On the Spread of Viruses on the Internet

Noam Berger [*] Christian Borgs [†] Jennifer T. Chayes [†] Amin Saberi[‡]

Abstract

We analyze the contact process on random graphs generated according to the preferential attachment scheme as a model for the spread of viruses in the Internet. We show that any virus with a positive rate of spread from a node to its neighbors has a non-vanishing chance of becoming epidemic. Quantitatively, we discover an interesting dichotomy: for a virus with effective spread rate λ, if the infection starts at a typical vertex, then it develops into an epidemic with probability $\lambda^{\Theta\left(\frac{\log(1/\lambda)}{\log\log(1/\lambda)}\right)}$, but on average the epidemic probability is $\lambda^{\Theta(1)}$.

1 Introduction

There is compelling evidence that many self-engineered networks, notably the Internet, have scale-free structures in the sense that the degree distributions of these networks have power-law tails [11]. Motivated by these observations, there has been a great deal of study, both non-rigorous and rigorous, of the detailed structural properties of so-called preferential attachment models and other models with power-law degree distributions; see [1], [4] and references therein for some of the non-rigorous and rigorous work, respectively. However, thus far, there has been much less work on the impact of these structures on *processes* occurring on these networks.

In this paper, we give a rigorous analysis of processes which model the spread of viral infections on scale-free structures, and show how these processes differ markedly from epidemics on more conventional structures. Since there are also observations which indicate that the network of human sexual contacts follows a power-law degree distribution [14], this work is relevant both in the context of the spread of computer viruses on the Internet, and the spread of sexually transmitted diseases (STD).

The standard model used in the study of viral infections is called the *contact process* or the *susceptible-infected-susceptible (SIS) model.* In this model, every vertex is either infected or healthy (but susceptible). An infected vertex becomes healthy with rate 1 independently of the status of its neighbors. A healthy vertex becomes infected at a rate equal to the propagation ratio of the disease, λ, times the number of its infected neighbors.

In our context, this model is describing the spread of viruses in a network in the presence of a particular class of antivirus software. Computers with the software installed are not permanently immune from the virus, but they are regularly scanned for the presence of the virus, and the software removes the virus if the computer is found to be infected. A computer can be infected by the same virus more than once, and each time it remains infected until the next scan by the antivirus software. Alternatively, the contact process also approximately describes the spread of epidemics in the presence of regularly updated antivirus software which confers permanent immunity, but where viruses mutate. In this case, the antivirus software prevents any given computer from being reinfected with the same virus, but does not prevent it from being reinfected with all mutated variants.

The contact process has been studied extensively in the probability community [13], but it is usually studied on bounded-degree or homogenous graphs. The most important general result in that context is the existence of epidemic thresholds. For infinite graphs it has been shown that there exist two epidemic thresholds $\lambda_1 \leq \lambda_2$. If $\lambda > \lambda_2$, then with positive probability the can spread and survive at any point of the graph. If $\lambda_1 < \lambda < \lambda_2$, the infection survives with positive probability, but every vertex heals eventually almost surely. If $\lambda < \lambda_1$, the infection dies out almost surely. As it turns out, $\lambda_1 = \lambda_2$ for $\mathbf{Z}^d$, whereas $\lambda_1 < \lambda_2$ for regular trees (see [13] and [20, 19]).

It is easy to see that, in a finite graph, the infection will eventually die out with probability 1. However, there is still a natural definition of epidemics in the finite case, as can be seen by considering finite subsets of well-studied infinite graphs, such as $\mathbf{Z}^d$. It turns out that, for the cube $[-n, n]^d$, there is a λ_c such that if

[*] California Institute of Technology, Pasadena, CA. Email: berger@its.caltech.edu. This work was done while the first and last authors were visiting Microsoft Research.

[†] Microsoft Research, Redmond, WA . E-mails: {borgs, jchayes}@microsoft.com.

[‡] Georgia Institute of Technology, Atlanta, GA. E-mail: saberi@cc.gatech.edu.

$\lambda > \lambda_c$ then with probability bounded away from zero the infection survival time is exponential in n^d, while if $\lambda < \lambda_c$ the infection dies out before time $\log(n)$ with probability $1 - o(n)$. Moreover, this λ_c is equal to the epidemic threshold for $\mathbf{Z}^d$. (See [13] for proofs of these statements.) Therefore, it is natural to say that the infection becomes an epidemic if the time that it takes for the infection to die out is super-polynomial in the number of vertices of the graph.

Using the epidemiologic models such as the SIS model for analyzing the spread of viruses has been suggested more than a decade ago by Kephart and White [12]. Pastor-Satorras and Vespignani [17, 16] were the first group to study the contact process on scale-free graphs in the Barabási-Albert model [2]. Using simulation and (non-rigorous) mean-field equations, they argued that the epidemic threshold λ_c in scale-free networks is 0. They also studied the actual data and found supporting evidences for their observation. Other recent work on the spread of computer viruses on the Internet includes [15, 21, 8].

In this paper, we present what is, to the best of our knowledge, the first rigorous analysis of the contact process on scale-free graphs in preferential attachment models.

The contribution of this paper is two-fold. First, we introduce a new representation of the preferential attachment model which we call the Pólya urn representation. Our representation, which we believe to be of independent interest, is a generalization of Bollobás and Riordan's random pairing representation [3]. It gives a new proof of the main result of [3] and enables us to analyze a natural generalization of their representation in which the vertices can also choose their neighbors uniformly at random with some probability; see also [18, 10, 6, 7] for other models with combinations of uniform and preferential attachment. We believe this representation will also be useful in rigorous analysis of many other structural and dynamical properties of preferential attachment graphs.

Second, we use our new representation to analyze the contact process on the preferential attachment model. We show that, as predicted by Pastor-Satorras and Vespignani [17, 16], the epidemic threshold is zero. The importance of this observation is that it shows that even viruses with very small propagation rate have a positive chance of becoming epidemic. We also provide much more detailed estimates yielding matching upper and lower bounds, as functions of λ, on the probability for an epidemic to occur – both for an epidemic beginning at a typical starting vertex and on average. Interestingly, it turns out that these two probabilities are quite different. In particular, the epidemic probability for an infection beginning at a typical vertex is a rather complicated function of λ, which would therefore have been quite difficult to ascertain by empirical means:

$$\lambda^{\Theta\left(\frac{\log(\lambda^{-1})}{\log\log(\lambda^{-1})}\right)}, \tag{1.1}$$

whereas the average epidemic probability is simply $\lambda^{\Theta(1)}$.

1.1 Strategy of the proof

We end the introduction by giving an intuitive description of the proof of (1.1), without delving into the rather tortuous technical details. The proof breaks into two relatively independent parts, the first dealing with the contact process and the second dealing with the structure of the graph.

The behavior of the contact process depends strongly on the degrees in the graph. In particular, we show that if all degrees in a graph G are significantly smaller than λ^{-1}, then the disease will die out very quickly. If, on the other hand, the virus has reached a vertex of degree significantly larger than λ^{-2}, then the disease is very likely to survive for very long time in the neighborhood of this vertex.

Therefore, we want to get an understanding of the degrees in a neighborhood of a vertex. We show, using our Pólya urn representation of the scale-free graph and Bollobás–Riordan's expanding environment method [3] that for a typical vertex v, the largest degree of a vertex in a ball of radius k around v is, with high probability, $(k!)^{\Theta(1)}$.

In view of this, the closest vertex of degree $\lambda^{-\Theta(1)}$ is at distance $\Theta(\log(\lambda^{-1})/\log\log(\lambda^{-1}))$, and the question of survival of the disease boils down to whether the infection manages to arrive at a vertex of degree $\lambda^{-\Theta(1)}$. Therefore the survival probability is the probability that the infection manages to arrive at distance $\Theta(\log(\lambda^{-1})/\log\log(\lambda^{-1}))$, and this probability is given in equation (1.1).

The analysis above is useful in understanding the behavior if we start at a typical starting point. However, if we start at a point of degree higher than λ^{-2}, then the process has a very good chance of surviving for a long time. The power-law degree distribution of the Barabási-Albert graphs tells us that $\lambda^{\Theta(1)}$ of the vertices have this degree, and therefore the average survival probability is $\lambda^{\Theta(1)}$.

1.2 Structure of the paper

In Section 2, we precisely define the model and state our results. In Section 3, we present our Pólya urn representation of the scale-free graph, and give a number of technical lemmas that enable easy analysis of the model. In Section 4, we use the construction of Section 3 to give estimates on the maximum degree in a neighborhood of a randomly chosen vertex. The main tool we use is the method of rapidly expanding neighborhoods, first introduced in [3]. In Section 5 we prove a few simple facts on the contact process, and in the last section we give some details the proof of Theorem 2.1. Most of the more technical estimates are relegated to Sections 7, 8, and 9.

2 Definition of the Model and Statements of Results

The scale-free graph we define generalizes the model suggested by Barabási and Albert [2] and made rigorous in [3]. Fix an integer $m \geq 2$ and a real number $0 \leq \alpha < 1$. Let $\{v_i\}$ be a sequence of vertices, and let G_i be the graph at time i. Then, G_1 contains the vertex v_1 and no edges, and G_2 contains v_1 and v_2 and m edges connecting them. Given G_{n-1}, we create G_n the following way:

We add the vertex v_n to the graph, and choose m vertices $w_1, ..., w_m$, possibly with repetitions, from G_{n-1}. Then we draw edges between v_n and each of $w_1, ..., w_m$. Repetitions in the sequence $w_1, ..., w_m$ result in multiple edges in the graph G_n.

The vertices $w_1, ..., w_m$ are chosen inductively as follows: With probability α, w_1 is chosen uniformly, and with probability $1 - \alpha$, w_1 is chosen according to the preferential attachment rule, i.e., for every $i = 1, \ldots, n-1$, we take $w_1 = v_i$ with probability $(\deg_{n-1}(v_i))/Z$ where Z is the normalizing constant

$$Z = \sum_{i=1}^{n-1} (\deg_{n-1}(v_i)) = 2m(n-2).$$

Then we proceed inductively, applying the same rule, but when determining w_k, instead of the degree $\deg_{n-1}(v_i)$, we use

$$\deg'_{n-1}(v_i) = \deg_{n-1}(v_i) + \#\{1 \leq j \leq k-1 | w_j = v_i\}.$$

It should be noted that the $\alpha = 0$ case of our model differs slightly from the model of and Bollobás and Riordan [3] in that they allow (self-)loops, while we do not. Both [3] and the model defined above allow multiple edges. One might argue that the most natural — though mathematically harder — case is that without multiple edges, i.e., when the w_i are all conditioned to be different (for $n > m$) and are all determined according to the rule described for w_1. It turns out that we can provide Pólya urn representations of any of these three variants for general α. Here we will consider only the variant defined above, without loops but with multiple edges. In the full version of this paper, we will also give the more natural variant without multiple edges, and show that it does not change the final results.

Our main results are the following:

THEOREM 2.1. *For every $\lambda > 0$, there exists N such that for a typical sample of the scale-free graph of size $n > N$, if we choose a uniform vertex v, then with probability $1 - O(\lambda^2)$, v is such that an infection starting at v will survive with probability bounded from below by*

$$\lambda^{C_1 \frac{\log(1/\lambda)}{\log\log(1/\lambda)}} \tag{2.2}$$

and from above by

$$\lambda^{C_2 \frac{\log(1/\lambda)}{\log\log(1/\lambda)}} \tag{2.3}$$

where C_1 and C_2 are constants not depending on λ or n.

The $O(\lambda^2 n)$ vertices left out in Theorem 2.1 turn out to have a dramatic effect on the average survival probability, as demonstrated in the next theorem:

THEOREM 2.2. *For every $\lambda > 0$, there exists N such that for a typical sample of the scale-free graph of size $n > N$, if we choose a uniform vertex v and start the infection at v, then the infection will survive with probability bounded from below by*

$$\lambda^{C_3} \tag{2.4}$$

and from above by

$$\lambda^{C_4} \tag{2.5}$$

where C_3 and C_4 are constants not depending on λ or n.

It is interesting to mention that the survival probability of the contact process is much higher than the density of the percolation cluster which was proved in [5] to be between $\exp(-\Theta(\lambda^{-2}))$ and $\exp(-\Theta(\lambda^{-1}))$.

Another interesting comparison is with recent nonrigorous results of Pastor-Satorras and Vespignani [17] who calculate the percentage of infected nodes in the metastable state, where they implicitly condition on the event of survival. Their calculation yields that the density of infected nodes is of the order of $\exp(-(m\lambda)^{-1})$. The comparison reveals another aspect of the inhomogeneity of the scale-free network: In more homogeneous graphs we expect these two quantities (the survival probability and the density of infected nodes in the metastable state) to be similar to each other.

3 Pólya Urn Representation of the Barabási-Albert Graph

In early twentieth century, Pólya proposed and analyzed the following model known as the Pólya urn model [9]. We have a number of urns, each holding a number of balls, and at each step, a new ball is added to one of the urns. The probability that the ball is added to urn i is proportional to $N_i + u$ where N_i is the number of balls in the i-th urn and u is a predetermined parameter of the model.

Pólya showed that this model is equivalent to another process as follows. For every i, choose a parameter (which we call "strength" or "attractiveness") p_i, and at each step, *independently* of our decision in previous steps, put the new ball in urn i with probability p_i. Pólya specified the distribution (as a function of u and the initial number of balls in each urn) for which this mimics the urn model. A particularly nice example is the case of two urns, each starting with one ball and $u = 0$. Then p_1 is a uniform $[0, 1]$ variable, and $p_2 = 1 - p_1$. He showed that for general values of u and $\{N_i(0)\}$, the values of $\{p_i\}$ are determined by the β-distribution with appropriate parameters.

It is not hard to see that there is a close connection between the preferential attachment model of Barabási and Albert and the Pólya urn model in the following sense: every new connection that a vertex gains can be represented by a new ball added in the urn corresponding to that vertex. We use this idea to give an equivalent description of the scale-free graph which is easy to analyze. We will see throughout the paper the properties of this description that make it useful for understanding the graph.

3.1 Formal description We describe an equivalent representation of the n-vertex Barabási-Albert graph with m connections and probability α of uniform connection. Let u be s.t. $\alpha = u/(1 + u)$. We take $\psi_1 = 1$, and for every $2 \le k \le n$, we take ψ_k to be distributed according to $\beta(m + mu, 2km + kmu)$ (We say that $X \sim \beta(a, b)$ if the density of X is $\frac{x^{a-1}(1-x)^{b-1}}{Z}$ with Z being the appropriate normalization. See [22] for the properties of the β distribution). For $1 \le k \le n$, we take

$$\varphi_k = \psi_k \prod_{j=k+1}^{n} (1 - \psi_j).$$

It is easy to see that $\sum_{k=1}^{n} \varphi_k = 1$. Let

$$l_k = \sum_{j=1}^{k} \varphi_k.$$

For every $a \in [0, 1]$, we define $\kappa(a) = \min\{k : l_k \ge a\}$. Let $\{U_{i,k}\}_{1 \le i \le m, 1 \le k \le n}$ be independent random variables, uniform on $[0, 1]$. For $k > j$, we draw an edge between k and j if for some $1 \le i \le m$ we have

$$j = \kappa(U_{i,k} l_{k-1}). \tag{3.6}$$

We allow multiple edges — the number of edges connecting k to j is the number of values of i such that (3.6) is satisfied. The next lemma follows immediately from the theory of Pólya urns.

LEMMA 3.1. *The random graph described above has the same distribution as the n-vertex Barabási-Albert graph with m connections and probability α of uniform connection.*

Lemma 3.1 gives us a representation of the Barabási-Albert graph with much more independence that the original description, thus enabling us to do rigorous calculations.

In order to use Lemma 3.1 effectively, we need to have a few estimates on the values of l_k, φ_k and $\kappa(a)$. These estimates are deferred to Section 7.

4 Maximum Degree in a Neighborhood of a Vertex

In this section we state the main two propositions controlling the structure of the graph. These propositions say that, with high probability, all of the vertices in the ball H_t of radius t around a uniform vertex have degree smaller than $(t!)^{100}$, but there exists some vertex in H_t of degree $(t!)^{\Theta(1)}$.

The proofs of these propositions use the Pólya urn representation and the methods of expanding neighborhoods. The details are presented in Sections 8 and 8.2.

PROPOSITION 4.1. *Let a be chosen uniformly in $[0, 1]$, and let $k = \kappa(a)$. For every ϵ there exists T such that with probability larger than $1 - \epsilon$, for every $t > T$, every vertex in H_t has degree smaller than $(t!)^{100}$.*

PROPOSITION 4.2. *Let a be chosen uniformly in $[0, 1]$, and let $k = \kappa(a)$. There exists $C > 0$, depending only on χ, such that for every ϵ there exists T such that with probability larger than $1 - \epsilon$, for every $t > T$, there exist a vertex in H_t with degree larger than $(t!)^C$.*

5 The Contact Process

The contact process is often studied as a model for the spread of infections. It has been the subject of intensive research, both rigorous work within the mathematics community [13, 20, 19], and numerical and simulation analysis in the networking, social sciences and physics literature. An excellent reference for the mathematical background is Liggett [13].

In this model, every computer or individual is represented by a vertex in a graph. A vertex is either healthy or infected. An infected vertex becomes healthy after an exponential time with mean 1, independently of the status of its neighbors. A healthy vertex becomes infected at a rate that is proportional to the number of its infected neighbors. More formally:

DEFINITION 5.1. *The contact process with infection parameter λ on a graph $G(V,E)$ is a continuous time Markov process η_t which can be identified at any time t by a subset $A = \{v \in V : \eta_t(v) = 1\}$ of vertices. The vertices in A are regarded as infected and the rest of the vertices are thought of as being healthy. The transition rates for η_t are given by*

$A \to A \setminus \{v\}$, for $v \in A$ at rate 1 and

$A \to A \cup \{v\}$, for $v \notin A$ at rate $\lambda|\{u \in A : \{u,v\} \in E\}|$.

We assume that at $t = 0$ one of the vertices of the graph is infected. This vertex is usually called the *root* or *origin*. In an infinite graph, the disease might *survive* in the graph for an infinite time. However, it is easy to see that in a finite graph the disease will eventually *die out*, i.e., A becomes empty and remains empty afterwards.

In finite graphs, we study the time that it takes for the graph to become healthy. In particular, we say a disease becomes an epidemic if and only if the time that it takes to die out is exponential in the number of vertices.

We will show that in a scale-free graph of size n, there is a λ_n such that with high probability, any disease with infection rate $\lambda > \lambda_n$ has a constant probability of becoming epidemic, and $\lambda_n \to 0$ as n tends to infinity. This is in contrast to bounded-degree graphs in which with high probability the disease dies out exponentially fast if $\lambda < 1/(2d)$; see [13].

LEMMA 5.1. *Let G be a graph with maximum degree d. Let S be the set of vertices ever to be infected in G, then $\mathbf{P}(|S| > k) < (4d\lambda)^k$ for every k.*

Proof. We may assume without loss of generality that $\lambda d \leq 1/4$. Define X to be the random variable indicating $|A|$ at any time. The probability that two events (either a healthy node becoming infected or vice versa) happen at the same time is zero. Therefore, the transition rates for X are given by

$X \to X - 1$, at rate X and

$X \to X + 1$, at rate $\lambda|c(A, \bar{A})|$,

where $c(A, \bar{A}) = \{\{u, v\} \in E : u \in A, v \in \bar{A}\}$.

Clearly, $|c(A, \bar{A})| \leq Xd$. Therefore, at any time, the next event increments X with probability at most

$$\frac{\lambda X d}{X + \lambda X d} = \frac{\lambda d}{1 + \lambda d} < \lambda d$$

and decrements X with probability at least

$$\frac{1}{1 + \lambda d} > 1 - \lambda d.$$

In order to infect more than k vertices , we will need at least k increments among the first $2k$ events, the probability of which is bounded above by $2^{2k}(\lambda d)^k = (4d\lambda)^k$, as desired.

As a corollary of the proof, we get the following result:

COROLLARY 5.1. *Let G be a graph. Let $v \in G$ and let l be a positive integer. Assume that in the ball of radius l around v, all of the degrees are bounded by d. Start a contact process with parameter $\lambda < d^{-1}/2$ at $\{v\}$. For $T > 0$, let $S(T)$ be the event that $A_T \neq \emptyset$, and let $B(l)$ be the event that the infection never leaves the ball of radius l around v. Then, for every T,*

$$\mathbf{P}(S(T)|B(l)) < (2\lambda d)^T.$$

In the next lemma, we will study the survival time of the contact process in a star. This lemma is crucial for the proof of our main theorem. We will show that with high probability, the disease survives in a star for an exponential time in the number of vertices.

The idea of the proof is as follows: When the center of the star becomes infected, it starts infecting the leaves at a very high rate. The number of leaves infected before the center becomes healthy again is high enough to ensure that the disease will survive in the graph until the center becomes infected again. The proof of this lemma is in Section 9.

LEMMA 5.2. *Let G be a star graph, with center x and leaves $y_1, \ldots, y_k$. Let A_t be the set of vertices infected at time t. There exists C such that if $A_0 = \{x\}$ then $\mathbf{P}(A_{\exp(Ck\lambda^2)} \neq \emptyset) = 1 - o(k)$.*

6 Proof of Theorem 2.1

In this section we prove Theorem 2.1. The theorem breaks into two propositions, each of which is a simple corollary of the results of previous sections. Let G_n be the (random) Barabási-Albert graph, and let v_n be a uniformly chosen vertex in G_n.

The proof of Theorem 2.2 is very similar to that of Proposition 6.1 below.

Proposition 6.1. *For every n there exists λ_n, with $\lambda_n \to 0$ as n tends to ∞, such that for every $\lambda_{G_n,v_n} > \lambda > \lambda_n$, if we start an infection with parameter λ at v_n, it will survive with probability bounded from below by*

$$\lambda^{C_1 \frac{\log(1/\lambda)}{\log\log(1/\lambda)}},$$

where C_1 is a universal constant, and

$$\mathbf{P}(\lambda_{G_n,v_n} < x) 1/10 \log(1/x) \tag{6.7}$$

i.e., λ_{G_n,v_n} stochastically dominates a variable that does not depend on n.

Conversely, we have:

Proposition 6.2. *For every n there exists λ_n, with $\lambda_n \to 0$ as n tends to ∞, such that for every $\lambda_{G_n,v_n} > \lambda > \lambda_n$, if we start an infection with parameter λ at v_n, it will survive with probability bounded from above by*

$$\lambda^{C_2 \frac{\log(1/\lambda)}{\log\log(1/\lambda)}}$$

where C_2 is a universal constant and λ_{G_n,v_n} is as in (6.7).

Note that the difference between the two propositions is that Proposition 6.1 bounds the survival probability from below, whereas Proposition 6.2 bounds the survival probability from above.

Proof. [Proof of Proposition 6.1] Fix λ. Let

$$k_0 = 10C^{-1} \frac{\log(1/\lambda)}{\log\log(1/\lambda)}$$

where C is as in Proposition 4.2. By Lemma 8.2 and Proposition 4.2, with probability as in (6.7), G_n and v_n are so that the k-neighborhood of v_n contains a vertex $u^{(1)}$ of degree larger than

$$(k_0!)^C > \left(\frac{1}{\lambda}\right)^{10}$$

such that

$$l_{u^{(1)}} < 2^{-0.5\log(k_0!)} < \lambda^D$$

for some $D = D(m, u) > 0$. Now, let $u^{(2)}$ be a parent of $u^{(1)}$, let $u^{(3)}$ be a parent of $u^{(2)}$, and continue up to $u^{(\log(n)/100)}$. Then, $l_{u^{(j)}} = U_j l_{u^{(j-1)}}$ where $\{U_j\}$ are i.i.d. variables, uniform on $[0, 1]$. With probability larger than $1/2$,

$$l_{u^{(j)}} < \left(\frac{9}{10}\right)^j l_{u^{(1)}}$$

for all $j = 2, \ldots, \log(n)/100$. Therefore, using Lemmas 7.2 and 7.3, with probability larger than $1/4$, for every $j = 2, \ldots, \log(n)/100$, the degree of $u^{(j)}$ is larger than

$$1.05^{j(\chi^{-1}-1)} \left(\frac{1}{\lambda}\right)^5.$$

Thus far, we have the following: There exists a vertex $u^{(1)}$ of distance k_0 from v_n, and a sequence of vertices $u^{(j)}$, $j = 2, \ldots \log(n)/100$ such that:

1. For every j, the degree of $u^{(j)}$ is bounded from below by $1.05^{j(\chi^{-1}-1)} \left(\frac{1}{\lambda}\right)^5$, i.e., the degrees of $u^{(j)}$ grow exponentially with j.

2. The vertex $u^{(j)}$ is a neighbor of $u^{(j-1)}$.

Let $v^{(1)} = v_n, v^{(2)}, v^{(3)}, \ldots, v^{(k_0)} = u^{(1)}$ be a path starting at v_n and reaching $u^{(1)}$. With probability

$$\left(\frac{\lambda}{1+\lambda}\right)^{k_0} \geq \lambda^{C_1 \frac{\log(1/\lambda)}{\log\log(1/\lambda)}}$$

the infection reaches $u^{(1)}$. By iterative applications of Lemma 5.2, conditioned on the event that the infection reaches $u^{(1)}$, with probability bounded away from zero, the infection will reach $u^{(\log(n)/100)}$, and by another application of Lemma 5.2, the infection will survive up to time at least

$$\exp\left(C\lambda^2 \cdot 1.05^{\log(n)/100}\right) = \exp(n^\nu)$$

for some $\nu = \nu(m, \alpha, \lambda)$.

Proof. [Proof of Proposition 6.2] Proposition 6.2 follows immediately from Lemma 8.1 and Proposition 4.1, and Lemma 5.1 and Corollary 5.1.

7 Estimates for the Pólya Urn Representation

In this section we complete the work started in Section 3 by providing estimates for the quantities defined in that section. Let

$$\chi = \frac{m + mu}{2m + mu}.$$

Then the following hold:

Lemma 7.1. *l_k converges uniformly in probability to $\left(\frac{k}{n}\right)^\chi$, i.e., for every ϵ there exist N such that if $n > N$, then with probability larger than $1-\epsilon$, for every $1 \leq k \leq n$, we have $|l_k - (k/n)^\chi| < \epsilon$.*

From Lemma 7.1 we get that:

Lemma 7.2. *For every ϵ there exist N such that if $n > N$ then with probability larger than $1-\epsilon$, for every $a \in [0,1]$, we have $|\kappa(a) - a^{1/\chi}n| < \epsilon n$.*

For φ_k, which is the (random) strength of the k-th vertex, the estimate is as follows:

Lemma 7.3. *Let $\{\varphi''_k\}_{k=1}^{\infty}$ be i.i.d. variables distributed $\Gamma(m+mu)$. and let $\varphi'_k = \varphi''/(2m+mu)$. For every ϵ there exist N and K such that for every $n > N$ there exists a coupling between*

$$\left\{\frac{k^{\chi-1}}{n^{\chi}}\varphi'_k\right\}_{k=K}^{n}$$

and $\{\varphi_k\}_{k=K}^{n}$ so that with probability larger than $1-\epsilon$,

$$(1-\epsilon)\varphi_k \le \frac{k^{\chi-1}}{n^{\chi}}\varphi'_k \le (1+\epsilon)\varphi_k$$

for every $K \le k \le n$.

Recall that the Γ-distribution with parameter a is the distribution with density $\frac{x^{a-1}\exp(-x)}{Z}$, with Z being the proper normalization. In particular, if a is an integer, then the Γ-distribution with parameter a is the distribution of the sum of a independent exponentials with parameter 1.

8 Expanding Neighborhood Calculation

We want to estimate the maximum degree of a vertex in a neighborhood of radius k around a random vertex v. This has already been done by Bollobás and Riordan [3] for the (looped) version of model without uniform connections. In this section we show that the ideas of Bollobás and Riordan, when applied to the Pólya urn description of the graph instead of the random pairing description, give good estimates for the maximum degree of a vertex in a neighborhood of radius k around a random vertex v in the more general setting (i.e., $\alpha > 0$).

We start from a uniformly chosen vertex v. Let Θ_j be the set of vertices at distance exactly j from v. We take

$$H_j = \cup_{i=1}^{t}\Theta_i.$$

Assume

$$n > e^{t^2}. \tag{8.8}$$

Let

$$G_t(i) = \#\{k \in \Theta_t : 2^{-i} < l_k \le 2^{-i+1}\}.$$

8.1 Evolution of $G_t(\cdot)$

Fix n large, let $a \in [0,1]$ and let $k = \kappa(a)$. Let i be so that $a \in [2^{-i}, 2^{-i+1}]$. We want to understand the distribution of the neighbors of k. k has two types of neighbors: the m connections that k made when it joined the graph, and the connections that newer vertices made to k when they arrived.

For the first type, let $\{U_i\}_{i=1}^{m}$ be m independent $U([0,1])$-s. The m connections are $\{\kappa(a'U_i) : i = 1,\dots,m\}$ where $a' = l_{\kappa(a)} = a + O(n^{-\chi})$. Therefore, for each $j > i$, the number of neighbors of k in $[2^{-j}, 2^{-j+1}]$ is bounded from below and from above by constants times $\mathrm{Bin}(m, 2^{i-j-1})$.

For the second type, fix $j < i$. The number of connections from $[2^{-j}, 2^{-j+1}]$ is

$$\sum_{h | l_h \in [2^{-j}, 2^{-j+1}]} X_h$$

where $X_h \sim \mathrm{Bin}(m, w_k/l_h)$. Therefore, the number of neighbors of k in $[2^{-j}, 2^{-j+1}]$ is bounded from below and from above by constants times

$$\mathrm{Poi}\left(2^{-j}\frac{w_k}{\mathbf{E}(w_{\kappa(2^{-j})})}\right). \tag{8.9}$$

From (8.9) and Lemmas 7.1, 7.2, and 7.3, we get that there exist constants $0 < C_1, C_2 < \infty$ such that for every t and j,

$$G_{t+1}(j) \preceq \mathrm{Poi}\left(C_1\left[\sum_{i \le j} 2^{i-j}G_t(i) + \sum_{i \ge j} 2^{\beta(i-j)}G_t(i)\right]\right) \tag{8.10}$$

and

$$G_{t+1}(j) \succeq \mathrm{Poi}\left(C_2\left[\sum_{i < j} 2^{i-j}G_t(i) + \sum_{i > j} 2^{\beta(i-j)}G_t(i)\right]\right) \tag{8.11}$$

where $\preceq$ and $\succeq$ denote stochastic domination and $\beta = \chi^{-1} - 1$ satisfies $0 < \beta \le 1$. From (8.10) and (8.11) we get (8.12) and (8.13) below, which are slightly weaker but are much more convenient to use:

$$G_{t+1}(j) \preceq \mathrm{Poi}\left(C_u\left[\sum_{i=1}^{\infty} 2^{i-j}G_t(i)\right]\right) \tag{8.12}$$

and

$$G_{t+1}(j) \succeq \mathrm{Poi}\left(C_l\left[\sum_{i=1}^{\infty} 2^{\beta(i-j)}G_t(i)\right]\right) \tag{8.13}$$

with $0 < C_u, C_l < \infty$.

8.2 Proofs of the Upper and Lower Bounds In this subsection we will show that with high probability, all of the vertices in H_t have degree smaller than $(t!)^{100}$, but there exists a vertex of degree higher than $(t!)^{100}$. First we show the upper bound. This will be done using induction. For every $t > 1$, let $B_t = \lceil 20 \log(t!) \rceil < 20t^2$. The induction step is the following lemma:

Lemma 8.1. *Let $E_t^{(\ell)}$ be the event that $G_{\ell+t}(j) < 10 \cdot 2^{-j}(t!)^4$ for every j. Then*

(8.14)
$$\mathbf{P}(E_{t+1}^{(\ell)} | E_t^{(\ell)}) \geq 1 - \frac{1}{t^2} - \sum_{j=B_t}^{\infty} 2^{-j}(t!)^4 = 1 - o\left(t^{-2}\right).$$

Proof. Since $G_{\ell+t}(j)$ is integer, if we condition on $E_t^{(\ell)}$, then $G_t(j) = 0$ for every $j > B_t$. Therefore, using (8.12), $G_{\ell+t+1}(j)$ is stochastically dominated by a Poisson variable with parameter

$$2^{-j} B_t (t!)^4 < 2^{-j} \frac{((t+1)!)^4}{t^2}$$

for every j. Therefore, by Markov's inequality, the probability that there exists $j \leq B_t$ such that $G_{\ell+t+1}(j) > 10 \cdot 2^{-j}((t+1)!)^4$ is bounded by

$$\frac{B_t}{t^4} < \frac{1}{t^2}. \tag{8.15}$$

For $j > B_t$, the probability that $G_{\ell+t+1}(j) > 10 \cdot 2^{-j}((t+1)!)^4$ is the probability that $G_{\ell+t+1}(j) \geq 1$, and by Markov's inequality this is bounded by

$$2^{-j}(t!)^4. \tag{8.16}$$

Equation(8.14) follows from (8.15) and (8.16).

We can now prove the upper bound:

Proposition 8.1. *Let a be chosen uniformly in $[0,1]$, and let $k = \kappa(a)$. For every ϵ there exists T such that with probability larger than $1 - \epsilon$, for every $t > T$, every vertex in H_t has degree smaller than $(t!)^{100}$.*

Proof. Let l be such that $a > 2^{-l}$ with probability $1 - \epsilon/4$, and let $\ell < -l$. Also, let ℓ be so large in absolute value that

$$\sum_{t=-\ell}^{\infty} \left(\frac{1}{t^2} + \sum_{j=B_t}^{\infty} 2^{-j}(t!)^4 \right) < \epsilon/2. \tag{8.17}$$

Notice that in (8.17) we are summing on the expression from (8.14). Let $T > 1 - \ell$, such that $(t!)^{10} > ((t+\ell)!)^4$ for all $t > T$. By the choice of ℓ, the probability of $E_{1-\ell}^{(\ell)}$ is larger than $1 - \epsilon/4$. Therefore, by Lemma 8.1, with probability larger than $1 - \epsilon/2$, for every $t > T$, the event $E_t^{(\ell)}$ occurs.

Now, condition on the occurrence of $\bigcap_{t=T}^{\infty} E_t^{(\ell)}$. Then for every $t > T$, the number of elements in H_t is no more than $(t!)^{10}$, and

$$\min\{l_k : k \in H_t\} > 2^{-B_t} > \frac{1}{(t!)^{20}}.$$

Therefore, using Lemmas 7.2 and 7.3,

$$\mathbf{P}\left(\exists\, k \in H_t \text{ such that } w_k > \frac{t^2 \cdot (t!)^{20(\chi-1)}}{n} \right) < \frac{1}{t!}.$$

The degree of k is dominated by m plus a Poisson process with rate nw_k/l_k. Since $l_k > (t!)^{-20}$, we get that the probability that there exists a vertex of degree larger than $(t!)^{100}$ is bounded by $(t!)^{-50}$. This gives the required result.

Now, we show that with high probability, there exists a vertex v in H_t of degree $(t!)^{\mu}$ where $\mu = \mu(\chi) > 0$. The proof is not much different from that of the upper bound. Let C_1 be so that

$$2^{-\beta C_1 \log(t!)} > (t!)^{-0.25} \tag{8.18}$$

for every t. Let $F_t = C_1 \log(t!)$. The induction step follows from the following lemma:

Lemma 8.2. *Let $D_t^{(\ell)}$ be the event that $G_{\ell+t}(j) > 10 \cdot 2^{-\beta j}(t!)^{1/2}$ for every $j < F_t$. Then*

$$\mathbf{P}(D_{t+1}^{(\ell)} | D_t^{(\ell)}) \geq 1 - e^{-t}. \tag{8.19}$$

Proof. Condition on the event $D_t^{(\ell)}$. $G_{\ell+t+1}(j)$ dominates a Poisson variable with parameter

$$\begin{aligned} 2^{-\beta j} \sum_{i=1}^{\infty} 2^{\beta i} &\geq 2^{-\beta j} = 10 F_t (t!)^{1/2} \\ \geq 2^{-\beta j} 10(t+1)(t!)^{1/2} &\geq 2^{-\beta j} 1000 = ((t+1)!)^{1/2} \\ &\geq 1000((t+1)!)^{1/4} \end{aligned}$$

for $j < F_{t+1}$. Therefore, for $j < F_{t+1}$,

$$\mathbf{P}\left(G_{\ell+t+1}(j) < 10 \cdot 2^{-\beta j}(t!)^{1/2} \right) < \exp\left(-\frac{((t+1)!)^{1/4}}{16} \right),$$

and summing up we get the desired result.

The following proposition is the main result in the subsection:

PROPOSITION 8.2. *Let a be chosen uniformly in $[0,1]$, and let $k = \kappa(a)$. There exists $C > 0$, depending only on χ, such that for every ϵ there exists T such that with probability larger than $1-\epsilon$, for every $t > T$, there exist a vertex in H_t with degree larger than $(t!)^C$.*

Proof. First we need to choose ℓ. Let k_i be a sequence of ancestors of k. Then, l_{k_i} has the distribution of the product of $i+1$ independent variables distributed $U([0,1])$. In particular, with probability exponentially close to 1, $l_{k_i} < 2^{-i}$ (this is because of the inequality of the means). Let T be such that $\sum_{t=T}^{\infty} < \epsilon/4$, and let ℓ be such that with probability larger than $1-\epsilon/4$, l_{k_i} is so small that with probability larger than $1-\epsilon/4$, for every $j < F_T$, the set $U = \{k' : k' \text{ connects to } k_i\}$ is of size larger than $10 \cdot 2^{-\beta j}(T!)^{1/2}$.

Then, by Lemma 8.2 we get that with probability larger than $1 - 3\epsilon/4$, for every $t > T$, there exists $v \in H_{t+\ell}$ with $l_v < 2^{-0.5\log(t!)}$. By Lemmas 7.2 and 7.3, with probability larger than $1-\epsilon e^{-t}$, the degree of this vertex is larger than

$$(t!)^{0.5\log 2\cdot(\chi^{-1}-1)}$$

and the proof of the proposition is complete.

9 Proof of the Star Lemma

Proof. [Proof of Lemma 5.2] First, we will show that the decrease in the number of infected leaves during a period in which the central vertex is healthy can be bounded by a Gamma variable with parameter $\frac{\lambda}{1+\lambda}$. Then, we will show that the number of infected vertices when the center is infected can be simulated by a simple biased random walk on a line. For the first part, suppose we are at the state in which the vertex in the center of the star is healthy. Define I to be the random variable of the number of infected leaves. I is decreasing by 1 at rate I. The center is becoming infected at rate λI. Therefore, at any moment, the probability that in the next event the center becomes infected is $\frac{\lambda}{1+\lambda}$ and the probability that I decreases by one is $\frac{1}{1+\lambda}$. Clearly, this shows that the number of infected vertices cured in a period in which the center is healthy is a random variable with the distribution $\text{Geom}(\frac{\lambda}{1+\lambda})$. Now, in the period in which the center is infected, the number of infected leaves X has the following transition rates:

$$X \to X-1, \text{ at rate } X \text{ and}$$
$$X \to X+1, \text{ at rate } \lambda|k-X|.$$

One can easily verify that X dominates the following process

$$\begin{array}{ll} Y \to Y-1, \text{ at rate } \frac{1}{4}\lambda k & \text{if } Y = \frac{1}{4}\lambda k \\ & \\ Y \to Y+1, \text{ at rate } \frac{3}{4}\lambda k & \text{if } Y < \frac{1}{4}\lambda k \\ Y \to Y-1, \text{ at rate } \frac{1}{4} = \lambda k & \end{array}$$

where the initial value of Y is the number of infected leaves in the beginning of each period. Merging this with the number of leaves that become healthy during the time in which the center is healthy, the following process will give a simple lower bound on the number of leaves infected in the contact process:

If $Y = \frac{1}{4}\lambda k$

$$\begin{array}{ll} Y \to Y-1 & \text{at rate } \frac{1}{4}\lambda k \\ Y \to Y - \text{Geom}(\frac{\lambda}{1+\lambda}) & \text{at rate } 1 \end{array}$$

If $Y < \frac{1}{4}\lambda k$

$$\begin{array}{ll} Y \to Y+1 & \text{at rate } \frac{3}{4}\lambda k \\ Y \to Y-1 & \text{at rate } \frac{1}{4}\lambda k \\ Y \to Y - \text{Geom}(\frac{\lambda}{1+\lambda}) & \text{at rate } 1 \end{array}$$

Therefore, the problem reduces to calculating the survival time of the system described above. This system is a factor $\lambda k + 1$ speedup of the following discrete time system:

If $Y = \frac{1}{4}\lambda k$

$$\begin{array}{ll} Y \to Y-1 & \text{with prob. } \frac{1}{4}(1-(\lambda k)^{-1}) \\ Y \to Y - \text{Geom}(\frac{\lambda}{1+\lambda}) & \text{with prob. } (\lambda k)^{-1} \end{array}$$

If $Y < \frac{1}{4}\lambda k$

$$\begin{array}{ll} Y \to Y+1 & \text{with prob. } \frac{3}{4}(1-(\lambda k)^{-1}) \\ Y \to Y-1 & \text{with prob. } \frac{1}{4}(1-(\lambda k)^{-1}) \\ Y \to Y - \text{Geom}(\frac{\lambda}{1+\lambda}) & \text{with prob. } (\lambda k)^{-1} \end{array}$$

and therefore it is enough to show that the survival time of the above system is, with high probability, exponential. To do that, it is sufficient to show that starting at $Y = \frac{1}{4}\lambda k$ the probability of hitting 0 before returning to $\frac{1}{4}\lambda k$ decays exponentially with $\lambda^2 k$. Let E be the event that Y changes by more than 1 at least $\lambda^2 k/100$ times before hitting 0 or returning to $\frac{1}{4}\lambda k$. The probability of this event is at most:

$$\mathbf{P}(E) + \mathbf{P}\left(\text{hitting 0 before returning to } \frac{1}{4}\lambda k \,\middle|\, E^c\right).$$

First we want to bound $\mathbf{P}(E)$: For every value $0 < x < \frac{1}{4}\lambda k$ the probability of reaching $\frac{1}{4}\lambda k$ before the next occurrence of a change larger than 1 is at least $1/4$ and therefore

$$\mathbf{P}(E) < 4^{-\lambda^2 k/100}.$$

Now we want to estimate

$$\mathbf{P}\left(\text{hitting 0 before returning to } \frac{1}{4}\lambda k \middle| E^c\right).$$

Let t_1 be the change in Y that is larger than 1, let t_2 be the second and so on. Let s_1 be the first change in Y of size 1, let s_2 be the second and so on. Notice that the s-s and the t-s are independent of each other, s_j is a Bernoulli variable with $\mathbf{P}(s_j = 1) = 1 - \mathbf{P}(s_j = -1) = 3/4$ and t_i is the negative of a geometric variable with parameter $\lambda/(1+\lambda)$ The process hits 0 before returning to $\frac{1}{4}\lambda k$ only if there exist $i < \lambda^2 k/100$ and j so that $t_1 + t_2 + \ldots t_i + s_1 + s_2 + \ldots s_j < -\frac{1}{4}\lambda k$. $\{s_j\}$ is biased random walk, and therefore

$$\mathbf{P}\left(\exists\, l : \sum_{i=1}^{l} s_i < -\frac{1}{8}\lambda k\right) < e^{-\lambda k/20}$$

and

$$\mathbf{P}\left(\exists\, l < \lambda^2 k/100 : \sum_{i=1}^{l} t_i < -\frac{1}{8}\lambda k\right) < e^{-C\lambda^2 k}.$$

The lemma now follows from the above equations.

Acknowledgment: We thank Milena Mihail, Bobby Kleinberg and Oliver Riordan for useful discussions.

References

[1] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–98, 2002.

[2] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] B. Bollobas and O. Riordan. The diameter of scale-free graphs. to appear in Combinatorica.

[4] B. Bollobás and O. Riordan. Mathematical results on scale-free random graphs. *Preprint*, 2003.

[5] B. Bollobás and O. Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Math.*, 1, 2003.

[6] P.G. Buckley and D. Osthus. Popularity-based random graph models leading to a scale free degree sequence. *Preprint.*

[7] C. Cooper and A. Frieze. A general model of web graphs. *Preprint*, 2002.

[8] Z. Dezso and A.-L. Barabasi. Halting viruses in scale-free networks. *Physical Review E*, 65, 2002.

[9] R. Durrett. *Probability: Theory and Examples.* Duxbury Press, 1996.

[10] M. Enachescu E. Drinea and M. Mitzenmacher. Variations on random graph models for the web. *Preprint*, 2001.

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. 1999.

[12] J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. In *IEEE Computer Security Symposium on research in Security and Privacy. Oakland, California*, 1993.

[13] T. M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes.* Springer-Verlag, 1999.

[14] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature*, 411, 2001.

[15] M. E. J. Newman. The spread of epidemic disease on networks. *Physical Review E*, 66, 2002.

[16] R. Pastor-Satorras and A. Vespignani. Epidemics and immunization in scale-free networks.

[17] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, pages 3200–3203, 2001.

[18] J.F.F. Mendes S.N. Dorogovtsev and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85:4633, 2000.

[19] A. Stacey. The existence of an intermediate phase for the contact process on trees. *Annal Prob.*, 1996.

[20] A. Stacey. The contact process on finite homogeneous tree. *Prob. Th. Rel. Fields*, 2001.

[21] Y. Wang and C. Faloutsos D. Chakrabarti, C. Wang. Epidemic spreading in real networks: an eigenvalue viewpoint, 2003.

[22] E. W. Weisstein. Beta distribution – mathworld–a wolfram web resource.