for a cycle graph the performance of (9) is much worse, showing that the choice of equal weights does not catch the real situation, see Figure 2. The reason of that can easily be seen intuitively. Namely, those states in $\mathcal{S}^k$ where the infected nodes are in one group are much probable than those where the infected nodes are scattered along the cycle graph.

Now we show a better performing weighting based on the leading eigenvector corresponding to the quasi steady state. The leading eigenvector of P in system (1) is $(1, 0, \ldots, 0)^T$ belonging to the leading eigenvalue $\lambda_0 = 0$. This eigenvector corresponds to the steady state when all nodes are of type S. The solution converges to the steady state, however, it takes extremely long time to get close to this steady state [8]. Therefore the second eigenvalue of P, denoted by λ_1, plays an important role in the long time behaviour of the system, that is it corresponds to the quasi steady state. We note that this eigenvalue is also close to zero, which is the reason of the fact that the corresponding eigenvector, v_1, yields a quasi steady state.

Instead of finding the second eigenvalue of P we approximate v_1 by the leading eigenvector of

$$\widetilde{P} = \begin{pmatrix} B_1 & C_2 & 0 & 0 & 0 \\ A_1 & B_2 & C_3 & 0 & 0 \\ 0 & A_2 & B_3 & C_4 & 0 \\ \vdots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & A_{N-1} & B_N \end{pmatrix},$$

that is obtained from P by omitting the first row and first coloumn and then modifying the matrix B_1 to get a diagonal matrix with main diagonal $-e_2 A_1$. This ensures that the sum of the entries in every coloumn is zero. Then zero is an eigenvalue of $\widetilde{P}$ and its corresponding eigenvector u gives a good approximation of v_1 once the first coordinate of v_1 is omitted. This approach can also be interpreted as the state $\mathcal{S}^0$, when all nodes are of type S, is omitted from the state space. The steady state of this new Markov process is the quasi steady state of the original one. Now the eigenvector u will determine the weights in $e_{SI}(k)$.

Proposition 2 *For an arbitrary graph let us construct the matrix P in the master equation (1) and let $\widetilde{P}$ be given as above. Let u be the eigenvector of $\widetilde{P}$ belonging to the eigenvalue 0, and let u_k be the part of u belonging to the states in class $\mathcal{S}^k$. That is $u = (u_1, u_2, \ldots, u_N)^T$ and u_k has N_k entries. Let w_k be the vector of weights given by u_k, that is*

$$w_k = \frac{u_k}{e_k u_k}, \quad k = 1, \ldots, N.$$

If $e_{SI}(k)$ is given by (15), then for all $k = 1, \ldots, N-1$ we have

$$e_{SI}(k) = (k+1) \cdot \frac{\gamma}{\tau} \cdot \frac{e_{k+1} u_{k+1}}{e_k u_k}.$$

PROOF The eigenvector equation $\widetilde{P}u = 0$ can be written as

$$\begin{aligned} B_1u_1 + C_2u_2 &= 0, \\ A_1u_1 + B_2u_2 + C_3u_3 &= 0, \\ A_2u_2 + B_3u_3 + C_4u_4 &= 0, \\ &\vdots \\ A_{N-1}u_{N-1} + B_Nu_N &= 0. \end{aligned}$$

The sum of the entries in every coloumn is zero, hence for $k = 2, \ldots, N-1$ we have $e_kB_k = -(e_{k-1}C_k + e_{k+1}A_k)$, moreover, $e_1B_1 = -e_2A_1$ and $e_NB_N = -e_{N-1}C_N$ hold for $k = 1$ and $k = N$. Let us multiply the k-th equation of the above system by e_k and substitute these expressions for e_kB_k. This leads to

$$\begin{aligned} -e_2A_1u_1 + e_1C_2u_2 &= 0, \\ e_2A_1u_1 - (e_1C_2 + e_3A_2)u_2 + e_2C_3u_3 &= 0, \\ &\vdots \\ e_NA_{N-1}u_{N-1} - e_{N-1}C_Nu_N &= 0. \end{aligned}$$

Adding the first k equations we get

$$e_kC_{k+1}u_{k+1} = e_{k+1}A_ku_k.$$

Now we can make use of formulas $e_{k-1}C_k = k\gamma e_k$ and $N_{SI}(\mathcal{S}^k) = \frac{1}{\tau} \cdot e_{k+1}A_k$ that were already used in the beginning of this section and were first proved in [10]. Using the first one the above equation takes the form

$$(k+1)\gamma e_{k+1}u_{k+1} = e_{k+1}A_ku_k,$$

then using the second one

$$e_{SI}(k) = N_{SI}(\mathcal{S}^k) \cdot w_k = \frac{1}{\tau} \cdot e_{k+1}A_k \cdot \frac{u_k}{e_ku_k} = \frac{1}{\tau} \cdot \frac{(k+1)\gamma e_{k+1}u_{k+1}}{e_ku_k} = (k+1) \cdot \frac{\gamma}{\tau} \cdot \frac{e_{k+1}u_{k+1}}{e_ku_k}.$$

□

Let us now apply this proposition to cycle graphs with $N = 5$ and $N = 10$ nodes, in order to reduce their master equation to equation (9). In the case $N = 5$ we get

$$\begin{aligned} e_{SI}(1) &= 2, \\ e_{SI}(2) &= 3 \cdot \frac{\gamma}{\tau} \cdot \frac{2\tau(3\gamma^2 + 8\tau\gamma + 6\tau^2)}{\gamma(9\gamma^2 + 21\tau\gamma + 14\tau^2)}, \\ e_{SI}(3) &= 4 \cdot \frac{\gamma}{\tau} \cdot \frac{\tau(3\gamma^2 + 10\tau\gamma + 8\tau^2)}{2\gamma(3\gamma^2 + 8\tau\gamma + 6\tau^2)}, \\ e_{SI}(4) &= 2. \end{aligned}$$

Substituting these values into (14) we get the coefficients in the master equation (9). It can be proved that for any value of N we can express $e_{SI}(k)$ as

$$e_{SI}(k) = (k+1) \cdot \frac{1}{r} \cdot \frac{p_k(r)}{q_k(r)},$$

where p_k and q_k are polynomials and $r = \frac{\tau}{\gamma}$. However, the eigenvector u is used during the calculation, hence it is computationally impossible to determine the polynomials for large N. We developed an algorithm that computes the coefficients of the polynomials, and implemented it in MATLAB. Using this code the coefficients $e_{SI}(k)$ and hence a_k were determined for a cycle graph with $N = 10$ nodes. (The formulas are too long to be presented here.) Then the solution of the master equation (9) was compared to the solution of the exact system (1), which consists of 2^{10} equations, see Figure 3. The number of infected nodes in the steady state is given accurately by the approximating system (9), since the coefficients $e_{SI}(k)$ were determined by the eigenvector corresponding to the steady state. This also explains why the approximation is less accurate in the initial stage of the spread of the epidemic.

3.2 Approximation of the expected value of the SI edges by combinatorial arguments

In the previous subsection the derivation for $e_{SI}(k)$ was based on the eigenvector corresponding to the quasi steady state. Now, we concentrate more on the expanding stage of the epidemic and start from the observation that the infected nodes form a more or less connected group along the cycle graph with a few possibly susceptible nodes in it. This observation is based on the fact that at the initial instant the infected nodes are localized to a few nodes close to each other. The number of susceptible nodes inside the group is determined by the balance of reinfection and recovery, that is $\tau \overline{e}_{SI}(k) = \gamma k$ holds, where $\overline{e}_{SI}(k)$ denotes the number if SI edges inside the group, and k is the number of infected nodes. This implies that $e_{SI}(k)$ depends linearly on k and the slope of this linear function is γ/τ. Moreover, it is obvious that for one infected node the number of SI edges is 2, that is $e_{SI}(1) = 2$, leading to

$$e_{SI}(k) = 2 + \frac{\gamma}{\tau}(k-1) \quad \text{and} \quad a_k = \tau(2 + \frac{\gamma}{\tau}(k-1)). \tag{17}$$

In the second stage of the epidemic spread, when the number of infected nodes gets closer to the steady state the balance of reinfection and recovery is violated. In this stage $e_{SI}(k)$ is more accurately approximated by (16) with $n = 2$, which is based on the assumption that there is a large group of infected nodes with susceptible ones scattered randomly. Hence we have

$$a_k = \tau \frac{2}{N-1} \cdot k \cdot (N-k),$$

when $k > k_0$, where k_0 is given by the intersection point of this parabola and the above line, i.e.

$$2 + \frac{\gamma}{\tau}(k_0 - 1) = \frac{2}{N-1} \cdot k_0 \cdot (N - k_0).$$

In Figure 4 we plotted $e_{SI}(k)$ in terms of k as it is given by (16) and (17) and also as it is measured from simulation, see the left panel. (More exactly, the average number of SI edges as a function of the expected value $[I]$ is measured from simulation. This causes that the theoretical curve in the right panel does not fit simulation well, that is explained in more detail in the Concluding section.) In the right panel of the figure the expected value of the number of infected nodes is plotted from simulation and as it is obtained from the master equation (9) with coefficients given in (16) and in (17).

4 Master equation for the length of the front

In this section a new approach is applied that is based on the observation that the epidemic spreads along the cycle graph as a front if at the initial instant the infected nodes are localized to a few nodes close to each other, that will be assumed in the following. First, we consider the front as a connected part of the cycle graph full of infected nodes. The front propagates at its two ends, by infecting a new node, or the end points can recover. The recoveries inside the front will be accounted for in a different way later. We introduce a continuous time Markov chain describing the evolution of the length of the front. The state space of this Markov chain is the set $\{1,\dots,N\}$. (The front of length 0 is neglected since those simulations, for which the epidemic does not spread (i.e. the initial infected node recovers before infecting its neighbours) are discarded.) Let us denote by $q_k(t)$, $k=1,\dots,N$ the probability that the length of the front at time t is k, meaning that the largest distant between two infected nodes is k. The system can move from state k either to state $k-1$ by recovery at the two end points of the front with rate 2γ, or to state $k+1$ by infection with rate 2τ (as it was mentioned above, the recovery of the inner points will be considered later). The master equations can be then formulated as follows.

$$\begin{aligned}
\dot{q}_1(t) &= -2\tau q_1(t) + 2\gamma q_2(t),\\
&\vdots\\
\dot{q}_k(t) &= 2\tau q_{k-1} - 2(\tau+\gamma)q_k(t) + 2\gamma q_{k+1}(t),\\
&\vdots\\
\dot{q}_N(t) &= 2\tau q_{N-1} - 2\gamma q_N(t).
\end{aligned} \tag{18}$$

Once this system is solved the expected value of the length of the front can be determined as $m(t)=\sum_{k=1}^{n} k q_k(t)$. Now, let us consider the recovery of the nodes inside the front. Based on simulation results we can claim that the expected value of the number of infected nodes is proportional to the length of the front, that is there is a constant α, such that

$$[I](t) = \alpha m(t). \tag{19}$$

This can be explained by the fact that inside the front a certain part of the infected nodes recovers. The recovery of infected nodes and the reinfection of susceptible nodes is in steady state inside the front, yielding that $\tau[SI]=\gamma[I]$, where $[SI]$ is the average number of SI edges inside the front. The simplest approximation of $[SI]$ in terms of $[I]$ is $[SI]=2[I](m-[I])/m$,

where m is the length of the front. Then equation $\tau[SI] = \gamma[I]$ yields

$$[I](t) = \left(1 - \frac{\gamma}{2\tau}\right) m(t),$$

that is $\alpha = 1 - \frac{\gamma}{2\tau}$. As we will see in Section 5 the constant α can be determined more accurately, by simply saying that $\alpha = I_s/m_s$, where I_s is the steady state value of $[I]$ obtained by solving (25) and m_s is the steady state value of m that will be determined below. For example, in the case of $\tau = 5$, $\gamma = 1$ we get $1 - \frac{\gamma}{2\tau} = 0.9$, while $\alpha = I_s/m_s = 0.88$. We note that the steady state I_s can also be obtained by determining the equilibrium of the pair approximation (5) - (8). Thus the theoretical curve in Figure 5 is obtained as follows. First, we solved (18), then the expected value of the length of the front was determined as $m(t) = \sum_{k=1}^{n} kq_k(t)$, and finally $[I]$ was computed from (19) with $\alpha = I_s/m_s$.

Starting from system (18) we can determine the stationary value of the front length analytically. The largest eigenvalue of the matrix in the right hand side of system (18) is zero, the remaining eigenvalues have negative real part. Hence the steady state of the system is determined by the eigenvector u belonging to the zero eigenvalue. It is determined by system

$$\begin{pmatrix} -2\tau & 2\gamma & 0 & \dots & & 0 \\ 2\tau & -(2\gamma+2\tau) & 2\gamma & 0 & \dots & 0 \\ & & \vdots & & & \\ 0 & \dots & 0 & 2\tau & -(2\gamma+2\tau) & 2\gamma \\ 0 & & \dots & & 2\tau & -2\gamma \end{pmatrix} \cdot u = 0.$$

Adding the first k equations we get that the coordinates of u satisfy $-2\tau u_k + 2\gamma u_{k+1} = 0$, implying $u_{k+1} = u_k \frac{\tau}{\gamma}$. Moreover, since u is a probability distribution, the sum of its coordinates is 1, hence

$$1 = \sum_{k=1}^{N} u_k = u_1 \sum_{k=1}^{N} \left(\frac{\tau}{\gamma}\right)^{k-1} = u_1 \frac{\left(\frac{\tau}{\gamma}\right)^N - 1}{\frac{\tau}{\gamma} - 1}.$$

Thus the steady state is given by

$$u_k = u_1 \left(\frac{\tau}{\gamma}\right)^{k-1}, \quad u_1 = \frac{\frac{\tau}{\gamma} - 1}{\left(\frac{\tau}{\gamma}\right)^N - 1}. \tag{20}$$

This enables us to determine explicitly the expected value of the length of the front in the steady state as follows.

$$m_s = \sum_{k=1}^{N} k u_k = u_1 \cdot \sum_{k=1}^{N} k \left(\frac{\tau}{\gamma}\right)^{k-1} = \frac{1 - (1+N)\left(\frac{\tau}{\gamma}\right)^N + N\left(\frac{\tau}{\gamma}\right)^{N+1}}{\left(\left(\frac{\tau}{\gamma}\right)^N - 1\right)\left(\frac{\tau}{\gamma} - 1\right)},$$

where we used that

$$\sum_{k=1}^{N} k x^{k-1} = \frac{1 - (1+N)x^N + Nx^{N+1}}{(x-1)^2}.$$

The above expression for m_s can be easily transformed to

$$m_s = N - \frac{\gamma}{\tau - \gamma} + O\left(\left(\frac{\gamma}{\tau}\right)^N\right).$$

The performance of this approximation is justified by numerical simulation.

Now our aim is to derive a single (at least approximating) differential equation for m that yields the length of the front without solving system (18) of N ODEs. Differentiating the function m we get

$$\dot{m}(t) = \sum_{k=1}^{n} k\dot{q}_k(t) = 2(\tau-\gamma)\sum_{k=1}^{n} q_k(t) + 2\tau q_1(t) - 2\gamma q_N(t) = 2(\tau-\gamma) + 2\gamma q_1(t) - 2\tau q_N(t). \quad (21)$$

Thus in order to derive a self-contained differential equation for m we need to express the probabilities q_1 and q_N in terms of m. We will use the functional forms

$$q_1 = L_1(m), \quad q_N = L_N(m).$$

Based on observations obtained by simulations these functions can be approximated by piece-wise linear functions satisfying

$$L_1(1) = 1,\ L_1(h_1 N) = 0,\ L_N(h_2 N) = 0,\ L_N(N) = u_N,$$

where h_1 and h_2 are artificial parameters that can be estimated by using the simulation, and u_N is the steady state value of q_N given in (20). That is

$$L_1(m) = \begin{cases} \frac{m-1}{1-h_1 N} + 1, & \text{if } 1 \le m < h_1 N, \\ 0, & \text{if } h_1 N \le m < N. \end{cases}$$

and

$$L_N(m) = \begin{cases} 0, & \text{if } 1 < m < h_2 N, \\ \frac{u_N(m-h_2 N)}{N-h_2 N}, & \text{if } h_2 N \le m < N, \end{cases}$$

Thus equation (21) takes the form

$$\dot{m} = 2(\tau-\gamma) + 2\gamma L_1(m) - 2\tau L_N(m). \quad (22)$$

Using again (19) with $\alpha = I_s/m_s$ we obtain $[I]$ as a function of time, and then it can be compared to simulation. This comparison is shown also in Figure 5, with the parameter values $h_1 = 0.1$, $h_2 = 0.9$ that were determined based on simulation.

5 Subsystem approximation at the level of pairs

In this section we apply the subsystem approach developed by Sharkey in [9] in the case of a cycle graph. Now the state place consists of the possible states of each edge. If the k-th node is in state A and the $k+1$-th node is in state B, then we say the k-th edge (linking the k-th and

the $k+1$-th nodes) is in state AB. The probability that this edge is in state AB is denoted by $y_{AB}^{k,k+1}$, where AB can be IS, SI, SS or II. Obviously,

$$y_{IS}^{k,k+1} + y_{SI}^{k,k+1} + y_{SS}^{k,k+1} + y_{II}^{k,k+1} = 1 \tag{23}$$

holds for $k = 1, \ldots, N$, where the edge $(N, N+1)$ is by definition the edge $(N, 1)$. Using the general subsystem method in [9] the master equations for the probabilities of the states can be written as

$$\begin{aligned}
\dot{y}_{SI}^{k,k+1} &= -(\gamma+\tau)y_{SI}^{k,k+1} + \gamma y_{II}^{k,k+1} + \tau y_{SSI}^{k,k+1,k+2} - \tau y_{ISI}^{k-1,k,k+1}, \\
\dot{y}_{IS}^{k,k+1} &= -(\gamma+\tau)y_{IS}^{k,k+1} + \gamma y_{II}^{k,k+1} + \tau y_{ISS}^{k-1,k,k+1} - \tau y_{ISI}^{k,k+1,k+2}, \\
\dot{y}_{SS}^{k,k+1} &= \gamma(y_{SI}^{k,k+1} + y_{IS}^{k,k+1}) - \tau y_{SSI}^{k,k+1,k+2} - \tau y_{ISS}^{k-1,k,k+1}, \\
\dot{y}_{II}^{k,k+1} &= -(\dot{y}_{SI}^{k,k+1} + \dot{y}_{IS}^{k,k+1} + \dot{y}_{SS}^{k,k+1}),
\end{aligned} \tag{24}$$

where $y_{ABC}^{k-1,k,k+1}$ denotes the probability that the triple $k-1, k, k+1$ is in state ABC. Note that the last equation is the result of the property (23), thus it can be omitted.

To form a self-contained system of differential equations at the pair level we apply the Kirkwood closure for the probabilities of the triples [9]:

$$y_{ABC}^{k-1,k,k+1} = y_{AB}^{k-1,k}\frac{y_{BC}^{k,k+1}}{y_{BC}^{k,k+1} + y_{B\neg C}^{k,k+1}} = \frac{y_{AB}^{k-1,k}}{y_{AB}^{k-1,k} + y_{\neg AB}^{k-1,k}} y_{BC}^{k,k+1},$$

where $y_{B\neg C}^{k,k+1}$ denotes the probability that the k-th node is in state B, and the $k+1$-th node is not in state C and similarly, $y_{\neg AB}^{k-1,k}$ denotes the probability that the $k-1$-th node is not in state A, and the k-th node is in state B.

$y_{A\neg B}^{k,k+1}$ denotes the probability that the k-th node is in state A, and the $k+1$-th node is not in state B.

Substituting these closure relations into equations (24), we obtain

$$\begin{aligned}
\dot{y}_{SI}^{k,k+1} &= -(\gamma+\tau)y_{SI}^{k,k+1} + \gamma y_{II}^{k,k+1} + \tau y_{SS}^{k,k+1}\frac{y_{SI}^{k+1,k+2}}{y_{SI}^{k+1,k+2} + y_{SS}^{k+1,k+2}} - \tau\frac{y_{IS}^{k-1,k}}{y_{IS}^{k-1,k} + y_{SS}^{k-1,k}} y_{SI}^{k,k+1}, \\
\dot{y}_{IS}^{k,k+1} &= -(\gamma+\tau)y_{IS}^{k,k+1} + \gamma y_{II}^{k,k+1} + \tau\frac{y_{IS}^{k-1,k}}{y_{IS}^{k-1,k} + y_{SS}^{k-1,k}} y_{SS}^{k,k+1} - \tau y_{IS}^{k,k+1}\frac{y_{SI}^{k+1,k+2}}{y_{SI}^{k+1,k+2} + y_{SS}^{k+1,k+2}}, \\
\dot{y}_{SS}^{k,k+1} &= \gamma(y_{SI}^{k,k+1} + y_{IS}^{k,k+1}) - \tau y_{SS}^{k,k+1}\frac{y_{SI}^{k+1,k+2}}{y_{SI}^{k+1,k+2} + y_{SS}^{k+1,k+2}} - \tau\frac{y_{IS}^{k-1,k}}{y_{IS}^{k-1,k} + y_{SS}^{k-1,k}} y_{SS}^{k,k+1}.
\end{aligned} \tag{25}$$

We solved this system numerically by using a MATLAB ODE solver. The probabilities of the pairs $y_{II}^{k,k+1}$ were obtained from $y_{II}^{k,k+1} = 1 - y_{SI}^{k,k+1} - y_{IS}^{k,k+1} - y_{SS}^{k,k+1}$. To calculate the probabilities y_I^k at each node (y_I^k denotes the probability that the k-th node is infected), we applied the formula $y_I^k = \frac{1}{2}(y_{SI}^{k-1,k} + y_{IS}^{k,k+1} + y_{II}^{k-1,k} + y_{II}^{k,k+1})$. Then the expected value of the number of infected nodes was determined by $[I] = \sum_{k=1}^{n} y_I^k$. The result is compared to simulation in Figure 6. The method overestimates the speed of the spread of the disease, but gives a good approximation for the ratio of the infected nodes of the total population in the steady state. The steady state value I_s was used in Section 4 to determine α in (19) as $\alpha = I_s/m_s$.

6 Conclusion

The goal of this paper was to derive low dimensional ODE approximations that capture the main characteristics of SIS-type epidemic propagation along a cycle graph. This research is motivated by the fact that usual ODE approximations, like the mean-field equation and the pair approximation, fail to work for a graph with a special structure. We introduced three different methods to get the number of infected nodes as a function of time.

The first method is based on the derivation of a master equation for the number of infected nodes. This uses $e_{SI}(k)$, the average number of SI edges for a given number, k, of the infected nodes. In order to define formulas for this quantity we introduced two different methods. The first method starts from the leading eigenvector of the matrix in the right hand side of the master equation of the full system. Hence it is computationally demanding, since the dimension of this vector is 2^N. However, in exchange this gives very accurate approximations for small values of N. On the other hand, a simple formula (17) was derived for $e_{SI}(k)$ based on combinatorial arguments. This can be easily determined for large N, but it is less accurate. The accuracy of this method can be increased if we measure $e_{SI}(k)$ as a function of k from simulation and then derive a better approximating $e_{SI}(k)$ curve. The curve shown in the left panel of Figure 4 shows the average number of SI edges as a function of the expected value $[I]$, not as a function of the exact number of infected nodes, k.

Our second approach was based on the observation that the epidemic spreads along the cycle graph as a front. We introduced a continuous time Markov chain describing the evolution of the length of the front. The state space of this Markov chain contained N elements, hence the master equation can be solved for large values of N. This way we got a very accurate approximation for the length of the front, but then the number of infected nodes was determined as a constant multiple of the frontlength. The accuracy of the approximation depends on the accuracy of this constant that is hard to estimate theoretically. We showed two different estimations for the value of this constant.

The third method we applied was the subsystem approximation using the edges as subsystems. This can be carried out for large values of N since the number of equations is of order N. This method gives a good approximation for the steady state but does not perform well in the first stage of the spread of the epidemic.

As a comparison of the different methods we computed the steady state value of the number of infected nodes in five different ways: Monte-Carlo simulation, mean-field equation, pair approximation, using master equation (9) with coefficients given in (17) and finally using the subsystem approach by solving (25). The steady state values obtained by these methods are given in Table 1 for different values of τ (and fixing $\gamma = 1$). We can observe that the mean-field equation and the master equation (9) yield similar results, the reason of which can be that both models are given at the level of nodes and the closure is at the level of pairs. On the other hand, the pair approximation and the subsystem approach yield also similar values, because both of them are closed at the level of triples.

References

[1] A. Barrat, M. Barthelemy, A. Vespignani, Dynamical Processes on Complex Networks, Cambridge University Press, Cambridge, 2008.

[2] Bollobás, B., *Random graphs*, Cambridge University Press, 2001.

[3] Brauer, F., van den Driessche, P. & Wu, J. Mathematical epidemiology, In Lecture Notes in Mathematics, Springer-Verlag Berlin Heidelberg, 2008.

[4] L. Danon, A.P. Ford, T. House, C.P. Jewell, M.J. Keeling, G.O. Roberts, J.V. Ross, M.C. Vernon 2011. Networks and the Epidemiology of Infectious Disease, Interdisciplinary Perspectives on Infectious Diseases 2011:284909 special issue "Network Perspectives on Infectious Disease Dynamics".

[5] Gleeson JP, High-accuracy approximation of binary-state dynamics on networks, *Phys. Rev. Letters* **107** (2011), 068701.

[6] T. House, M. J. Keeling, Insights from unifying modern approximations to infections on networks, *J. Roy. Soc. Interface* **8** (2011), 67-73.

[7] M.J. Keeling, K.T.D. Eames, Networks and epidemic models, *J. Roy. Soc. Interface* **2** (2005), 295-307.

[8] Nåsell I., The quasi-stationary distribution of the closed endemic SIS model, *Adv. Appl. Probab.* **28** (1996), 895 – 932.

[9] K. J. Sharkey, Deterministic epidemic models on contract networks: Correlations and unbiological terms, *Theor. Popul. Biol.* **79** (2011), 115-29.

[10] Simon, P.L., Taylor, M., Kiss, I.Z., Exact epidemic models on graphs using graph automorphism driven lumping, *J. Math. Biol.* **62** (2010), 479-508.

[11] Taylor, M., Simon, P. L., Green, D. M., House, T., Kiss, I. Z., From Markovian to pairwise epidemic models and the performance of moment closure approximations, *J. Math. Biol.* (2012) DOI: 10.1007/s00285-011-0443-3.

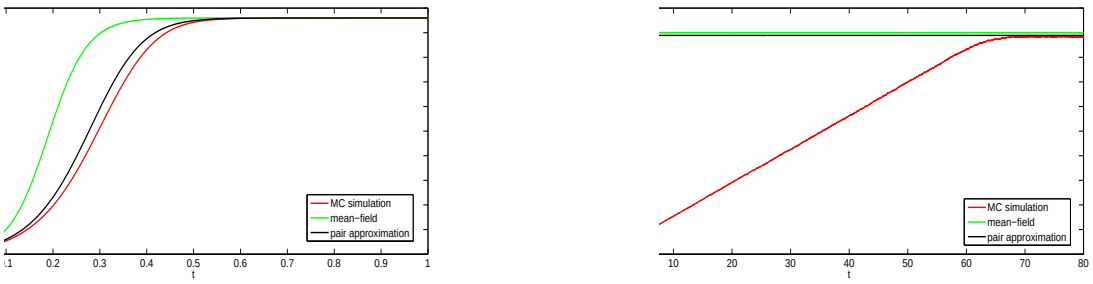

Figure 1: The expected value of the infected nodes obtained from Monte-Carlo simulation, from the mean-field approximation (4) and from the pair approximation (5) - (8). The comparison is shown for a regular random graph (left panel) with $N = 500$ nodes, average degree $n = 5$, $\gamma = 1$, $\tau = 5$, and for the cycle graph (right panel) with $N = 500$, $\gamma = 1$, $\tau = 5$.

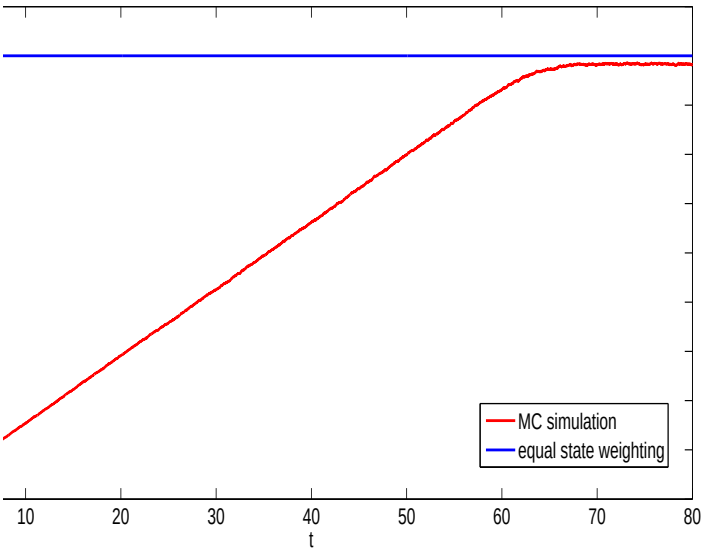

Figure 2: The expected value of the infected nodes obtained from Monte-Carlo simulation and from the approximation (9) with coefficients (14) given by (16). The comparison is shown for the cycle graph with $N = 500$, $\gamma = 1$, $\tau = 5$.

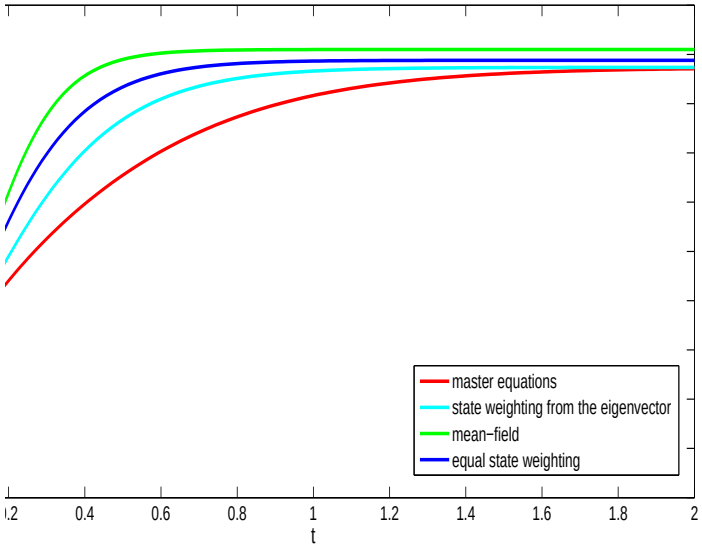

Figure 3: The expected value of the infected nodes given by the exact master equation (1) (red), by the approximating master equation (9) with coefficients obtained by state weighting (light blue) and with coefficients given by equal weighting (dark blue), and by the mean-field equation (4) (green), for a cycle graph with $N = 10$ nodes, $\gamma = 1$, $\tau = 5$.

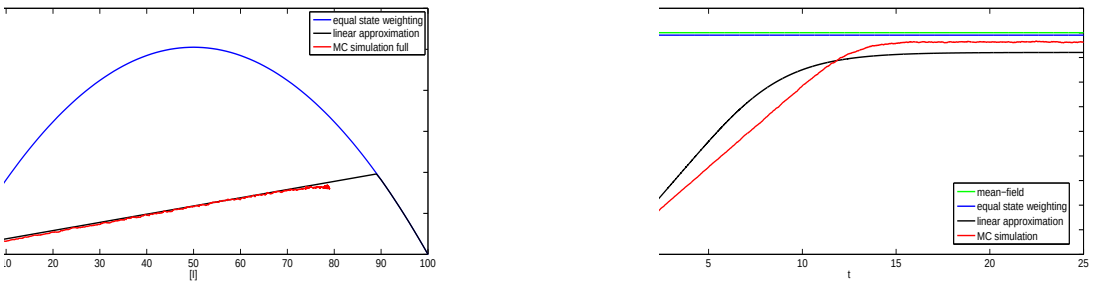

Figure 4: The average number of SI edges in terms of $[I]$, i.e. $e_{SI}(k)$ is shown in the left panel, as it is obtained from simulation, from (16) and from (17), for a cycle graph with $N = 100$, $\gamma = 1$, $\tau = 5$. The right panel shows the expected value of the infected nodes given by simulation and by the approximating master equation (9) with coefficients given in (16) and in (17).

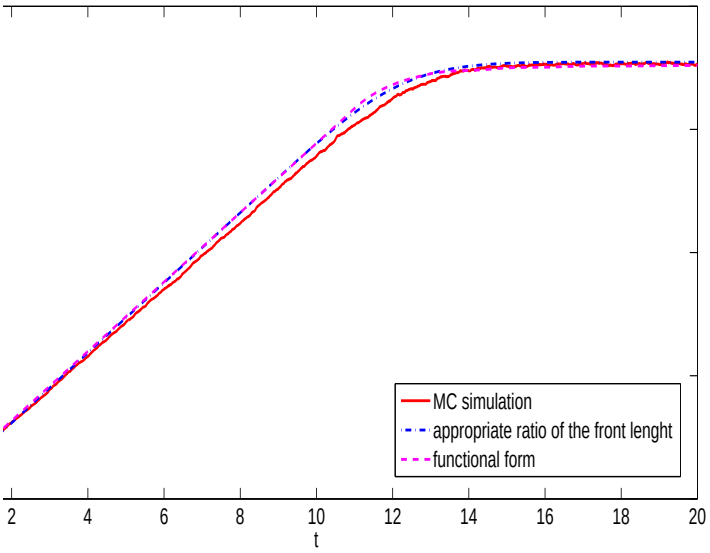

Figure 5: The expected value of the infected nodes obtained from simulation (continous red line) and from the length of the front $m(t)$ using (19) with $\alpha = I_s/m_s$. The length of the front is determined in two different ways: by solving (18) (dashed-dotted blue line) and by solving (22) with $h_1 = 0.1$, and $h_2 = 0.9$ (dashed magenta line), for a cycle graph with $N = 100$, $\gamma = 1$, $\tau = 5$.

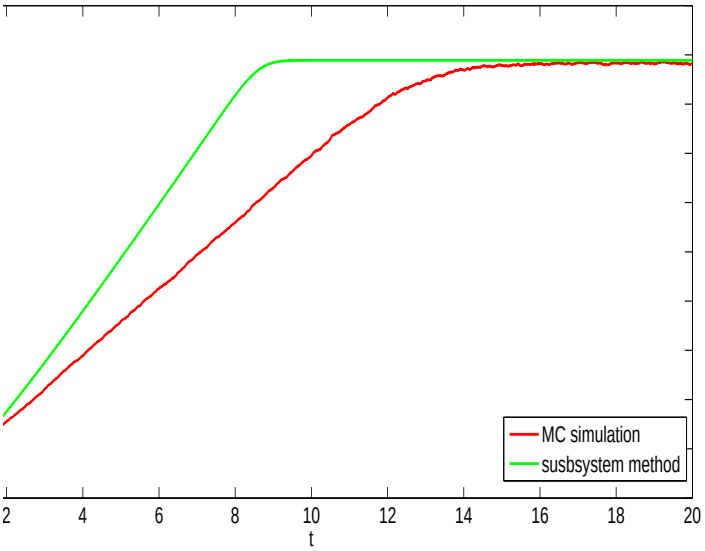

Figure 6: The expected value of the infected nodes obtained from simulation and by solving (25), for a cycle graph with $N = 100$, $\gamma = 1$, $\tau = 5$.

Table 1: Steady state value of the number of infected nodes for different values of τ as it is obtained in the following five different ways: Monte-Carlo simulation, mean-field equation (4), pair approximation (5) - (8), using master equation (9) with coefficients given in (17) and finally using the subsystem approach by solving (25).

τ	Steady state				
	MC simulation	mean-field	pair approximation	master equation	subsystem method
2	0.6127	0.7512	0.6666	0.7496	0.6667
5	0.8763	0.9005	0.8889	0.8999	0.8889
10	0.9445	0.9503	0.9474	0.95	0.9474