

# EPIDEMIC SPREADING IN NETWORKS - VARIANCE OF THE NUMBER OF INFECTED NODES

J. OMIC AND P. VAN MIEGHEM\*

**Abstract.** The Susceptible Infected Susceptible (SIS) model is one of the basic models and it is applied for different networks and services in telecommunications. For more detailed prediction of the epidemic, it is necessary to examine the higher order moments, namely the variance of the number of infected nodes. Also, the predictability of mean-field models depends on the variations around the mean. However, the variance of epidemic spread on networks has so far received insufficient attention. Epidemics spread in significantly different topologies - from power law to complete graphs, thus the model should be independent of the underlying topology. We use the  $N$ -intertwined model which captures the topology influences of the finite graph defined by adjacency matrix  $A$  to determine the variance. We also determine upper and lower bounds on the variance as a function of effective spreading rate  $\tau$  and show that for some spreading conditions, the variance is highly dependent on the degree distribution of the underlying network and not on other topological properties. Further, we apply our findings to two types of graphs: complete and complete bipartite graphs. For the complete bipartite graph, we derive the probability distribution function of the number of infected nodes. Finally, we provide deeper understanding of the structural properties expressed via the second smallest eigenvalue of the Laplacian matrix and link this parameter to our  $N$ -intertwined model.

**Key words.** Epidemic theory, graph theory, variance of epidemics, Markov theory

**AMS subject classifications.** 00A71 Theory of mathematical modeling

**1. Introduction.** Although the modeling of diseases is over three hundred years old [1], the epidemic theory was first applied in telecommunication to the spread of email viruses, worms and other computer malware by Kephart and White (KW) [2]. The KW model belongs to the homogeneous models of the *SIS* (Susceptible Infected Susceptible) type and the network influence was represented by a fixed degree of each node. Later, the regular graph assumption of the KW model was shown to be inadequate for malware spreading. Pastor-Satorras and Vespignani [3] discuss discrepancy between the data of virus spread on the Internet and theoretical results of the *SIS* model for homogeneous networks. They introduce a model that takes into account degree distribution heterogeneity of the underlying power-law networks, which degree distribution follows a power law. At that point, the influence of topology in the application of epidemic theory in networks became an important issue. Ganesh *et al.* determine the influence of topological characteristic on the mean epidemic lifetime for the *SIS* model [4] and the number of removed nodes in the *SIR* (Susceptible infected Removed) model [5].

Except for virus spread in computer networks, SIS model is applied for distributed systems [7]. Distributed and scalable algorithms based on the epidemic paradigm of spreading were employed in order to retain scalability and reduce overload [6]. The Erdős-Rényi graph or a hypercube were often used as topologies to represent networks. An Erdős-Rényi graph is a graph in which an edge between two nodes exists with probability  $p$ .

The adequate functioning of the epidemic algorithm in sensor and P2P networks is responsible for the resilience of the network [7]. Propagation of faults and failures is also one of the possible applications for epidemic models in telecommunication

---

\*Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, P.O Box 5031, 2600 GA Delft, The Netherlands. Email: J.Omic@tudelft.nl, P.F.A.VanMieghem@tudelft.nl.

networks. One of the examples is the work of Coffman *et al.* [8], who models cascading BGP failures on a fully connected topology .

Since the range of topologies appearing in applications varies significantly, an epidemic model should not depend on a specific underlying topology, but should be capable of modeling an epidemic process on *any* given finite graph. An extensive overview of different epidemic models used in biology as well as in telecommunication is given in [1], [9] and [10]. We concentrate on the *SIS* epidemic model, which is one of the standard computer virus models. We will especially emphasize the topology influence by using the *N*-intertwined model [11], that incorporates any topology adequately.

In a *SIS* model, a node can be in one of two states: susceptible (*S*) or infected (*I*). Infected nodes can infect other neighboring nodes, and each node can be randomly cured. All cured nodes can be infected again. A peculiar aspect of the *SIS* model is the oscillation of a nodal state between the infected and the susceptible states as long as the epidemic persists. Clusters of infected nodes are appearing and disappearing, expanding and shrinking. The persistence of an epidemic is another interesting aspect of the *SIS* model. For a finite graph, the state with all healthy nodes is the absorbing state - in which the system will eventually end. However, a large portion of nodes can be infected for a very long, though finite time. The lifetime of this so-called *metastable* state depends on the effective spreading rate  $\tau = \frac{\beta}{\delta}$ , the infection rate  $\beta$  per link divided by the curing rate  $\delta$  per node. We can observe critical point behavior for different effective spreading rates with very short lifetime on one side and very long on the other.

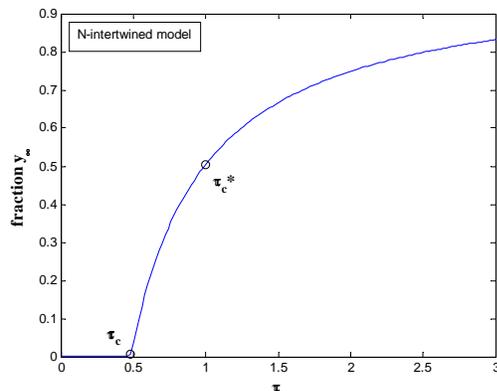


FIG. 1.1. The number of infected nodes in the metastable state for *SIS* model as a function of  $\tau$ .

The fraction  $y_\infty$  of infected nodes in the metastable state as a function of effective spreading rate  $\tau$  is computed in Figure 1.1 based on the *N*-intertwined model, which was proposed and analyzed in [11] and reviewed in Section 2. The critical effective spreading rate is denoted with  $\tau_c$ . Below this threshold  $\tau < \tau_c$ , the epidemic will extinct and above the threshold  $\tau > \tau_c$ , a certain percentage of the nodes will stay infected infinitely long. The infinite lifetime is an artefact of the mean-field theory applied in the *N*-intertwined model as explained in Section 2. The real epidemic also shows a threshold behavior, but with two thresholds  $\tau_c$  and  $\tau_c^*$ . Below the first one  $\tau < \tau_c$ , the lifetime of the epidemic is of the order  $O(\log(N))$ , where  $N$  is the number of nodes in the network [4]. Above the second threshold  $\tau > \tau_c^*$ , the lifetime of the

metastable state is of the order<sup>1</sup>  $\Omega(e^{N^\alpha})$ , for  $\alpha > 0$  [4]. For example, suppose that we are above the threshold  $\tau_c^*$ , that the units of  $\beta$  and  $\delta$  are  $sec^{-1}$  and  $N = 1000$ , then the lifetime of an epidemic on the network would be of order  $\Omega(e^{1000})$  seconds =  $\Omega(10^{326})$  years which is considerably more than the life time of the universe ( $4 \cdot 10^9$  years).

Between the two thresholds  $\tau_c$  and  $\tau_c^*$  the lifetime of an epidemic increases. This time depends on the effective spreading rate  $\tau$  and is closely related to the variation of the fraction of infected nodes around the mean value  $y_\infty$ . If the variations are large and the mean is small, the time to extinction decreases. The rate with which the states change are of order of curing and infecting rate  $\delta$  and  $\beta$ . The average number of infected nodes, the variance of this number and the expected life time of an epidemic are closely related. We show this using the second smallest eigenvalue of the Laplacian matrix, which influences the lifetime of an epidemic [4] and the variance of the number of infected nodes. Although the variance of the number of infected nodes plays an important role in determining the lifetime and the survival of an epidemic as well as in the predictability of mean-field models, this subject was not studied extensively in literature. Crepey *et al.* in [12] examine this problem by means of event driven simulations.

We study the variance of the number of infected nodes as a function of the nodal degree and of the effective spreading rate  $\tau$ . For large  $\tau$ , we show that the degree distribution alone determines the steady-state. Further, we apply our findings on two type of graphs: complete and complete bipartite graphs. For important structure in telecommunication networks, namely the complete bipartite graph, we derive the probability distribution function of the number of infected nodes.

**2.  $N$ -intertwined continuous Markov chains with 2 states.** This section reviews the homogeneous  $N$ -intertwined model [11].

By separately observing each node, we will model the virus spread in a bidirectional (connected) network specified by a symmetric adjacency matrix  $A$ . Every node  $i$  at time  $t$  in the network can be in two states: infected with probability  $\Pr[X_i(t) = 1]$  and healthy with probability  $\Pr[X_i(t) = 0]$ . At each moment  $t$ , a node can only be in one of two states, thus  $\Pr[X_i(t) = 1] + \Pr[X_i(t) = 0] = 1$ . The state of a node  $i$  is specified by a Bernoulli random variable  $X_i \in \{0, 1\}$ :  $X_i = 0$  for a susceptible node and  $X_i = 1$  for a infected node. We assume that the curing process is a Poisson process with rate  $\delta$ , and that the infection rate per link is a Poisson process with rate  $\beta$ .

If we apply Markov theory, the infinitesimal generator  $Q_i(t)$  of this two-state continuous Markov chain is,

$$Q_i(t) = \begin{bmatrix} -q_{1;i} & q_{1;i} \\ q_{2;i} & -q_{2;i} \end{bmatrix}$$

with  $q_{2;i} = \delta_i$  and

$$q_{1;i} = \sum_{j=1}^N \beta_j a_{ij} \mathbf{1}_{\{X_j(t)=1\}}$$

where the indicator function  $\mathbf{1}_x = 1$  if the event  $x$  is true else it is zero. The coupling of node  $i$  to the rest of the network is described by an infection rate  $q_{1;i}$  that is a random

---

<sup>1</sup> $\Omega$  is a lower asymptotic bound with definition: for a given real function  $g(x)$ , we denote  $\Omega(g(x))$  as the set of functions such that  $\Omega(g(x)) = \{f(x), \exists x_0, \exists c > 0 \mid 0 \leq cg(x) \leq f(x), \forall x \in \mathbb{R}, x > x_0\}$ .

variable, which essentially makes the process doubly stochastic. This observation is crucial. For, using the definition of the infinitesimal generator [16, p. 181],

$$\Pr[X_i(t + \Delta t) = 1 | X_i(t) = 0] = q_{1;i}\Delta t + o(\Delta t)$$

the continuity and differentiability shows that this process is not Markovian anymore. The random nature of  $q_{1;i}$  is removed by an additional conditioning to all possible combinations of rates, which is equivalent to conditioning to all possible combinations of the states  $X_j(t) = 1$  (and their complements  $X_j(t) = 0$ ) of the neighbors of node  $i$ . Hence, the number of basic states dramatically increases. Eventually, after conditioning each node in such a way, we end up with the exact  $2^N$ -state Markov chain, studied in [11].

Instead of conditioning, we replace the actual, random infection rate by an effective or average infection rate, which is basically a mean field approximation,

$$E[q_{1;i}] = E\left[\sum_{j=1}^N \beta_j a_{ij} \mathbf{1}_{\{X_j(t)=1\}}\right] \quad (2.1)$$

In general, we may take the expectation over the rates  $\beta_i$ , the network topology via the matrix  $A$  and the states  $X_j(t)$ . Since we assume that both the infection rates  $\beta_i = \beta$  (homogeneous setting) and the network are constant and given, we only average over the states. Using  $E[\mathbf{1}_x] = \Pr[x]$  (see e.g. [16]), we replace  $q_{1;i}$  by

$$E[q_{1;i}] = \beta \sum_{j=1}^N a_{ij} \Pr[X_j(t) = 1]$$

which results in an effective infinitesimal generator,

$$\overline{Q_i(t)} = \begin{bmatrix} -E[q_{1;i}] & E[q_{1;i}] \\ \delta_i & -\delta_i \end{bmatrix}$$

The effective  $\overline{Q_i(t)}$  allows us to proceed with Markov theory. Denoting  $v_i(t) = \Pr[X_i(t) = 1]$  and recalling that  $\Pr[X_i(t) = 0] = 1 - v_i(t)$ , the Markov differential equation for state  $X_i(t) = 1$  turns out to be nonlinear

$$\frac{dv_i(t)}{dt} = \beta \sum_{j=1}^N a_{ij} v_j(t) - v_i(t) \left( \beta \sum_{j=1}^N a_{ij} v_j(t) + \delta \right) \quad (2.2)$$

Each node obeys a nonlinear differential equation as (2.2),

$$\begin{cases} \frac{dv_1(t)}{dt} = \beta \sum_{j=1}^N a_{1j} v_j(t) - v_1(t) \left( \beta \sum_{j=1}^N a_{1j} v_j(t) + \delta \right) \\ \frac{dv_2(t)}{dt} = \beta \sum_{j=1}^N a_{2j} v_j(t) - v_2(t) \left( \beta \sum_{j=1}^N a_{2j} v_j(t) + \delta \right) \\ \vdots \\ \frac{dv_N(t)}{dt} = \beta \sum_{j=1}^N a_{Nj} v_j(t) - v_N(t) \left( \beta \sum_{j=1}^N a_{Nj} v_j(t) + \delta \right) \end{cases}$$

Written in matrix form, with  $V(t) = [v_1(t) \ v_2(t) \ \cdots \ v_N(t)]^T$ , we arrive at

$$\frac{dV(t)}{dt} = \beta AV(t) - \text{diag}(v_i(t)) (\beta AV(t) + \delta u) \quad (2.3)$$

where  $u$  is the all-one vector and  $\text{diag}(v_i(t))$  is the diagonal matrix with elements  $v_1(t), v_2(t), \dots, v_N(t)$ .

The time-dependent evolutionary equation (2.3) explicitly shows the dependence on the network via the adjacency matrix  $A$ . The nonlinearity introduced by the mean field approximation (2.1) gives rise to the so-called metastable state, where the probability distribution of the number of infected nodes is of temporary stability [11], [4]. One of the main results in [11] is the rigorous establishment of the threshold for the effective spreading rate  $\tau_c = \frac{1}{\lambda_1(A)}$ , where  $\lambda_1(A)$  is the largest eigenvalue of the adjacency matrix  $A$ . In the metastable state and above this threshold  $\tau > \tau_c$ , the infection persists and a nonzero fraction of the nodes remains infected, whereas below the threshold  $\tau < \tau_c$ , all network nodes are healthy.

The steady-state infection probabilities, defined by  $v_{i\infty} = \lim_{t \rightarrow \infty} v_i(t)$  and where  $\frac{dv_i(t)}{dt} = 0$ , follow for each node  $1 \leq i \leq N$  from (2.2) as

$$v_{i\infty} = \frac{\beta \sum_{j=1}^N a_{ij} v_{j\infty}}{\beta \sum_{j=1}^N a_{ij} v_{j\infty} + \delta}$$

The fraction of infected nodes  $y(t)$  at any given time  $t$  is defined by

$$y(t) = \frac{1}{N} \sum_{j=1}^N v_j(t) \quad (2.4)$$

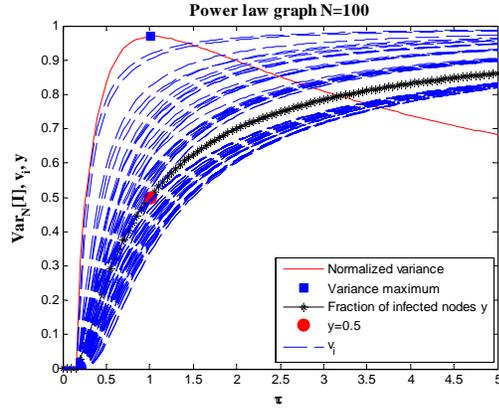


FIG. 2.1. Normalized variance and fraction of infected nodes as a function of effective spreading rate  $\tau$  for the  $N$ -intertwined model. Dashed line represents the variance and full line is fraction of infected nodes. Square marks the point for which  $\tau = \frac{2}{d_{av}}$ .

**2.1. Beyond the mean field approximation.** The time-dependent Bernoulli random variable  $X_i(t) \in \{0, 1\}$  can change from 1 to 0 with curing rate  $\delta$  and from 0 to 1 with infecting rate  $\beta \sum_{j=1}^N a_{ij} X_j(t)$ , that depends on the other Bernoulli random variables. We can write

$$\frac{X_i(t + \Delta t) - X_i(t)}{\Delta t} = \beta(1 - X_i(t)) \sum_{j=1}^N a_{ij} X_j(t) - \delta X_i(t)$$

After taking the expectation of both sides with  $E[X_i(t)] = \Pr[X_i(t) = 1] = v_i(t)$  and the limit  $\Delta t \rightarrow 0$ , we obtain

$$\frac{dv_i(t)}{dt} = \beta \sum_{j=1}^N a_{ij} v_j(t) - \beta \sum_{j=1}^N a_{ij} E[X_j(t)X_i(t)] - \delta v_i(t)$$

The existence of an infected node in the neighborhood cannot decrease the probability that a node is infected, hence

$$c_{ij}(t) = \Pr[X_j(t) = 1 | X_i(t) = 1] \geq \Pr[X_j(t) = 1] = v_j(t)$$

For dependent Bernoulli random variables  $X_j(t)$  and  $X_i(t)$ , the expectation is

$$\begin{aligned} E[X_j(t)X_i(t)] &= \Pr[X_j(t) = 1, X_i(t) = 1] \\ &= \Pr[X_j(t) = 1 | X_i(t) = 1] \Pr[X_i(t) = 1] \\ &= c_{ij}(t)v_i(t) \geq v_i(t)v_j(t) \end{aligned} \quad (2.5)$$

Finally,

$$\frac{dv_i(t)}{dt} = \beta \sum_{j=1}^N a_{ij} v_j(t) - v_i(t) \left( \beta \sum_{j=1}^N a_{ij} c_{ij}(t) + \delta \right)$$

Comparison with (2.2) shows that the only approximation of  $N$ -intertwined model lies in  $c_{ij} = v_j$ , which implies that the random variables  $X_j$  and  $X_i$  are implicitly assumed to be independent. Therefore, the  $N$ -intertwined model upper bounds the exact probability  $v_i(t)$  of an infection.

**3. The variance of the number of infected nodes in the  $N$ -intertwined model.** For the Bernoulli random variable  $X_i(t)$ , we have

$$\begin{aligned} E[X_i(t)] &= v_i(t) \\ \text{Var}[X_i(t)] &= v_i(t)(1 - v_i(t)) \end{aligned} \quad (3.1)$$

Although the states of nodes  $X_1, X_2, \dots, X_N$  are not independent random variables, the only approximation of the  $N$ -intertwined model leads to  $\Pr[X_i(t) = 1, X_j(t) = 1] = \Pr[X_i(t) = 1] \Pr[X_j(t) = 1]$  which means that the random variables  $X_i, X_j$  are assumed to be independent. Within this independence approximation, the variance of a sum of independent random variables  $I(t) = \sum_{i=1}^N X_i(t)$ ,

$$\text{Var}[I(t)] = \sum_{i=1}^N \text{Var}[X_i(t)] = \sum_{i=1}^N (1 - v_i(t))v_i(t) \quad (3.2)$$

Since  $0 \leq v_i \leq 1$ , the maximum in (3.2) is obtained when  $v_i(t) = \frac{1}{2}$ , for each node  $i$ , such that the variance of the number of infected nodes in the  $N$ -intertwined model is bounded by  $\text{Var}[I(t)] \leq \frac{N}{4}$ . In the worst case, the  $N$ -intertwined model can only predict a maximum variance not higher than  $0.25N$ .

The consequence of the mean field approximation underestimates the real variance. In case the Bernoulli random variables  $X_i(t)$  and  $X_j(t)$  are dependent, the variance of the number of infected nodes in the steady state is [16, p. 30]

$$\text{Var}[I_\infty] = \sum_{i=1}^N \text{Var}[X_i(t)] + 2 \sum_{i=1}^N \sum_{j=1}^{i-1} \text{Cov}[X_{i\infty}, X_{j\infty}] \quad (3.3)$$

where, using (2.5),

$$\begin{aligned} Cov[X_{i\infty}, X_{j\infty}] &= E[X_{i\infty}X_{j\infty}] - E[X_{i\infty}]E[X_{j\infty}] \\ &= c_{ij\infty}v_{i\infty} - v_{i\infty}v_{j\infty} \end{aligned}$$

The joint probability  $\Pr[X_{i\infty} = 1, X_{j\infty} = 1] = E[X_{i\infty}X_{j\infty}]$  can be calculated by introducing the correlation model also known as the moment closure model or the pair approximation model. The survey of the pair approximation model is given by Rand [21]. However, the correlation model is numerically more demanding.

The covariance of two random variables is bounded by

$$\begin{aligned} Cov[X_{i\infty}, X_{j\infty}] &\leq \sqrt{\text{Var}[X_{i\infty}] \text{Var}[X_{j\infty}]} \\ &= \sqrt{v_{i\infty}(1-v_{i\infty})v_{j\infty}(1-v_{j\infty})} \end{aligned}$$

The covariance thus tends to zero for very large and very small  $\tau$ , namely when  $v_{i\infty}$  tends to zero, or  $v_{j\infty}$  tends to one. The process of curing is independent of the underlying topology and a node to be cured is randomly chosen among all infected nodes, such that the states of nodes are not correlated. By analogy and excluding the cases of short living epidemics, one may claim that the covariance of the epidemic with a small number of infected nodes should be equal to the case with a small number of susceptible nodes. However, the problem is not symmetric because the infection process is graph dependent while the homogeneous curing process is not. This implies that the largest covariance can be expected in the region above the threshold  $\tau_c$ , where a small to medium portion of nodes is infected.

The covariance of two nodes is smaller than the probabilities of infection of each node and therefore the total covariance is never larger than the squared number of infected nodes in the network.

$$\text{Var}[I_\infty] \leq N/4 + I^2 \quad (3.4)$$

This bound is relevant only for a small number of infected nodes  $I$ , because the covariance and variance tends to zero for large  $\tau$ .

### 3.1. The variance on the number of infected nodes as a function of $\tau$ .

In this section, we will estimate variance as a function of  $\tau$  using upper and lower bounds on probability of infection of a node.

**THEOREM 3.1.** *For  $\tau > \tau_c$  in the metastable state, the variance of the number of infected nodes is bounded as*

$$\left\{ \begin{array}{ll} 0 \leq \text{Var}[X_i] < b_i(1-b_i), & \tau \leq \frac{1}{d_i} \\ 0 < \text{Var}[X_i] \leq 0.25, & \frac{1}{d_i} \leq \tau \leq \frac{1}{d_{\min}} \\ a_i(1-a_i) < \text{Var}[X_i] \leq 0.25, & \frac{1}{d_{\min}} \leq \tau \leq \frac{1}{2d_{\min}} \left(1 + \frac{\sqrt{d_i^2 + 4d_{\min}^2}}{d_i}\right) \\ b_i(1-b_i) < \text{Var}[X_i] \leq 0.25, & \frac{1}{2d_{\min}} \left(1 + \frac{\sqrt{d_i^2 + 4d_{\min}^2}}{d_i}\right) \leq \tau \leq \frac{1}{d_i} + \frac{1}{d_{\min}} \\ b_i(1-b_i) \leq \text{Var}[X_i] \leq a_i(1-a_i), & \tau \geq \frac{1}{d_i} + \frac{1}{d_{\min}} \end{array} \right. \quad (3.5)$$

where

$$\begin{aligned} a_i &= 1 - \frac{1}{1 + \frac{d_i}{d_{\min}}(\tau d_{\min} - 1)} \\ b_i &= 1 - \frac{1}{1 + \tau d_i} \end{aligned}$$

*Proof.* If  $\tau > \tau_c$ , the variable  $v_{i\infty}$  is bounded by  $a_i$  and  $b_i$

$$a_i \leq v_{i\infty} \leq b_i \quad (3.6)$$

where the upper and lower bounds  $a_i, b_i$  for the infection probability  $v_{i\infty}$  are derived in [11].

The lower bound does not belong to the probability range  $a_i \notin [0, 1]$  for values  $\tau < \frac{1}{d_{\min}}$ . The variance of the number of infected nodes is a concave function of  $v_{i\infty}$  with a maximum at  $v_{i\infty} = 0.5$ . For  $b_i \leq 0.5$  the variance of the number of infected nodes is upper-bounded by  $b_i(1 - b_i)$ . After the maximum of  $b_i$  is reached, the variance is upper bounded by the total maximum 0.25. Because the variance is a concave function, the lower bound of  $v_i$  will become the upper bound of the variance.

When the two curves intersect for  $\tau_p = \frac{1}{2d_{\min}} \left(1 + \frac{\sqrt{d_i^2 + 4d_{\min}^2}}{d_i}\right)$ , the variance is lower-bounded by  $b_i(1 - b_i)$ . The variance is upper bounded by the lower bound of  $v_i$  after the maximum of the lower bound is reached for  $\tau = \frac{1}{d_i} + \frac{1}{d_{\min}}$ . Therefore we can write (3.5).  $\square$

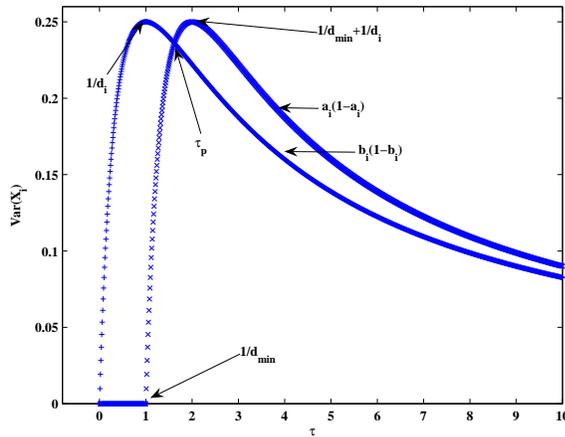


FIG. 3.1. Upper and lower bound of the variance of node infection as a function of  $\tau$ .

In Figure 3.1, the variance of node infection is calculated using upper and lower bounds for the infection probability  $v_i$ . It was deduced in [11] that the  $N$ -intertwined model is accurate for large values of  $\tau$ . We also have that  $a_i \rightarrow b_i$  if  $\tau \rightarrow \infty$ , which implies that the variance, for a large  $\tau$ , mainly depends on the degree distribution, not on other structural properties of the network like the diameter, or the clustering coefficient. Because the curing process is not topology dependent, for large  $\tau$ , there is a small percentage of susceptible nodes and they are isolated. In this region of  $\tau$ , the probability of infection depends on the degree of randomly chosen nodes.

**COROLLARY 3.2.** For the effective spreading rate  $\tau \geq \frac{2}{d_{\min}}$  the variance of the number of infected nodes of the  $N$ -intertwined model is tightly bounded by

$$\sum_{i=1}^N \frac{d_i \tau}{(1 + d_i \tau)^2} \leq \text{Var}[I(t)] \leq \sum_{i=1}^N \frac{\frac{d_i}{d_{\min}} (\tau d_{\min} - 1)}{\left(1 + \frac{d_i}{d_{\min}} (\tau d_{\min} - 1)\right)^2}$$

*Proof.* Using the last inequality from Theorem 3.1 and summing over all  $i$ , we obtain the bounds.  $\square$

**COROLLARY 3.3.** *For the effective spreading rate  $\tau \geq \frac{2}{\mu_{N-1}}$  and large  $N$ , where  $\mu_{N-1}$  is the second smallest eigenvalue of the Laplacian matrix, the infection probability for individual nodes, as well as the variance of infection, is tightly bounded with the nodal degree.*

*Proof.* The bounds from [11] and Corollary 3.2 imply that, for  $\tau \geq \frac{2}{d_{\min}}$ , the infected fraction and the variance of infection depend mostly on the node degree. Using Grone and Merris' bound [19] on the second smallest eigenvalue of the Laplacian matrix,  $\mu_{N-1} \leq d_{\min}$ , we have  $\frac{2}{d_{\min}} \leq \frac{2}{\mu_{N-1}}$ .  $\square$

Corollary 3.3 provides a deeper understanding of the structural properties of a graph. A graph with a larger  $\mu_{N-1}$  is more strongly interconnected (see [20]) and a lower value of the effective spreading rate is enough for the infection probability of a node to depend more on the degree than on the hopcount or the structure of a graph. Similarly, for a small  $\mu_{N-1}$ , it takes a larger effective spreading rate  $\tau$  to enlarge the probability of uniformly spreading the infection among all separable clusters. Since the second smallest eigenvalue of the Laplacian determines the bound on the standard isoperimetric constant  $\frac{1}{\eta(G)} \leq \frac{2}{\mu_{N-1}(Q)}$ , it determines the life-time of an epidemic [4]. To summarize, for  $\tau \geq \frac{2}{\mu_{N-1}}$  the life-time of an epidemic is of order  $\Omega(e^N)$  and the variance can be expressed as a function of the node degrees and effective infection rate.

The variance of the number of infected nodes as a function of  $\tau$  is given in Figure 3.2. Numerical results from  $N$ -intertwined model indicate that the maximum of the variance of the number of infected nodes is reached approximately for  $\tau \simeq \frac{2}{d_{av}}$  for considered graphs (complete bipartite, power-law, regular graphs, random graphs (Erdős-Rényi)). This observation implies that the graphs with the same link density <sup>2</sup>  $p = \frac{L}{\binom{N}{2}}$  will have a maximal variance for the same effective spreading rate  $\tau$ . We ought to emphasize that only the epidemics in the metastable state are considered. For two structurally different graphs with the same  $d_{av}$ , namely the star topology and the line topology the maximum of the variance of the number of infected nodes is reached for the same value of  $\tau$ , although the thresholds  $\tau_c$  are different, as shown in Figure 3.2. Among all graphs, the line graph has the largest hopcount and the star the shortest. In the steady-state, the  $N$ -intertwined model for these two special graphs, shows that the variance of the number of infected nodes is indeed maximal for  $\tau \simeq \frac{2}{d_{av}} \approx 1$ .

If we use the KW model [2] with  $k = d_{av}$ , half of the nodes will be infected for  $\tau = \frac{2}{d_{av}}$ . Numerical calculation of the  $N$ -intertwined model in Figure 2.1 suggests that the maximum of the variance is reached for  $\tau \simeq \frac{2}{d_{av}}$  and that the fraction of infected nodes is close to half.

**LEMMA 3.4.** *For the effective spreading rate  $\tau \leq \frac{1}{d_{av}}$ , the fraction of infected nodes is bounded by  $y_\infty \leq \frac{1}{2}$*

*Proof.* For every node  $i$ , it holds that  $v_i \leq 1 - \frac{1}{1+\tau d_i}$ . After summing over all  $i$  and dividing by  $N$  we have  $y_\infty \leq 1 - \frac{1}{N} \sum_{i=1}^N \frac{1}{1+\tau d_i}$ . Using the harmonic-geometric-arithmetic

---

<sup>2</sup>In any graph, the average degree  $d_a = \frac{2L}{N}$ , where  $L$  is the number of links in the graph.

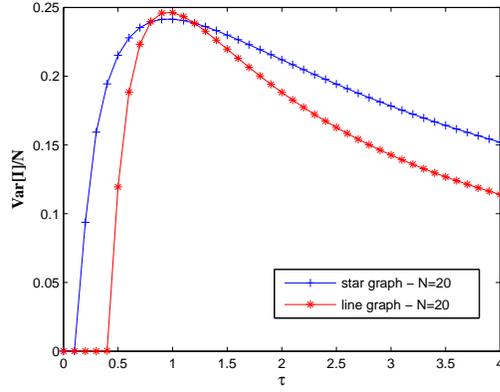


FIG. 3.2. Variance as a function of  $\tau$ . Dashed line denotes a star topology and the full line denotes a line graph.  $N = 20$ .

mean inequality valid for  $\forall a_i \geq 0$

$$\frac{N}{\sum_{i=1}^N \frac{1}{a_i}} \leq \sqrt[N]{\prod_{i=1}^N a_i} \leq \frac{1}{N} \sum_{i=1}^N a_i \quad (3.7)$$

we have

$$1 - \frac{1}{N} \sum_{i=1}^N \frac{1}{1 + \tau d_i} \leq 1 - \frac{N}{\sum_{i=1}^N (1 + \tau d_i)} \leq 1 - \frac{1}{1 + \tau d_{av}}$$

We can now deduce that, if  $\tau \leq \frac{1}{d_{av}}$ , the fraction of infected nodes is  $y_\infty \leq \frac{1}{2}$ .  $\square$

The fraction of infected nodes (2.4) is estimated [11] as  $y_\infty \approx 1 - \frac{1}{\tau N} \sum_{i=1}^N \frac{1}{d_i}$ .

Invoking (3.7), we have that  $y_\infty = 1/2$  is reached for

$$\tau \approx \frac{2}{N} \sum_{i=1}^N \frac{1}{d_i} \geq \frac{2}{d_{av}}$$

#### 4. Examples: Complete bipartite graph $K_{m,n}$ and complete graph.

In this section, we will consider complete graph and complete bi-partite graph as an examples of the model presented. We will first discuss the complete graph. Due to the symmetry, the set of  $N$  equations in the steady state reduces to only one equation.

$$v_\infty = \frac{\tau(N-1) - 1}{(N-1)\tau} \quad (4.1)$$

The variance as a function of  $\tau$  (eq. 3.2)

$$\text{Var}[I] = \frac{(\tau(N-1) - 1)N}{(N-1)^2\tau^2} \quad (4.2)$$

The maximum variance is reached for  $\tau = \frac{2}{d_{av}}$ . For  $\tau = \frac{2}{d_{av}}$ , the average number of infected nodes is  $\frac{N}{2}$ . We compare this formula with the event driven simulations in section (5).

A complete bi-partite graph  $K_{M,N}$  consists of two disjoint sets  $S_1$  and  $S_2$  containing respectively  $M$  and  $N$  nodes, such that all nodes in  $S_1$  are connected to all nodes in  $S_2$ , while within each set no connections occur. Figure 4.1 gives an example of a complete bi-partite graph on 6 nodes.

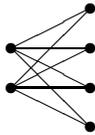


FIG. 4.1. Complete bi-partite graph  $K_{2,4}$

Notice that (core) telecommunication networks often can be modeled as a complete bi-partite topology. For instance, the so-called double-star topology (i.e.  $K_{M,N}$  with  $M = 2$ ) is quite commonly used because it offers a high level of robustness against link failures. For example, the Amsterdam Internet Exchange (see [www.amstix.net](http://www.amstix.net)), one of the largest public Internet exchanges in the world, uses this topology to connect its four locations in Amsterdam to two high-density Ethernet switches. Sensor networks are also often designed as complete bi-partite graphs.

Due to the symmetry, for complete bi-partite graph the set of  $N$  equations for  $N$ -intertwined model reduces to only two [17], [11].

$$v_N = \frac{\tau^2 MN - 1}{N\tau(M\tau + 1)}; v_M = \frac{\tau^2 MN - 1}{M\tau(N\tau + 1)} \quad (4.3)$$

Applying (3.2) to the complete bi-partite graph yields

$$\text{Var}[I] = \frac{(\tau^2 MN - 1)(1 + M\tau)}{M\tau^2(N\tau + 1)^2} + \frac{(\tau^2 MN - 1)(1 + N\tau)}{N\tau^2(M\tau + 1)^2} \quad (4.4)$$

By substituting  $M = \frac{rN_{tot}}{r+1}$  and  $N = \frac{N_{tot}}{r+1}$ , where  $N_{tot}$  is the total number of nodes in bipartite graph and  $r \in (\frac{1}{N_{tot}}, 1)$ , the expression for the number of infected nodes simplify and it can be shown that for  $\tau = \frac{2}{d_{av}}$  the percent of infected nodes is in the range (0.5, 0.52). The variance is also close to the maximum  $\text{Var}[I] \in (0.25N_{tot}, 0.023N_{tot})$ .

Because the steady state probabilities depend only on the effective spreading rate the steady state probability  $\Pr[I_N, I_M]$  satisfies [17]

$$\Pr[I_N = x, I_M = y] = \binom{N}{x} i_\infty^x (1 - i_\infty)^{N-x} \binom{M}{y} j_\infty^y (1 - j_\infty)^{M-y} \quad (4.5)$$

From the joined probability distribution (4.5), we can derive the probability distribution of the number of infected nodes for the whole network  $\Pr[I_\infty]$ . We will compare this formula (4.5) with the event driven simulations in section (5).

**5. Simulation results.** In this section, we will compare the theoretical results for epidemic variance from previous sections and event driven simulations. We will calculate the variance as a function of the effective spreading rate using the  $N$ -intertwined model, equation (3.2) and test its accuracy with event driven simulations of a spreading process.

The results are presented for the complete bipartite, complete, and line graph with  $N = 100$  nodes. Small networks of  $N = 100$  are enough to demonstrate our results. The variance  $Var[I_\infty]$  as a function of  $\tau$  is shown in the Figure 5.1.

We have also conducted a set of event driven simulations for  $N = 100$  and compared them to the calculations. We have conducted  $10^4$  simulations with 2 initially infected nodes and  $\tau > \tau_c^* = \frac{2}{d_{av}}$ . For the values of  $\tau < \tau_c^*$ , a large portion of the epidemics dies out during the initial phase, which prevents a good estimation of the variance in the steady state.

The virus spread is a stochastic process and during the transition phase some of the infections can die out, even though the effective spreading rate is above the threshold  $\tau_c$ . These extinctions have been excluded from calculations for the expected number of infected nodes in the steady state.

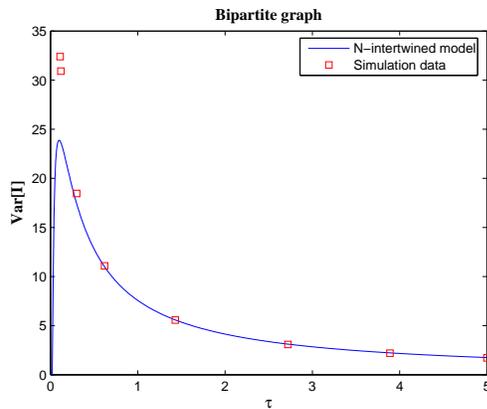


FIG. 5.1. Variance as a function of  $\tau$ . Model results compared with the event driven simulations for the bipartite graph ( $N = 100$ ,  $m = 10$ ), and the full mesh.  $N = 100$ .

The large variance observable in line graph for  $\tau \sim \frac{2}{d_{av}}$  can be explained by the fact that the infected nodes are forming clusters and the covariance of two infected nodes is large. In [11], it was shown that the  $N$ -intertwined model gives better predictions for large  $N$ .

In the interval  $\tau \in (\tau_c, \tau_c^*)$  the model predicts a decrease of variance as  $\tau$  decreases towards  $\tau_c$ , but in this interval the covariance increases which implies that the variance can also increase. The fact that the variance is increasing means that an extinction may take place more often. If the variations are larger than the mean value in the steady state, the infection will extinct very quickly. The event driven simulations also show that the variance grows larger and the extinction is more probable as  $\tau \downarrow \tau_c$ .

The maximal variance for the bipartite graph is smaller than for the line and complete graph. The reason is a significant degree diversity in the bipartite graph. The degree diversity has as a consequence that not all the nodes will have  $v_{i\infty} = 0.5$  for the same  $\tau$ .

The figure 5.3 shows that the worst prediction is given for the line graph which

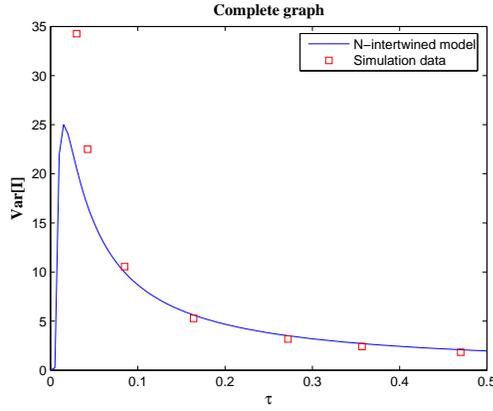


FIG. 5.2. Variance as a function of  $\tau$ . Model results compared with the event driven simulations for the complete graph.  $N = 100$ .

has the largest diameter. A small number of neighbors induces high dependency on the state of a neighbor. The  $N$ -intertwined model gives worst prediction in the region of  $\frac{2}{d_{av}} < \tau \leq \frac{2}{d_{min}}$  (Figure 5.3).

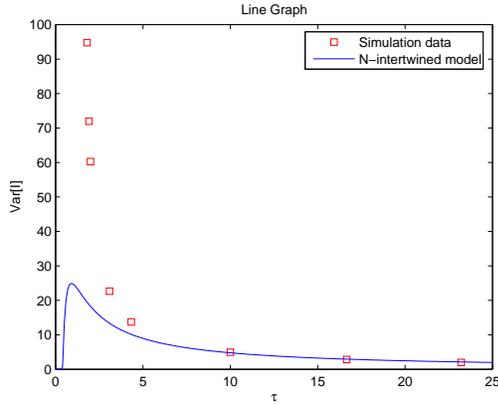


FIG. 5.3. Variance as a function of  $\tau$ . Model results compared with the event driven simulations for the line graph with correction for covariance.  $N = 100$ .

We conducted event driven simulations for the complete bipartite graphs  $K_{10,990}$ ,  $K_{250,750}$  and  $K_{500,500}$  with different effective spreading rates  $\tau$ . We have assumed that the system is in the steady state from  $t = 6000$  time units onwards. The probability distribution for the number of infected nodes in steady state is compared with the probability distribution given by Eq. (4.5). In Figures 5.4, 5.5 and 5.6 dashed lines represent event driven simulations, full lines represent theoretical predictions. It is possible that virus dies out during the evolution of the epidemic and the probabilities for different effective spreading rates are given.

We conclude from the simulations that Eq. (4.5) predicts the probability distribution of the number of infected nodes in steady-state very well for large values of the effective spreading rate  $\tau$ .

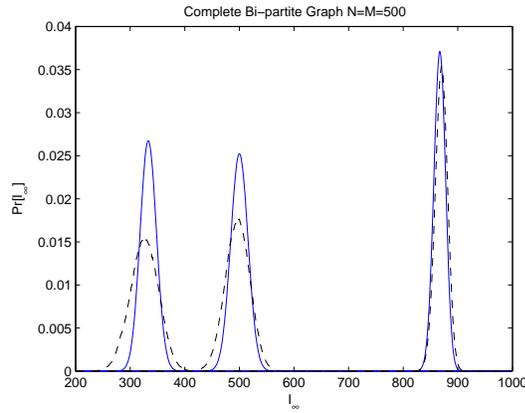


FIG. 5.4. *Probability distribution of the number of infected nodes in the steady state for  $K_{500,500}$ ,  $\tau \in \{0.003, 0.004, 0.015\}$  with the average number of infected  $I_\infty \in \{333.33, 500, 866.66\}$  respectively. The probability of extinction during the initial phase of simulations  $Pr[I_\infty = 0] \in \{0.14, 0.038, 0\}$  respectively.*

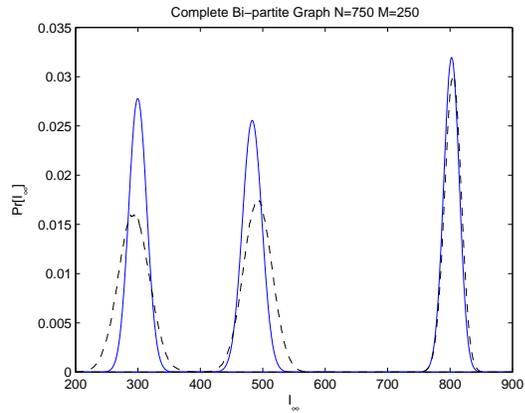


FIG. 5.5. *Probability distribution of the number of infected nodes in the steady state for  $K_{250,750}$ ,  $\tau \in \{0.0035, 0.005, 0.015\}$  with the average number of infected  $I_\infty \in \{299.83, 483.04, 802.21\}$  respectively. The probability of extinction during the initial phase of simulations  $Pr[I_\infty = 0] \in \{0.2, 0.034, 0.002\}$  respectively.*

**6. Conclusions.** The *SIS* model is especially interesting for its so-called metastable state and dynamic behavior. Motivated by the importance of the underlying interaction structure on epidemic spreading, we have explored the variance of the number of infected nodes in *SIS* model for a fixed topology defined by a symmetric adjacency matrix  $A$ . Using the  $N$ -intertwined model, we have calculated the variance of the number of infected nodes and shown the model limitations for variance estimation. The implicit formula and application regions are also given. We have shown that for  $\tau > \frac{1}{d_i} + \frac{1}{d_{\min}}$  the variance is highly dependent on the degree distribution and not on the distance between nodes or other topological properties. Further, we have exemplified the general theory by two specific graphs, namely complete and complete bipartite graphs. For complete bipartite graphs, we have derived the probability distribution of the number of infected nodes.

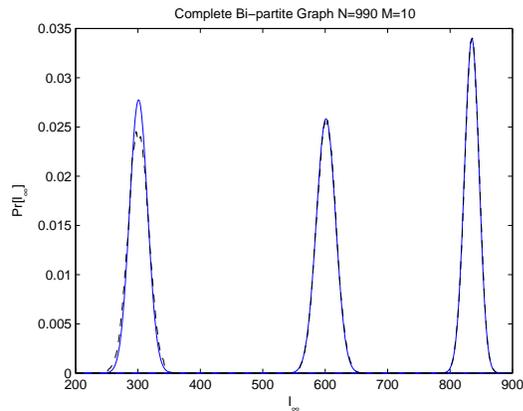


FIG. 5.6. Probability distribution of the number of infected nodes in the steady state for  $K_{10,990}$ ,  $\tau \in \{0.045, 0.15, 0.5\}$  with the average number of infected  $I_\infty \in \{301.2, 601.22, 834.63\}$  respectively. The probability of extinction during the initial phase of simulations  $Pr[I_\infty = 0] \in \{0.16, 0.006, 0\}$  respectively.

In the previous work [11], we have established the relation between the largest eigenvalue of the adjacency matrix  $A$  and the threshold. In this paper, we have related the spreading process to the second smallest eigenvalue of the Laplacian matrix (Corollary 3.3). Regarding the second smallest eigenvalue of the Laplacian matrix, we have discussed the relation between the variance of the number of infected nodes and the lifetime of an epidemic.

Finally, we have conducted event driven simulations that confirm results and indicate that the maximum variance for the  $N$ -intertwined model is reached for  $\tau \approx \frac{2}{d_{av}}$ .

**7. Acknowledgement.** This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 643.000.503, by the European project ResumeNet (FP7-224619) and by Next Generation Infrastructures (Bsik).

#### REFERENCES

- [1] D. J. DALEY AND J. GANI, *"Epidemic modelling: An Introduction"*, Cambridge University Press, 1999.
- [2] J. O. KEPHART AND S. R. WHITE, *"Direct-graph epidemiological models of computer viruses"*, Proc. 1991 IEEE Computer Society Symposium on Research in Security and Privacy, 343-359, May 1991.
- [3] R. PASTOR-SATORRAS AND A. VESPIGNANI, *"Epidemic Spreading in Scale-Free Networks"*, Phys. Rev. Lett., Vol. 86, No. 14, 3200-3203, April 2001.
- [4] A. GANESH, L. MASSOULIÉ AND D. TOWSLEY, *"The Effect of Network Topology on the Spread of Epidemics"*, Proc. IEEE INFOCOM2005, 2005.
- [5] M. DRAIEF, A. GANESH AND L. MASSOULIÉ, *"Threshold for virus spread on Networks"*, Ann. Appl. Probab. Volume 18, Number 2 (2008), 359-378.
- [6] P. TH. EUGSTER, R. GUERRAOUI, A.-M. KERMARREC, L. MASSOULIÉ, *"From Epidemics to Distributed computing"*, IEEE Computer, Vol. 37, No.5, 60-67, May 2004.
- [7] D. CHAKRABARTI, J. LESKOVEC, C. FALOUTSOS, S. MADDEN, C. GUESTRIN, M. FALOUTSOS, *"Information Survival Threshold in Sensor and P2P Networks"*, Proc. IEEE INFOCOM2007, 2007.
- [8] F. G. COFFMAN JR., Z. GE, V. MISRA AND D. TOWSLEY, *"Network resilience: exploring cas-*

- cading failures within BGP*", Proc. 40th Annual Allerton Conference on Communications, Computing and Control, October 2002.
- [9] N. T. J. BAILEY, "*The Mathematical Theory of Infectious Diseases and its Applications*", Charlin Griffin & Company, London, 2nd ed., 1975.
- [10] S. BOCCALETTI, V. LATORA, Y. MORENO, M. CHAVEZ, D. U. HWANG, "*Complex networks: Structure and dynamics*", Phys. Rev., Vol. 424, No. 4-5. (February 2006), pp. 175-308.
- [11] P. VAN MIEGHEM, J. OMIĆ AND R. KOOLIJ, "*Virus Spread in Networks*", IEEE Transactions on Networking, Vol. 17, No. 1, Feb. 2009, pp. 1-14.
- [12] P. CREPEY, F. P. ALVAREZ AND M. BARTHELEMY, "*Epidemic variability in complex networks*", Phys. Rev. E: Stat. Nonlin. Soft. Matter Phys. 73, Apr 2006.
- [13] R. PASTOR-SATORRAS AND A. VESPIGNANI, "*Evolution and Structure of the Internet, A Statistical Physics Approach*", Cambridge University Press, 2004.
- [14] Y. WANG, D. CHAKRABARTI, C. WANG AND C. FALOUTSOS, "*Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*", In 22nd International Symposium on Reliable Distributed Systems (SRDS'03), pp. 25-34. IEEE Computer, October 2003.
- [15] F. R. K. CHUNG, "*Spectral Graph Theory*", CBMS Conference on Recent Advances in Spectral Graph Theory, California State University of Fresno, June 6-10, 1994.
- [16] P. VAN MIEGHEM, "*Performance Analysis of Communication Systems and Networks*", Cambridge University Press, 2006.
- [17] J. OMIĆ, R. KOOLIJ AND P. VAN MIEGHEM, "*Virus Spread in Complete Bi-partite Graphs*", Proc. BIONETICS, 2007.
- [18] N. ALON AND V. D. MILMAN, " $\lambda_1$  isoperimetric inequalities for graphs and supercontractors", J. Combin. Theory Ser. B, 38, 73-88, 1985.
- [19] R. GRONE AND R. MERRIS, "The Laplacian spectrum of a graph  $\pi^*$ ", SIAM Journal on Discrete Mathematics, Vol. 7, No. 2, May, pp. 221-229, (1994).
- [20] B. BOLLOBAS, "*Modern Graph Theory*", Graduate Texts in Mathematics, Springer, New York, 1998.
- [21] D. A. RAND, "*Correlation equations and pair approximation for spatial ecologies*", Advanced Ecological Theory, Principles and Applications, Oxford, Blackwell Science, 100-142, 1999.