Evaluation of Epidemic Seeding Strategies under Variable Node Costs*

Ronald Chiesse[1,3], Daniel R. Figueiredo[2], Antonio A. de A. Rocha[1], Artur Ziviani[3]

[1]Instituto de Computação, UFF
Niterói, RJ – Brasil

[2]Prog. de Engenharia de Sistemas e Computação, UFRJ
Rio de Janeiro, RJ – Brasil

[3]Laboratório Nacional de Computação Científica
Petrópolis, RJ – Brasil

Abstract. *Network epidemics is a general modeling framework useful to represent dynamic processes over networks, such as the spread of a virus on a computer network or information dissemination on a social network. Within this framework, network seeding is the problem of determining which network nodes should be selected to start an epidemic. Intuitively, the success of an epidemic under some performance metric (ex. number of reached nodes) largely depends on the set of initially infected nodes. Prior approaches to effective network seeding have treated nodes identically in terms of their respective cost to start an epidemic. However, we argue that such assumption is inadequate in many cases and thus consider network seeding under variable node cost. We propose a degree-based cost function and evaluate the performance of four different network seeding strategies over two different network models. Our results show that no seeding strategy is consistently superior under identical budgets. In particular, we identify a tradeoff between strategies that select (hire) a larger number of cheap nodes (low degree) and strategies that select few expensive nodes (high degree). Our results shed light on the importance of taking into account variable node cost, a more realistic assumption in many applications.*

1. Introduction

Epidemics on networks is a powerful model to capture the spreading dynamics of various phenomena on structured context, such as a computer virus spreading through a computer network, influenza on human population, and information in social networks [Figueiredo 2011, Newman 2010]. In the classical approach to network epidemics, network nodes have a state such as "infected" (I) and "susceptible" (S), coupled with some specific model to determine how states of nodes change over time (e.g., how does a node become infected). An important criterion in such models is the set of infected nodes when the epidemic starts at time zero, known as the *seeding nodes*. Intuitively, the set of seeding nodes can determine if an epidemic will die out quickly, infecting very few nodes, or be long lasting, infecting a very large number of nodes. Figure 1 illustrates this intuition: seeding nodes 1, 2, 3 is likely to generate a larger epidemic than seeding nodes nodes 10, 11, 12, under a reasonable infection model.

*This work was funded in part by research project grants from CAPES, CNPq and FAPERJ.

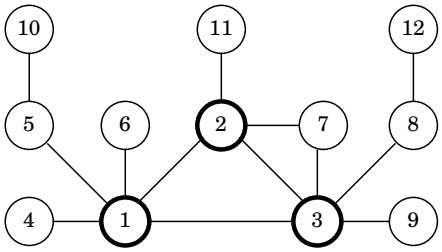


Figure 1. Example of a network and a seeding strategy that selects three nodes.

A natural and fundamental problem emerges from this scenario: which nodes should be seeded in order to maximize some property of an epidemic? This problem, known as *epidemic seeding*, has applications in various contexts, such as opinion formation, spread of innovation, and marketing. Although many approaches to tackle epidemic seeding have been proposed in the recent literature (see Section 5), the problem is still unsolved in many practical formulations, mostly due to its combinatorial nature and strong dependence on various parameters, most notably the network structure.

In general, epidemic seeding is constrained by some resource, such as a fixed budget in terms of number of nodes. Thus, only a limited number of network nodes can be selected for seeding, that is, can be set to infected at time zero. For example, in Figure 1, only three nodes can be chosen for seeding. Intuitively, this constraint represents the fact that there is an intrinsic cost associated with infecting a node at time zero, a cost to start an epidemic.

Prior works have considered seeding budgets in terms of number of nodes (e.g., three nodes as in Figure 1), implying that every network node has a fixed and equivalent cost. Thus, the cost of starting the epidemic is the same, regardless of nodes chosen. We argue that in many contexts this assumption is inadequate, since network nodes are different and would naturally have different costs due to different reasons. For example, consider an online social network and a marketing campaign for a given product. Intuitively, more popular individuals on the network are likely to charge much more to start a campaign than less popular ones. In fact, this is exactly the case today with Twitter (and other social media services), where popular individuals (celebrities) are paid differently to tweet a marketing message to their followers [Kornowski 2013]. Thus, a more realistic seeding problem formulation should consider nodes with variable costs, in particular costs proportional to their popularity.

In this paper, we consider the epidemic seeding problem when nodes have variable costs. We use node degree as a proxy for node popularity and consider different cost functions, in terms of degree dependence. We evaluate three different degree-based seeding strategies (and one random strategy) over two models for network topologies. Intuitively, given a fixed budget, the largest degree strategy can seed the epidemic with a small number of high degree nodes, while the smallest degree strategy can seed the epidemic with a large number of small degree nodes. Which strategy is better? By considering a fixed budget (in terms of average node costs), we show that no single strategy is consistently superior than others in terms of epidemic spread. In particular, the largest degree strategy, which usually shows good performance when nodes have identical costs, can have

a worse performance than smallest degree strategy. Our results thus indicate the variable node costs play a fundamental role when determining an effective seeding strategy.

The remainder of this paper is organized as follows. In Section 2, we present the network and epidemic models under consideration. Section 3 presents the seeding strategies that will be evaluated. The evaluation of various scenarios is presented and discussed in Section 4. Related work is presented and discussed in Section 5, while final considerations are given in Section 6.

2. Network and Epidemic Models

The dynamics of classical network epidemics require the characterization of two fundamental aspects: (i) the network topology over which the epidemic unfolds; and (ii) the epidemic model which determines the transition rules between epidemic states for network nodes. In this section, we present models for each of such aspects that are investigated in this paper.

2.1. Network Models

Networks are represented as graphs consisting of a set of nodes (or vertices) denoted by V and a set of edges denoted by E, thus $G = (V, E)$. We assume the set of edges E represents a symmetric relationship (such as co-authorship in papers), thus leading to undirected graphs. Moreover, let $n = |V|$ and $m = |E|$ denote, respectively, the number of nodes and edges of the network. Note that the average degree of the network is given by $\overline{d} = 2m/n$.

We consider two random network models in our study: Erdős-Rényi model (aka., $G(n, p)$ model) and Barabási-Albert model (aka. preferential attachment model) [Figueiredo 2011, Newman 2010]. The classical $G(n, p)$ model consists of a graph with n labeled vertices where each possible edge is present with probability p, independently of all other edges. The degree of a randomly chosen node follows a Binomial distribution, $Bi(n-1, p)$. Thus, the degree of nodes tends to occur around its average value, given by $\overline{d} = (n-1)p$. Note that the probability that a node has degree more than five times the average is negligible.

The preferential attachment is a random network growth model where nodes are added in sequence to the network. At each time step, a single node is added and s edges incident to the arriving node are added to other nodes already present in the network. Therefore, each node brings s edge points that are randomly placed in the existing network according to the following rule. Consider a node u present in the network at time t and let $d_u(t)$ denote the degree of this node at time t. The probability the arriving node v at time t chooses node u to receive an edge point is given by $p_u(t) = d_u(t) / \sum_{v \in V(t)} d_v(t)$. Note that the probability of selecting a node is thus proportional to its degree. As a consequence, larger degree nodes are more likely to receive incoming edges, making them even more likely to receive future edges, giving rise to the name "preferential attachment". The growth process starts with a small clique network at time $t = 0$. Note that the average degree is $\overline{d} = 2s$ for t large enough.

Unlike $G(n, p)$, the degree distribution of the Barabási-Albert model follows a power-law of the form $p_k \approx k^{-2}$ (where p_k denotes the probability that a randomly chosen

node in a large network has degree k). Under this distribution, nodes with very large degree occur in the network with non-negligible probability, which is the case for several real networks, such as the AS Graph (Internet) and online social networks (e.g., Facebook or Twitter).

In our evaluation, we will consider networks generated from these two well-known models as they lead to fundamentally different structures while being employed as primitives for many real network models. Furthermore, since such networks have very different degree distributions, intuitively they will lead to different behaviors in terms of seeding and epidemic spread. Indeed, we will soon illustrate these behaviors.

2.2. Epidemic Model

A network epidemic model determines the rules for transitions between epidemic states of network nodes. Two epidemic states that compose any epidemic model are susceptible (S) and infected (I). The most simple epidemic model has a single possible transition: a network node in the susceptible state can become infected, giving rise to what is known as the SI model [Newman 2010]. In this work, we consider a discrete time SI model and a simple transition rule that depends only on the state of neighboring nodes. Thus, infection occurs over the network edges.

In particular, consider a network $G = (V, E)$ and let $S(t)$ and $I(t)$ denote, respectively, the set of nodes in the susceptible (S) and infected (I) state at time t. Note that each node $v \in V$ must be either in $S(t)$ or $I(t)$ for all $t \geq 0$. A transition from S to I occurs as follows: if node $v \in S(t)$ has $\theta > 0$ or more infected neighbors at time t, then node v becomes infected at time $t+1$, and thus $v \in I(t+1)$. Note that θ represents the threshold parameter of the epidemic model, which is identical for every node in the network. More formally, for all $t > 0$ we have:

$$\forall v \in V, v \in \begin{cases} I(t) & \text{if } \ v \in I(t-1), \\ I(t) & \text{if } \ |N(v) \cap I(t-1)| \geq \theta, \\ S(t) & \text{otherwise}, \end{cases} \tag{1}$$

where $N(v)$ denotes the set of neighbors of node v. Note that θ and $I(0)$, i.e. the set of nodes infected at time zero, are the sole parameters of the model. In particular, this epidemic model is known as the *threshold model* [Kempe et al. 2003, Newman 2010].

Figure 2 illustrates the spread of an epidemic under this model with $\theta = 2$ and two (seeding) nodes infected at time 0 (shown in yellow). Note that at each time step, one or more susceptible nodes (shown in white) may become infected (indicated by red edges). Moreover, the epidemic stops when there are no changes to the set $I(t)$ in consecutive time steps. Finally, note that the epidemic may stop without all network nodes becoming infected, as illustrated. This occurs because the remaining susceptible nodes have less than θ infected neighbors, and thus do not become infected.

Finally, Algorithm 1 shows the SI epidemic dynamics for the considered threshold model. Note that epidemic state of nodes at time t influences the states of nodes only at time $t + 1$. The algorithm returns the set of nodes infected at each time step, namely, the vector of sets $I(\cdot)$.

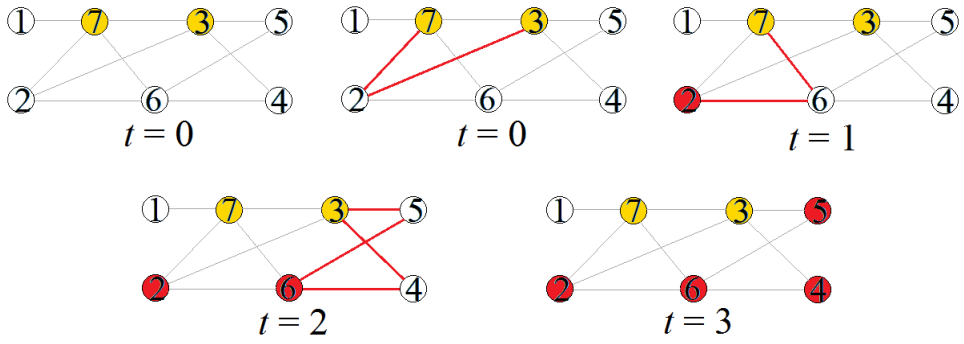


Figure 2. Example of an SI epidemic using threshold model with $\theta = 2$: Yellow indicates nodes infected at time zero (seeding nodes); white indicates susceptible (S) nodes; red indicates infected (I) nodes. Red edges indicate nodes that meet the threshold and will thus become infected in the next time step.

Algorithm 1 SI epidemic dynamics on a network using the threshold model.

Require: $G = (V, E)$, $I(0) \subseteq V$, $\theta > 0$
  $S(0) \leftarrow V - I(0)$
  $t \leftarrow 0$
  repeat
    $I(t+1) \leftarrow I(t)$
    $S(t+1) \leftarrow \emptyset$
    for all $v \in S(t)$ **do**
      if $|N(v) \cap I(t)| \geq \theta$ **then**
        $I(t+1) \leftarrow I(t+1) \cup \{v\}$
      else
        $S(t+1) \leftarrow S(t+1) \cup \{v\}$
      end if
    end for
    $t \leftarrow t+1$
  until $I(t-1) = I(t)$
  return $I(\cdot)$

3. Seeding Strategies

A *seeding strategy* consists of a policy to determine the set of seeding nodes, i.e. the set of nodes that will be infected at time zero to start the epidemic process, denoted by $I(0)$ in Section 2.2. Of course, the seeding strategy is usually constrained in the sense that not all nodes of the network can be selected as infected at time zero. A classical approach is to constrain the seeding strategy by imposing a fixed limit on the number of nodes that can be selected. This approach implicitly assumes that every node in the network has the same cost to be infected at time zero. We consider a different approach, where nodes have a variable cost to be infected at time zero. In particular, we assume that the cost of selecting (or hiring) node u to start an epidemic is proportional to its degree, d_u. This seems more reasonable in various contexts, including advertising in online social networks, such as Twitter where node dependent costs are already happening [Kornowski 2013].

Thus, we assume that the cost of selecting (hiring) node u to start an epidemic is given by:

$$c(u) = (d_u)^\alpha \tag{2}$$

where $\alpha > 0$ is a constant that controls the cost dependency on node degree. Note that this

function is fairly generic as α controls the sensitivity on degree, a feature that intuitively represents *node popularity*. As $\alpha \to 0$ the cost depends less and less on degree, and in the limit becomes constant, independent of node degree. In contrast, as α increases beyond one, node cost depends more and more on degree, to the point that the highest degree node may cost more than the entire network. In Section 4 we present and discuss results with various values for α.

The seeding strategy will be constrained by a budget that determines the amount of wealth available to cover the cost of nodes to start the epidemic. In order to allow comparison between different networks and different node cost functions, the budget is given as the cost of a fraction of nodes assuming average node cost. Thus, let k be a fraction of nodes, such as 1%. Note that the average node cost is given by the cost of the average degree, and thus, $(\overline{d})^{\alpha}$. The budget is then given by:

$$b = kn(\overline{d})^{\alpha} \tag{3}$$

where n is the number of network nodes. Note that the budget increases with any of its parameters: network size n, average degree $\overline{d}$, or sensitivity to node degree α.

Finally, we consider three different degree-based seeding strategies and one random strategy. For each strategy, the budget is spent until it becomes insufficient to cover the cost of any other network node. In what follows, each seeding strategy is presented, where O denotes the set of nodes selected by the strategy:

- *Largest degree (LA)*: This seeding strategy selects nodes by their descending order of degree. Thus, the largest degree of the network is selected first, budget permitting, then the second largest, and so on, as long the budget permits. If a node cannot be selected due to its high cost compared to the remaining budget, the node is simply skipped and the process continues. Algorithm 2 presents this seeding strategy.
- *Smallest degree (SH)*: This seeding strategy selects nodes by their ascending order of degree. Thus, the smallest degree is selected first, then the second smallest, and so on, until the remaining budget can no longer cover the cost of a node. Note that the cost increases as nodes are considered in increasing order, and thus the process can stop as soon as a node cannot be selected. This seeding strategy is described in Algorithm 3.
- *Median degree (ME)*: This seeding strategy selects nodes around the median degree distribution. In particular, the median node degree is first selected, budget permitting. The node following the median in the ordering is then selected, budget permitting, and then the node prior to the median, budget permitting, and so on. Thus, the algorithm selects nodes alternately, after and before the median. A node is simply skipped if its cost is larger than the remaining budget. This seeding strategy is described by Algorithm 4. Note that index j plays the role of alternating around the median, while nodes are sorted in ascending order.
- *Random degree (RA)*: This seeding strategy selects nodes at random, budget permitting. The order considered by the strategy is random and disconsiders node degrees. As before, a node that has a cost higher than the remaining budget is simply skipped. Algorithm 5 presents this seeding policy.

Algorithm 2 Largest degree seeding strategy.

Require: $G = (V, E),\ \ b > 0$
  $V \leftarrow sortDescendingBasedOnDegree(V)$
  $O \leftarrow \emptyset$
  for $i = 1, \ldots, |V|$ **do**
    if $c(v_i) \leq b$ **then**
      $O \leftarrow O \cup \{v_i\}$
      $b \leftarrow b - c(v_i)$
    end if
  end for
  return O

Algorithm 3 Smallest degree seeding strategy.

Require: $G = (V, E),\ \ b > 0$
  $V \leftarrow sortAscendingBasedOnDegree(V)$
  $O \leftarrow \emptyset$
  for $i = 1, \ldots, |V|$ **do**
    if $c(v_i) > b$ **then**
      break
    end if
    $O \leftarrow O \cup \{v_i\}$
    $b \leftarrow b - c(v_i)$
  end for
  return O

Algorithm 4 Median degree seeding strategy.

Require: $G = (V, E),\ \ b > 0$
  $V \leftarrow sortAscendingBasedOnDegree(V)$
  $O \leftarrow \emptyset$
  for $i = 1, \ldots, |V|$ **do**
    $j \leftarrow \left\lceil \frac{|V|}{2} \right\rceil + \left(\left\lfloor \frac{i}{2} \right\rfloor \times (-1^{i-1}) \right)$
    if $c(v_j) \leq b$ **then**
      $O \leftarrow O \cup \{v_j\}$
      $b \leftarrow b - c(v_j)$
    end if
  end for
  return O

Algorithm 5 Random seeding strategy.

Require: $G = (V, E),\ \ b > 0$
  $V \leftarrow$*shuffle*(V) //{Random permutation of the nodes}
  $O \leftarrow \emptyset$
  for $i = 1, \ldots, |V|$ **do**
    if $c(v_i) \leq b$ **then**
      $O \leftarrow O \cup \{v_i\}$
      $b \leftarrow b - c(v_i)$
    end if
  end for
  return O

4. Evaluation

In this section, we present the evaluation of the four different network seeding strategies using extensive simulations. Different scenarios have been considered but due to space limitation in the following we present and discuss only the most interesting ones.

We consider networks generated by the $G(n, p)$ and Barabási-Albert models of size $n = 10000$ with different values for average degree ($\overline{d}$). We also consider scenarios with a different fraction of nodes to determine initial budget (k), epidemic threshold parameter (θ), and cost dependency on node degree (α), as defined in Section 3. Each parameter instance determines a *simulation scenario* for which the four seeding strategies are evaluated. For each scenario, we perform 30 independent simulation runs and report on the sample average and 95% confidence interval for the different statistics (error bars on plots). Note that both network models are random and each run considers a realization of the model, thus yielding a different network structure. Moreover, this is the only random component in our evaluation since the epidemic threshold model is deterministic, given the set of nodes initially infected.

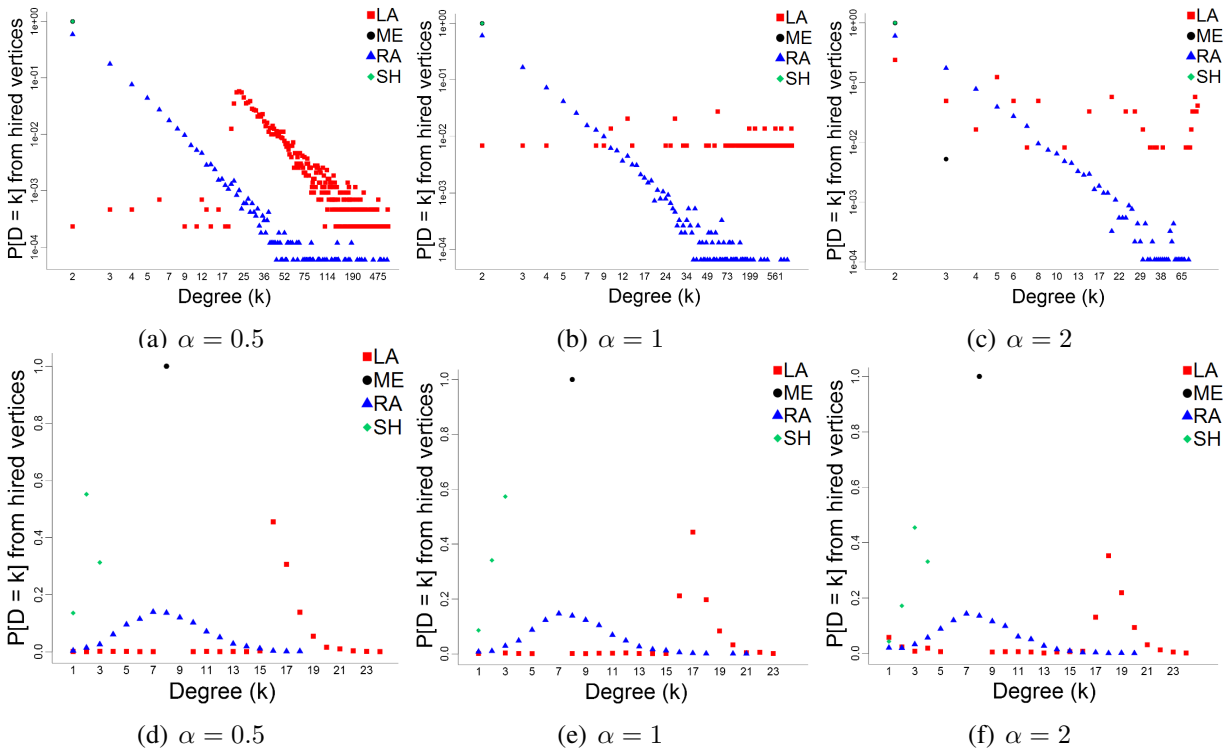
We consider the values of $\theta = \{2, 3, 4\}$ and $\alpha = \{0.5, 1.0, 2.0\}$ which have been selected to illustrate their strong influence on the epidemic. Note that α controls the node cost dependency on its degree, and the three chosen values represent a sub-linear, linear and quadratic dependency, respectively. For Barabási-Albert networks we consider $\overline{d} = 4$ and $k = 0.05$ while for $G(n, p)$ networks $\overline{d} = 8$ and $k = 0.01$, unless otherwise stated. Our goal is not to compare the two network models directly, but to understand how an epidemic unfolds within them with respect to some parameters. Thus, the parameters considered for average degree ($\overline{d}$) and fraction of nodes to determine the initial budget (k) for the two models are different. They were chosen to better illustrate for each network representative behaviors for each seeding strategy.

We start by reporting the *degree distribution of seeders*. Note that each seeding strategy will select a different set of nodes which in turn will induce a degree distribution. This allows us to understand what kind of nodes, with respect to their degree, the seeding strategy is selecting as seeders. Results are shown in Figure 3 for both network models with different plots corresponding to different values for α. Each plot presents the degree distribution for each seeding strategy. Note that for Barabási-Albert networks (Figure 3(a)-(c)), the SH and ME seeding strategies always select the minimum node degree (which is 2), independent of α. This occurs because this model generates networks that follow a power-law degree distribution with the vast majority of its nodes exhibiting the minimum degree (about 80%). Moreover, note that the number of nodes selected depends on α, since a larger α yields a larger budget (see Equation 3). Thus, since only nodes of minimum degree are selected, the exact number of nodes selected by these strategies is given by $b/d_0^\alpha = kn(\overline{d}/d_0)^\alpha$, where d_0 denotes the minimum degree. For the scenarios shown, this corresponds to 707, 1000 and 2000 nodes, for each value of α, respectively. Note this number is smaller than network size, $n = 10000$. Table 1 shows the number of nodes selected by each seeding strategy obtained through simulation results.

In the RA seeding strategy, since nodes are selected at random (budget permitting), the degree distribution of seeders should be similar to the degree distribution of the network (with a bias towards smaller degrees, due to budget restrictions) which does not depend on α, as observed in Figure 3(a)-(c). However, the LA seeding strategy shows a

Table 1. Size of seeding set ($|I(0)|$) for different scenarios.

	$G(n,p)$			Barabási-Albert		
Strategy	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
LA	69.03	46.70	22.13	144.13	4.97	3.80
SH	193.97	319.47	631.70	707.00	1000.0	1999.00
ME	100.00	100.00	100.00	707.00	1000.0	1988.57
RA	101.57	99.83	92.30	564.10	509.93	268.77



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 2$

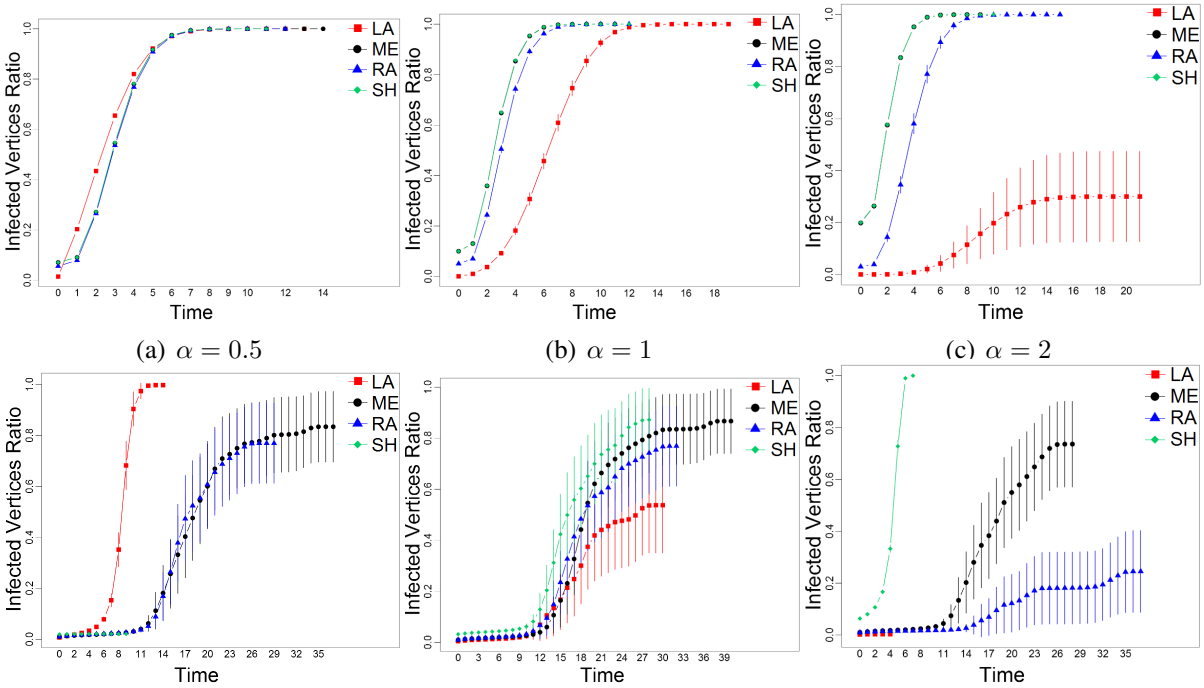
(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 2$

Figure 3. *Degree distribution of seeders*: Barabási-Albert (top) e $G(n,p)$ (bottom) networks. Parameters: $n = 10000$; $\theta = 2$.

very interesting behavior, since the set of selected nodes is greatly influenced by α. Note that when $\alpha = 0.5$, LA selects a bunch of high degree nodes, as well as some small degree ones (144.13 nodes on average, see Table 1). This occurs because under such circumstance ($\alpha = 0.5$), high degree nodes can be considered *cheap* if compared to $\alpha = 1.0$ or $\alpha = 2.0$. As α increases, LA selects very few high degree nodes and shows a mostly uniform distribution. This occurs because high degree nodes become much more expensive and a large fraction of the budget is spent on them. As a result, a very small number of nodes is selected, on average 3.80 for $\alpha = 2$ (see Table 1). Intuitively, this will have a strong influence on the epidemic, as we soon illustrate.

The degree distribution of seeders for the $G(n,p)$ networks for different seeding strategies are shown in Figure 3(d)-(f). Table 1 also shows the number of nodes selected by each strategy. Note that SH and ME are quite different, since ME selects only nodes with the same degree (in this case 8, which is the average), while SH selects nodes of degree smaller than 4. In fact, assuming only nodes with average degree are selected by the ME strategy, the number of nodes selected is given by $b/(\bar{d})^{\alpha} = kn$, independently of

α. This corresponds to 100 nodes in the scenario considered (as confirmed by simulation results, see Table 1). This assumption is reasonable because the network nodes follow a Binomial degree distribution, which is greatly concentrated around its average. The RA seeding strategy shows a degree distribution that is similar to that of the entire network (also with a bias, but smaller since nodes follow Binomial degree distribution), and is not dependent on α. Finally, the LA strategy selects most of high degree nodes of the network and then drops to a few small degree nodes, due to budget constraint. Note that most high degree nodes are selected, since they are not many and not very expensive (high degree is not very high compared to average degree). In any case, the number of nodes selected also drops significantly as α increases, as illustrated in Table 1.



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 2$

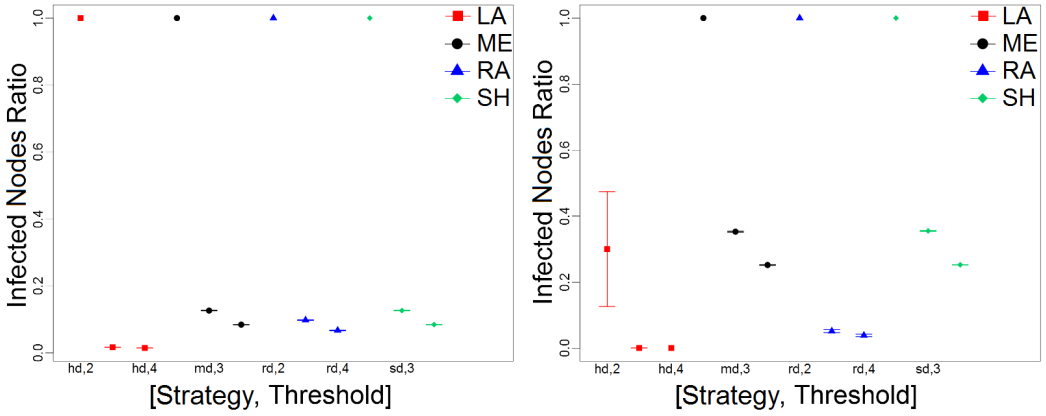
(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 2$

Figure 4. *Fraction of infected nodes over time*: Barabási-Albert (top) and $G(n, p)$ (bottom) networks. Parameters: $n = 10000$; $\theta = 2$.

In order to evaluate the epidemic, we consider the *fraction of infected nodes over time*, namely $I(t)/n$, as defined by Algorithm 1. Figure 4 illustrates the results for the different scenarios considered. The differences in performance for the different seeding strategies as a function of α is striking, for both network models considered. Note that LA is superior to all other strategies when $\alpha = 0.5$, for both network models. This occurs because when node costs are relatively inexpensive, selecting more high degree nodes yields a stronger epidemic. However, the story is quite different when node costs are high, in the case $\alpha = 2$. Despite the larger budget, the LA strategy exhibits worst performance than all other strategies for both network models. In the $G(n, p)$ model the epidemic spread to just a very small fraction of nodes, stopping after just a few iterations. Interestingly, in such regime the SH strategy is superior to all other strategies in both models. When nodes are expensive, selecting a very large number of low degree nodes pays off. Note that the ME strategy has identical performance to SH strategy in Barabási-Albert networks since these two strategies select the same set of nodes, as shown in Figure 3(a)-(c). Moreover,

note that the fraction of nodes infected at time zero increases significantly for SH and ME strategies in Barabási-Albert networks (from 5% to 20%) for different α values, as discussed above. Finally, note that the SH strategy on $G(n, p)$ networks for $\alpha = 0.5$ infects a negligible fraction of the network stopping very early, but is the only strategy to infect the entire network for $\alpha = 2$ and very fast. Also interestingly, no strategy is capable of consistently infecting 100% of nodes for $G(n, p)$ networks when $\alpha = 1$. In this case, SH and ME have comparable performance with SH terminating earlier, but infecting around 85% of the nodes. Note that results for this scenario have very large confidence intervals (compared to other results) indicating the sensitivity of the seeding strategies to specific network structure (which is random realization of the $G(n, p)$ model). Again, we observe the tradeoff between fraction of nodes selected and degree of nodes selected.

Figure 5 shows the fraction of infected nodes at the end of the epidemic for different seeding strategies and different epidemic thresholds θ in Barabási-Albert networks. For each strategy, three values for θ are shown. Note that varying θ does not affect the budget or the set of nodes selected as seeders by any of the strategies. Intuitively, a larger θ induces a more constrained epidemic and therefore should yield a smaller fraction of infected nodes when it terminates. The results clearly indicate that the epidemic is very sensitive to θ values while also depending on seeding strategy and initial budget. Note that when $\alpha = 0.5$ less than 15% of nodes are infected when $\theta > 2$, for any strategy. When $\alpha = 2$ less than 40% of nodes are infected when $\theta > 2$. Note that a larger budget allows for selecting a larger number of nodes for the SH and ME strategies, which explains their better performance for $\alpha = 2$. Again, we observe that SH and ME strategies have identical performance, as discussed. Moreover, their performances are superior or identical to other strategies for all cases shown.

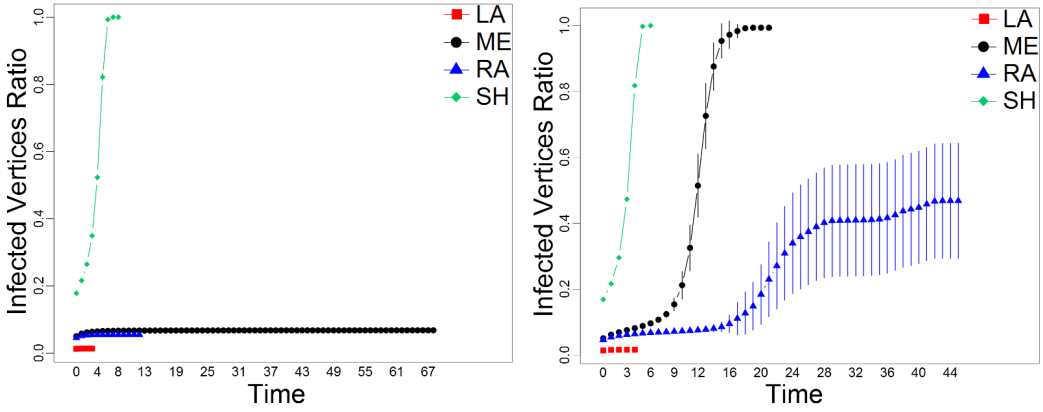


(a) $\alpha = 0.5$ (b) $\alpha = 2$

Figure 5. *Fraction of infected nodes* for different strategies and epidemic thresholds θ for Barabási-Albert networks.

Figure 6 illustrates an interesting phenomenon known as epidemic phase transition, where a small change in network parameters can yield very large change on epidemic contagion [Newman 2010]. We show that this phenomenon also depends on seeding strategies when node costs are variable. The figure shows the fraction of infected nodes over time for $G(n, p)$ networks for two different average degrees: $\overline{d} = 8$ (Figure 6(a)) and $\overline{d} = 9$ (Figure 6(b)). While the SH strategy already induces a supercritical epi-

demic (100% nodes infected) in both cases, the story is quite different for other strategies. For ME we see a change from less than 10% to almost 100% infected, and at the same time the epidemic duration reduced by more than half. It is worth to note that the epidemic does not reach 100% of susceptible nodes because there is a small fraction of nodes in the $G(n, p)$ network which has degree smaller than $\theta = 2$. The RA strategy changed from less than 10% to around 45% infected, with a large confidence interval. Thus, ME has clearly reached the epidemic phase transition threshold while the RA has not. Interestingly, the LA strategy has very poor performance in both cases, and also in terms of duration (very short epidemic). This indicates the epidemic phase transition threshold for LA requires a much larger average degree. Again, this indicates the importance of seeding strategies when nodes have variable costs, as the epidemic phase transition thresholds also depend on this characteristic as illustrated.



(a) $\overline{d} = 8$ (b) $\overline{d} = 9$

Figure 6. *Epidemic phase transition* in $G(n, p)$ networks. *Fraction of infected nodes over time* for two different average degrees keeping constant all other parameters ($\alpha = 2.0; \theta = 2; k = 0.05$).

5. Related Work

The problem of epidemic seeding (or network targeting) consists in determining the relatively small set of nodes to start an epidemic that will most successfully spread over the network. Under different formulations, finding an optimal set of nodes that maximizes some epidemic criteria is an NP-hard problem [Kempe et al. 2003, Arthur et al. 2009]. Thus, several heuristics have been proposed over the past decade targeting both different models for network structure and different epidemic models [Kempe et al. 2003, Arthur et al. 2009, Kitsak et al. 2010, Chen et al. 2009, Hinz et al. 2011, Aral et al. 2013]. For example, in the seminal work of [Kempe et al. 2003], a greedy heuristic is shown to approximate the optimal solution to a constant factor and is empirically superior to the high degree heuristic when evaluated with real networks. In another example, [Chen et al. 2009] provide degree-based heuristics that are computationally inexpensive, exhibiting low execution time and good performance. More recent works have considered other network characteristics (such as node homophily) and more realistic epidemic models to determine the influence of the seeding node set [Aral et al. 2013, Kostka et al. 2008]. All these previous works, however, implicitly assume that the cost of selecting a given node for seeding is the same for

all network nodes, since the constraint is simply the number of nodes that can be selected for seeding.

The work of [Arthur et al. 2009] is more related to ours in the sense that node costs are not fixed or identical across the network. Nevertheless, they consider a very different problem formulation where nodes purchase a given product at a node-dependent price and receive cashback for making recommendations to neighbors. The goal is to set prices and cashback rewards in order to maximize revenue with the offered product.

Finally, epidemic spread is far from being well understood in real social networks [Leskovec et al. 2007, Aral et al. 2013, Cha et al. 2010]. For example, [Cha et al. 2010] explore various indicators for the size of epidemic cascades in Twitter, indicating that degree is not necessarily the best predictor. This suggests that cost policies adopted by network nodes should take into consideration other network characteristics, a topic we leave for future research.

6. Conclusion

Network seeding is a fundamental problem in network epidemics since the set of nodes selected to start the epidemic strongly determines its outcome. We have considered network seeding under the assumption that nodes have costs that are proportional to their degrees, a more reasonable assumption in many applications, in particular viral marketing in online social networks, where popular nodes tend to be celebrities. We extensively evaluate the performance of a SI epidemic driven by a simple threshold model considering four different seeding strategies. Although the seeding strategies considered are not optimal, they are adopted either directly or as part of other heuristics when determining the seeding set, and serve to illustrate the importance of explicitly considering node costs.

Our main results show a fundamental tradeoff between selecting a larger number of cheap nodes (low degree) and a smaller number of expensive nodes (high degree). While both features are important, neither extreme is desirable: having many but smallest degree nodes or having very few high degree nodes may hinder an epidemic. In particular, we show that no single seeding strategy considered is consistently superior in terms of inducing a larger epidemic. Such results hold for the Erdős-Rényi and Barabási-Albert network models, despite having network structures that are fundamentally different. Finally, our work indicates the need of considering variable node costs in network seeding as this plays a fundamental role while also being a more reasonable assumption regarding many real networks. On such scenarios, the seeding strategy behavior may vary significantly according to both the cost function and the network structure.

As future work, we plan to investigate optimal (or near optimal) seeding strategies under variable node costs, in particular, strategies that automatically explore the tradeoff we have identified and can fare well independent of node costs and network structure.

References

Aral, S., Muchnik, L., and Sundararajan, A. (2013). Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science*, 1:125–153.

Arthur, D., Motwani, R., Sharma, A., and Xu, Y. (2009). Pricing strategies for viral marketing on social networks. In *Internet and Network Economics*, volume 5929 of *Lecture Notes in Computer Science*, pages 101–112.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 10–17.

Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 199–208.

Figueiredo, D. R. (2011). Introdução a redes complexas. Atualizações em Informática, chapter 7, pages 303–358. PUC-Rio.

Hinz, O., Skiera, B., Barrot, C., and Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75:55–71.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 137–146.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, 6:888–893.

Kornowski, L. (2013). Celebrity sponsored tweets: What the stars get paid for advertising in 140 characters. The Huffington Post.

Kostka, J., Oswald, Y., and Wattenhofer, R. (2008). Word of mouth: Rumor dissemination in social networks. In *Structural Information and Communication Complexity*, volume 5058 of *Lecture Notes in Computer Science*, pages 185–196.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Trans. Web*, 1(1).

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.