Math. Model. Nat. Phenom.
Vol. ?, No. ?, 2013, pp. ??

Approximate master equations for dynamical processes on graphs

Noémi Nagy[a], **Istvan Z. Kiss** [b], **Péter L. Simon**[a 1]

[a] Institute of Mathematics, Eötvös Loránd University Budapest, and
Numerical Analysis and Large Networks Research Group, Hungarian Academy of Sciences, Hungary
[b] School of Mathematical and Physical Sciences,
Department of Mathematics, University of Sussex, Falmer, Brighton BN1 9QH, UK

Abstract. We extrapolate from the exact master equations of epidemic dynamics on fully connected graphs to non-fully connected by keeping the size of the state space $N + 1$, where N is the number of nodes in the graph. This gives rise to a system of approximate ODEs (ordinary differential equations) where the challenge is to compute/approximate analytically the transmission rates. We show that this is possible for graphs with arbitrary degree distributions built according to the configuration model. Numerical tests confirm that: (a) the agreement of the approximate ODEs system with simulation is excellent and (b) that the approach remains valid for clustered graphs with the analytical calculations of the transmission rates still pending. The marked reduction in state space gives good results, and where the transmission rates can be analytically approximated, the model provides a strong alternative approximate model that agrees well with simulation. Given that the transmission rates encompass information both about the dynamics and graph properties, the specific shape of the curve, defined by the transmission rate versus the number of infected nodes, can provide a new and different measure of network structure, and the model could serve as a link between inferring network structure from prevalence or incidence data..

Key words: SIS epidemic, ODE approximation, network process
AMS subject classification: 05C82, 37N25, 60J28, 90B15

[1]Corresponding author. E-mail: simonp@cs.elte.hu

1. Introduction

Concerted efforts of the analysis of different ODE (ordinary differential equation) models of various dynamics on networks has led to a better understanding of how these models relate to each other [7, 15, 16], what are the assumptions that these rely on, and whether these models can serve as the limiting case of stochastic or exact models in some well defined limit [1, 3, 14]. The specific limits may typically depend on the size and type of the network or the time horizon over which agreement is sought. The main candidate models include the pairwise [8], edge-based compartmental models [12], as well as effective-degree type models [10, 11] which have mainly originated from epidemic models such as the SIS and SIR (S - susceptible, I - infected and infectious and R - recovered or removed). For all these models, the primary test of their performance, is in terms of the agreement between the time evolution of some expected quantity (e.g. the expected number of infectious nodes/vertices as a function of time) from the exact model or simulation and the equivalent quantity based on the ODE model.

It is well known that the derivation of such models poses less of a challenge when attempted for and SIR model. This is due to the SIR dynamics avoiding the problem of having to account for stronger dependence of nodes on their neighbourhood, where accounting for repeated re-infections amongst neighbouring nodes cascades into complicated and un-tractable descriptions. However, the SIS dynamics still remains attractive due to its simplicity, where for a fully connected network with N nodes, the exact system of master equations consists of $N+1$ equations given as

$$\dot{x}_k(t) = a_{k-1}x_{k-1}(t) - (a_k + c_k)x_k(t) + c_{k+1}x_{k+1}(t) \tag{1.1}$$

where $x_k(t)$ is the probability of observing k infectious nodes at time t, $a_k = \tau k(N-k)$, $c_k = \gamma k$ for $k = 0, 1, \ldots, N$ and $a_{-1} = c_{N+1} = 0$. Furthermore, it is assumed that each node is either susceptible (S) or infected (I) and susceptible nodes become infected at rate τ across any link to an infectious node, while infected nodes recover at rate γ and become susceptible again, with all events occurring independently of each other. It is known that, in this case, the 2^N-dimensional system of master equations can be lumped to the $(N+1)$-dimensional system above. This extreme reduction is possible due to the symmetry of the network where all nodes are topologically equivalent and thus the process is driven by the number of infected nodes without knowing their precise location. This leads to being able to simply write $a_k = \tau k(N-k)$ which effectively means that, in the presence of k infected nodes, the number of potentially disease transmitting edges/links, (SI), is $k(N-k) = a_k/\tau$.

Given that a SIS dynamics on an arbitrary network leads to a state space with 2^N equations, the reduced system above, raises the question whether the lumping or collapsing of the state space above can be extended to networks other than fully connected, even if the reduced system may not be exact. This problem breaks down into two important sub-questions. First, is it possible to count or approximate the expected number of (SI) edges in the presence of k infected nodes, and the second, whether the master equation above (1.1) with the expected rates can give a reasonable agreement with the full system or simulation? The counting procedure is non-trivial since the rate of infection in the presence of k infected nodes has to take into account the special placement of these as determined by the epidemic transmission process and implicitly influenced by the struc-

ture of the network. Even if this first step is successful, there is no guarantee that the reduced system will give good agreement with the exact model. In this paper, we will explore the viability of collapsing the full state space with 2^N elements to a state space with $N+1$ elements, i.e. $\{0, 1, \ldots, N\}$, for graphs other than fully connected. In particular we will explore the potential of systems such as the one given above to approximate results from full, exact systems or their counterpart based on simulation. The focus of the paper will be on networks of classic type such as, homogeneous random, Erdős-Rényi, bimodal and Barabási-Albert graphs with heterogenous degree distribution. For a fuller exploration, clustered networks generated based on the so called "Big-V" [5, 6] rewiring will be also considered, but only numerically. The main idea of the present paper is based around estimating the infection rate a_k, where a_k/τ can be interpreted as the expected number of (SI) edges in the presence of k infected nodes. The approach proposed in the paper revolves around various semi-heuristic and combinatorial arguments to determine these values.

The paper is structured as follows. First, we introduce our approximate master equations and show that once the transmission rates are accurate (determined from simulations), then the model agrees very well with simulation results, i.e. the feasibility of the model is shown. In Section 3 we present our combinatorial method to determine the number of II edges when recovery is neglected (i.e. for high values of τ). This method can predict the average number of II edges for configuration random graphs. The average number of SI edges is determined in Section 4 also for small values of τ based on semi-heuristic arguments using the pairwise model. The steps of our method are summarized in Section 5, where a case study is also shown as an illustration of the applicability of the method. For a regular random graph the coefficients of our master equation are determined analytically and the results are compared to simulations.

2. Model formulation: feasibility and transmission rates

In this paper a simplified Markov chain model for SIS epidemic propagation on a network is formulated and studied. The main idea of simplification is to use the state space $\{0, 1, \ldots, N\}$, denoting the number of infected nodes in the network, and introduce $x_k(t)$ as the probability that the system is in state k 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 $k-1$ (an infected node recovers) or to state $k+1$ (a susceptible node becomes infected), the master equations of the Markov chain take the form

$$\dot{x}_k = a_{k-1}x_{k-1} - (a_k + c_k)x_k + c_{k+1}x_{k+1}, \quad k = 0, \ldots, N. \tag{2.1}$$

2.1. Models feasibility

As we noted above, this state space is too small to describe the system exactly, since the full state space contains 2^N elements. The key finding of this paper is that the epidemic process on the network can be approximated with high accuracy by using this much simpler state space. To

support our statement, in Fig. 1 we plot the time evolution of the expected prevalence from simulation and from the master equations (2.1) with a_k taken as an average from simulation, namely $a_k = \tau e_{[SI]}(k)$, where $e_{[SI]}(k)$ is the expected number of (SI) pairs in the presence of k infected nodes. The excellent agreement for graphs with heterogenous degree distributions and even clustered ones shows that if transmission rates can be computed based on some analytic or semi-analytic approaches, then the reduced system can produce excellent agreement with simulation. All graphs, except the clustered ones, have been generated using the configuration model [2].

Thus, the main challenge is to specify the infection rates a_k. These will of course depend on the parameters of the model, i.e. τ and γ, as well as the topology of the graph. The most straightforward way of doing this is to recover these from simulation, as it is shown in Fig. 1. This amounts to recording the exact number of SI-type links whenever infection is present and use that $a_k = \tau e_{[SI]}(k)$, where we ignore that a_k itself is a random variable with some distribution and use simply its expected value. For a fixed value of k, the count of the SI-type links, and thus a_k, will take different values as different arrangements of k infectious nodes on the graph, as given by the process, will generate different counts. If the process is simulated based on a Gillespie-type approach then recording a_k raises the question of whether the counts have to be weighted by the time that the system spends in a given state. If synchronous updating is used, a change during each iteration is not guaranteed and thus some configurations will have a higher recording frequency. This will not be the case if the simulation is of Gillespie type, as in this case, a change is always guaranteed. Thus the synchronous updating naturally captures the longer or shorter time spent in a given state, and this is the simulation approach that we adopt, but always making sure that the simulation step δt is always small enough to guarantee not more than one single update per iteration step.

2.2. Determining the transmission rates

The most frequently used theoretical approximation for a_k is

$$a_k = \tau n k \frac{N-k}{N-1},$$

where n denotes the average degree of the nodes. This approximation ignores the natural correlations that will develop and operates on the assumption of random mixing. This means that although the natural disease transmission process will lead to different arrangements of k infectious individuals on the network, and thus different counts for the various pair types, the model above takes an average over all configurations and assumes that Ss and Is are randomly distributed on the network. The performance of this approximation is shown in Fig. 2 where this is compared to values computed from simulations on homogeneous and bimodal random graphs, on a clustered graph and on a Barabási-Albert graph. One can observe that heterogeneity shifts the maximum point of the curve to left and up, since highly connected nodes cause an increase in the number of SI edges. On the other hand, clustering shifts the maximum point of the curve to right and down, since I nodes are more likely to connect to I nodes making the number of SI edges smaller.

The poor accuracy of the theoretical approximation motivated our research to derive improved approximations that give better accuracy/agreement, and can better reflect the graph structure.

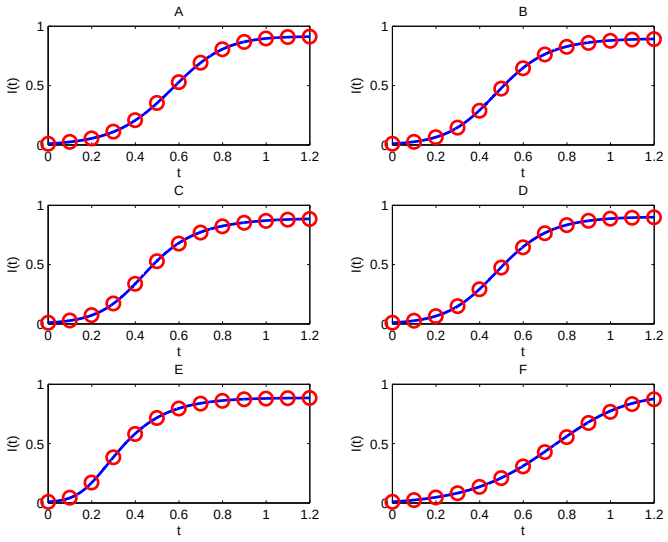

Figure 1: Time evolution of the expected prevalence from simulation ($\circ$ markers) and from master equations (2.1) with a_k taken as an average from simulation (continuous curve) for (A) homogeneous random graph, (B) Erdős-Rényi random graph, (C) bimodal random graph, (D) negative binomial random graph, (E) Barabási-Albert graph, (F) clustered random graph with clustering coefficient 0.4. The parameters are $N = 1000$, $\tau = 2$, $\gamma = 1$, average degree 6, number of initially infected nodes 10. The simulation results were obtained as the average of 250 simulations.

It is important to note that the above approximation gives the same a_k values for any graphs with average nodal degree n. The improved approximation can be achieved by deriving the transmission rates a_k ($k = 0, 1, \ldots, N$), be it based on some combinatorial or random walk argument, in a way in which a_k captures and conserves at least some of the natural features of the true dynamics on graphs. The theory that we develop is based on a combinatorial derivation of the number of II edges when the number of infected nodes is given. This derivation is presented in the next section.

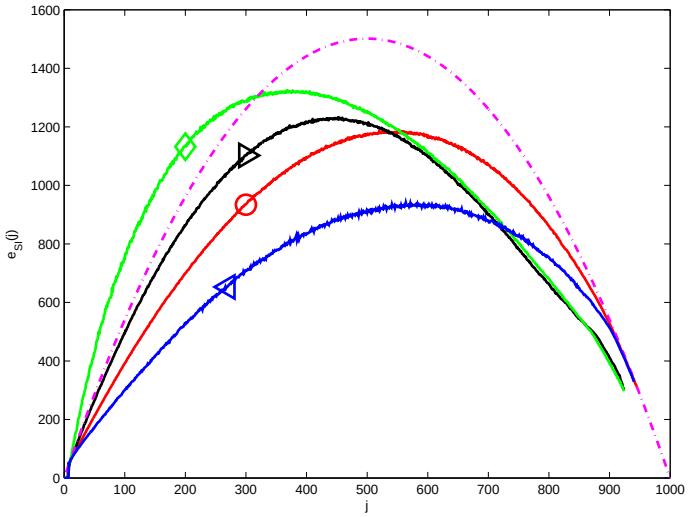

Figure 2: The $a_k/\tau = e_{[SI]}(k)$ values from simulations on a homogeneous random graph ($\circ$), a bimodal random graph ($\triangleright$), a clustered random graph with clustering coefficient 0.4 ($\triangleleft$) and a Barabási-Albert graph ($\diamond$) and the theoretical value $a_k/\tau = nk\frac{N-k}{N-1}$ (dash-dotted line) with $N = 1000$, $\tau = 2$, $\gamma = 1$, $n = 6$.

3. Theoretical approximation of the number of II edges

In this section we neglect the effect of recovery (assume that the value of τ is large) and derive a recursive formula for $e_{II}(j)$ which denotes the average number of II edges when the number of infected nodes is j. Our derivation is valid for a graph with arbitrary degree distribution and constructed by using the configuration model [2]. In the first subsection, the derivation is presented for the simplest case, i.e. for a homogeneous random graph where all nodes have degree n. Then, in the second subsection, the derivation is generalised to graphs with arbitrary degree distribution. Although the first, simpler derivation is a special case of the second, we use it to convey our ideas and arguments in a clearer and more concise way.

3.1. The case of regular random graphs

Let us consider a regular random graph in which every node has degree n. Further, we assume that there are j infected nodes and $e_{II}(j)$ denotes the number of II edges (doubly counted). The derivation is based on determining an explicit formula for the average number of new II edges that are created when a new infection happens, i.e. we derive a formula for $e_{II}(j+1) - e_{II}(j)$. The approach that we use is based on constructing the graph concurrently with the disease transmission.

If there are j infected nodes then these have $m := nj - e_{II}(j)$ free stubs that can become connected to the free stubs of susceptible nodes, the number of which is $M := n(N-j)$, since there are $N-j$ susceptible nodes and each of these has n stubs. Let us choose m stubs randomly out of the M free stubs of the susceptible nodes, and connect these to the m free stubs of the infected nodes. By construction, a susceptible node with multiple available stubs can and will connect to a number of infected nodes and this needs to be captured. Let us denote by p_k the proportion of susceptible nodes that connect to k distinct infected nodes, $k = 0, 1, \ldots, n$. (Here, we ignore that multiple links between the same susceptible and infected nodes are possible.) First, we determine the probability p_k by noting that, for the ease of calculation, this can be interpreted in an equivalent but, more convenient way as follows. p_k can be thought of being the probability that a given susceptible node has k links to infected nodes. This interpretation implies that p_k has hypergeometric distribution with the following parameters. We choose m stubs out of M stubs, and n of them belongs to the given (fixed) node. p_k is the probability that k of the chosen m stubs belong to the given node. Hence we have

$$p_k = \frac{\binom{n}{k}\binom{M-n}{m-k}}{\binom{M}{m}}, \qquad k = 0, 1, \ldots, n.$$

Thus the average number of susceptible nodes having k infected links is $(N-j)p_k$. We note that using that the expected value of the hypergeometric distribution is $\frac{nm}{M}$, we get that the total number of newly created SI links is

$$(N-j)\sum_{k=0}^{n} kp_k = (N-j)\frac{nm}{M} = (N-j)\frac{nm}{n(N-j)} = m$$

that is exactly the number of free stubs of the infected nodes. However, in our calculation we wish to consider the newly gained II links upon the birth of a single additional infected node. Thus, we require to identify which of the newly created SI links will lead to one additional infected node. Let us now associate a number to each SI edge as follows. To an edge we associate the number k if it is connected to a susceptible node that has k infected neighbours. This means that a susceptible node with k infected neighbours will have each of its individual edges to infected nodes labeled with k. Then the number of edges to which the number k is associated is $k(N-j)p_k$, since the average number of susceptible nodes having k infectious links is $(N-j)p_k$ and each of them has k SI type edges. Hence, if we choose an SI edge randomly (along which the next infection will happen), then the probability that there will be k new II edges, after this specific infection happens, is the probability that we choose an edge to which the number k is associated. This probability is

$$q_k = \frac{k(N-j)p_k}{m}, \quad k = 1, 2, \ldots, n.$$

Thus the average number of new II edges that are created when the next infection happens is the expected value $\sum kq_k$, and since a new $I-I$ connection creates two II edges, we get that

$$e_{II}(j+1) - e_{II}(j) = 2\sum_{k=1}^{n} kq_k = 2\sum_{k=1}^{n} \frac{k^2(N-j)p_k}{m} = 2\frac{N-j}{m}\sum_{k=1}^{n} k^2 p_k.$$

This sum can be obtained from the variance V and the expected value $\frac{nm}{M}$ as

$$\sum_{k=1}^{n} k^2 p_k = V + \left(\frac{nm}{M}\right)^2.$$

The variance of the hypergeometric distribution (using our parameters M, n, m) is

$$V = \frac{nm}{M}\frac{M-n}{M}\frac{M-m}{M-1}.$$

Hence

$$\sum_{k=1}^{n} k^2 p_k = \frac{nm}{M}\frac{M-n}{M}\frac{M-m}{M-1} + \left(\frac{nm}{M}\right)^2,$$

yielding

$$e_{II}(j+1) - e_{II}(j) = 2\frac{N-j}{m}\frac{nm}{M}\left(\frac{M-n}{M}\frac{M-m}{M-1} + \frac{nm}{M}\right) = 2\frac{M-m-n+mn}{M-1},$$

where $M = n(N-j)$ was used.

This way we get the value of $e_{II}(j)$ recursively starting from $e_{II}(1) = 0$ by using that

$$e_{II}(j+1) - e_{II}(j) = 2\frac{n(N-j) + (n-1)(nj - e_{II}(j)) - n}{n(N-j) - 1}, \quad j = 1, 2, \ldots, N-1. \quad (3.1)$$

In Fig. 3 this theoretical value of $e_{II}(j)$ is compared to the $e_{II}(j)$ values obtained from simulation. This figure highlights the excellent agreement between the recursion and results obtained from simulation. This motivates us to extend our combinatorial argument to networks with heterogeneous degree distributions. This more complete recursion for networks with arbitrary degree distributions is presented in the next subsection.

3.2. The case of configuration random graphs with arbitrary degree distribution

Consider a random graph with a given degree distribution $\{d_l : \ l = 1, 2, \ldots, L\}$. We denote by N_l the number of nodes of degree n_l, hence $d_l = N_l/N$ and $N_1 + N_2 + \ldots + N_L = N$. We will need an approximation for the number of infected nodes of degree n_l when the total number of infected

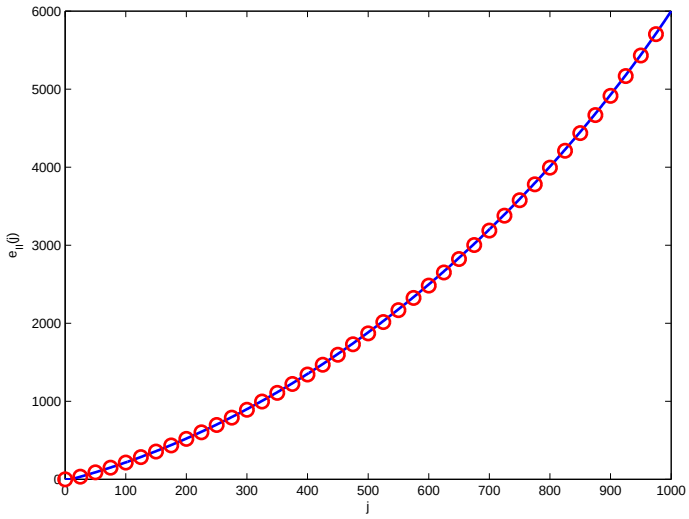

Figure 3: The $e_{II}(j)$ values from simulation ($\circ$) on a homogeneous random graph and their theoretical value obtained from recursion (3.1) (continuous curve) with $N = 1000$, $n = 6$, $\tau = 10$, $\gamma = 1$. 250 simulations, started with 10 infected nodes, were averaged.

nodes is given. Let this number be j and denote by $I_l(j)$ the expected value of infected nodes of degree n_l. Then we obviously have

$$\sum_{l=1}^{L} I_l(j) = j.$$

First, we derive a recursive relation for $I_l(j)$. Assume that the value of $I_l(j)$ is known for a given j and for all $l = 1, 2, \ldots, L$. Then the probability that the next infected node is of degree n_l is denoted by $P_l(j)$. Since the next infected node has degree n_l with probability $P_l(j)$ the expected value of degree n_l infected nodes will be

$$I_l(j+1) = I_l(j) + P_l(j).$$

The probability $P_l(j)$ will be determined later, now we determine a formula for the average number of new II edges that are created when the next infection happens.

To get a formula for $e_{II}(j+1) - e_{II}(j)$, we have to notice that the average number of new II edges depends on the degree of the $(j+1)$-th new infected node. Thus we compute first the conditional expected value of $e_{II}(j+1) - e_{II}(j)$ given that the new infected node has degree n_l and then the law of total probability is applied to get $e_{II}(j+1) - e_{II}(j)$.

The derivation will be analogous to that in the previous section. Suppose that there are j infected nodes, then these have

$$m = \sum_{l=1}^{L} n_l I_l(j) - e_{II}(j)$$

free stubs, and the $\sum_{l=1}^{L}(N_l - I_l(j))$ susceptible nodes have

$$M = \sum_{l=1}^{L} n_l(N_l - I_l(j))$$

stubs. Supposing that the next infected node has degree n_l let p_k^l denote the probability that an arbitrary susceptible node with degree n_l has k links to infected nodes, $k = 0, 1, \ldots, n_l$. Thus p_k^l has hypergeometric distribution

$$p_k^l = \frac{\binom{n_l}{k}\binom{M-n_l}{m-k}}{\binom{M}{m}}, \qquad k = 0, 1, \ldots, n_l.$$

We get that the average number of susceptible nodes with degree n_l having k infected neighbours is $(N_l - I_l(j))p_k^l$. Using the expected value of the hypergeometric distribution, the total number of SI links from the susceptible nodes having degree n_l is

$$(N_l - I_l(j)) \sum_{k=0}^{n} k p_k^l = (N_l - I_l(j)) \frac{n_l m}{M}.$$

Similarly to the previous subsection, let us associate a number to each SI edge as follows. To an edge we associate the number k if it is connected to a susceptible node that has k infected neighbours. This means that a susceptible node with k infected neighbours will have each of its individual edges to infected nodes labeled with k. Then the number of edges to which the number k is associated is $k(N_l - I_l(j))p_k^l$, since the average number of susceptible nodes having k infectious links is $(N_l - I_l(j))p_k^l$ and each of them has k SI type edges. Hence, if we choose an SI edge randomly (along which the next infection will happen), then the probability that there will be k new II edges, after this specific infection happens, is the probability that we choose an edge to which the number k is associated. This probability is

$$q_k^l = \frac{k(N_l - I_l(j))p_k^l}{(N_l - I_l(j))\frac{n_l m}{M}} = \frac{M}{n_l m} k p_k^l, \quad k = 1, 2, \ldots, n_l.$$

Thus the average number of new II edges that are created when the next infection happens (given that the degree of the newly infected node is n_l) is the conditional expected value

$$E_l(j) = \sum_{k=1}^{n_l} k q_k^l = \frac{M}{n_l m} \sum_{k=1}^{n_l} k^2 p_k^l = \frac{M - n_l}{M} \frac{M - m}{M - 1} + \frac{n_l m}{M},$$

where we used from the previous subsection that

$$\sum_{k=1}^{n_l} k^2 p_k^l = \frac{n_l m}{M} \frac{M - n_l}{M} \frac{M - m}{M - 1} + \left(\frac{n_l m}{M}\right)^2.$$

Let us now turn to the derivation of $P_l(j)$. This can be simply given as the ratio of the number of SI links starting from S nodes of degree n_l and the total number of SI links. That is

$$P_l(j) = \frac{(N_l - I_l(j))\frac{n_l m}{M}}{\sum_{l=1}^{L}(N_l - I_l(j))\frac{n_l m}{M}}.$$

Finally, the recursion for $e_{II}(j)$ can be obtained by applying the law of total probability

$$e_{II}(j+1) - e_{II}(j) = 2\sum_{l=1}^{L} P_l(j)E_l(j). \tag{3.2}$$

As an application we determine $e_{II}(j)$ for a bimodal random graph, in which half of the nodes have degree 3 and the other half has degree 9, i.e. $L = 2$, $n_1 = 3$, $n_2 = 9$, $d_1 = \frac{1}{2} = d_2$. Then the (3.2) determines $e_{II}(j)$ that we compared to $e_{II}(j)$ values obtained from simulation. Fig. 4 shows that the theory and the simulation are in excellent agreement.

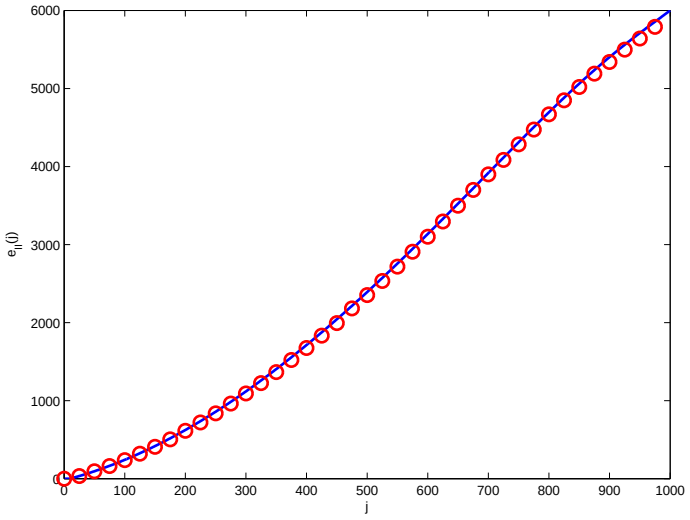

Figure 4: The $e_{II}(j)$ values from simulation ($\circ$) on a bimodal random graph and their theoretical value obtained from the recursion (continuous curve) with $N = 1000$, $N_3 = 500$, $N_9 = 500$, $\tau = 10$, $\gamma = 1$. 250 simulations, started with 10 infected nodes, were averaged.

4. Theoretical approximation of the number of SI edges

In this section we derive a formula for e_{SI}, the average number of SI edges, based on the approximation of the number of II edges. Simulations show that the points $(j, e_{SI}(j))$ (for $j =$

$0, 1, \ldots, N$) lie on a parabola-like curve, see Fig. 5. The shape of the parabola (for simplicity we call it parabola) depends on the structure of the graph, while the value of the parabola's maximum depends on τ. As the value of τ increases, the maximum value decreases. For large values of τ the parabola-like curve converges to a limiting curve that will be denoted by $e_{SI}^{\infty}(j)$, since it corresponds to the case $\tau \to \infty$. This phenomenon is shown in Fig. 5.

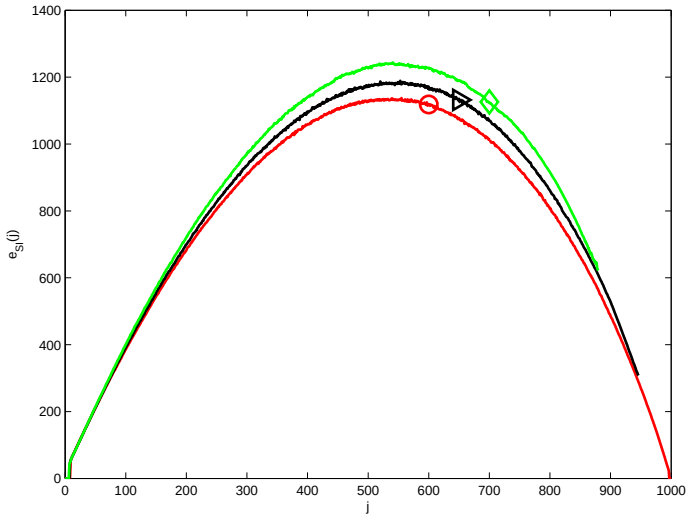

Figure 5: The $(j, e_{SI}(j))$ curves in the case of a regular random graph for different values of τ. ($\tau = 1$ ($\diamond$), $\tau = 2$ ($\triangleright$), $\tau = 10$ ($\circ$).)

To account for this special features of the $(j, e_{SI}(j))$ parabola, our theoretical model is built up in two steps. First, we derive a formula for the shape of the limiting curve $e_{SI}^{\infty}(j)$. In this step we assume that τ is large, hence neglect recovery. Then, we investigate the effect of τ on the value of the maximum of the $(j, e_{SI}(j))$ parabola.

The first step can be simply reduced to the result of the previous section by observing that the total number of SI and II edges is equal to the number of stubs starting from infected edges. Hence we have

$$e_{SI}^{\infty}(j) = \sum_{l=1}^{L} n_l I_l(j) - e_{II}(j),$$

where the degrees of the nodes are denoted by n_l for $l = 1, 2, \ldots, L$. In the case of a regular random graph, when the degree of every node is n, this takes the simple form

$$e_{SI}^{\infty}(j) = nj - e_{II}(j).$$

Thus the recursion for $e_{II}(j)$ in the previous section gives a theoretical value for $e_{SI}^{\infty}(j)$ directly. We note that the superscript ∞ was not used in the case of $e_{II}(j)$, because these numbers were introduced only for large τ, while $e_{SI}(j)$ will be computed also for small values of τ.

The recursion for $e_{II}(j)$ can be easily converted to a recursion for $e_{SI}^{\infty}(j)$. The recursions obviously do not give an explicit formula for $e_{SI}^{\infty}(j)$ or for $e_{II}(j)$ in terms of j. An approximate explicit formula can be derived for large N by transforming the difference equation given by the recursive relation into a differential equation. This will be presented in the next subsection in the case of a regular random graph.

4.1. Explicit formula for $e_{SI}^{\infty}(j)$ in the large N limit

In this subsection we assume that the graph is regular, the degree of each node is n. Based on $e_{SI}^{\infty}(j) = nj - e_{II}(j)$ the recursion for $e_{SI}^{\infty}(j)$ takes the form

$$n - (e_{SI}^{\infty}(j+1) - e_{SI}^{\infty}(j)) = e_{II}(j+1) - e_{II}(j). \tag{4.1}$$

Let us introduce $x = \frac{j}{N} \in [0,1]$ as a continuous variable corresponding to j, and let $f : [0,1] \to \mathbb{R}$ be a differentiable function, for which $f(\frac{j}{N}) = e_{SI}^{\infty}(j)$ holds. Then

$$e_{SI}^{\infty}(j+1) - e_{SI}^{\infty}(j) = f(\frac{j+1}{N}) - f(\frac{j}{N}) \approx \frac{1}{N} f'(\frac{j}{N}),$$

by using the definition of the derivative. Hence the recursion (3.1) and (4.1) lead to the differential equation

$$n - \frac{1}{N} f'(x) = 2\frac{nN(1-x) + (n-1)f(x) - n}{nN(1-x) - 1},$$

by using $m = nj - e_{II}(j) = f(x)$ and $n(N-j) = nN(1-x)$. Rearranging the equation it takes the form

$$\frac{1}{N} f'(x) + 2f(x)\frac{n-1}{nN(1-x)-1} = \frac{(n-2)nN(1-x) + n}{nN(1-x)-1}, \tag{4.2}$$

showing that this is a linear differential equation for f the general solution of which can be easily derived. Using the standard method of solving linear ODEs, first the homogeneous equation is solved then applying the method of variation of constants the solution can be given in the form

$$f(x) = (nN(1-x) - 1)^{2-2/n} K(x)$$

with an unknown function K. Substituting this form into the differential equation (4.2), the function K has to satisfy the following equation

$$-\frac{1}{N} K'(x) = (2-n)(nN(1-x)-1)^{2/n-2} - (2n-2)(nN(1-x)-1)^{2/n-3}.$$

Integrating this equation and after some algebra we obtain

$$K(x) = (nN(1-x)-1)^{2/n-1} + (nN(1-x)-1)^{2/n-2} + c$$

with a constant c. Thus the function f takes the form

$$f(x) = nN(1-x) + c(nN(1-x)-1)^{2-2/n}.$$

The constant c can be determined from the initial condition $f(\frac{1}{N}) = n$. This is equivalent to $e_{SI}^{\infty}(1) = n$ (expressing the simple fact that if there is a single infected node, then the number of SI edges is n). Hence we get

$$f(x) = nN(1-x) - n(N-2)\left(\frac{nN(1-x)-1}{n(N-1)-1}\right)^{2-2/n}. \tag{4.3}$$

This yields an explicit expression for $e_{SI}^{\infty}(j)$ as $e_{SI}^{\infty}(j) = f(\frac{j}{N})$. In Fig. 6 this approximation is compared to the result of recursion (3.1) and $e_{SI}^{\infty}(j) = nj - e_{II}(j)$.

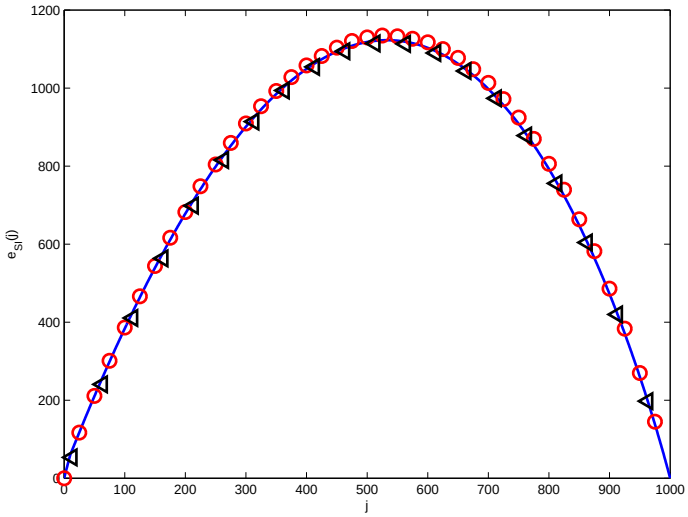

Figure 6: The $(j, e_{SI}(j))$ curve obtained from simulation ($\circ$), from recursion (3.1) via $e_{SI}^{\infty}(j) = nj - e_{II}(j)$ (continuous curve) and given by $e_{SI}^{\infty}(j) = f(\frac{j}{N})$ ($\triangleleft$) in case of a homogeneous random graph with $N = 1000$, $\tau = 10$, $\gamma = 1$, $n = 6$.

4.2. The dependence of the maximum number of SI edges on τ

Now we show an analytic method to derive the $e_{SI}(j)$ values based on the known values of $e_{SI}^{\infty}(j)$. As it is shown in Fig. 5, the value of τ effects the maximum value of the parabola-like curves. Hence we will derive a formula yielding this maximum value. The main idea is to use the pairwise equations, from which this maximum value, $[SI]_{max}$, can be obtained in terms of the location of the maximum, $[I]_{max}$. For a network with a given degree distribution the heterogeneous pairwise model is given in [4]. For ease of calculation we present the derivation for a homogeneous random

graph when the pairwise equations take the form

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

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 closed by the moment closure [9]

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

The number of SI edges is maximal when $[\dot{SI}] = 0$, that is when

$$\gamma([II] - [SI]) = \tau([SI] + [ISI] - [SSI]).$$

Using the closure relations and the formulas

$$[II] + [SI] = n[I], \quad [SS] + [SI] = n[S]$$

this equation leads to

$$\gamma n[S](n[I] - 2[SI]) = \tau[SI]\left(n[S] + (n-1)[SI] - (n-1)(n[S] - [SI])\right).$$

Since $[S] = N - [I]$, we get the following quadratic equation for the maximal value of SI edges

$$2\tau(n-1)[SI]^2 + n(N-[I])(2\tau - n\tau + 2\gamma)[SI] - n^2\gamma(N-[I])[I] = 0. \tag{4.4}$$

The positive solution of this equation for $[SI]$ is the maximal value of SI that will be denoted by $[SI]_{max}$. Here the value $[I]$ denotes the location of the maximum of the $e_{SI}^{\infty}(j)$ curve that can be obtained as follows. First, the maximum of $e_{SI}^{\infty}(j)$ has to be determined. Let us denote that by $[SI]_{max}^{\infty}$. This maximum is achieved at a certain value of j denoted by j_{max}, then $[I]_{max} = j_{max}$. We note that a more accurate value of $[I]_{max}$ can be obtained if a differentiable function $f : [0,1] \to \mathbb{R}$ can be given, for which $f(\frac{j}{N}) = e_{SI}^{\infty}(j)$ holds, as it was done in the previous subsection. Then determining the maximum of f that is taken at x_{max}, we get $[I]_{max} = Nx_{max}$. This way we get $[SI]_{max}$ from (4.4) in terms of τ. Finally, the average number of SI edges for the given value of τ can be expressed in the form

$$e_{SI}(j) = e_{SI}^{\infty}(j)\frac{[SI]_{max}}{[SI]_{max}^{\infty}}. \tag{4.5}$$

The result of this derivation is compared to simulation in Fig. 7 for a homogeneous random graph.

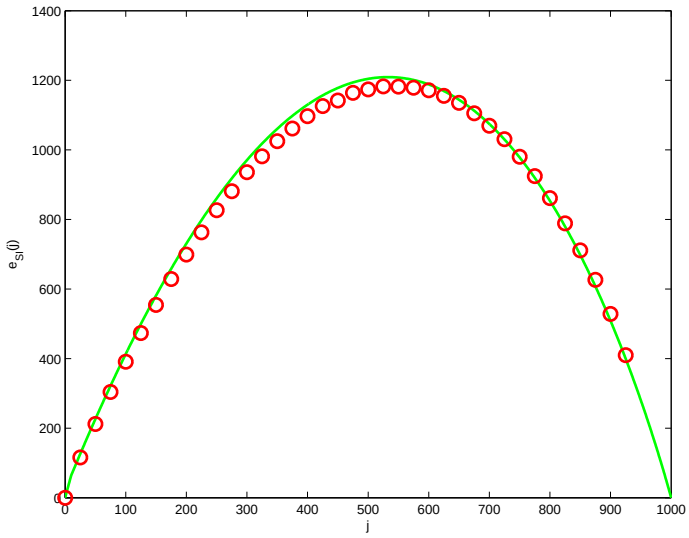

Figure 7: The $(j, e_{SI}(j))$ curve obtained from simulation ($\circ$) and from recursion (4.5) (continuous curve) in the case of a homogeneous random graph with $N = 1000$, $\tau = 2$, $\gamma = 1$, $n = 6$.

5. Summary of results

Here, we summarise our findings by comparing simulation results with results based on Eq. (1.1) with the transmission rate given by the recurrence relations in Eq. (3.2) and equation (4.5). For all networks considered, the agreement is excellent and this highlights that the heavily reduced state space and the resulting equations can capture the most significant components of the dynamics.

The steps of our method are summarised below. We start from a configuration random graph with given degree distribution, and from given values of τ and γ. We showed an analytical method that allows us to determine the coefficients a_k and c_k in equation (1.1). It is obvious that $c_k = \gamma k$, hence we show only the steps of determining the coefficients a_k.

1. Based on the degree distribution we determine the average number of II edges, $e_{II}(j)$ by using the recurrence relation (3.2), see Section 3.. (We note that this quantity, at this stage, is independent of τ, since it corresponds to the large τ limit case.)

2. The average number of SI edges (belonging to the case of large τ) can be given as $e^{\infty}_{SI}(j) = \sum_{k=1}^{K} kI_k(j) - e_{II}(j)$, see Section 4.

3. The average number of SI edges belonging to the given finite value of τ is determined from (4.5), see Section 4.2. Then the desired coefficients can be obtained as $a_k = \tau e_{SI}(k)$, $k = 0, 1, \ldots, N$.

4. Using the theoretically derived values of a_k and c_k equation (1.1) is solved numerically and the prevalence $I(t) = \sum_{k=0}^{N} kx_k(t)$ is compared to the prevalence obtained from simulation.

As an illustration of the performance of the method we carried out the above algorithm for a homogeneous random graph with $n = 6$. The comparison of the prevalence curves obtained from theory and from simulation is shown in Fig. 8.

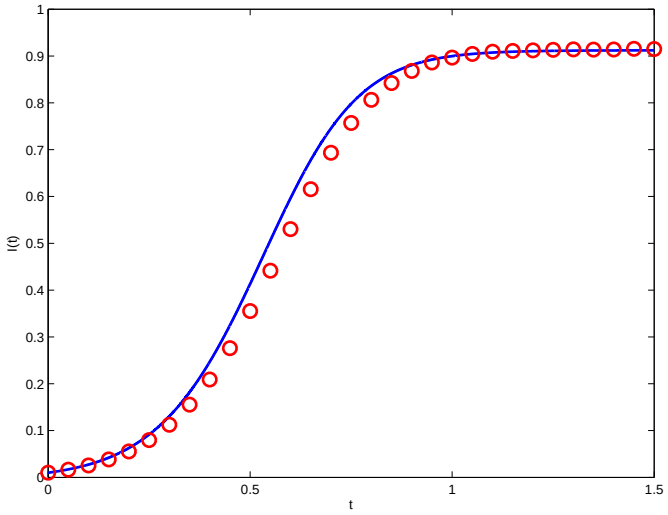

Figure 8: The time dependence of prevalence for a homogeneous random graph with parameters $N = 1000$, $n = 6$, $\tau = 2$, $\gamma = 1$ from theory (continuous curve) and from simulation ($\circ$). 250 simulations, started with 10 infected nodes, were averaged.

6. Discussion

In this paper we have formulated a new type of approximate ODE model for simple SIS dynamics on graphs with arbitrary degree distributions connected up according to the configuration model. This new model is inspired by the exact master equations corresponding to a fully connected graph with N nodes and $N + 1$ equations. It turns out that reducing the state space in the spirit of the fully connected graph leads to a viable approach to derive an approximate system provided that the transmission rates can be computed or approximated based on some analytic/combinatorial arguments. While in the paper, the analytical calculations are for configuration graphs, numerical results confirm that our approach is extendable to clustered networks (see Fig. 1) provided that analytical calculations for the transmission rates can be completed.

The results are somehow surprising as the transmission rates themselves are in fact random variables with some distribution since k infected nodes on a graph corresponding to different realisations of the simulation model will lead to a distribution of values in the number of SI edges. Thus, if one wants to improve the performance of the method, then the model would be a set of ODEs with transmission rates specified by random variables (instead of using their expected value)

with some distribution that are also correlated with previous and future states with less and more infected case, respectively.

The approach here can potentially be extended to clustered networks and this is a direction that we will explore in future work. More importantly, given the very specific shape of the the transmission rates curve, we can ask the question whether it is possible to infer important network characteristics, such as the degree distribution, from it. The transmission rates (i.e. transmission rates versus number of infected nodes) is a signature or footprint of the combined properties of epidemic and graph properties. While data on this quantity may be difficult or impossible to collect, it is still possible to use prevalence or incidence data from a real epidemic. Such data could be fitted with the proposed ODE model and then, the numerically inferred transmission rates could be compared to a class of typical parabolas describing the most common network types, see Fig. 2. Thus, the proposed model, could serve as a link between prevalence data and the process of inferring the underlying network structure.

In Subsection 4.1. we derived a functional form that relates the average number of SI edges to the average number of infected nodes, see equation (4.3). The most common functional form, based on the assumption of random mixing is $[SI] = n[I]\frac{N-[I]}{N-1}$, where n denotes the average degree of the nodes. In the formalism of Subsection 4.1. (where $x = [I]/N$) this approximation corresponds to the functional form $f(x) = \frac{nN^2}{N-1}x(1-x)$. This approximation ignores the natural correlations that develop during the process, hence it was generalised to $f(x) = kx^p(1-x)^q$ with some phenomenological parameters k, p and q in [13]. Our functional form yields an alternative to that with the advantage that it can be derived theoretically. The two functional forms are close to each other numerically, their theoretical relation can be the subject of future work.

Acknowledgements

Péter L. Simon acknowledges support from OTKA (grant no. 81403).

References

[1] A. Bátkai, I.Z. Kiss, E. Sikolya & P.L. Simon 2012. Differential equation approximations of stochastic network processes: an operator semigroup approach. Networks and Heterogeneous Media 7, 43-58.

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

[3] L. Decreusefond, J.-S. Dhersin, P. Moyal & V. C. Tran 2012. Large graph limit for an SIR process in random network with heterogeneous connectivity. Ann. Appl. Probab. 22, 541575.

[4] K.T.D. Eames, M.J. Keeling 2002. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. PNAS 99, 13330-13335.

[5] D. M. Green & I. Z. Kiss 2010. Large-scale properties of clustered networks: Implications for disease dynamics. Journal of Biological Dynamics 4, 431-445.

[6] T. House and M.J. Keeling 2010. The impact of contact tracing in clustered populations. PLoS Comput. Biol. 6, e1000721.

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

[8] M.J. Keeling 1999. The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. Lond. B 266, 859-867.

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

[10] J. Lindquist, J. Ma, P. van den Driessche & F.H. Willeboordse 2011. Effective degree network disease models. J. Math. Biol. 62, 143-164.

[11] V. Marceau, P.-A. Noël, L. Hébert-Dufresne, A. Allard & L. J. Dubé 2010. Adaptive networks: coevolution of disease and topology. Phys. Rev. E 82, 036116.

[12] J. C. Miller, A. C. Slim & E. M. Volz 2012. Edge-based compartmental modelling for infectious disease spread. J. R. Soc. Interface 9, (70) 890906.

[13] M. Roy, M. Pascual 2006. On representing network heterogeneities in the incidence rate of simple epidemic models. Ecol. Complexity 3, 80-90.

[14] P. L. Simon & I. Z. Kiss 2012. From exact stochastic to mean-field ODE models: a new approach to prove convergence results. IMA J. Appl. Math. doi: 10.1093/imamat/hxs001.

[15] P.L. Simon, M. Taylor & I.Z. Kiss 2011. Exact epidemic models on graphs using graph-automorphism driven lumping. J. Math. Biol. 62, 479-508.

[16] M. Taylor, P. L. Simon, D. M. Green, T. House & I. Z. Kiss 2012. From Markovian to pairwise epidemic models and the performance of moment closure approximations. J. Math. Biol. 64, 1021-1042.