Epidemic Thresholds in Real Networks

DEEPAYAN CHAKRABARTI, YANG WANG, CHENXI WANG, JURE LESKOVEC AND CHRISTOS FALOUTSOS
Carnegie Mellon University

How will a virus propagate in a real network? How long does it take to disinfect a network given particular values of infection rate and virus death rate? What is the single best node to immunize? Answering these questions is essential for devising network-wide strategies to counter viruses. In addition, viral propagation is very similar in principle to the spread of rumors, information and "fads", implying that the solutions for viral propagation would offer insights into these other problem settings too.

We answer these questions by developing a Non-Linear Dynamical System (*NLDS*) that accurately models viral propagation in any *arbitrary* network, including real and synthesized network graphs. We propose a general epidemic threshold condition for the *NLDS* system: we prove that the epidemic threshold for a network is exactly the inverse of the largest eigenvalue of its adjacency matrix. Finally, we show that below the epidemic threshold, infections die out at an exponential rate.

Our epidemic threshold model subsumes many known thresholds for special-case graphs (e.g., Erdös-Rényi, BA power-law, homogeneous). we demonstrate the predictive power of our model with extensive experiments on real and synthesized graphs, and show that our threshold condition holds for arbitrary graphs. Finally, we show how to utilize our threshold condition for practical uses: It can dictate which nodes to immunize; it can assess the effects of a throttling policy; it can help us design network topologies so that they are more resistant to viruses.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications – Data Mining

General Terms: Measurement, Theory

Additional Key Words and Phrases: Viral Propagation, Epidemic threshold, Eigenvalue

1. INTRODUCTION

Computer viruses remain a significant threat to today's networks and systems. Existing defense mechanisms typically focus on local scanning of virus signatures. While these mechanisms can detect and prevent the spreading of known viruses, they prove to be of little help in designing globally optimal defenses. The recent proliferation of Distributed Denial of Service attacks in conjunction with viral spread exacerbates the problem [23]. The magnitude of viral propagation means that DDoS attacks riding on the tail of viruses and worms can occur in an unprecedented scale and are therefore especially detrimental. With the exception of a few specialized modeling studies [18; 19; 28; 31; 34], much still remains unknown about the propagation characteristics of computer viruses and the factors that influence them. The key question in all these cases is: *Will the virus/worm "linger for ever" the entire network, or will it die out?*

In this paper, we investigate epidemiological modeling techniques to answer this question. Specifically, we are interested in the following:

—*How does a virus spread over a network?* Answering this requires a general analytic model of viral propagation, which can capture the impact of the underlying topology without being limited by any assumptions about the topology.

- —*Does an epidemic threshold exist, and if so, how can we find it?* The epidemic threshold condition is a condition linking the characteristics of the virus and the network topology such that, if the condition is satisfied, then a viral infection dies out over time. We want to derive this threshold, based on the propagation model above.
- —*If the epidemic threshold condition is satisfied, how fast will the network get disinfected?*
- —*Which node is best to immunize?* Should it be the one with the highest degree, as the "targeted" immunization policy suggests? Or maybe the one with the highest PageRank score? We would like the resulting graph to be as resistant as possible, that is, with the highest possible threshold.

In this paper, we propose answers to exactly these questions. We develop the *NLDS* (Non-Linear Dynamical Systems) approach which accurately models viral propagation, and find the epidemic threshold in this *NLDS* system. Our solutions hold for arbitrary graphs and render surprisingly simple yet accurate predictions.

The layout of this paper is as follows. Section 2 gives a background review of previous models. In Section 3, we describe our proposed model. We show that our model conforms better to simulation results than previous models over real networks. In Section 4, we compute the epidemic threshold and present a surprising new result—the epidemic threshold of a given network is related intrinsically to (and depends *only* on) the first eigenvalue of its adjacency matrix. In Section 5, we demonstrate the accuracy of our results by experiments on several real and synthetic datasets. We conclude in Section 7. Appendices A and B provide detailed proofs of our theorems, and some discussions on the complimentary nature of some very recent follow-up work on this topic [13].

2. RELATED WORK

The problem of virus propagation has attracted huge interest. Here we survey some of the related work, focusing mainly on (a) epidemic thresholds for real and realistic graphs (b) immunization policies and (c) related results from epidemiology.

Among the many proposed models for viral propagation, two have garnered wide acceptance. The first, called the SIS model, considers individuals as being either susceptible (S) or infective (I); a susceptible individual can become infective on contact with another infective individual, then heal herself with some probability to become susceptible again. The second, called the SIR model, is similar with the only difference being that once healed, an individual is considered removed (R) from the population and immune to further infection. Intuitively, SIS models the flu, while SIR models mumps. While both are important, we focus on the SIS model in this paper.

2.1 Earlier epidemic thresholds and limitations

The class of epidemiological models that are most widely used are the so-called *homogeneous models* [2; 24; 1]. A homogeneous model assumes that every individual has equal contact to others in the population, and that the rate of infection is largely determined by the density of the infected population. Kephart and White [18; 19] were among the first to propose epidemiology-based models to analyze the propagation of computer viruses. In their model, the communication among individuals is modeled as a directed graph: a directed edge from node i to node j denotes that i can directly infect j. A rate of infection, called the birth rate β, is associated with each edge. A virus death rate δ (also called the node curing rate), is associated with each infected node.

If we denote the size of the infected population at time t as η_t, a deterministic time evolution of η_t in the Kephart-White model (hereafter referred to as the KW model) can be represented as:

$$\frac{d\eta_t}{dt} = \beta\langle k\rangle\eta_t(1-\frac{\eta_t}{N}) - \delta\eta_t \tag{1}$$

where $\langle k\rangle$ is the average connectivity (degree) and N is the total number of nodes. The steady state solution for Equation 1 is $\eta = N\left(1-\frac{\delta}{\beta\langle k\rangle}\right)$.

An important prediction of Equation 1 is the notion of an *epidemic threshold.* Intuitively, the epidemic threshold τ is a value such that a viral outbreak dies out quickly if

$$\beta/\delta < \tau \tag{2}$$

For the KW model, the proposed epidemic threshold was:

$$\tau_{KW} = \frac{1}{\langle k\rangle} \tag{3}$$

where $\langle k\rangle$ is the average connectivity **[18]**.

The KW model provides a good approximation of virus propagation in networks where the contact among individuals is sufficiently homogeneous. However, there is overwhelming evidence that real networks (including social networks [32], router and AS networks [12], and Gnutella overlay graphs [33]) deviate from such homogeneity—they follow a power law structure instead. In other words, if $P(k)$ is the probability of a node having degree k, then $P(k) \propto k^{-\gamma}$, where γ is called the power-law exponent. In such a structure, there exist a few nodes with very high connectivity, but the majority of the nodes have low connectivity. The high-connectivity nodes are expected to often get infected and then propagate the virus, making the infection harder to eradicate.

Pastor-Satorras and Vespignani studied viral propagation for such power-law networks [26; 28; 30; 31]. They developed an analytic model for the Barabási-Albert (BA) power-law topology [3]. Their steady state prediction is:

$$\eta = 2Ne^{-\delta/m\beta} \tag{4}$$

where m is the minimum connectivity in the network. However, this derivation depends critically on that fact that $\gamma = 3$ in the BA model, which does not hold for many real networks [20; 12]. Pastor-Satorras et al. [30] also proposed an epidemic threshold condition, but this uses the "mean-field" approach, where all graphs with a given degree distribution are considered equal. There is no particular reason why all such graphs should behave similarly in terms of viral propagation. The proposed epidemic threshold for this MFA (for Mean-Field Approximation) model is:

$$\tau_{MFA} = \frac{\langle k\rangle}{\langle k^2\rangle} \tag{5}$$

where $\langle k \rangle$ is the expected connectivity and $\langle k^2 \rangle$ signifies the connectivity divergence (i.e., sum of squared degrees). However, we observe experimentally that this model yields less than accurate predictions for many networks, as we will show later.

Several follow-up attempts focus on analyzing even more realistic graph models. Eguiluz and Klemm [10] derive a more accurate epidemic threshold for real graphs, namely $1/(< k > -1)$. The derivation assumes their earlier model of *highly clustered* power law graphs, which have more realistic behavior than power law graphs with random wiring. Their simulation results on several Internet graphs show that their threshold is more accurate.

Boguñá and Satorras considered the spread of a virus in correlated networks where the connectivity of a node is related to the connectivity of its neighbors [5]. These *correlated networks* include Markovian networks where, in addition to $P(k)$, a function $P(k|k')$ determines the probability that a node of degree k is connected to a node of degree k'.

While some degree of correlations may exist in real networks, it is often difficult to characterize connectivity correlations with a simple $P(k|k')$ function. Prior studies on real networks [12; 27] have not found any conclusive evidence to support the type of correlation as defined in [5]. Hence, we will not discuss models for correlated networks further in this paper.

In the following sections, we will present a new analytic model (called *NLDS*) that makes *no assumptions* about the network topology. In other words, *NLDS* does not depend on the presence of any specific structure in the topology. We will show later that *NLDS* performs as well or better than some previous models that are tailored to fit special-case graphs (homogeneous, BA power-law, etc.) Then, we shall derive the epidemic threshold condition for *NLDS*.

In a recent follow-up to our original paper on this model [35], Ganesh et al. [13] also obtained the same epidemic threshold result (along with other results). Their approach is complimentary to ours; they derive an exact bound on the viral propagation equations, whereas we use a point estimate. Appendix B gives details.

2.2 Immunization

Briesemeister et al [6] focus on immunization of power law graphs. They focus on the random-wiring version (that is, standard preferential attachment), versus the “highly clustered” power law graphs of Klemm and Eguiluz (KE). Their simulation experiments on such synthetic graphs show that KE graphs can be more easily defended against viruses, while random-wiring ones are typically overwhelmed, despite identical immunization policies.

Cohen et al [9] studied the *acquaintance* immunization policy, and showed that it is much better than random, for both the SIS as well as the SIR model. The “acquaintance” immunization policy works as follows: pick a random person, and immunize one of its neighbors at random (which will probably be a ’hub’). For power law graphs (with no rewiring), they also derived formulas for the critical immunization fraction f_c, above which the epidemic is arrested. Madar et al [22] continued along these lines, mainly focusing on the SIR model for scale-free graphs. They linked the problem to bond percolation, and they derived formulas for the effect of several immunization policies, showing that the “acquaintance” immunization policy is the best. Both works were analytical, without studying any real graphs.

Hayashi et al [14] study the case of a growing network, and they derive analytical formulas for such power law networks (no rewiring). They introduce and study the SHIR model (Susceptible, Hidden, Infectious, Recovered), to model computers under e-mail virus attack. and they derive the conditions for extinction under random and under targeted immunization, always for power law graphs with no rewiring.

2.3 Analytical results

Berger et al [4] did an analytical study of graphs generated by the preferential-attachment method, and specifically focused on SIS epidemics, and the effect that the starting node has. Chung et al [8; 7] study the behavior of eigenvalues for power law graphs (random wiring). Among the several interesting results, they prove that, under mild assumptions, the eigenvalues of the adjacency matrix follows a power law, while the spectrum of the Laplacian matrix follows the semi-circle law.

2.4 Survey of epidemiology results

Hethcote [15] gives an overview of the analysis of epidemics, with the typical SIS and SIR models, several of their extensions, the differential equations that model the population evolution, and policies and effects of immunization. In all cases, the topology is not considered, implicitly taken to be a complete clique, or a homogeneous graph, that is, all nodes have similar degrees. The only exception is the non-homogeneous model in [16], where the topology is a collection of cliques, with the behavior of every clique member being identical to the behavior of the rest clique-members. The model is a continuous time model, and, using a theorem by Yorke and Lajmanovich, they show that the epidemic threshold is related to the first eigenvalue of the appropriate group-to-group interaction matrix. Our upcoming model is related, but more general, because it needs no assumption about the topology; moreover, it uses a more elaborate discrete-time model, and it explicitly states its assumption (the independence assumption of Eq. 6).

In conclusion, none of the earlier methods focuses on epidemic thresholds for arbitrary, real graphs, with only exceptions our earlier conference paper [35], and its follow-up paper by Ganesh et al. [13].

3. PROPOSED MODEL

In this section, we describe our model of virus propagation (called *NLDS*), which does not rely on the presence of homogeneous connectivity or any particular topology. We assume a connected undirected network $\mathcal{G} = (N, E)$, where N is the number of nodes in the network and E is the set of edges. We assume a universal infection rate β for each edge connected to an infected node, and a virus death rate δ for each infected node. Table I lists the symbols used.

3.1 Model

Our model works with small discrete time-steps Δt, with $\Delta t \to 0$. This is only for ease of exposition; the same results hold for the continuous case too. During each time interval Δt, an infected node i tries to infect its neighbors with probability β. At the same time, i may be cured with probability δ.

Symbol	Description
$\mathcal{G}$	An undirected connected graph
N	Number of nodes in $\mathcal{G}$
E	Number of edges in $\mathcal{G}$
$\mathbf{A}$	Adjacency matrix of $\mathcal{G}$
$\mathbf{A}'$	The transpose of matrix $\mathbf{A}$
β	Virus birth rate on a link connected to an infected node
δ	Virus death rate on an infected node
t	Time stamp
$p_{i,t}$	Probability that node i is infected at t
$\vec{\mathbf{P}}_{\mathbf{t}}$	$\vec{\mathbf{P}}_{\mathbf{t}} = (p_{1,t}, p_{2,t}, \ldots, p_{N,t})'$
$\zeta_{i,t}$	Probability that node i does not receive infections from its neighbors at t
$\lambda_{i,A}$	The i-th largest eigenvalue of $\mathbf{A}$
$\vec{\mathbf{u}}_{\mathbf{i},\mathbf{A}}$	Eigenvector of $\mathbf{A}$ corresponding to $\lambda_{i,A}$
$\vec{\mathbf{u}}'_{\mathbf{i},\mathbf{A}}$	Transpose of $\vec{\mathbf{u}}_{\mathbf{i},\mathbf{A}}$
$\mathbf{S}$	The 'system' matrix describing the equations of infection
$\lambda_{i,S}$	The i-th largest eigenvalue of $\mathbf{S}$
s	Score $s = \beta/\delta \cdot \lambda_{1,A}$
η_t	Number of infected nodes at time t
$\langle k \rangle$	Average degree of nodes in a network
$\langle k^2 \rangle$	Connectivity divergence (sum of squared degrees)

Table I. Table of Symbols

This process can be modeled as a Markov chain with 2^N states. Each state in the Markov chain corresponds to one particular system configuration of N nodes, each of which can be in one of two states (Susceptible or Infective), which leads to 2^N possible configurations. Also, the configuration at time-step $t+1$ depends only on that at time-step t; thus, it is a Markov chain.

This Markov chain also has an "absorbing" state, when all nodes are uninfected (i.e., Susceptible). This absorbing state can be reached from any starting state of the Markov chain, implying that this state will be reached with probability 1 over a long period of time. However, this state could be reached very quickly, or it could take time equivalent to the age of the universe (in which case, the viral epidemic practically never dies).

The obvious approach of solving the Markov chain becomes infeasible for large N, due to the exponential growth in the size of the chain. To get around this limitation, we use the "independence" assumption and replace the problem with Equation 7 (discussed below), which is solvable.

Let the probability that a node i is infected at time t by $p_{i,t}$. Let $\zeta_{i,t}$ be the probability that a node i will not receive infections from its neighbors in the next time-step. This happens if each neighbor is either uninfected, or is infected but fails to spread the virus with probability $(1-\beta)$. Since we consider infinitesimal timesteps ($\Delta t \to 0$), the probability of multiple events within the same Δt is negligible compared to first-order effects, and can be ignored.

$$\begin{aligned}\zeta_{i,t} &= \prod_{j:neighbor\ of\ i} (p_{j,t-1}(1-\beta) + (1-p_{j,t-1})) \\ &= \prod_{j:neighbor\ of\ i} (1-\beta * p_{j,t-1}) \qquad (6)\end{aligned}$$

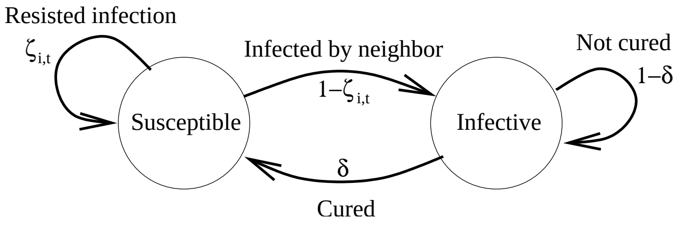


Fig. 1. *The SIS model, as seen from a single node:* Each node, at each time step t, is either Susceptible (S) or Infective (I). A susceptible node i is currently healthy, but can be infected (with probability $1-\zeta_{i,t}$) by receiving the virus from a neighbor. An infective node can be cured with probability δ; it then goes back to being susceptible. Note that $\zeta_{i,t}$ depends on the both the virus birth rate β and the network topology around node i.

This is the *independence assumption:* we assume that probabilities $p_{j,t-1}$ are independent of each other.

A node i is healthy at time t if it did not receive infections from its neighbors at t *and* i was uninfected at time-step $t-1$, or was infected at $t-1$ but was cured at t. Denoting the probability of a node i being infected at time t by $p_{i,t}$:

$$1 - p_{i,t} \; = \; (1 - p_{i,t-1})\zeta_{i,t} + \delta p_{i,t-1}\zeta_{i,t} \quad i = 1 \dots N \tag{7}$$

This equation represents our *NLDS* (Non-Linear Dynamical System). Figure 1 shows the transition diagram[1].

Given a network topology and particular values of β and δ, we can solve Equation 7 numerically to obtain the time evolution of the infected population size η_t, where $\eta_t = \sum_{i=1}^{N} p_{i,t}$.

3.2 Accuracy of *NLDS*

Having described our model, we will now show that it closely models the propagation of viruses over different networks. The datasets used for these experiments are:

—*Oregon:* This is a real network graph collected from the Oregon router views. It contains 32,730 links among 11,461 AS peers. More information can be found from
`http://topology.eecs.umich.edu/data.html`

—*BA power-law:* These are graphs with power-law degree distributions, as described in [3]. We synthesized 1000-node power-law graphs using BRITE [25], with the parameters $m_0 = 10$ and $m = 2$.

—*Homogeneous:* We used 1000-node graphs generated by the Erdős-Rényi model [11].

Unless otherwise specified, each simulation plot is averaged over 15 individual runs. Different values of β and δ were used to check robustness, but all that matters (as we show later) is the ratio β/δ.

[1] The $\zeta_{i,t}$ factor eventually makes no difference to the final threshold result, whether we keep it or not. We decided to keep it for backward compatibility with the conference paper in SRDS [35], but it makes no difference. The reason is that the time-step Δt is so small that the ζ factor eventually vanishes.

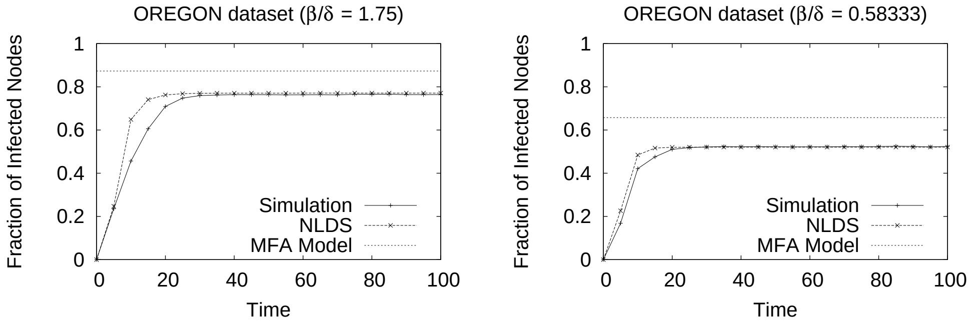


Fig. 2. *Experiments on the real-world Oregon graph:* The plots show the time evolution of infection in the *Oregon* network. Both simulations were performed with fixed β, but varying δ. In both cases, our model conforms more precisely to the simulation results than the model in[30].

We begin each simulation with a set of randomly chosen infected nodes on a given network topology[2]. Simulation proceeds in steps of one time unit. During each step, an infected node attempts to infect each of its neighbors with probability β. In addition, every infected node is cured with probability δ. An infection attempt on an already infected node has no effect.

3.2.0.1 *Oregon graph.* Figure 2 shows the time evolution of η as predicted by our model (see Equation 7) on the Oregon AS graph, plotted against simulation results and the steady state prediction of the model proposed in [30] (see Equation 4). The parameter values were $\beta = 0.14, \delta = 0.08$ for plot (a), and $\beta = 0.14, \delta = 0.24$ for plot (b). Since this model in [30] does not estimate the transients, we plot only the steady state for it. As shown, our model yields results very close to the simulation. The steady state prediction of [30] is not accurate, exactly because the Oregon AS graph violates the homogeneity assumption.

3.2.0.2 *BA power-law graph.* Figure 3 compares the predictions of our model against the simulation. It also shows the steady state prediction of the mean-field analysis model MFA [30] for Barabási-Albert networks (see Equation 4). The topology used in Figure 3 is a synthesized 1000-node BA network, with $m_0 = 10$ and $m = 2$. The parameter values used were $\delta = 0.8$ for all plots, and $\beta = 0.175, 0.15, 0.125$ for plots (a)-(c) respectively. As shown, our model nicely tracks the simulation results. The steady-state prediction of the MFA model is reasonably accurate (although consistently lower), since the dataset obeys its assumptions.

3.2.0.3 *Erdős-Rényi graph.* Figure 4 shows simulation results of epidemic spreading on a synthesized 1000-node Erdös-Rényi random graph [11], plotted against our model. We also show the KW model [18], since it is designed for such graphs. The parameter values used were $\beta = 0.2$ for all plots, and $\delta = 0.24, 0.48, 0.72$ for plots (a)-(c) respectively. The results in Figure 4 suggest that again, our model matches

[2] The number of initially-infected nodes does not affect the equilibrium of the propagation.

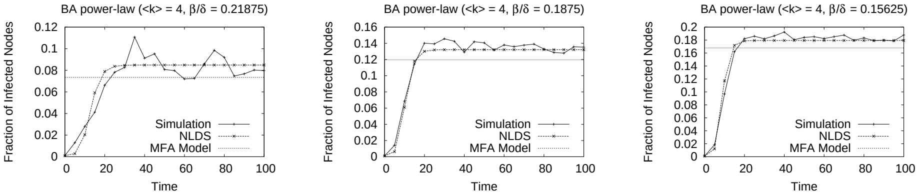


Fig. 3. *Experiments on BA power-law topology:* We compare our model and the model in [30] to the simulation results for several choices of β, keeping δ fixed. The plots show time evolution of infected population in a 1000-node BA power-law network. Our model outperforms the other model in steady state predictions by a slight margin.

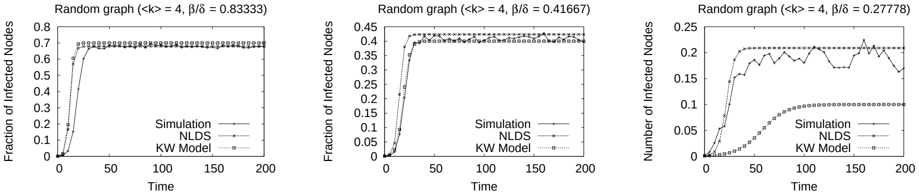


Fig. 4. *Experiments on random graph topology:* We compare our model and the KW model to the simulation results for several choices of δ, keeping β fixed. The plots show time evolution of infected population in a 1000-node Erdös-Rényi random graph, which is the topology that the KW model is specifically aimed at. Our model generally yields similar predictions as the KW model, but outperforms it when δ is high.

the simulation results very well. The KW model is reasonably close, since it is designed specifically for such graphs.

The experiments shown above, conducted on a real network, a synthesized BA power-law network, and an Erdős-Rényi network, illustrate the predictive power of our NLDS model, as well as the accuracy of our *independence assumption.* In all cases, our predictions match the simulation extremely well, and often better than older models (KW, MFA), even on graphs that specifically obey the assumptions of those models.

4. EPIDEMIC THRESHOLD AND EIGENVALUES

As shown in the previous section, *NLDS* matches simulated viral infections very closely, irrespective of the underlying network topology. We will now use this model to derive results on viral infections. Specifically, under what conditions does a viral outbreak become an epidemic?

The epidemic threshold condition codifies this very notion. An informal definition was presented in Equation 2; we will restate this more formally for *NLDS*:

Definition 1 *NLDS* Epidemic threshold. *The epidemic threshold τ for* NLDS *is* *a value such that*

$$\beta/\delta < \tau \Rightarrow \textit{infection dies out over time}$$
$$\beta/\delta > \tau \Rightarrow \textit{infection survives and becomes an epidemic}$$

Previous models have derived threshold conditions only for special-case graphs. For instance, the epidemic threshold for a homogeneous network is the inverse of the average connectivity $\langle k \rangle$ [18]. Similarly, Pastor-Satorras and Vespignani derived a threshold of zero infinite power-law networks [29]. However, a unifying model for *arbitrary*, *real* graphs has not appeared in the literature. The closest model thus far is the MFA model [30] (see Equation 5), but that uses the mean-field assumption. We show later that the MFA model is not as accurate as *NLDS* for arbitrary graphs.

The challenge is to capture the essence of an *arbitrary* graph in as few parameters as possible, making no assumptions about its topology. In the following paragraphs, we derive the epidemic threshold for our model. Our theory is surprisingly simple yet accurate. We show that the epidemic threshold depends only on a *single* parameter—the largest eigenvalue of the adjacency matrix of the graph. We demonstrate later in Section 5 that this new threshold condition subsumes prior models for special-case graphs. The main idea behind the proofs is that our NLDS approach translates the problem of virus survival into the problem of stability of a non-linear dynamical system. The system is *stable* if the first eigenvalue of the appropriate matrix is small, and *unstable* otherwise. A stable NLDS implies that a small perturbation (i.e., a few initially infected nodes) will eventually return to all-nodes-healthy state, while an unstable system will move away.

Theorem 1 Epidemic Threshold. *In* NLDS*, the epidemic threshold τ for an undirected graph is*

$$\boxed{\tau = \frac{1}{\lambda_{1,A}}} \tag{8}$$

where $\lambda_{1,A}$ is the largest eigenvalue of the adjacency matrix $\mathbf{A}$ of the network.

Proof. We will prove this in two parts: the necessity of this condition in eliminating an infection, and the sufficiency of this condition for wiping out *any* initial infection. The corresponding theorem statements are shown below; the proofs are shown in Appendix A. Following this, we will see how quickly an infection dies out if the epidemic threshold condition is satisfied. □

Theorem 2 Part A: Necessity of Epidemic Threshold. *In order to ensure that over time, the infection probability of each node in the graph goes to zero (that is, the epidemic dies out), we must have $\frac{\beta}{\delta} < \tau = \frac{1}{\lambda_{1,A}}$.*

Proof. Proved in Appendix A. □

Theorem 3 Part B: Sufficiency of Epidemic Threshold. *If $\frac{\beta}{\delta} < \tau = \frac{1}{\lambda_{1,A}}$, then the epidemic will die out over time (the infection probabilities will go to zero), irrespective of the size of the initial outbreak of infection.*

Proof. Proved in Appendix A. □

Definition 2 Score. *The* score s *of a virus on a graph is defined as*

$$s = \frac{\beta}{\delta} \cdot \lambda_{1,A} \tag{9}$$

where β and δ are the virus attack and virus death probability, and $\lambda_{1,A}$ is the first eigenvalue of the adjacency matrix of the graph.

Theorem 1 provides the conditions under which an infection dies out ($s < 1$) or survives ($s \geq 1$) in our dynamical system. To visualize this, consider the spread of infection as a random walk on the graph. The virus spreads across one hop according to $\beta\mathbf{A}$, and thus it spreads across h hops according to $(\beta\mathbf{A})^h$, which grows as $(\beta\lambda_{1,A})$ every hop. On the other hand, the virus dies off at a rate δ. Thus, the "effective" rate of spread is approximately $\beta\lambda_{1,A}/\delta$, which is exactly the "score" s. Thus, to have any possibility of an epidemic, the score s must be greater than 1. This is exactly the epidemic threshold condition that we find.

We can ask another question: if the system is below the epidemic threshold, how *quickly* will an infection die out?

Theorem 4 Exponential Decay. *When an epidemic is diminishing (therefore $\beta/\delta < \frac{1}{\lambda_{1,A}}$ and $s < 1$), the probability of infection decays at least exponentially over time.*

Proof. Proved in Appendix A. □

We can use Theorem 1 to compute epidemic thresholds for many special cases, as detailed below. All of these are proved in Appendix A.

Corollary 1. NLDS *subsumes the KW model for homogeneous or random Erdös-Rényi graphs.*

Corollary 2. *The epidemic threshold τ for a star topology, is exactly $\frac{1}{\sqrt{d}}$, where $\sqrt{d}$ is the square root of the degree of the central node.*

Corollary 3. *Below the epidemic threshold (score $s < 1$), the expected number of infected nodes η_t at time t decays exponentially over time.*

5. EXPERIMENTS

Theorems 1 and 3 allow us to calculate the *NLDS* epidemic threshold condition for an arbitrary undirected graph. In the following paragraphs, we explore the application of these results to specific topologies, and demonstrate with simulations that our threshold predictions either subsume or outperform models aimed at specific network topologies.

Specifically, we perform experiments to answer the following questions:

—(*Q1*) How accurate is our epidemic threshold condition when applied to different topologies? Does a viral infection indeed die out when our epidemic threshold condition is satisfied?

—(*Q2*) How do our predictions regarding the epidemic threshold compare to previous work?

—(*Q3*) When the epidemic threshold condition is satisfied, does the infection die out exponentially fast, as predicted by Theorem 4?

Dataset	**Nodes**	**Edges**	**Largest** Eigenvalue
RANDOM	256	982	8.691
POWER-LAW	3,000	5,980	11.543
STAR-10K	10,001	10,000	100
OREGON	11,461	32,730	75.241
ENRON	33,696	361,622	118.417
EMAIL	1,049	37,012	83.460

Table II. Dataset characteristics.

The datasets we used were:

—*RANDOM*: An Erdös-Rényi random graph of 256 nodes and 982 edges.

—*POWER-LAW*: A graph of 3,000 nodes and 5,980 edges, generated by the popular Barabási-Albert process [3]. This generates a graph with a power-law degree distribution of exponent 3.

—*STAR-10K*: A "star" graph with one central hub connected to 10,000 "satellites."

—*OREGON*: A real-world graph of network connections between Autonomous Systems (AS), obtained from `http://topology.eecs.umich.edu/data.html`. It has 11,461 nodes and 32,730 edges. We have already seen this graph before in Section 3.2.

—*ENRON*: We created a graph from the Enron email dataset, which contains 517,431 emails from 151 Enron employees. Every email address is a node; and there is an undirected link between two nodes if they had one or more e-mail exchanges . The graph contains 33,696 nodes and 361,622 edges.

—*EMAIL*: A second real-world email graph comes from a large research institution. We only considered email addresses coming from the research institution, and exclude emails sent outside the institution. This gives us a network with 1,049 nodes and 37,012 edges.

For each dataset, all nodes were initially infected with the virus, and then its propagation was studied in a simulator. All simulations were run for 10,000 timesteps, and were repeated 100 times with different seeds, reporting the mean. We do not plot the standard deviation error-bars, for clarity. Table II provides more details.

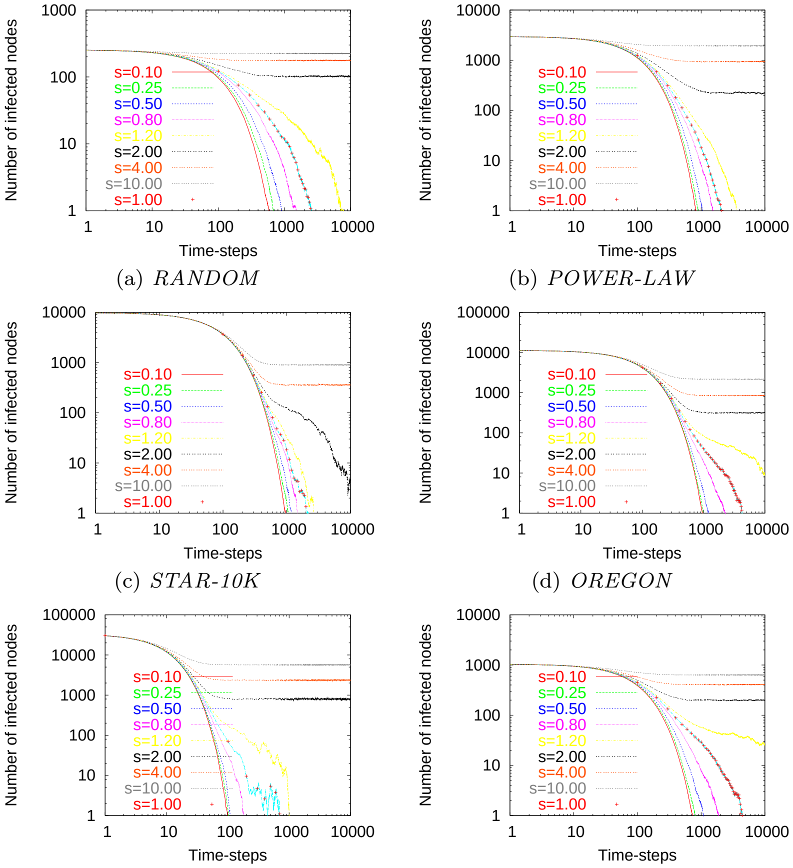
5.1 *(Q1)* Accuracy of our epidemic threshold condition

Figure 5 shows the number of infected nodes over time for various values of the score s, in log-log scales. The dotted line shows the case for $s = 1$. We observe a clear trend: below the threshold ($s < 1$), the infection dies off, while it survives above the threshold ($s > 1$). This is exactly as predicted by Theorem 1, and justifies our formula for the threshold.

Thus, *our epidemic threshold condition is accurate: infections become extinct below the threshold, and survive above it.*

5.2 *(Q2)* Comparison with previous work

In figure 6, we compare the predicted threshold of our model against that of the MFA model. For several values of the score s, we plot the number of infected



(a) *RANDOM*

(b) *POWER-LAW*

(c) *STAR-10K*

(d) *OREGON*

(e) *ENRON*

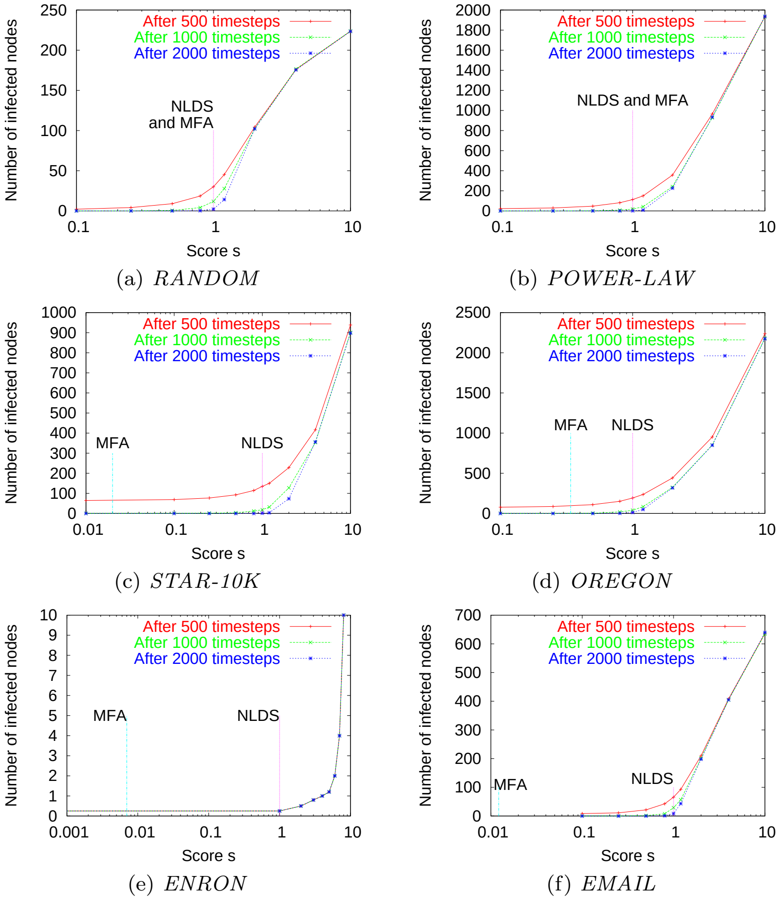
(f) *EMAIL*

Fig. 5. *Accuracy of our epidemic threshold:* The number of infected nodes is plotted versus time for various values of the score s (log-log scales). The case when score $s = 1$ is shown with the dotted line. There is a clear distinction between the cases where $s < 1$ and $s > 1$: below 1, the infection dies out quickly, while above 1, it survives in the graph. This is exactly our proposed epidemic threshold condition.

nodes left after a "long" time (specifically $500, 1000$ and 2000 timesteps). Below the threshold, the infection should have died out, while it could have survived above the threshold. In all cases, we observe that this change in behavior (extinction to epidemic) occurs *exactly* at our predicted epidemic threshold. This is a strong indication that the only assumption of our NLDS model, the *independence assumption*, holds for several real and synthetic graphs.

Notice that the *NLDS* threshold is more accurate than that of the MFA model for the *STAR-10K* and the real-world *OREGON* graphs, while we subsume their predictions for *RANDOM* and *POWER-LAW*, which are the topologies the MFA model was primarily developed for.

Recalling also that our model subsumes the KW model on homogeneous networks



(a) *RANDOM* (b) *POWER-LAW* (c) *STAR-10K* (d) *OREGON* (e) *ENRON* (f) *EMAIL*

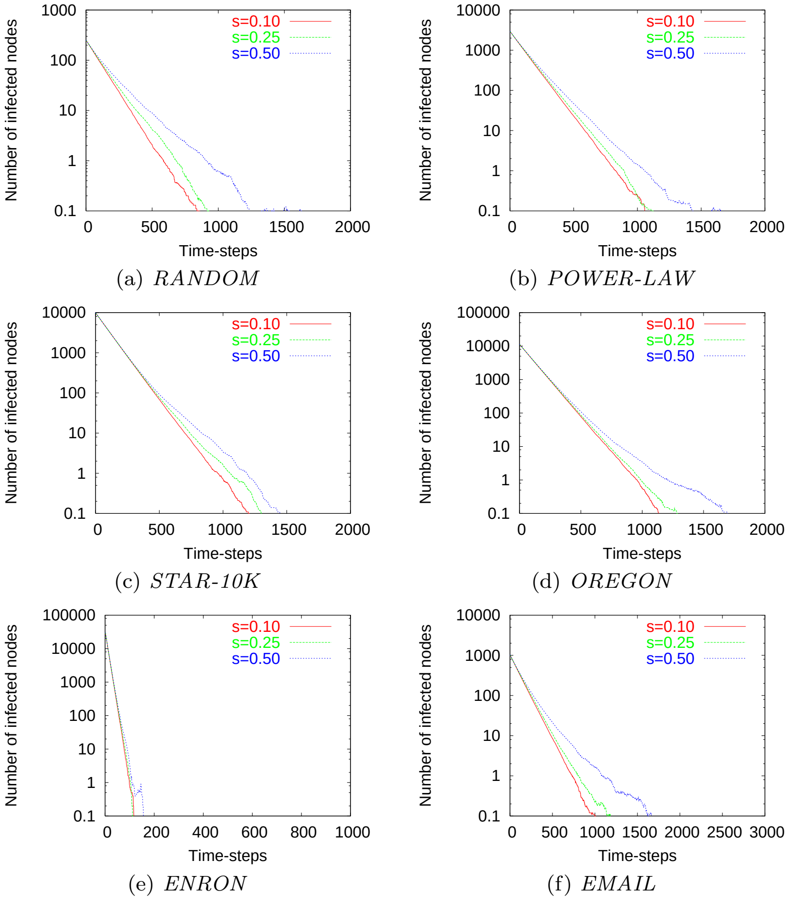
Fig. 6. *Comparison with the MFA model:* We plot number of infected nodes after a "long" time for various values of the score s, versus s. For each dataset, we show results after $500, 1000$ and 2000 timesteps. In each case, we observe a sharp jump in the size of the infected population at our epidemic threshold of $s = 1$. Note that our results are far more accurate than those of the MFA model.

(Corollary 1), we arrive at the following conclusion: *our epidemic threshold for* NLDS *subsumes or performs better than those for other models.*

5.3 *(Q3)* Exponential decay of infection under threshold

Figure 7 demonstrates the rate of decay of the infection when the system is below the threshold ($s < 1$). The number of infected nodes is plotted against time on a log-linear scale We see that in all cases, the decay is exponential (because the plot looks linear on a log-linear scale). This is exactly as predicted by Theorem 4.

Thus, *the infection dies out exponentially quickly below the threshold* ($s < 1$).



(a) *RANDOM*

(b) *POWER-LAW*

(c) *STAR-10K*

(d) *OREGON*

(e) *ENRON*

(f) *EMAIL*

Fig. 7. *Exponential decay below the threshold:* The number of infected nodes is plotted versus time for various values of the score $s < 1$ (log-linear scales). Clearly, the decay is exponentially quick (because it is linear on a log-linear plot). This matches the predictions of Theorem 4.

6. APPLICATIONS

How can we suppress the propagation of a virus? Our epidemic threshold results allow us to analyze many different virus suppression schemes under the same theoretical framework. Two of the most commonly used schemes are throttling and immunization; we discuss these below.

6.1 Throttling

Throttling is a method of slowing down the spread of a worm by limiting the maximum rate of transmission of every node. In our framework, this corresponds to capping the value of the birth rate β to a maximum β_{max}. Thus, by Theorem 1, the epidemic dies out if the worm can be eliminated at a rate of at least $\delta > \beta_{max} \cdot \lambda_{1,A}$, where $\mathbf{A}$ represents the adjacency matrix of the graph.

6.2 Immunization

If we have a budget of k nodes which can be immunized, which nodes should we pick? The "targeted" immunization policy would choose the nodes with the highest degrees. However, in light of Theorem 1, we should choose the ones that cause the maximal reduction in $\lambda_{1,A}$, because these are the ones that will cause maximal increase of the epidemic threshold τ. Let's refer to the above policy as " $max - \Delta\lambda$ ". Will " $max - \Delta\lambda$ " and "targeted" ever disagree? The answer is *yes*, and we illustrate it with an example.

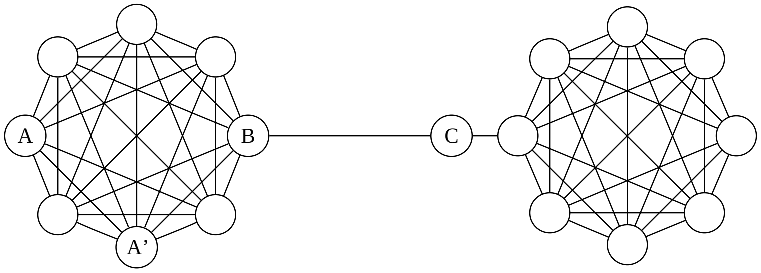


Fig. 8. The "barbel" graph. Two cliques of the same size connected with a bridge.

Consider the graph of Figure 8, a "bar-bell" graph consisting of two equal-size cliques connected with a bridge (node C). Note the "gateway" node B has one edge to the clique missing so that it has the same degree as other nodes in the clique (e.g. node A). For simplicity, assume that k=1, that is, we are allowed to immunize only one node.

On this graph, intuitively, the single best node to immunize would be the bridge node C. The "targeted" immunization policy will clearly fail, picking any one of the clique members, but not the bridge C. Let's see the eigenvalues: Without immunization, the original graph has the first eigenvalue of magnitude $\lambda_1 = 6.803$. Immunizing node A gives a drop of 0.0153, with a drop of 0.0155 for A', 0.0160 for B and 0.0315 for C.

This example demonstrates that the intuitive "targeted" immunization policy (and its popular approximation, the "acquaintance" immunization policy) is not necessarily the best. Instead, it is the " $max - \Delta\lambda$ " policy that creates the toughest obstacles for the virus.

6.3 SIS versus SIR models

Does our threshold condition extend to the SIR model? The SIR model stands for Susceptible-Infectious-Recovered, and models viruses like the mumps, where people obtain life-time immunity. This is an interesting research direction. Our preliminary analysis (not included here) shows that the same threshold condition governs the SIR model, too. Of course, in the SIR model, the virus will eventually become extinct; above threshold, it will infect a significant portion of the population, while below threshold it will not spread. The intuition is that, for the virus to spread further, an infected person should be able to infect at least one more person, before she recovers. The average recovery time is $1/\delta$, and thus the expected number of infected nodes would be $\beta * \lambda_1/\delta$. Thus, we conjecture that our threshold condition carries over to the SIR model, too.

7. CONCLUSIONS

How will a virus propagate in a real computer network? What is the epidemic threshold for a finite graph, if any? Which nodes should we immunize first? In this paper we answer these questions by providing a new analytical model that accurately models the propagation of viruses on *arbitrary graphs*. The primary contributions of this paper are:

—We propose a new model for virus propagation in networks (Equation 7), and show that our model is more precise and general than previous models. We demonstrate the accuracy of our model on both real and synthetic networks.

—We show that we can capture the virus-propagation properties of an arbitrary graph in a single parameter, namely the largest eigenvalue $\lambda_{1,A}$ of the adjacency matrix A. We propose a precise epidemic threshold, $\tau = 1/\lambda_{1,A}$, which holds irrespective of the network topology; an epidemic is prevented when $\delta > \delta_c = \beta * \lambda_{1,A}$. We show that our epidemic threshold is more general and more precise than previous models for special-case graphs (e.g., Erdös-Rényi, homogeneous, e.t.c.).

—We show that, below the epidemic threshold, the number of infected nodes in the network decays exponentially.

—We show how our threshold condition can be used to guide anti-virus policies In particular, we show how to predict the effectiveness of throttling mechanisms, as well as we propose a new immunization policy (" $max - \Delta\lambda$ "), so as to maximally slow down the spread of a virus.

Future research directions abound, both for theoretical as well as experimental work. One extension is to study the SIR model, where we conjecture that the threshold condition is still the same. Another direction is to derive an analytical formula for the size of the epidemic η_t at time t, or at the steady state; in either case, we conjecture that the first eigenvalue should be involved, again.

Acknowledgments

An earlier version of this paper appeared in SRDS 2003. This work is partially supported by the National Science Foundation under Grants No. CCR-0208853, ANI-0326472, IIS-0083148, IIS-0209107, IIS-0205224, INT-0318547, IIS-0326322, CNS-0433540, IIS-0534205, and under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No.W-7405-ENG-48. This work is also partially supported by the Pennsylvania Infrastructure Technology Alliance (PITA), an IBM Faculty Award, a Yahoo Research Alliance Gift, with additional funding from Intel, NTT and Hewlett-Packard. Jure Leskovec was supported in part by Microsoft Research Graduate Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

REFERENCES

Roy M. Anderson and Robert M. May. *Infectious diseases of humans: Dynamics and control.* Oxford Press, 2002.

Norman Bailey. *The Mathematical Theory of Infectious Diseases and its Applications.* Griffin, London, 1975.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 15 October 1999.

Noam Berger, Christian Borgs, Jennifer T. Chayes, and Amin Saberi. On the spread of viruses in the internet. In *Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithm (SODA)*, pages 301–310, Vancouver, BC, Canada, 2005.

Marián Boguñá and Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66:047104, 2002.

Linda Briesemeister, Patric Lincoln, and Philip Porras. Epidemic profiles and defense of scale-free networks. *WORM 2003*, Oct. 27 2003.

Fan Chung, Lincoln Lu, and Van Vu. Eigenvalues of random power law graphs. *Annals of Combinatorics*, 7:21–33, 2003.

Fan Chung, Linyuan Lu, and Van Vu. Spectra of random graphs with given expected degrees. *PNA*, 100(11):6313–6318, May 27 2003.

Reuven Cohen, Shlomo Havlin, and Daniel ben Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91(24), December 2003.

Victor M. Eguiluz and Konstantin Klemm. Epidemic threshold in structured scale-free networks. *arXiv:cond-mat/02055439*, May 21 2002.

Paul Erdös and Alfred Rényi. On the evolution of random graphs. In *Publication 5*, pages 17–61. Institute of Mathematics, Hungarian Academy of Sciences, Hungary, 1960.

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationship of the internet topology. In *Proceedings of ACM Sigcomm 1999*, September 1999.

A. Ganesh, L. Massoulié, and D. Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM*, 2005.

Yukio Hayashi, Masato Minoura, and Jun Matsukubo. Recoverable prevalence in growing scale-free networks and the effective immunization. *arXiv:cond-mat/0305549 v2*, Aug. 6 2003.

Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000.

Herbert W. Hethcote and James A. Yorke. *Gonorrhea Transmission Dynamics and Control*, volume 56. Springer, December 1984. Lecture Notes in Biomathematics.

M. W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra.* Academic Press, 1974.

Jeffrey O Kephart and Steve R White. Directed-graph epidemiological models of computer viruses. In *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, pages 343–359, May 1991.

Jeffrey O Kephart and Steve R White. Measuring and modeling computer virus prevalence. In *Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy*, pages 2–15, May 1993.

S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493, 1999.

C. R. MacCluer. The many proofs and applications of Perron's theorem. *SIAM Review*, 42(3):487–498, 2000.

Nilly Madar, Tomer Kalisky, Reuven Cohen, Daniel ben Avraham, and Shlomo Havlin. Immunization and epidemic dynamics in complex networks. *Eur. Phys. J. B*, 38(2):269–276, 2004.

Helen Martin, editor. *The Virus Bulletin: Independent Anti-Virus Advice.* World Wide Web, http://www.virusbtn.com, 2002. Ongoing.

A G McKendrick. Applications of mathematics to medical problems. In *Proceedings of Edin. Math. Society*, volume 14, pages 98–130, 1926.

Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. Brite: Universal topology generation from a user's perspective. Technical Report BUCS-TR2001-003, Boston University, 2001. World Wide Web, http://www.cs.bu.edu/brite/publications/.

Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B*, 26:521–529, 4 February 2002.

Mark E J Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66:035101(R), 10 September 2002.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63:066117, 2001.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2 April 2001.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics in finite size scale-free networks. *Physical Review E*, 65:035108, 2002.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemics and immunization in scale-free networks. In Stefan Bornholdt and Heinz Georg Schuster, editors, *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH, Berlin, May 2002.

M. Richardson and P. Domingos. Mining the network value of customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, pages 57–66, San Francisco, CA, 2001.

M. Ripeanu, I. Foster, and A. Iamnitchi. Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6(1), 2002.

Chenxi Wang, John C Knight, and Matthew C Elder. On computer viral infection and the effect of immunization. In *Proceedings of the* 16^{th} *ACM Annual Computer Security Applications Conference*, December 2000.

Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *SRDS*, 2003.

A. DETAILS OF THE PROOFS

Theorem 1 Epidemic Threshold. *In* NLDS*, the epidemic threshold* τ *for an undirected graph is*

$$\boxed{\tau = \frac{1}{\lambda_{1,A}}} \tag{10}$$

where $\lambda_{1,A}$ *is the largest eigenvalue of the adjacency matrix* $\mathbf{A}$ *of the network.*

Proof. This follows from Theorems 1 and 3 proved below. □

Theorem 2 Part A: Necessity of Epidemic Threshold. *In order to ensure that over time, the infection probability of each node in the graph goes to zero (that is, the epidemic dies out), we must have* $\frac{\beta}{\delta} < \tau = \frac{1}{\lambda_{1,A}}$*, where* β *is the birth rate,* δ *is the death rate and* $\lambda_{1,A}$ *is the largest eigenvalue of the adjacency matrix* $\mathbf{A}$.

Proof. We have modeled viral propagation as a *discrete dynamical system*, with the following non-linear dynamical equation:

$$\begin{aligned} 1 - p_{i,t} &= (1 - p_{i,t-1})\zeta_{i,t} \\ &\quad + \delta p_{i,t-1}\zeta_{i,t} \qquad \text{(from Eq 7)} \\ \text{or, } \vec{\mathbf{P}}_{\mathbf{t}} &= g(\vec{\mathbf{P}}_{\mathbf{t-1}}) \\ \text{where, } \quad g_i(\vec{\mathbf{P}}_{\mathbf{t-1}}) &= 1 - (1 - p_{i,t-1})\zeta_{i,t} - \delta p_{i,t-1}\zeta_{i,t} \end{aligned} \tag{11}$$

where $g_i(.)$ is the i-th element of the vector $g(.)$.

The infection dies out when $p_i = 0$ for all i. We can easily check that the $\vec{\mathbf{P}} = \vec{\mathbf{0}}$ vector is a *fixed point* of the system; when $p_{i,t-1} = 0$ for all i (all nodes healthy), the equation above results in $p_{i,t} = 0$ for all i, and so all nodes remain healthy forever. The question we need to answer is: If the infection probabilities of all nodes in the graph come close to zero, will the dynamical system push them even closer to zero? In other words, is the $\vec{\mathbf{P}} = \vec{\mathbf{0}}$ fixed point *asymptotically stable?* If yes, the infection probabilities will go to zero and the infection will die out, but if not, the infection could survive and become an epidemic.

Using dynamical systems theory, the required condition is:

Lemma 1 Asymptotic stability. *The system is asymptotically stable at* $\vec{\mathbf{P}} = \vec{\mathbf{0}}$ *if the eigenvalues of* $\nabla g(\vec{\mathbf{0}})$ *are less than* 1 *in absolute value, where* $\left[\nabla g(\vec{\mathbf{0}})\right]_{i,j} = \left.\frac{\partial g_i}{\partial p_j}\right|_{\vec{\mathbf{P}}=\vec{\mathbf{0}}}$ *A proof is shown in [17] (pages* $278 - 281$*).*

From Eq 11, we can calculate $\nabla g(\vec{\mathbf{0}})$:

$$\begin{aligned}\left[\nabla g(\vec{\mathbf{0}})\right]_{i,j} &= \begin{cases} \beta \mathbf{A}_{j,i} & \text{for} \quad j \neq i \\ 1-\delta & \text{for} \quad j = i \end{cases} \\ \text{Thus,} \quad \nabla g(\vec{\mathbf{0}}) &= \beta \mathbf{A}' + (1-\delta)\mathbf{I} \\ &= \beta \mathbf{A} + (1-\delta)\mathbf{I} \qquad (12)\end{aligned}$$

where the last step follows because $\mathbf{A} = \mathbf{A}'$ (since the graph is undirected).

This matrix describes the behavior of the virus when it is very close to dying out; we call it the system matrix $\mathbf{S}$:

$$\mathbf{S} = \nabla g(\vec{\mathbf{0}}) = \beta \mathbf{A} + (1-\delta)\mathbf{I} \qquad (13)$$

As shown in Lemma 2 following this proof, the matrices $\mathbf{A}$ and $\mathbf{S}$ have the same eigenvectors $\vec{\mathbf{u}}_{\mathbf{i},\mathbf{S}}$, and their eigenvalues, $\lambda_{i,A}$ and $\lambda_{i,S}$, are closely related:

$$\lambda_{i,S} = 1 - \delta + \beta \lambda_{i,A} \quad \forall i \qquad (14)$$

Hence, using the stability condition above, the system is asymptotically stable when

$$|\lambda_{i,S}| < 1 \quad \forall i \qquad (15)$$

that is, all eigenvalues of $\mathbf{S}$ have absolute value less than one.

Now, since $\mathbf{A}$ is a real symmetric matrix (because the graph is undirected), its eigenvalues $\lambda_{i,A}$ are real. Thus, the eigenvalues of $\mathbf{S}$ are real too. Also, since the graph $\mathcal{G}$ is a connected undirected graph, the matrix $\mathbf{A}$ is a real, nonnegative, irreducible, square matrix. Under these conditions, the Perron-Frobenius Theorem [21] says that the largest eigenvalue is a positive real number and also has the largest magnitude among all eigenvalues. Thus,

$$\lambda_{1,S} = |\lambda_{1,S}| \geq |\lambda_{i,S}| \quad \forall i \qquad (16)$$

Using this in Equation 15:

$$\begin{aligned}\lambda_{1,S} &< 1 \\ \text{that is,} \quad 1 - \delta + \beta \lambda_{1,A} &< 1\end{aligned}$$

which means that an epidemic is prevented if $\beta/\delta < 1/\lambda_{1,A}$. Also, if $\beta/\delta > 1/\lambda_{1,A}$, the probabilities of infection may diverge from zero, and an epidemic could occur. Thus, the epidemic threshold is $\boxed{\tau = \frac{1}{\lambda_{1,A}}}$ □

Lemma 2 Eigenvalues of the system matrix. *The $i-th$ eigenvalue of* $\mathbf{S}$ *is of the form $\lambda_{i,S} = 1 - \delta + \beta\lambda_{i,A}$, and the eigenvectors of* $\mathbf{S}$ *are the same as those of* $\mathbf{A}$.

Proof. Let $\mathbf{u_{i,A}}$ be the eigenvector of $\mathbf{A}$ corresponding to eigenvalue $\lambda_{i,A}$. Then, by definition, $\mathbf{A\vec{u}_{i,A}} = \lambda_{i,A} \cdot \mathbf{\vec{u}_{i,A}}$ Now,

$$\begin{aligned}\mathbf{S\vec{u}_{i,A}} &= (1-\delta)\mathbf{u_{i,A}} + \beta\mathbf{A\vec{u}_{i,A}} \\ &= (1-\delta)\mathbf{\vec{u}_{i,A}} + \beta\lambda_{i,A}\mathbf{\vec{u}_{i,A}} \\ &= (1-\delta+\beta\lambda_{i,A})\mathbf{\vec{u}_{i,A}} \qquad (17)\end{aligned}$$

Thus, $\mathbf{\vec{u}_{i,A}}$ is also an eigenvector of $\mathbf{S}$, and the corresponding eigenvalue is $(1-\delta+\beta\lambda_{i,A})$.

Conversely, suppose $\lambda_{i,S}$ is an eigenvalue of $\mathbf{S}$ and $\mathbf{u_{i,S}}$ is the corresponding eigenvector. Then,

$$\begin{aligned}\lambda_{i,S}\mathbf{\vec{u}_{i,S}} &= \mathbf{S\vec{u}_{i,S}} \\ &= (1-\delta)\mathbf{\vec{u}_{i,S}} + \beta\mathbf{A\vec{u}_{i,S}} \\ \Rightarrow \left(\frac{\lambda_{i,S}+\delta-1}{\beta}\right)\mathbf{\vec{u}_{i,S}} &= \mathbf{A\vec{u}_{i,S}}\end{aligned}$$

Thus, $\vec{u}_{i,S}$ is also an eigenvector of $\mathbf{A}$, and the corresponding eigenvalue of $\mathbf{A}$ is $\lambda_{i,A} = (\lambda_{i,S} + \delta - 1)/\beta$. □

Theorem 3 Part B: Sufficiency of Epidemic Threshold. *If $\frac{\beta}{\delta} < \tau = \frac{1}{\lambda_{1,A}}$, then the epidemic will die out over time (the infection probabilities will go to zero), irrespective of the size of the initial outbreak of infection. Here β is the birth rate, δ is the death rate and $\lambda_{1,A}$ is the largest eigenvalue of the adjacency matrix* $\mathbf{A}$.

Proof.

$$\begin{aligned}\zeta_{i,t} &= \prod_{j:neighbor\ of\ i} (1 - \beta * p_{j,t-1}) \quad \text{(from Eq. 6)} \\ &\geq 1 - \beta * \sum_{j:neighbor\ of\ i} p_{j,t-1} \qquad (18)\end{aligned}$$

where the last step follows because all terms are positive.

Now, for $i = 1 \dots N$,

$$\begin{aligned}& 1 - p_{i,t} \\ =\ & (1-p_{i,t-1})\zeta_{i,t} + \delta \cdot p_{i,t-1} \cdot \zeta_{i,t} \\ =\ & (1 - (1-\delta)\,p_{i,t-1})\,\zeta_{i,t} \\ \geq\ & (1 - (1-\delta)\,p_{i,t-1}) \times \left(1 - \beta \sum_{j:neighbor\ of\ i} p_{j,t-1}\right)\end{aligned}$$

$$\begin{aligned}
&\quad\text{(using Eq. 18)}\\
&\geq\ (1-(1-\delta)\,p_{i,t-1})\times\left(1-\beta\sum_{j=1}^{N}\mathbf{A}_{j,i}\cdot p_{j,t-1}\right)\\
&\quad\text{(because } \mathbf{A}_{j,i}=1 \text{ for neighbors only)}\\
&\geq\ 1-(1-\delta)p_{i,t-1}\\
&\quad -\beta\sum_{j}\mathbf{A}_{j,i}\cdot p_{j,t-1}+\beta(1-\delta)p_{i,t-1}\sum_{j}\mathbf{A}_{j,i}\cdot p_{j,t-1}\\
&\geq\ 1-(1-\delta)p_{i,t-1}-\beta\sum_{j}\mathbf{A}_{j,i}\cdot p_{j,t-1} \qquad (19)
\end{aligned}$$

No assumptions were required in this.

Thus,

$$p_{i,t}\ \leq\ (1-\delta)p_{i,t-1}+\beta\sum_{j}\mathbf{A}_{j,i}\cdot p_{j,t-1} \qquad (20)$$

Writing this in vector form, we observe that this uses the same system matrix **S** from Equation 13:

$$\begin{aligned}
\mathbf{P}\ &\leq\ \mathbf{S}\vec{\mathbf{P}}_{\mathbf{t-1}} \quad \text{(using the definition of } \mathbf{S} \text{ from Eq. 13)}\\
&\leq\ \mathbf{S}^2\vec{\mathbf{P}}_{\mathbf{t-2}}\leq\ldots\\
&\leq\ \mathbf{S}^t\vec{\mathbf{P}}_{\mathbf{0}}\\
&\leq\ \sum_{i}\lambda_{i,S}^{t}\ \vec{\mathbf{u}}_{\mathbf{i},\mathbf{S}}\ \vec{\mathbf{u}}'_{\mathbf{i},\mathbf{S}}\ \vec{\mathbf{P}}_{\mathbf{0}} \qquad (21)
\end{aligned}$$

where the last step is the spectral decomposition of $\mathbf{S}^t$. Using Eq. 14,

$$\begin{aligned}
\lambda_{i,S}\ &=\ 1-\delta+\beta\lambda_{i,A}\\
&<\ 1-\delta+\beta\frac{\delta}{\beta} \quad \text{(the sufficiency condition)}\\
&<\ 1\\
\text{and so,}\quad \lambda_{i,S}^{t}\ &\approx\ 0 \quad \text{for all } i \text{ and large } t
\end{aligned}$$

Thus, the right-hand side of Eq. 21 goes to zero, implying that

$$\mathbf{P}\rightsquigarrow 0 \text{ as } t \text{ increases}$$

implying that the infection dies out over time. □

THEOREM 4 EXPONENTIAL DECAY. *When an epidemic is diminishing (therefore $\beta/\delta<\frac{1}{\lambda_{1,A}}$), the probability of infection decays at least exponentially over time.*

PROOF. We have:

$$\begin{aligned}
\mathbf{P}_t\ &\leq\ \sum_{i}\lambda_{i,S}^{t}\ \vec{\mathbf{u}}_{\mathbf{i},\mathbf{S}}\ \vec{\mathbf{u}}'_{\mathbf{i},\mathbf{S}}\ \vec{\mathbf{P}}_{\mathbf{0}} \quad \text{(from Eq 21)}\\
&\leq\ \lambda_{1,S}^{t}*\mathbf{C} \qquad (22)
\end{aligned}$$

where $\mathbf{C}$ is a constant vector. Since the value of $\lambda_{1,S}$ is less than 1 (because the epidemic is diminishing), the values of $p_{i,t}$ are decreasing exponentially over time. □

Corollary 1. NLDS *subsumes the KW model for homogeneous or random Erdös-Rényi graphs.*

Proof. According to the KW model, the epidemic threshold in a random Erdös-Rényi graph is $\tau_{KW} = 1/\langle k \rangle$, where $\langle k \rangle$ is the average degree [18]. It is easily shown that, in a homogeneous or random network, the largest eigenvalue of the adjacency matrix is $\langle k \rangle$. Therefore, our model yields the same threshold condition for random graphs, and thus, our *NLDS* model subsumes the KW model. □

Corollary 2. *The epidemic threshold τ for a star topology, is exactly $\frac{1}{\sqrt{d}}$, where $\sqrt{d}$ is the square root of the degree of the central node.*

Proof. The eigenvalue of the adjacency matrix, λ_1, is simply $\sqrt{d}$. Thus, the epidemic threshold is $\tau = \frac{1}{\sqrt{d}}$. □

Corollary 3. *Below the epidemic threshold (score $s < 1$), the expected number of infected nodes η_t at time t decays exponentially over time.*

Proof.

$$\begin{aligned}\eta_t &= \sum_{i=1}^{N} p_{i,t} \\ &= \sum_i \lambda_{1,S}^t * C_i \quad \text{(from Theorem 4)} \\ &= \lambda_{1,S}^t * \sum_i C_i\end{aligned}$$

where C_i are the individual elements of the matrix $\mathbf{C}$ in Equation 22. Since $\sum_i C_i$ is a constant and $\lambda_{1,S} < 1$ (from Theorem 1), we see that n_t decays exponentially with time. □

B. OTHER METHODS

In a recent paper, Ganesh et al. [13] found the same threshold condition for fast extinction of the virus, but without needing the independence assumption that we use. Instead, they found an upper bound for the expected number of infected nodes using a *linearized* dynamical system, which was easy to analyze. How do these two approaches compare to each other?

In the following paragraphs, we will show that the two approaches are complimentary to each other. In a nutshell, Ganesh et al. [13] use a (perhaps weak) upper bound, while we use a *point estimate*. It should be noted, however, that both approaches give the same result, giving more confidence in its accuracy.

A system of random variables: We will first model the viral propagation solely using binary random variables; this requires no assumptions. Let $I_i(t)$, $D_i(t)$ and $B_{ji}(t)$ be 1/0 random variables; $I_i(t)$ is 1 if node i is infected at time t, $D_i(t)$ is 1 if node i has a "virus death" event between time-steps $t-1$ and t, and $B_{ji}(t)$ is

1 if node j is successful in transmitting the virus to node i in between time-steps $t-1$ and t.

The equation for infection is:

$$\begin{aligned} I_i(t) \;&=\; I_i(t-1)\,(1-D_i(t)) + (1-I_i(t-1)) \times \\ &\quad [B_{1i}(t)\cdot A_{1i}\cdot I_1(t-1)\ \underline{\text{OR}} \\ &\quad B_{2i}(t)\cdot A_{2i}\cdot I_2(t-1)\ \underline{\text{OR}}\ \ldots \\ &\quad B_{Ni}(t)\cdot A_{Ni}\cdot I_N(t-1)] \end{aligned}$$

Taking expectations on both sides, and using $E[A\ \underline{\text{O}}\text{R}\ B] \leq E[A+B] = E[A]+E[B]$:

$$\begin{aligned} E[I_i(t)] \;&\leq\; E\left[I_i(t-1)\,(1-D_i(t))\right] \\ &\quad + E\left[(1-I_i(t-1))\cdot\sum_{j=1}^{N} B_{ji}(t)A_{ji}I_j(t-1)\right] \\ &=\; E[I_i(t-1)]\,(1-\delta) \\ &\quad + E\left[\left(\sum_{j=1}^{N} B_{ji}(t)A_{ji}I_j(t-1)\right)\right. \\ &\quad \left. - \left(I_i(t-1)\cdot\sum_{j=1}^{N} B_{ji}(t)A_{ji}I_j(t-1)\right)\right] \\ &=\; E[I_i(t-1)]\,(1-\delta) + \beta\sum_{j=1}^{N} A_{ji}E\left[I_j(t-1)\right] \\ &\quad -\beta\sum_{j=1}^{N} A_{ji}E\left[I_i(t-1)I_j(t-1)\right] \end{aligned} \tag{23}$$

Thus,

$$\begin{aligned} E[I_i(t)] \;&\leq\; \mathcal{X}+\mathcal{Y}-\mathcal{Z} \qquad (24) \\ \text{where, } \mathcal{X} \;&=\; E[I_i(t-1)]\,(1-\delta) \\ \mathcal{Y} \;&=\; \beta\sum_{j=1}^{N} A_{ji}E\left[I_j(t-1)\right] \\ \mathcal{Z} \;&=\; \beta\sum_{j=1}^{N} A_{ji}E\left[I_i(t-1)I_j(t-1)\right] \end{aligned}$$

Note that $\mathcal{X},\mathcal{Y},\mathcal{Z} \geq 0$, and that only $\mathcal{Z}$ contains the non-linear terms.

Linearized system: Here, we neglect $\mathcal{Z}$ to get an upper-bounding linear system:

$$\begin{aligned} E[I_i(t)] \;&\leq\; \mathcal{X}+\mathcal{Y}-\mathcal{Z} \quad \text{(from Eq. 24)} \\ &\leq\; \mathcal{X}+\mathcal{Y} \\ &=\; E[I_i(t-1)]\,(1-\delta) \end{aligned}$$

$$+\beta \sum_{j=1}^{N} A_{ji} E\left[I_j(t-1)\right] \tag{25}$$

which is a linear equation. Summing over all i, we get an upper-bound on the expected number of infected nodes. This is exactly the method being used in [13] to find the epidemic threshold below which the infection dies out.

Independence assumption: Instead of neglecting $\mathcal{Z}$, we *approximate* it using a point estimate:

$$\begin{aligned} \mathcal{Z} &= \beta \sum_{j=1}^{N} A_{ji} E\left[I_i(t-1) I_j(t-1)\right] \\ &\approx \beta \sum_{j=1}^{N} A_{ji} E\left[I_i(t-1)\right] \cdot E\left[I_j(t-1)\right] \end{aligned} \tag{26}$$

Summing over all i gives exactly the dynamical system that we studied in the preceding sections (that is, Equation 7).

We can say more about this point estimate. Consider the correlation between the random variables $I_i(t-1)$ and $I_j(t-1)$:

$$\rho = \frac{E[I_i(t-1)I_j(t-1)] - E[I_i(t-1)] \cdot E[I_j(t-1)]}{Var(I_i(t-1)) \cdot Var(I_j(t-1))}$$

Now,

$$\begin{aligned} & -1 \leq \quad \rho \quad \leq 1 \\ \Rightarrow \quad & -1 \leq \tfrac{E[I_i(t-1)I_j(t-1)] - E[I_i(t-1)] \cdot E[I_j(t-1)]}{Var(I_i(t-1)) \cdot Var(I_j(t-1))} \leq 1 \\ \Rightarrow \quad & E[I_i(t-1)I_j(t-1)] \geq E[I_i(t-1)] \cdot E[I_j(t-1)] \\ & \qquad -Var(I_i(t-1)) \cdot Var(I_j(t-1)) \\ \text{and} \quad & E[I_i(t-1)I_j(t-1)] \leq E[I_i(t-1)] \cdot E[I_j(t-1)] \\ & \qquad +Var(I_i(t-1)) \cdot Var(I_j(t-1)) \end{aligned}$$

However, since $I_i(t-1)$ is a 0/1 random variable, we have:

$$\begin{aligned} Var(I_i(t-1)) &= 1^2 \cdot P(I_i(t-1)) \\ &= 1 \cdot P(I_i(t-1)) \\ &= E[I_i(t-1)] \\ \text{and} \quad Var(I_j(t-1)) &= E[I_j(t-1)] \end{aligned}$$

Hence,

$$0 \leq E[I_i(t-1)I_j(t-1)] \leq 2 \cdot E[I_i(t-1)] \cdot E[I_j(t-1)]$$

Thus, the independence assumption approximates $E[I_i(t-1)I_j(t-1)]$ as $E[I_i(t-1)] \cdot E[I_j(t-1)]$, using the midpoint of this interval as a point estimate. Neglecting the $E[I_i(t-1)I_j(t-1)]$ term completely leads to the upper-bound. Thus, the independence assumption can be considered to be "closer" to the truth, while the linearized system gives the harder guarantees. Which approach is chosen depends on the needs of the user.