Just a Few Seeds More: Value of Network Information for Diffusion[*]

Mohammad Akbarpour[†]
Suraj Malladi[‡]
Amin Saberi[§]

First Draft: September 2017 This Draft: June 2018

Abstract

Identifying the optimal set of individuals to first receive information ('seeds') in a social network is a widely-studied question in many settings, such as the diffusion of information, microfinance programs, and new technologies. Numerous studies have proposed various network-centrality based heuristics to choose seeds in a way that is likely to boost diffusion. Here we show that, for some frequently studied diffusion processes, randomly seeding $s+x$ individuals can prompt a larger cascade than optimally targeting the best s individuals, for a small x. We prove our results for large classes of random networks, but also show that they hold in simulations over several real-world networks. This suggests that the returns to collecting and analyzing network information to identify the optimal seeds may not be economically significant. Given these findings, practitioners interested in communicating a message to a large number of people may wish to compare the cost of network-based targeting to that of slightly expanding initial outreach.

Keywords: Diffusion, seeding, social networks, targeting, word-of-mouth

JEL classification codes: D85, D83, O12, Z13

[*] This paper has been previously distributed with the title "Diffusion, Seeding, and the Value of Network Information." We thank Daron Acemoglu, Nava Ashraf, Stephane Bonhomme, Jeremy Bulow, Arun Chandrasekhar, Zoe Cullen, Darrell Duffie, Pascaline Dupas, Matthew Elliot, Ben Golub, Sanjeev Goyal, Matthew Jackson, Emir Kamenica, Alireza Khatami, Jon Kleinberg, Scott Kominers, Shengwu Li, Yucheng Liang, Greg MacNamara, Mihai Manea, Ilya Morozov, Michael Ostrovsky, Peter Reiss, Evan Sadler, Andy Skrzypacz, Alex Teytelboym, Christopher Tonetti, Carlos Varjao, as well as several other seminar participants, for valuable comments. We thank Jing Cai for helping us navigate their dataset. We also thank Maxime Rivet for coding some of the simulations reported here.

[†] Stanford Graduate School of Business. mohamwad@stanford.edu

[‡] Stanford Graduate School of Business. surajm@stanford.edu

[§] Department of Management Science and Engineering, Stanford University. saberi@stanford.edu

1 Introduction

How to identify individuals who are the best 'seeds' for maximizing the spread of information in a social network is a widely studied policy question in settings such as the diffusion of brand awareness for products (Richardson and Domingos, 2002), the propagation of microfinance programs (Banerjee et al., 2013), and the adoption of agricultural technologies in developing economies (Beaman et al., 2015). Since this problem is known to be computationally complex (Kempe et al., 2003), a large body of theoretical and empirical studies introduce heuristics such as 'degree centrality,' 'eigenvector-centrality,' 'diffusion-centrality,' and the 'k-shell' index as proxies for ranking candidate individuals to target[1]. While such heuristic approximations are computationally feasible, implementing them requires knowledge of the network structure, which can be extremely costly to acquire in field settings[2]. This is part of the motivation for studies such as Banerjee et al. (2014) or Breza et al. (2017), which develop methods for identifying central nodes or approximating the network structure without conducting a thorough census. Here, our goal is *not* to identify the central individuals, but instead to quantify the value of doing so. We are interested in questions such as: When is it important to target central individuals? What is the value of having access to the network information? And how does this value compare with the cost of seeding?

The main contribution of this paper is to recast the benefit of following a network-guided seeding heuristic in terms of the additional seeds required for a heuristic that ignores the network structure to perform just as well. For a widely studied model of information diffusion in networks, we show that seeding a slightly larger number of individuals randomly can prompt a larger cascade than seeding by optimizing over the network structure. We also show such results hold in simulations on some real-world network data and some alternative models of diffusion studied in the development economics literature. One can interpret our result as an upper bound on the value of network information and analysis for a policymaker attempting to spread information through word-of-mouth. This suggests that slightly expanding initial outreach may be more economical than network-guided targeting.

In our model, we consider a population of n individuals (or nodes) who are connected to each other through a social network. Individuals are either informed or uninformed about some product. The information percolates in the network according to a variant of the ubiquitous Susceptible-Infected-Recovered (SIR) diffusion model. In this model, all individuals (nodes) other than a small group (seeds) selected by the policymaker are initially uninformed. Once informed at time t, a node has one chance to speak to

[1] This problem has been studied in sociology, economics, marketing, computer science, medical sciences, physics, etc. See section section 1.1 for some references.

[2] Breza et al. (2017) estimate that conducting network surveys in 120 Indian villages would cost approximately $\$190,000$ and take over eight months.

each of its uninformed neighbors. This information sharing is successful with probability c independently for each neighbor, in which case the corresponding neighbors become informed by time $t + 1$. This cascade of information continues until no new individual has the opportunity to become informed.

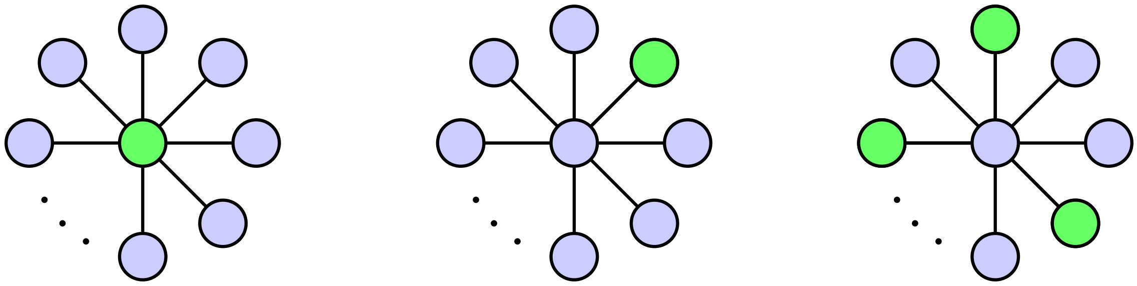
To quantify the value of network information in a policy-relevant way, we consider the following thought experiment: Suppose in one setting, we have access to full network data and unlimited computational power to optimally pick s individuals as initial seeds. In the second setting, we ignore the network and simply pick $s + x$ initial seeds uniformly at random. For what value of x will random seeding inform as many agents, in expectation, as the optimal seeding?[3]

In fact, we compare random seeding to a 'better than optimal' strategy, in the following sense. Suppose, in addition to the network structure, the policymaker has a perfect forecast of who would successfully share information with whom. She then picks the best s individuals to seed, equipped with this information. Comparing this 'omniscient' seeding with random seeding provides a generous upper bound for the value of network information, because for all realizations, the omniscient strategy will perform at least as well as the optimum, which itself performs better than computationally feasible heuristics.

Our main result shows that for a wide range of parameters, the random seeding strategy with $s+x$ seeds asymptotically performs as well as the omniscient strategy with s seeds, where x is vanishingly small relative to the size of the network. This result shows that, when the network is not too sparse or the communication probability is not too low, a policymaker interested in informing the greatest number of people should compare the cost of identifying the optimal seeds with the cost of seeding a few extra individuals. When these conditions fail and a sizable fraction of the population is not given the information, it is challenging to theoretically compare random and omniscient seedings. But in exactly this state of the world, we prove that even under the omniscient seeding, the fraction of informed population is vanishingly small compared to the total population. Moreover, our simulations show that even when the sufficiency conditions fail, except for a narrow range of parameters, extra seeds required by random to compete with the omniscient seeding is again small. This suggests that before considering network-based targeting strategies, a policymaker who believes that random seeding with a few additional agents is not a good strategy should first reconsider the efficacy of a word-of-mouth campaign.

We initially establish our result for 'sparse' Erdős-Rényi random graphs. Such graphs, however, do not admit highly central agents and have low 'clustering' coefficients. Indeed, all nodes in such networks are *ex ante* identical, so one may wonder whether the result is

[3] This thought experiment is analogous to the famous comparison of auctions and negotiations in Bulow and Klemperer (1994), and its generalization in Hartline and Roughgarden (2009). These results address how many additional bidders have to participate in a second-price auction, which requires no information on bidder valuations to implement, to generate as much revenue as an optimal auction with n bidders.



(a) Optimal seeding

(b) Random seeding

(c) Random seeding with additional seeds

Figure 1: A simple intuition for the main result: Consider a star network with n leaves, for some large n. Suppose an informed node passes information along to each of its neighbors independently with probability 0.5. 1(a): With a single seed, diffusion is maximized by picking the central node and in expectation $\frac{n}{2}$ of nodes will be informed. 1(b): Random seeding with a single seed will pick a non-central node with high probability. This means that half the time, diffusion ends immediately, and half the time, the central node becomes informed by the randomly chosen seed. Expected diffusion is approximately $\frac{n}{4}$, far below what optimal seeding achieves. 1(c): Now consider a scenario with $1 < x \ll n$ seeds. Random seeding will again pick x non-central nodes with high probability. However, the probability that a central seed is informed is $1 - (\frac{1}{2})^x$, so expected diffusion is nearly $\frac{n}{2}(1 - (\frac{1}{2})^x)$, which quickly converges to $\frac{n}{2}$ as x grows. For instance, random seeding with 5 additional seeds performs better than 97% of optimal seeding.

an artifact of the Erdős-Rényi structure. We therefore state similar results for a model of networks with power-law degree distributions and a generalized version of Erdős-Rényi graphs with high clustering. One may believe that informing one of the few highly central nodes in a power-law network can be very important. But the key here is that random seeding is likely to seed some of the connections of those highly central agents, precisely because such individuals have many connections. Therefore, central individuals will become informed through their neighbors. Figure 1 explains this intuition.

Our next result concerns the speed of diffusion. When considering the spread of a new technology through imitation, policymakers may also be concerned with the rate of adoption, rather than just the eventual reach. Indeed, insofar as imitation of neighbors' technology is a driving force of local economic growth, the speed of diffusion may be a primary concern (Alvarez et al., 2013; Perla and Tonetti, 2014). This is plausible, for instance, in agrarian village economies of the kind studied in Beaman et al. (2015), where the goal is to increase the influence of efficient agricultural technologies. Our second result considers a more general variant of the diffusion model in which the diffusion process ends after $T \geq 1$ periods, for graphs with n agents without very high number of connections (in a sense that we make precise). We then prove that after any number of periods, a random seeding strategy with a proportional increase of $o(\log(n))$ in the number of seeds performs, asymptotically, better than the omniscient strategy with s seeds.

After presenting our asymptotic results, we turn to the question of whether similar findings hold for finite, real-world networks. We study the Indian village household networks of Banerjee et al. (2013), the Chinese village rice farmer networks of Cai et al. (2015), and a small subnetwork of Facebook using data from Leskovec and Krevl (2014). We verify that random seeding competes well in the extent and speed of diffusion with both typically proposed and omniscient targeting strategies. For instance, in the Facebook subnetwork with nearly 4000 nodes, if each node speaks to her neighbors with probability 5%, random seeding with 10 seeds prompts a larger cascade than "diffusion-central" seeding with 11 seeds. Our simulations also show that the extra seeds required by random to beat network-guided heuristics in speed of diffusion is smaller than the theoretical $o(\log(n))$ multiplicative bound.

We go on to explore the robustness of our results to the specification of the diffusion model. The diffusion model considered so far is a workhorse model in the literature that studies information diffusion. We show that similar results hold for some of the more complex models estimated in the development economics literature. For example, for a version of the model of diffusion studied and estimated in Banerjee et al. (2013), random seeding with 11 seeds performs nearly as well as central seeding with 10. For the farmer social networks in Cai et al. (2015) and only in two periods[4], as another example, random seeding with 6 seeds performs nearly as well as central seeding with 5. It will be also clear from our proofs that similar results hold for the game-theoretic model of diffusion studied in Sadler (2018), where agents choose to adopt a product or not. Moreover, the model we studied exhibits undirected relationships and communication. This assumption may be questioned in some settings. Advice-giving, for example, is an inherently directed relationship where the involved parties are not necessarily equals. We then prove that our results hold in a model of random directed relationships and communication.

We close by stating a caveat: Network targeting might be essential in contexts different than ours. For instance, in a *threshold* model of diffusion where an agent is informed only if sufficiently many of his neighbors are informed, random seeding (even with a few additional seeds) performs poorly. Moreover, if the objective is to *minimize* diffusion by "vaccinating" individuals, targeting central nodes can be necessary. (Figure 6 illustrates this.) Therefore, our main finding should not be seen as an unambiguous advocacy for the superiority of expanded outreach to network targeting. Instead, the project of this paper is to identify practically relevant settings under which collecting and analyzing network data is not cost-effective. Whether network data is valuable for the problem in hand is, of course, context-dependent. That said, regardless of the diffusion model, 'extra number of seeds required for a network-agnostic seeding to compete with a prescribed network-based heuristic' is a statistic that can provide useful and easily interpretable information about the economic value of the results of a study beyond statistical significance. Table 1 shows that this statistic for Banerjee et al. (2013) and Cai et al. (2015) is smaller than 3.

[4] For more than two periods, random seeding will do even better.

1.1 Related Literature

How to identify a group of individuals within a social network who are most likely to adopt an idea or technology, or spread these things to others, are old questions in sociology. Lazarsfeld et al. (1948) is one of the firsts who discussed this, and the idea that certain individuals accelerate the spread of ideas or technologies by adopting these themselves gained popularity since.

The threshold model of collective behavior and word-of-mouth was introduced in Granovetter (1978). Domingos and Richardson (2001) is one of the firsts to introduce the influence maximization problem in the context of viral marketing. The independent cascade model was introduced in Goldenberg et al. (2001). Prominent among the next set of papers is Kempe et al. (2003), which considers two common diffusion models and asks how difficult it is to generally solve for the optimal size k set of initial targets when the objective is maximum contagion. They show that computing the optimal set is NP-hard.

The influence maximization spawned a large literature developing algorithms for influence maximization over networks in a wide rang of disciplines. To see such examples in computer science and operations, see Chen et al. (2009); Goyal et al. (2011); Asadpour and Nazerzadeh (2015); Chen et al. (2016); Wilder et al. (2017), in health-care, see Rice (2010); Rice et al. (2012); Kim et al. (2015); Yadav et al. (2016), in marketing, see Leskovec et al. (2007); Watts and Dodds (2007), and in the physics literature, see Kitsak et al. (2010); Chen et al. (2012). For more references, see a survey by Liu-Thompkins (2012). Unlike these papers, we formally quantify the value of network information as the extra seeds required by random seeing to beat the optimum, and identify conditions under which careful seeding may or may not matter. Our formulation of the value of seeding raises a policy-relevant comparison for researchers and practitioners.

In the development economics literature, Duflo and Saez (2003); Conley and Udry (2010); Dupas (2014) all point out to the importance of social networks as a means of learning in development economics.

Three development economics papers are particularly related to our paper: Banerjee et al. (2013) is one of the first papers to introduce the question of careful seeding to effectively harness such social learning. One of the main findings of this paper is that eigenvector-centrality and diffusion-centrality (a measure introduced by the authors, which captures something between degree and eigenvector centrality) of initial seeds is strongly correlated with total participation into the microfinance program, while degree centrality is not. Beaman et al. (2015) study technological adoption by farmers as they vary seeding rules over 200 independent village-networks in Malawi in an experimental setting. Their result suggest a 'threshold' kind of diffusion, although across different seeding strategies, they find very little diffusion beyond the initial seeds. In the instances where some non-seed nodes adopt, it is more likely to happen under network-guided seeding. This observation is in line with our theoretical result that if diffusion is vanishingly small, then network-based targeting can beat random seeding. Cai et al. (2015) also conduct a randomized experiment in which they seed certain individuals in Chinese village with information about a weather insurance program and observe take-up rates among their neighbors within a village social network.

Banerjee et al. (2013), Beaman et al. (2015) and Cai et al. (2015) all find that the effect of choosing central nodes on the size of diffusion is *statistically* significant. In contrast, the present paper asks when the magnitudes of the effects identified in those papers are *economically* significant enough to justify network targeting. Keeping fixed the number of seeds, central seeding strategies may compare favorably to random seeding. But this viewpoint ascribes too much value to having precise network information, in light of the alternative option of expanded outreach.

Banerjee et al. (2014) suggests that for a diffusion process in which nodes know the originator of information, asking individuals who they think are the biggest gossips is enough to identify diffusion central individuals. Our result, on the other hand, elucidates when it is important to target central individuals in the first place. Moreover, keeping the same diffusion process as in Banerjee et al. (2014), our comparison between omniscient and random seeding holds for any information structure.

In one of their results, Jackson and Storms (2017) study a heuristic for optimal seeding in a model of diffusion with threshold behavior. Building on our intuition about the threshold model, they show that random seeding requires many extra seeds to beat network-guided heuristics in that context.

A relatively recent theoretical literature in economics have also studied the optimal seeding under various diffusion processes the conditions for achieving widespread contagion, as well as competition in diffusion (Morris (2000); Galeotti and Goyal (2009); Young (2009); Goyal et al. (2014); Bloch et al. (2014); Lim et al. (2015); Mobius et al. (2015); Sadler (2018); Galeotti et al. (2017); Banerjee et al. (2018)). Meanwhile, other papers describe game theoretic foundations for the traditional measures of centrality (e.g., Ballester et al. (2006); Bloch et al. (2016)) or role of influential nodes (e.g., Galeotti and Goyal (2010)).

Organization of the paper. We introduce our network diffusion model in section 2. In section 3, we present our benchmark theoretical result for Erdős-Rényi graphs. Section 4 generalizes the benchmark result to power-law networks, as well as several real-world networks. In section 5, we study generalizations and limitations of the results with respect to the diffusion model. We extend the objective function to speed of diffusion in section 6. Main ideas behind the proofs are presented in section 7. Section 8 discusses some aspects of the model and results. Section 9 concludes.

2 Model

There are $n \geq 3$ individuals and we will refer to them as agents or nodes, with labels $i \in N = \{1, 2, \cdots, n\}$. Agents are connected in a *social network* represented by a simple graph $G = (N, E)$, where E is the set of unordered pairs of agents and $\{i, j\} \in E$ if agent i and agent j are *linked* or are *neighbors*. A node's *degree* in G is the number of its neighbors.

Diffusion process. Time passes in discrete periods $t = \{0, 1, 2, \ldots\}$. An agent is either *informed* or *uninformed*. Once an agent becomes informed, it remains informed forever after. Initially, a subset $A_0 \in N$ of individuals are informed. Once informed at time t, an agent has one chance to speak to each of its uninformed neighbors. We focus on the case that an informed individual has only one chance to speak to her neighbors, but this can be easily generalized to multiple (bounded) chances. This information sharing is successful with probability c independently for each neighbor, in which case the corresponding neighbors become informed by time $t+1$. Diffusion continues until no new individual has the opportunity to become informed. Later in the paper, we will consider cases where all communication ceases after some $T \geq 1$ periods. The case where T is finite is called a *bounded* diffusion process. Otherwise the diffusion process is called *unbounded*.

There is an alternative contrived but useful way to think about the unbounded diffusion: Suppose at time t, there is a coin flip for each link of the social network G, and with probability c that link is maintained in the network. Let us call this new constructed network, which is clearly a function of the original network, a *communication network* and denote it by $\mathcal{K}(G) \subseteq G$. The communication network is a way to think about the set of all pairs of agents who will speak to each other, once one of them becomes informed.

The diffusion process considered here is one in which communication is undirected. In particular, the event that node i talks to j if informed is coupled with the event that j talks to i if informed. Moving from undirected communication settings to directed communication requires addressing some technical issues. We postpone the discussion of this case to the section 5.2, where we show that results can be extended to settings with directed communication.[5]

Seeding strategies. A seeding strategy takes as input a network and a number of initial seeds $s \leq n$ and outputs a (random) set of s initial seeds to be informed at time $t = 0$. Formally, let $\mathcal{U}_n$ be the set of all node-labeled networks on n nodes and let $[n] = \{1, 2, \ldots, n\}$. A seeding strategy is a set-valued (random) function $f : \mathcal{U}_n \times [n] \to 2^N$, with the property that $|f(G, s)| = s$.

We say seeding strategy f is *feasible* if for all networks $G = (N, E) \in \mathcal{U}_n$ and

[5] In addition, simulations of appendix E and appendix F consider models of directed communication.

$s \leq |N| = n$, $f(G, s)$ and $\mathcal{K}(G)$ are independent. The communication network encodes the information of who would speak to whom, which is of course not available to a policymaker *a priori*. A seeding strategy that does not satisfy this property uses the realization of this information in determining the choice of seeds, and is therefore infeasible to implement. While in practice a policymaker with no knowledge beyond the network structure can only use feasible seeding strategies, infeasible strategies can be useful as theoretical benchmarks. Let $\mathcal{F}$ be the space of feasible seeding strategies for graphs on n nodes.

Network formation model. A model of network formation is a probability distribution $\mathbb{P}_n$ over the set of all networks $\mathcal{U}_n$. Let G_n be a (random) graph drawn from the distribution $\mathbb{P}_n$. Going forward, we will consider specific distributions $\mathbb{P}_n$. We will drop n from the notations when it is clear from the context. As will be clear shortly, in stating our theoretical results, we are interested in limit behavior of the networks as $n \to \infty$.

To start, we consider an Erdős-Rényi network formation process on a set of agents N. In an Erdős-Rényi random network, there is a link between a pair of agents $(i, j) \in N^2$ with probability p, independently of other agents and links. We use the notation $ER(n, p)$ to denote an Erdős-Renyi network on n nodes in which each link exists with probability p.

Erdős-Rényi structure is the simplest and most widely used network model for which we state our results. We then generalize the main result to broader classes of models that resemble features of real-world networks.

Goal. Let $A_t(G, s, f) \subseteq N$ denote the (random) set of informed nodes at time $1 \leq t \leq T$, as a function of the network G, number of seeds s, and the seeding strategy f.

Let $\mathbf{h}(G, s, f) \equiv \mathbb{E}[|A_T(G, s, f)|]$ be the expected number of informed agents at the end of the process. Here the expectation is taken over the diffusion process. Let $\mathbf{H}(f, s) \equiv \mathbb{E}_{G \sim \mathbb{P}_n}[\mathbf{h}(G, s, f)]$. The function $\mathbf{H}$ measures the performance of a seeding strategy by taking the strategy and number of seeds as inputs and producing the expected total number of informed agents as output, for a given network formation process. The goal of the planner is to choose a seeding strategy f to maximize $\mathbf{H}(f, s)$.

Relevant seeding strategies. We denote the optimal seeding strategy by OPT. For a fixed network, this strategy picks the set of s seeds that maximizes the expected diffusion, with an arbitrary selection when there are multiple optimal candidates:

$$\mathrm{OPT}(s) \in \operatorname*{argmax}_{f \in \mathcal{F}} \mathbf{H}(f, s).$$

It is known that computing this strategy is NP-hard (Kempe et al., 2003). In practice, instead, policymakers resort to heuristics such as seeding the s most central individuals in the network, according to various measures of centrality. We will introduce them in the empirical section of the paper.

We define two seeding strategies as theoretical benchmarks. Let RAND(s) be the strategy which picks s nodes uniformly at random in G. This strategy ignores all the information about the network structure.

On the other end, we analyze the *omnicient* seeding strategy, denoted by OMN(s), which for every realization of the communication network picks s initial seeds to maximize diffusion. Notice that this strategy is infeasible by construction because it knows who is going to speak to whom, and it performs better than any feasible strategy for any realization of the diffusion and network formation processes. In particular, for any initial number of seeds k:

$$\mathbf{H}(\mathrm{OMN}, s) \geq \mathbf{H}(\mathrm{OPT}, s) \geq \mathbf{H}(\mathrm{RAND}, s)$$

3 Benchmark Result: Erdős-Rényi Networks

To quantify the value of learning the network and identifying the optimal seeds, we would ideally like to compare the performances of OPT and RAND. Recall that OPT exploits the full knowledge of the structure of the network and solves a computationally hard optimization problem, while RAND ignores any information about the network. Therefore, the difference between these two can be interpreted as the value of network information and analysis. As noted earlier, however, computing OPT is an NP-hard problem. Instead, we measure the difference between the performances of OMN and RAND. Since for any realization of the diffusion process, OMN performs better than OPT, comparing RAND and OMN gives a generous upper bound on the value of network information and analysis.

To operationalize the difference between these two seeding strategies for the policymaker, we pose the following question: how many additional seeds are required in random seeding to compete with the omniscient strategy? Our first theorem, stated below, shows that in Erdős-Rényi networks, when diffusion is effective, this number is small. The results stated in this section will be for the unbounded diffusion process. The proofs make use of results from percolation theory and are relegated to the appendix. The ideas behind the proofs are presented in section 7.

To simplify the statement of our theoretical results, we define the following notation: We say that a function $f(n)$ asymptotically weakly dominates $g(n)$ if $\lim_{n\to\infty} |\frac{f(n)}{g(n)}| \geq 1$. We also say f is of $o(g)$, $\omega(g)$, and $O(g)$ if and only if this limit is zero, infinity, and any finite constant respectively. For example, any divergent increasing function of n is $\omega(1)$. We refer to $\omega(1)$ as a *super-constant*.

Theorem 1. *Consider an Erdős-Rényi network on n nodes with average degree d. Let c be the probability that an informed node speaks to a given neighbor and let $s = o(\frac{n}{\log(n)})$. If $cd > 1$, then for any super-constant $x(n)$,*

$$\lim_{n\to\infty} \frac{\mathbf{H}(\mathrm{RAND}, s + x(n))}{\mathbf{H}(\mathrm{OMN}, s)} \geq 1.$$

If $cd \leq 1$, then

$$\lim_{n\to\infty} \frac{\mathbf{H}(\mathrm{OMN}, s)}{n} = 0.$$

This result states that if $cd > 1$, which means that each person on average talks to at least one of their friends, then random seeding with super-constant extra seeds (asymptotically) performs better than the omniscient seeding. On the other hand, when $cd \leq 1$, the fraction of informed nodes even under the omniscient seeding strategy goes to zero as $n \to \infty$.

We point out that while a super-constant $\omega(1)$ grows with n, such 'limit results' should not be read literally. The key lesson here is how quickly random takes over the optimum. Asymptotic results help us to theoretically study complex networks in a tractable way. Importantly, the number of additional seeds needed by random to be competitive with omniscient can be made to be asymptotically small compared to the number of seeds; that is, x/s can quickly go to zero. In fact, even for finite networks, random requires only a few additional seeds. For instance, consider an Erdős-Rényi random network on 1000 nodes in which each informed node passes along the information to 1.5 neighbors on average (so $cd = 1.5$). Simulations show that randomly seeding 7 nodes achieves within 95% of omnisciently seeding 5 nodes. If nodes pass along information to 2 neighbors on average, then even 20 omnisciently selected nodes do no more than 5% better than 23 randomly selected nodes (with differences being even more negligible for fewer seeds).

Remark 1. *While $x(n) \in \omega(1)$ additional seeds are needed for random seeding to asymptotically dominate omniscient seeding, only a finite number of additional seeds, $h(\epsilon) \in \mathbb{N}$, are needed for random seeding to come within a ratio of $1 - \epsilon$ of omnisicnet*[6].

The restriction on the size of $s = o(\frac{n}{\log(n)})$ precludes, for example, the case when a constant fraction of the seeds in a network are targeted. Indeed, the motivation for using word-of-mouth as a vehicle for diffusion is that few initial seeds are needed to reach a large population, so this restriction is natural. But this is not to say that random seeding is not comparable to practically implementable seeding strategies when more seeds are used. To the contrary, intuition and simulations suggest that flooding many individuals with information makes careful selection of initial seeds less valuable. However, the

[6] This can be seen from the proofs.

performance of implementable strategies eventually flattens out due to redundancy in seeding choices, while the omniscient strategy delicately avoids seeding the same cluster twice, and so the omniscient ceases to be a useful theoretical benchmark for comparison. In fact, simulations of section 4.2 confirm that similar bounds will go through for relatively small networks and all values of s.

4 Generalized Networks

In this section, we will show that our main result holds for a variety of theoretical network models and real-world network structures.

Some real-world networks are characterized by degree distributions with fat tails, in the sense that they exhibit few nodes that have significantly greater degrees than others. For example, Barabasi and Albert (1999) describe a variety of social networks, such as the network of linked web pages or collaborating actors, exhibiting a power-law like degree distribution on its right tail. They also have a high degree of clustering (neighbors of neighbors tend to be neighbors themselves). Erdős-Rényi networks fail to capture either of these properties. Here we will show that Theorem 1 can be extended to more general network models that exhibits power-law degree distribution. In the appendix B.1, we show that the result can be extended to a model of network with clustering. Last but not the least, we will show that our results hold for a variety of real-world networks with non-random structure and clustering, including Indian village network of Banerjee et al. (2013) and a subnetwork of Facebook.

4.1 Power-law Chung-Lu networks

We will now consider network formation models that allow for more general degree distributions. In particular, "Chung-Lu" networks (Chung and Lu, 2002) or inhomogeneous random graph models are generalizations of Erdős-Rényi that support power-law.

Definition 1 (Chung-Lu network)**.** *Fix a sequence* $\boldsymbol{w} = (w_1, \dots, w_n) \in \mathbb{R}^n_+$. *A* Chung-Lu *(undirected) network on* n *nodes,* $CL(n, \boldsymbol{w})$*, is generated by including each edge* $\{i, j\}$ *independently with probability* $p_{ij} = \min(\frac{w_i w_j}{\sum_k w_k}, 1)$.

For any node i, the expected degree is equal to $\sum_j \frac{w_i w_j}{\sum_k w_k} = w_i \frac{\sum_j w_j}{\sum_k w_k} = w_i$, which means that the sequence of weights $\mathbf{w} = (w_1, \dots, w_n)$ doubles as the sequence of expected node degrees as well. Therefore, in order to capture the power-law degree distribution, we consider a parametric power-law functional form for the weights. In particular, we assume that for all i,

$$w_i = [1 - F]^{-1}(i/n), \text{ where } F(x) = 1 - (d/x)^b \text{ on } [d, \infty). \tag{1}$$

We call a Chung-Lu network with such a weight sequence a power-law Chung-Lu network on n nodes with minimal expected degree d and scale parameter b. The parameter b determines the thickness of the tail, and as b grows, the tail becomes thinner and thinner. The more permissive definition of power-law requires that the mass of the cumulative distribution function lying to the right of some large enough k is proportional to $k^{-\tau}$. Our model satisfies this condition for $b = \tau + 1$. Moreover, when the weight distribution follows a power law, the degree distribution of the corresponding random graph follows a power-law tail Van Der Hofstad (2016). Barabasi and Albert (1999) estimates the scale parameter for the tails of different real-world network degree distributions and find this lies in the $(1, 2]$ interval for many of their examples.

Theorem 2. *Consider a power-law Chung-Lu network on n nodes with scale parameter b and minimal expected degree d . Let c be the probability that an informed node speaks to a given neighbor and let $s = o(\frac{n}{\log(n)})$. If either (1) $b \in (0, 2]$ or (2) $b > 2$ and $cd > (b-1)(b-2)$, then for any super-constant $x(n)$,*

$$\lim_{n\to\infty} \frac{\mathbf{H}(\mathrm{RAND}, s + x(n))}{\mathbf{H}(\mathrm{OMN}, s)} \geq 1.$$

If $b > 2$ and $cd \leq (b-1)(b-2)$, then

$$\lim_{n\to\infty} \frac{\mathbf{H}(\mathrm{OMN}, s)}{n} = 0.$$

A power-law network can have nodes with extremely high degrees. The optimal seeding strategies can pick those nodes as seeds, whereas the random seeding strategy (even with a few additional seeds) will most likely never picks those nodes. Perhaps surprisingly, it is in the case where the tail of the degree distribution is sufficiently thick that no further assumptions on communication probability are needed to ensure the result of the theorem. This raises the question of how random seeding can compete in this case. The intuition, as depicted in Figure 1, is that random seeding is likely to pick one of the neighbors of the highly connected nodes. Hence, highly connected nodes are likely to become informed through their connections.

4.2 Real-world networks

So far, we have provided bounds on the value of seeding for two important theoretical classes of networks. We can consider other theoretical network models, but at the end no network model can match all moments of the real-world networks. Therefore, here we offer a (network formation) model-free perspective on the results of section 3 in an economically relevant context. We simulate the diffusion model studied here on the microfinance network data in Banerjee et al. (2013) as well as a subnetwork of Facebook,

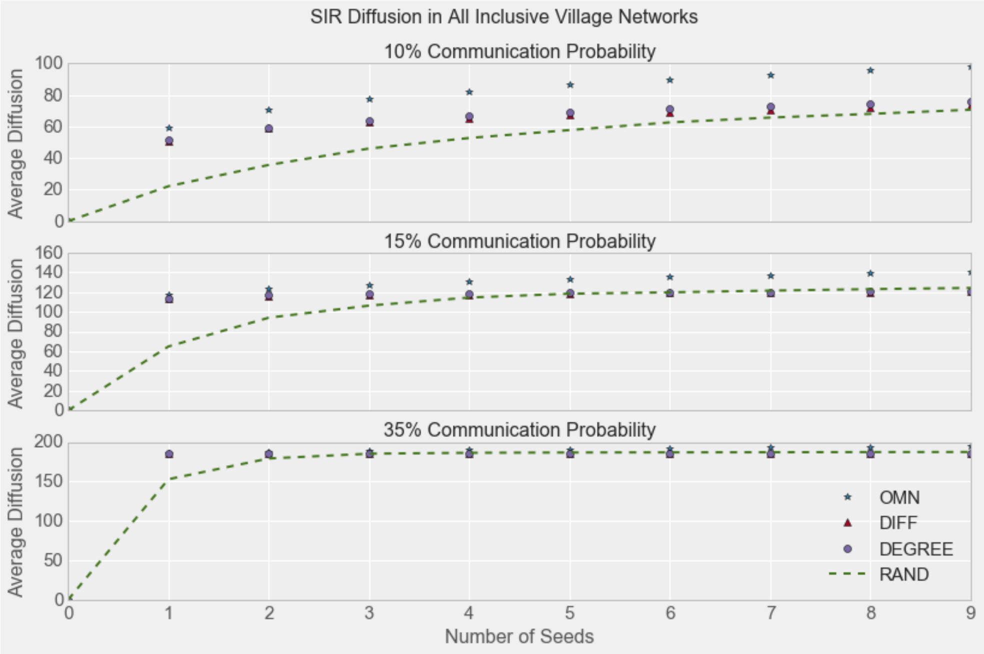


Figure 2: The average diffusion achieved by the various seeding strategies (omniscient, random, degree, diffusion) across 'all inclusive networks' in the village network data, for various levels of communication probabilities.

and compare the performance of various seeding strategies.

The networks in Banerjee et al. (2013) have households as nodes, with edges representing some sort of relationship. For example, in one network, the edges represent that members of the incident households go to temple, mosque or church together. In another network, the edges represent the fact that members of one household have borrowed or loaned money to those in the other or frequently give or take advice from the other, and so on. While some of these relationships are directed, the graph will be taken to be undirected. For information diffusion, it is not unreasonable to think that any sort of contact creates an opportunity to speak about the topic at hand.

Simulations in Figure 2 compare average performance of random, degree-central, diffusion-central[7], and omniscient seeding strategies on 'all inclusive' village networks, which includes an edge between two households whenever either party indicated some contact with the other group of any form. Results are included for different parameters of our diffusion process, which indicate the probability that two connected nodes communicate information to each other given that one of them is informed. Running the same simulations for sparser networks (e.g., the kerosene and rice lending networks within the

[7] Degree centrality is simply a ranking of nodes from those with the most neighbors to those with the least. Diffusion centrality for each node in a graph with adjacency matrix $\mathbf{g}$, diffusion probability q, and T periods of communication is given by $DC(\mathbf{g}, q, T) = [\sum_{t=1}^{T}(q\mathbf{g})^t] \cdot \mathbf{1}$ (Banerjee et al., 2013). At $T = 1$, this measure ranks nodes simply by degree, and as $T \to \infty$, depending on whether q is larger or smaller than the inverse of the largest eigenvalue of $\mathbf{g}$, the vector of diffusion centralities converges to a ranking proportional to Katz-Bonacich or eigenvector centrality respectively (these can be taken as the definitions of the latter measures).

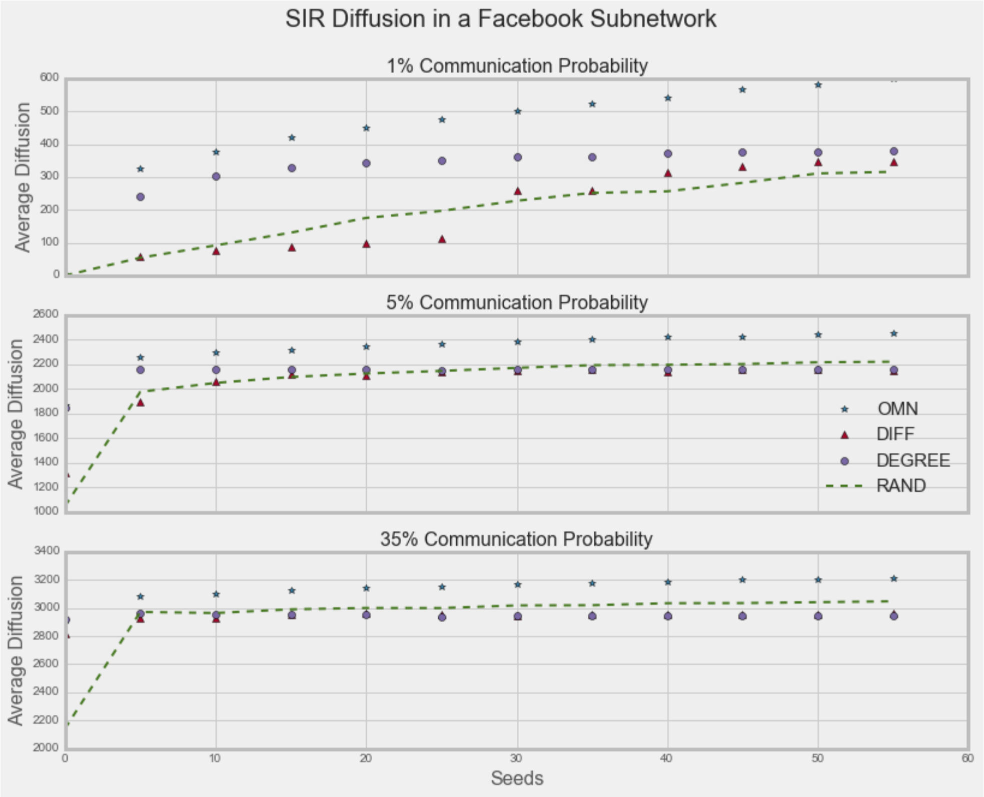


Figure 3: A comparison of the various seeding strategies on a Facebook subnetwork with approximately 4000 nodes, for different communication probabilities. The random seeding strategy competes well with OMN and other seeding strategies, when the communication probability is not too small. And when the number of seeds is not too small, it beats degree and diffusion central seeding strategies.

same Indian villages) does not qualitatively change the results.

Next, we replicate the comparison between diffusion strategies on a Facebook subnetwork in Figure 3 to show that the patterns observed for the Indian village data roughly bear out here as well. In comparison to the village data, the degree distribution for this network exhibits a fatter right tail.

5 Alternative Diffusion Models

So far, our theoretical results focused on the undirected SIR model of diffusion, which is used to study processes such as diffusion of information and ideas, rumors, or infectious diseases. We focused on the SIR model since this is a workhorse model, studied and estimated in several economic environments. We will now discuss alternative diffusion models under which our results will (or will not) hold.

5.1 Models from development economics

The diffusion models used in Banerjee et al. (2013) and Cai et al. (2015) are more complex, but still share the feature of the SIR model that an agent's neighbors are "substitutes", in the sense that having one informed neighbor ensures with sufficiently high probability that an agent will be subsequently informed. For instance, in Banerjee et al. (2013), once an agent gets informed, she may or may not participate in the microfinance program, and participants inform their neighbors with higher probability than non-participants. Cai et al. (2015), on the other hand, consider a linear probability model, where the chance that an agent gets informed is proportional to the number of its informed neighbors.

Our basic insight goes through for all diffusion models discussed above. To show this, we will consider the diffusion models and the social network data of Banerjee et al. (2013) and Cai et al. (2015) and compare centrality-guided and random seeding strategies. Simulations reported in appendix E (for the Microfinance model) and appendix F (for the weather insurance model) show that the number of additional seeds required for random to compete with centrality-guided heuristics is small.

When the diffusion process is such that neighbors are "complements", say when several of an agent's neighbors have to adopt a technology before he does the same, our results may fail to hold. For instance, in the *threshold* type models of diffusion, agents will only adopt a behavior if at least a certain number (or fraction) of their neighbors adopt, so there are complementaries in the inputs of propagation. Beaman et al. (2015) study technological adoption by farmers as they vary seeding rules in village-networks in Malawi in an experimental setting. Their result suggest a threshold-type diffusion process, although they observe little diffusion. Since random seeding is unlikely to inform multiple neighbors of the same node, if thresholds are uniformly high across all agents, random seeding will fail to prompt any diffusion. This intuition has been subsequently formalized in Jackson and Storms (2017). Typically, these models assume a uniform threshold across agents. But if thresholds are heterogeneous and sufficiently many agents have a threshold of 1, then results similar to our main theorems may continue to hold.

5.2 Directed communication

The models considered so far exhibit undirected relationships and communications. In particular, the event that node i talks to j if informed is coupled with the event that j talks to i if informed. The assumption of undirected relationships and communication may both be called into question. Indeed, it frequently happens in surveys that one individual names another as a close friend, without the other declaring in kind. In addition, even if relationships are undirected, it is not a foregone conclusion that just because one agent would have informed a friend of some information, that the reverse would have occurred had the latter party learned of the information first.

We will now consider a model of directed networks similar to Erdős-Renyi. $D(n,p)$ is a random directed network on n nodes in which directed edge (i,j) is drawn with probability $\frac{p}{n}$. In this setting, OMN observes a realization of the directed communication network and chooses the best nodes to seed using this information.

Theorem 3. *Consider an random directed network, $D(n,p)$. Let c be the probability that an informed node speaks to a given neighbor and let $s = o(\frac{n}{\log(n)})$. If $cp > 1$, then for any super-constant $x(n)$,*

$$\lim_{n\to\infty} \frac{\mathbf{H}(\text{RAND}, s + x(n))}{\mathbf{H}(\text{OMN}, s)} \geq 1.$$

If $cp \leq 1$, then

$$\lim_{n\to\infty} \frac{\mathbf{H}(\text{OMN}, s)}{n} = 0.$$

5.3 Game-theoretic models

The model of diffusion we study here, as well as those studied in Banerjee et al. (2013), Cai et al. (2015) and Beaman et al. (2015), are all "mechanical" models, in the sense that agents do not optimize any specific objective function. In our model, for instance, once an agent is informed, it will inform each one of its friends with some probability. In principle, one could micro-found this behavior. Doing so is not the focus of this paper. Instead, we will simply note here the implications for seeding in the game-theoretic diffusion model introduced in Sadler (2018). There, agents get informed, update their beliefs about their network position in a Bayesian fashion, and can *choose* to adopt a product. From the proof strategies of our paper, it is straightforward to see that our results hold in this context, since that paper (like ours) exploits the percolation results that if the diffusion process reaches a positive fraction of the population, there will be a giant component of informed individuals.

6 Speed of Diffusion

We will now discuss conditions under which a "similar" result can be extended to bounded diffusion processes, where all communication ceases after a fixed number of rounds. This result, then, shows that random seeding competes with omniscient seeding period by period, which can be interpreted as a statement on the relative *speeds* of diffusion in the unbounded case. This addresses the economically salient concern that while both seeding strategies eventually reach the same level of diffusion, network information allows policymakers to significantly accelerate the speed with which the information spreads. As an example, policymakers may be concerned with how quickly farmers adopt a new technology, so that the developing economies may grow at faster rates. Indeed, diffusion

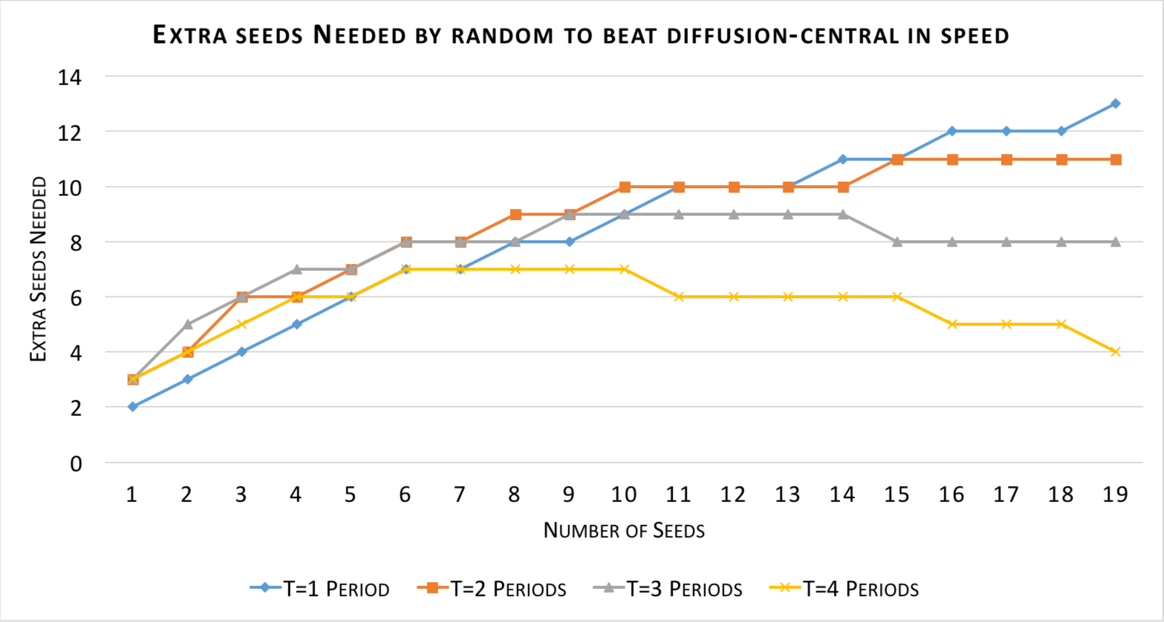


Figure 4: Average number of extra seeds required by random to outperform diffusion-centrality seeding in Indian village networks (in of *speed of diffusion*). If objective is diffusion in the first T=1 or T=2 periods, then extra seeds required is relatively high (still less than 15), but once total outreach in the first T=3, T=4 or more periods is the objective, less than 9 extra seeds is enough.

of technologies through imitation is a primary concern for the growth rate in a local economy (Perla and Tonetti, 2014), for instance in agrarian village economies of the kind studied in Beaman et al. (2015), where the policy-makers wish to expand the influence of modern, more efficient agricultural technologies.

Our main result in this section shows that for Erdős-Rényi networks, with $o(\log(n))$ times additional seeds, random seeding competes with the omniscient seeding even in the speed of diffusion.

Theorem 4. *Consider an Erdős-Rényi network on n nodes with average degree d and a bounded diffusion process that ends in $T \geq 1$ periods and let s be a non-negative integer. Then,* $\mathbf{H}(\text{RAND}, o(\log(n))s) \geq \mathbf{H}(\text{OMN}, s)$ *for n sufficiently large.*

In appendix B.2 we prove a generalized version of this theorem for a network model with clustering.

When the objective function is the speed of diffusion, our theoretical bounds are weaker: As opposed to $\omega(1)$ additional seeds for a broad class of network models, we provide $o(\log(n))$ multiplicative bound for Erdős-Rényi networks. This is because when it comes to bounded diffusion processes, it is easy to identify theoretical cases in which careful seeding is important. For instance, in a star network, when diffusion process ends after $T = 1$ period, seeding the central node is essential for obtaining any diffusion.

6.1 Speed of diffusion in microfinance setting

We now compare speed of diffusion under random and diffusion-central seeding using the microfinance model of diffusion and Indian village networks. Figure 4 depicts the extra number of seeds needed for random to beat diffusion-central seeding in microfinance setting. When the diffusion ends in $T = 1$ or $T = 2$, periods, the extra number of seeds required for random to catchup is between 3 to 13, depending on the number of seeds. When $T = 3$ and $T = 4$, the extra number of seeds needed for random is always less than 9 and 7, respectively.

7 Proof Ideas

In this section, we discuss some ideas behind our theoretical results. Rigorous proofs are presented in the appendix. Readers not interested in these techniques may skip this section. The exposition uses standard graph-theoretic terminology that can be found in standard references, e.g., Jackson (2010).

Theorem 1. Recall that the communication network $\mathcal{K}(G) \subseteq G$ is a way to think about the set of all pairs of agents who will speak to each other, once one of them becomes informed. We can consider the connected components of this communication network to better understand the behavior of random and omniscient seeding strategies. Note that in the SIR model, a node becomes informed if and only if one of the nodes in its connected components in $\mathcal{K}$ is seeded. This implies that an omniscient seeding strategy with s seeds would simply seed one node in each of the s largest connected components of $\mathcal{K}$. On the other hand, for each seed, the probability that the random strategy informs a given component is proportional to the component's size. This gives us a method of computing the expected diffusion for each of strategies, once we are given the distribution of component sizes for a communication network.

When n is sufficiently large and $cd > 1$, by the standard phase transition result for Erdős-Rényi random graphs, there exists a component in the communication network which contains a constant fraction of the total population. The remaining components, on the other hand, are vanishingly small ($O(log(n))$) in population size. So as long as the random seeding strategy informs the nodes in the large component, which it can with high probability once it is given a sufficiently large budget of seeds, omniscient seeding cannot do much better.

When $cd < 1$, then even the largest component is $O(log(n))$ in size, so the omniscient seeding strategy with $o(\frac{n}{log(n)})$ seeds can only inform a vanishingly small ($o(n)/n$) fraction of the population.

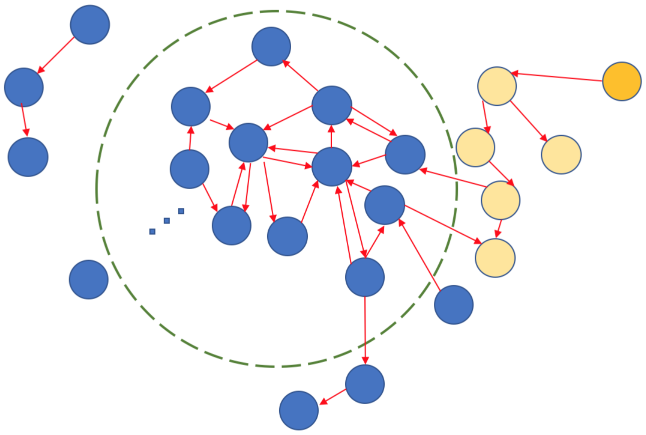


Figure 5: Above is an example communication network when communication is directed. The outgoing edges represent the nodes that a given node would inform if given information. The nodes within the dotted dashed circle represent the strongly connected giant component. The orange nodes, if informed, disseminate information to the SGC. In this example, OMN might choose to seed the dark orange node, given a single seed (and there could be many such useful entry points, though only one set of orange nodes is pictured above). In the proof of Theorem 3, we show that the size of the set of any cluster of orange nodes is $o(\log(n))$ so that OMN cannot significantly outperform RAND.

Theorem 2. The arguments are analogous and similar phase transition theorems hold for Chung-Lu graphs for some range of parameters (Van Der Hofstad (2016)). For our results, we need to extend these theorems over a range of parameters where standard techniques do not apply (see Appendix C for details).

Theorem 3. The idea of using the communication network applies also to the case of directed networks with directed communication. However, the nodes that ultimately become informed are those for which a *directed path* exists from a seed. The analog of the giant component is the unique *strongly connected* giant component (SGC), which the random seeding strategy reliably hits.

The trouble in this case, however, is that a seed which ultimately informs the SGC may not be a member of this component at all (see Figure 5). Consider such a node and the length of the shortest path leading from this node into the SGC. If the path length is long, an omniscient strategy would go choose this node as an entry point into informing the nodes in the SGC. But a random seeding strategy with any number of seeds would not hit such a node, other than through sheer luck. Results in Karp (1990) indicate that such paths are $o(\sqrt{n})$ in length, which is too generous of an upper bound for our results to hold. We establish that these paths are in fact $o(\log(n))$ in length, and can therefore be safely ignored.

Theorem 4. We need to show that the extent of diffusion that happens in T periods from any node must be $o(log(n))$ for large n. This is equivalent to bounding the size of a $T-$neighborhood of a random graph: with probability approaching 1, only $o(\log(n))$ nodes are reachable in T steps from any given node. As described in the appendix, this bound straightforwardly extends to a model of graphs with clustering inspired by Jackson and Rogers (2007), where friend of friends and friend of friend of friends (etc.) are likely to be one's direct neighbors as well. In this sense, clustering does not change the performance comparison between omniscient and random seeding in speed of diffusion.

8 Discussion

8.1 A policy-relevant statistic

Our results suggest a way to measure the operational value of a seeding heuristic for researchers looking at the benefits of network based targeting in settings beyond ours. In particular, consider a general network setting, where the goal of a research study is to identify optimal nodes of a network for maximizing diffusion, for a given diffusion model. This could be the diffusion model studied in this paper or generalizations thereof, or any other diffusion model of interest. Suppose the researchers identify a specific seeding heuristic to perform well. These researchers can report the following statistic as a policy-relevant quantity: *How many extra seeds are needed for the random seeding strategy to be within z% of their proposed strategy, for a small z?*

For example, for the diffusion model of Banerjee et al. (2013) and with $s = 10$ initial seeds, random seeding with 1 extra seed is within 95% of the seeding based on their proposed strategy (diffusion centrality), and for the weather insurance setting of Cai et al. (2015) with $s = 5$, random seeding with 1 additional seed performs within 95% of their prescribed strategy (eigenvector centrality).[8] Additional numbers are reported in Table 1.

8.2 The virtue of randomness

While simulations presented in this section are aligned with our theoretical findings, they also shed light on an important point: When the number of available seeds is not too small, random seeding can perform *better* than centrality-guided seeding heuristics. The intuition here is that centrality-guided seeding heuristics pick redundant agents, who are likely to be part of the connected core of the network. Seeding those individuals has

[8] For microfinance diffusion, for instance, we measure the expected diffusion of seeding s top degree-central agents, seed $s+x$ agents randomly, and measure the expected diffusion for $x \geq 0$ up to the point that we find some x for which the latter performs within a desired range of the former. Python code for computing such measures can be found on the authors' websites.

Extra seeds required by random to beat 95% of proposed heuristics				
Model	s (Number of seeds)	x (Extra seeds needed)	CENTRAL(s)	RAND(s+x)
Microfinance	5	3	165	159
Microfinance	10	1	175	169
Weather	2	2	12	13
Weather	5	1	20	19

Table 1: Calculating the statistic of extra seeds required by random to beat a network-guided heuristic for the Microfinance network of Banerjee et al. (2013) and the weather insurance network of Cai et al. (2015).

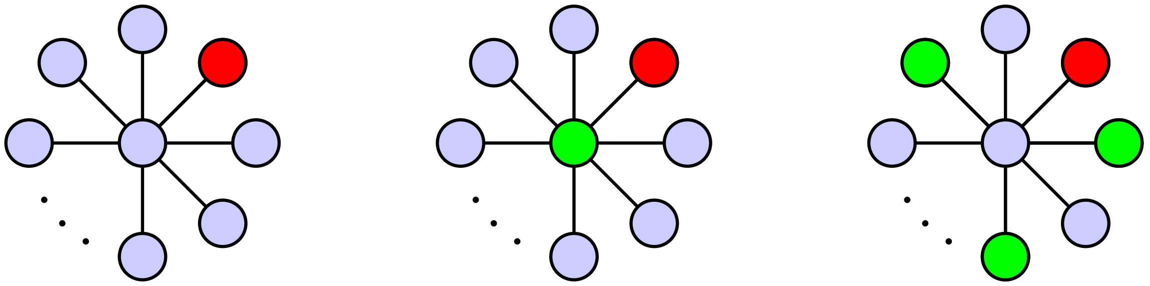
a decreasing marginal value. As the number of seeds increases, seeding an additional individual in the big component becomes less valuable than seeding individuals in the small components (that are disconnected from the big component). Random seeding, on the other hand, performs better because it is more likely to seed individuals in the small components as well.

8.3 Value of network targeting in vaccination

We will now show that network information can be highly valuable when a policymaker wishes to *halt* the spread of some diffusion. This is a relevant point for the diffusion of fake news (or an infections), where a policy-maker wants to inform individuals that the news is fake (or to vaccinate them) so that they stop spreading it. To fix ideas, suppose some random individual is infected with a disease, and the diffusion process is the diffusion model studied in this paper. A policymaker seeks to 'vaccinate' a group of individuals to *minimize* the extent of the diffusion. It is known that it is important to pick the optimal individuals for vaccination (Bollobás and Riordan, 2004; Drakopoulos et al., 2016). In fact, we conjecture that the number of additional individuals that we need in order for random vaccination to beat the optimum can be as large as a constant fraction of all agents. In a star network, for instance, vaccinating the central agent would fully stop the diffusion, while random needs to vaccinate most agents to be likely to stop the diffusion. (For a concrete example, see Figure 6.) Therefore, while we show that in accelerating diffusion, network information can be of low value, in 'vaccinating' individuals to hinder the diffusion, network information can be highly valuable.

8.4 On Asymptotics

One may question the relevance of the asymptotic results in this section to small networks, such as those in village network studies. As we have already pointed out, while $\omega(1)$ and $log(n)$ grow with n, these limits should not be read literally. The key lesson here is how



(a) No vaccination

(b) Optimal vaccination of 1 agent

(c) Random vaccination of several agents

Figure 6: Unlike in diffusion maximization, random strategy with a few additional individuals can perform poorly when the goal is to 'vaccinate' individuals to halt the diffusion. Consider a star network with n leaves, for some large n. Suppose some random individual gets infected with some disease (the red node), and any infected node infects its neighbors with probability $c = 0.5$. The goal is to vaccinate a single individual to minimize diffusion. 6(a): Without vaccination, the central node will be infected with probability 0.5, and thus $\frac{n}{4}$ of agents get infected in expectation. 6(b): Optimal vaccination picks the central node, which stops the diffusion completely. 6(c): Randomly vaccinating $x = o(n)$ individuals will *not* pick the central node with high probability. Thus, the central node will be infected with probability 0.5, and nearly $\frac{(n-x)}{4}$ of agents will be infected in expectation.

quickly additional seeds help random take over the optimum. Precisely at which point these limit results kick in (e.g., for what value of n, random with $\log(n)$ additional seeds beats omniscient?) is a question we answer through simulations.

Moreover, while there are several ways of expressing our theorems for large graphs, we have attempted to state the results in a manner most reflective of the relevant trade-offs in small networks. For example, the statement of Theorem 4 can be strengthened and combined with Theorem 5 to say that for large enough graphs, RAND with $s + log(n)$ seeds (as opposed to $o(log(n))s$ seeds) has the same eventual diffusion *and* speed of diffusion[9] as an omniscient seeding strategy with s seeds, since $\frac{s+log(n)}{o(log(n))s}$ will eventually be greater than one. However, this result will only emerge in exponentially large graphs, so stating a theorem in this manner is not practically relevant. The number of additional seeds needed for a random strategy to be competitive in terms of speed will turn out to grow multiplicatively in s, but as our results suggest, this multiple may be small in theory, and yet smaller in simulations.

On the other hand, one may measure the success of a seeding strategy by how small is the fraction of uninformed individuals to the population of the network. In this case, the relevant measure of value of network information is the ratio of the 'loss' of omniscient seeding to that of random seeding with $\omega(1)$ additional seeds. It is readily seen that

[9] Recall this means that for a fixed t, in a large enough network, the amount of diffusion that OMN achieves with s seeds in t rounds of communication will be no more than what RAND achieves in t rounds with $s + log(n)$ seeds.

this ratio converges to 1 in the size of the graph for any of the models studied above. Importantly, this does not depend on the communication probability being sufficiently large, as do the earlier theorems. In particular, when random seeding with additional seeds informs a vanishing fraction of individuals in the network, omniscient will do no better.

9 Concluding Remarks

Our theoretical bounds for omniscient seeding provide generous bounds for any practical network-based seeding strategy. We assumed that the omniscient strategy 'knows who communicates with whom' *a priori*. We also assumed that there are no network measurement errors and that individuals report their relationships truthfully. In practice, of course, policymakers do not have access to the *true* communication network. Moreover, we described the value of network information as the difference in diffusion between optimum and random seeding. But even with full network data, there is still an NP-hard problem to be solved to execute the optimum seeding algorithm. Studying the optimum as opposed to the omniscient or analyzing various seeding strategies when networks are imperfectly observed remain open questions.

It would be remiss not to restate that our results should not be read as an unqualified endorsement for expanding outreach as opposed to network targeting. To the contrary, our analysis suggests that network targeting could be valuable under several circumstances. Nevertheless, the current paper shows that the economic value of network information is necessarily context-dependent. In particular, we show that there are practically plausible conditions under which network position of seeds is not a first-order concern. Whether those conditions are satisfied in a specific context is an inherently empirical question. Much remains to be done to quantify the value of network information in other environments.

References

Alvarez, F. E., Buera, F. J., and Lucas Jr, R. E. (2013). Idea flows, economic growth, and trade. Technical report, National Bureau of Economic Research.

Asadpour, A. and Nazerzadeh, H. (2015). Maximizing stochastic monotone submodular functions. *Management Science*, 62(8):2374–2391.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144):1236498.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2014). Gossip: Identifying central individuals in a social network. Technical report, National Bureau of Economic Research.

Banerjee, A. V., Breza, E., Chandrasekhar, A. G., and Golub, B. (2018). When less is more: Experimental evidence on information delivery during india's demonetization.

Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509.

Beaman, L., BenYishay, A., Magruder, J., and Mobarak, A. M. (2015). Can network theory based targeting increase technology adoption. *Unpublished Manuscript*.

Bloch, F., Demange, G., and Kranton, R. (2014). Rumors and social networks.

Bloch, F., Jackson, M. O., and Tebaldi, P. (2016). Centrality measures in networks.

Bollobás, B. and Riordan, O. (2004). Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., and Pan, M. (2017). Using aggregated relational data to feasibly identify network structure without network data. *arXiv preprint arXiv:1703.04157*.

Bulow, J. and Klemperer, P. (1994). Auctions vs. negotiations. Technical report, National Bureau of Economic Research.

Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., and Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, 391(4):1777–1787.

Chen, W., Lin, T., Tan, Z., Zhao, M., and Zhou, X. (2016). Robust influence maximization. *arXiv preprint arXiv:1601.06551*.

Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145.

Conley, T. and Udry, C. (2010). Learning about a new technology: Pineapple in Ghana. *The American Economic Review*, 100(1):35–69.

Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM.

Drakopoulos, K., Ozdaglar, A., and Tsitsiklis, J. N. (2016). When is a network epidemic hard to eliminate? *Mathematics of Operations Research*, 42(1):1–14.

Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics*, 118(3):815–842.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Galeotti, A., Golub, B., and Goyal, S. (2017). Targeting interventions in networks. *arXiv preprint arXiv:1710.06026*.

Galeotti, A. and Goyal, S. (2009). Influencing the influencers: a theory of strategic diffusion. *The RAND Journal of Economics*, 40(3):509–532.

Galeotti, A. and Goyal, S. (2010). The law of the few. *American Economic Review*, 100(4):1468–92.

Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223.

Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2011). A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment*, 5(1):73–84.

Goyal, S., Heidari, H., and Kearns, M. (2014). Competitive contagion in networks. *Games and Economic Behavior*.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443.

Hartline, J. D. and Roughgarden, T. (2009). Simple versus optimal mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 225–234. ACM.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Jackson, M. O. and Rogers, B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *The American economic review*, 97(3):890–915.

Jackson, M. O. and Storms, E. C. (2017). Behavioral communities and the atomic structure of networks. *Available at SSRN: https://ssrn.com/abstract=3049748*.

Karp, R. M. (1990). The transitive closure of a random digraph. *Random Structures & Algorithms*, 1(1):73–93.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., and Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153.

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. *arXiv preprint arXiv:1001.5285*.

Lazarsfeld, P. F., Berelson, B., and Gaudet, H. (1948). The peoples choice: how the voter makes up his mind in a presidential campaign.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5.

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data.

Lim, Y., Ozdaglar, A., and Teytelboym, A. (2015). A simple model of cascades in networks. Technical report, mimeo.

Liu-Thompkins, Y. (2012). Seeding viral content. *Journal of Advertising Research*, 52(4):465–478.

Mobius, M., Phan, T., and Szeidl, A. (2015). Treasure hunt: Social learning in the field. Technical report, National Bureau of Economic Research.

Morris, S. (2000). Contagion. *Review of Economic Studies*, 67 (1):57–78.

Perla, J. and Tonetti, C. (2014). Equilibrium imitation and growth. *Journal of Political Economy*, 122(1):52–76.

Rice, E. (2010). The positive role of social networks and social networking technology in the condom-using behaviors of homeless young people. *Public health reports*, 125(4):588–595.

Rice, E., Tulbert, E., Cederbaum, J., Barman Adhikari, A., and Milburn, N. G. (2012). Mobilizing homeless youth for hiv prevention: a social network analysis of the acceptability of a face-to-face and online social networking intervention. *Health education research*, 27(2):226–236.

Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM.

Sadler, E. D. (2018). Diffusion games. *Working paper.*

Van Der Hofstad, R. (2016). Random graphs and complex networks.

Watts, D. J. and Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458.

Wilder, B., Yadav, A., Immorlica, N., Rice, E., and Tambe, M. (2017). Uncharted but not uninfluenced: Influence maximization with an uncertain network. In *Proceedings of AAMAS, 2017*, pages 1305–1313.

Yadav, A., Chan, H., Xin Jiang, A., Xu, H., Rice, E., and Tambe, M. (2016). Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty. In *Proceedings of AAMAS 2016*, pages 740–748.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review.*

A Proof of Theorem 1

In this appendix, we prove Theorem 1.

Let us start with a lemma on the performance of RAND and OMN on the communication graph $\mathcal{K}(G)$ for an arbitrary G. We will be using this lemma multiple times:

Lemma 1. *Let $\mathcal{K} = \mathcal{K}(G)$ denote the communication graph of a given graph G. Denote by CC the number of connected components of $\mathcal{K}$, and $\mathcal{C}_i$ the size of the i'th largest component in $\mathcal{K}$. Then,*

$$\boldsymbol{h}(G, s, OMN) = E[\sum_{i=1}^{\min\{s,CC\}} \mathcal{C}_i] \tag{2}$$

and

$$\boldsymbol{h}(G, s, RAND) = E[\sum_{i=1}^{cc} \mathcal{C}_i(1 - (1 - \frac{\mathcal{C}_i}{n})^s)] \tag{3}$$

Proof. The proof immediately follows the observation that in the SIR model a node becomes informed, if and only if one of the nodes in its connected components in $\mathcal{K}$ is seeded. In order to see equation (2), note that OMN maximizes the spread of the diffusion by informing one agent from each of the largest s connected components. Equation (3) captures the fact that the random policy hits a component with probability proportional to its size. □

Proof of Theorem 1. If $cd > 1$, by the standard phase transition result for Erdős-Rényi random graphs Van Der Hofstad (2016), there exists an $\alpha \in (0, 1]$ such that with high probability $\mathcal{C}_1 \geq \alpha n$ and $\mathcal{C}_i \in O(\log(n))$ for all $2 \leq i \leq CC$. By the above lemma, $\mathbf{h}(G, s, \text{OMN}) \leq \alpha n + s\mathcal{C}_2 = \alpha n + o(n)$, where the last equality uses the assumption that $s = o(n/\log n)$.

On the other hand, the probability that a node in the largest component is randomly seeded is at least $(1 - (1 - \alpha)^{\omega(1)+k})$ which goes to 1 as n goes to infinity, implying that $\mathbf{h}(G, s + \omega(1), \text{RAND})$ is at least αn. Therefore, $\mathbf{H}(\text{RAND}, \omega(1) + s)/\mathbf{H}(\text{OMN}, s) \geq \alpha n/(\alpha n + o(n))$, which is equal to 1 in the limit if $\omega(1)$ is $o(n)$ (the interesting case) and weakly greater otherwise.

When $cd < 1$, then even $C_1 \in O(log(n))$, so $\mathbf{H}(\text{OMN}, s) \in log(n)o(\frac{n}{log(n)}) = o(n)$, which shows the second part of the theorem. □

B Clustering and Speed of Diffusion

This appendix shows that the main insight goes through for a model of networks with clustering (Theorem 5). We will subsequently prove Theorem 4.

B.1 Higher clustering: k-level random networks

We will now consider a network formation model that allows for higher clustering coefficients. While there are many ways of generating such networks, we opt for a new model that is reminiscent of the one in Jackson and Rogers (2007). As in their model, nodes meet each other randomly at first and then make a few random friendships with the neighbors of their initial neighbors, which for the reasons given in Jackson and Rogers (2007), can be thought of as a natural model of how clustered relationships arise. To that end, we define a k-level random network to be an Erdős-Rényi network layered with additional random links to friend of friends, friend of friend of friends, and so on.

Definition 2 (k-Level Random Network)**.** *Let $\phi = (\lambda, q_1, \ldots, q_k) \in [0,1]^{k+1}$. A k-level network on n nodes, denoted $L_n(\phi)$, is constructed by drawing a graph X_n from $ER(n, \lambda)$ and including for every node, a link with one of its neighbors of neighbors with probability $1 - \sqrt{1-q_1}$, a link with one of its neighbors of a neighbor of a neighbor with probability $1 - \sqrt{1-q_2}$ and so on up to k.*

An Erdős-Rényi network is a special case of a k-level random network for $q_1 = \cdots = q_k = 0$, while other values of q_i allow for higher clustering coefficient. We will refer to X_n in the definition of k-level random graphs as the *base random graph* and $\lambda \cdot n$ as the *base-level average degree.*

Theorem 5. *Consider a k-level random network with base-level average degree d. Let c be the probability that an informed node speaks to a given neighbor and $s \in \mathbb{N}$. If $cd > 1$, then for any super-constant $x(n)$,*

$$\lim_{n\to\infty} \frac{\mathbf{H}(\text{RAND}, s + x(n))}{\mathbf{H}(\text{OMN}, s)} \geq 1.$$

This result suggests that the presence of high clustering coefficient does not hinder the performance of random seeding with a few more seeds than omniscient. To show that these asymptotic are relevant in finite networks, we simulate random and omniscient seedings in k-level random networks and compare their performances in Appendix G.[10]

To show theorem 5, we need the following additional lemma:

Lemma 2. *Fix $n \in \mathbb{N}$, $q_1, \ldots, q_k \in [0,1]$, and let $\phi_n = (\frac{p}{n}, q_1, \ldots, q_k)$. Let $L_n(\phi_n)$ be a k-level random network and let $K_n(\phi_n) = L_n(\phi_n) \cap Z_n$ where Z_n is independently drawn from $ER(n, c)$. Finally let $\mathcal{C}_i(G)$ denote the ith largest component of a network G. Then if $p \cdot c > 1$, there exists some $\alpha \in (0,1]$ such that with high probability $|\mathcal{C}_1(K_n(\phi_n))| \geq \alpha n$.*

[10] Moreover, studying the case when $cd \leq 1$ or when s is not constant is not very analytically tractable in our model of random graph with clustering. Nevertheless, in Appendix H we conduct simulations that suggest that even under omniscient seeding, diffusion is vanishingly small when d or c are sufficiently small. We also provide simulations that suggest that s can also grow at rate $o(n/\log(n))$.

Proof. Let X_n be the base random graph of $L_n = L_n(\phi_n)$. The law of $X_n \cap Z_n$ is the distribution of $ER(n, pc)$. If $pc > 1$, by the standard phase transition result for Erdős-Rényi random graphs, there exists an $\alpha \in (0, 1]$ such that with high probability $|\mathcal{C}_1(X_n \cap Z_n)| \geq \alpha \cdot n$. Since every edge in X_n is also present in L_n, $\mathcal{C}_1(X_n \cap Z_n) \subseteq \mathcal{C}_1(L_n \cap Z_n)$, implying that with high probability, $|\mathcal{C}_1(L_n \cap Z_n)| \geq n$.

□

Proof of Theorem 3. Fix $q_1, \dots, q_k \in [0, 1]$, and let $\phi_n = (\frac{d}{n}, q_1, \dots, q_k)$ and fix $s \in \mathbb{N}$. Let $L_n(\phi_n)$ denote a k-level random network on n nodes.

Case 1: $|\mathcal{C}_2| = o(n)$. By Lemma 1, there is a sequence $\{\alpha_i^{\mathcal{C}_1}\}_{i=1}^{\infty}$ bounded below by some $\alpha^{\mathcal{C}_1} \in (0, 1]$ such that, w.h.p., the sum of the $s \in \mathbb{N}$ largest component sizes in $\mathcal{K}_n(L_n(\phi_n))$ is $\alpha_n^{\mathcal{C}_1} \cdot n + (s-1)o(n) = \alpha_n^{\mathcal{C}_1} n + o(n)$. By Lemma 2, this is the expected diffusion achieved by OMN.

Let P_n be the probability that the largest component in $L_n(\phi_n)$ exceeds $\alpha^{\mathcal{C}_1} n$ in size. We know $P_n \to 1$ as $n \to \infty$. Therefore probability that a node in the largest component is randomly seeded is at least $P_n(1 - (1 - \alpha^{\mathcal{C}_1})^{\omega(1)+k}) \to 1$ as $n \to \infty$. Therefore, $\mathbf{H}(\text{RAND}, \omega(1) + s)/\mathbf{H}(\text{OMN}, s) = (1 - o(1))\alpha_n^{\mathcal{C}_1} n/(\alpha_n^{\mathcal{C}_1} n + o(n)) \to 1$ as $n \to \infty$.

Case 2: $|\mathcal{C}_2|$ *is not* $o(n)$ *but* $|\mathcal{C}_3| = o(n)$. Let D be the limiting distribution of $\frac{|\mathcal{C}_2|}{n}$ (which will have support contained in $[0, \alpha^{\mathcal{C}_1}]$). If $X \sim D$, the expected diffusion of of OMN is $(\alpha_n^{\mathcal{C}_1} + E[X]) \cdot n + o(n)$. On the other hand, for any $\alpha^{|\mathcal{C}_2|} \in (0, \alpha^{|J_1|}]$, the probability that both the largest component and the second largest component are seeded, conditional on $|\mathcal{C}_2| > \alpha^{|\mathcal{C}_2|}$, approaches one in probability as the number of seeds increases. So for any $\alpha^{|\mathcal{C}_2|}$, $\mathbf{H}(\text{RAND}, \omega(1) + s)/\mathbf{H}(\text{OMN}, s) \geq (1 - o(1))(\alpha_n^{\mathcal{C}_1} + E[X\mathbf{1}_{\frac{|\mathcal{C}_2|}{n} > \alpha^{\mathcal{C}_2}}])n/(\alpha_n^{\mathcal{C}_1} + E[X])n + o(n)$ as $n \to \infty$. Since this holds for all $\alpha^{\mathcal{C}_2} \in (0, \alpha^{\mathcal{C}_1}]$, the expression converges to 1.

$|\mathcal{C}_i|$ *is not* $o(n)$ *but* $|\mathcal{C}_{i+1}| = o(n)$, *for* $i \leq s$. This case follows analogously to *Case* 2.

□

B.2 Proof of Theorem 4

We now show a stronger result that encompasses the original statement of the Appendix B.2.

Theorem 6. *Consider a k-level random network and a bounded diffusion process that ends in $T \geq 1$ periods and let s be a non-negative integer. Then,* $\mathbf{H}(\text{RAND}, o(\log(n))s) \geq \mathbf{H}(\text{OMN}, s)$ *for n sufficiently large.*

Proof of Theorem 6. We want to show that $\mathbf{H}_T^c(\text{RAND}, sf(n)) \geq \mathbf{H}_T^c(\text{OMN}, s)$ for n sufficiently large, where $f(n) = o(log(n))$. By the correspondence between the existence of edges and the diffusion process described in the proof of Lemma 1, it will suffice to show that the largest T neighborhood of $\mathcal{K}(L_n(\phi))$ is of size $o(log(n))$.

Since for every realization of the $L_n(\phi)$ and its communication network, the k neighborhood of a node in $\mathcal{K}(L_n(\phi))$ is smaller than the corresponding k neighborhood in $L_n(\phi)$, it suffices to show the latter is $o(log(n))$ in size. We start by showing this holds for $ER(n, \frac{p}{n})$, the base random graph.

Let $Bin(n, \frac{p}{n}, 1)$ denote a binomial distribution with n draws and success probability $\frac{p}{n}$. We say $Y_i \sim Bin(n, \frac{p}{n}, t)$ if $Y_i = \sum_{i=0}^{Z_i} X_i$, where $Z_i \sim Bin(n, \frac{p}{n}, 1)$ and $X_i \sim Bin(n, \frac{p}{n}, t-1)$ with all variables being independently distributed.

The following lemma can be derived from classic results on branching processes (see for example Theorem 3.2 of Van Der Hofstad (2016)) but we include the proof for completeness.

Lemma 3. *For all $t \in \mathbb{N}$ and $\lambda \in \mathbb{R}$, there is a $C_{\lambda,t} > 0$ such that for any positive integer n and $B_t \sim Bin(n, \frac{p}{n}, t)$, $E[e^{\lambda B_t}] < C_{\lambda,t}$.*

Proof. Using the formula for the moment generating function of a binomial distribution, $E[e^{\lambda B_1}] = (1 - \frac{p}{n}(e^{\lambda} - 1))^n \to e^{p(e^{\lambda}-1)}$ as $n \to \infty$. This shows that the statement is true for the base case $t = 1$.

Now suppose for any n and λ, $E[e^{\lambda B_t}]$ is bounded from above by $C_{\lambda,t}$. Then

$$\begin{aligned} E[e^{\lambda B_{t+1}}] &= E[e^{\lambda \sum_{i=0}^{B_1} B_t^i}] \\ &= E[\prod_{i=0}^{B_1} e^{\lambda B_t^i}] \\ &\leq E[\prod_{i=0}^{B_1} C_{\lambda,t}] \\ &= E[e^{log(C_{\lambda,t})B_1}], \end{aligned}$$

where the step before the last follows from the law of iterated expectations and the inductive hypothesis. The last term is bounded from the above by the base case. □

For a given graph, let $N_t(i)$ be the set of vertices distance t from node i. It is easy to see that $N_t(i)$ is first order stochastically dominated by B_t; this fact along with the above lemma are used to prove the following:

Lemma 4. *For any node $i \in ER(n, \frac{p}{n})$ and $c > 0$, $Pr(|N_t(i)| \geq c\log(n)) = o(\frac{1}{n})$.*

Proof. By the previous lemma, let $C_{\lambda,t} > 0$ be such that for any n, $C_{\lambda,t} > E[e^{\lambda B_t}]$. Now:

$$\begin{aligned} Pr(N_t(i) \geq c\ \log(n)) &= Pr(B_t \geq c\ \log(n)) \\ &\leq Pr(e^{\lambda B_t} \geq n^{\lambda c}) \\ &\leq \frac{E[e^{\lambda B_t}]}{n^{\lambda c}} \\ &= \frac{C_{\lambda,t}}{n^{\lambda c}}. \end{aligned}$$

$C_{\lambda,t}$ is independent of n, so the above is $o(1/n)$ as long as $\lambda c > 1$. □

The above lemma shows for any c, the probability that a given node has a t-neighborhood of size exceeding $c \log(n)$ is $o(1/n)$. By union bound, the probability that a vertex has t-neighborhood of size exceeding $c \log(n)$ is $o(1)$. Therefore, with high probability, the largest t-neighborhood in $ER(n, \frac{p}{n})$ is smaller than $c \log(n)$ for any c.

Finally, note that for a k-level random network with a base $ER(n, \frac{p}{n})$ random graph, the maximum size of a t-neighborhood is no more than the maximum size of tk- neighborhoods in the base random graph [11]. Therefore, the largest $t-$neighborhood in $L_n(\phi)$ is of size $o(\log(n))$ as well.

□

C Proof of Theorem 2

In this section, we prove Theorem 2. Let $CL_n(w)$ be the power-law Chung-Lu network with scale parameter b and minimum expected degree d i.e., $w_i = [1 - F]^{-1}(i/n)$, where $F(x) = 1 - (\frac{d}{x})^b$. By Lemma 1 and the same arguments used in the proof of Theorem 1, it suffices to show that $\mathcal{K}(CL_n(w))$ has a linear sized giant component and $O(log(n))$ sized smaller components.

Let us use the notation $\wedge$ for taking the minimum of two numbers. The first observation is that the probability that nodes i and j are connected in $\mathcal{K}(CL_n(w))$ is

$$cp_{ij} \wedge c = c\frac{w_i w_j}{\sum_k w_k} \wedge c = \frac{(cw_i)(cw_j)}{\sum_k cw_k} \wedge c = \frac{w'_i w'_j}{\sum_k w'_k} \wedge c,$$

where $w'_i = cw_i$. Second, $[1 - F]^{-1}(x) = \frac{d}{x^{1/b}}$, so $w'_i = cw_i = \frac{cd}{(i/n)^{1/b}}$.

Consider the case when $b > 2$. We will show that $\mathcal{K}(CL_n(w))$ is also a power-law network with scale parameter b and minimum expected degree cd. This is because

[11] For example, if the shortest path between two nodes has 10 links in the base random graph, then in a 5-level random graph layered on top of this, it may be that the first node is connected to the 6th node and the 6th node to the 11th, which is distance minimizing scenario. This means a 2-neighborhood of the 5 level random graph is at most a 10 neighborhood of the base random graph.

for large enough n, $\frac{w'_i w'_j}{\sum_k w'_k} < c$, so the probability two nodes i and j are connected is $\frac{w'_i w'_j}{\sum_k w'_k} \wedge c = \frac{w'_i w'_j}{\sum_k w'_k} \wedge 1$. Equivalently, $w'_i = [1 - F']^{-1}(i/n)$, where $F'(x) = 1 - (\frac{cd}{x})^{1/b}$.

Define the random variable W'_n to denotes the weight of a node selected uniformly at random from the first n nodes and define F'_n to be its CDF.

Lemma 5. *Let $b > 2$ and $cd > (b-1)(b-2)$. The following conditions are satisfied.*

(C1) *There exists some $W' \sim F'$ such that $W'_n \to W'$ in distribution*

(C2) *$E[W'_n] \to E[W'] > 0$ as $n \to \infty$*

(C3) *$\frac{E[{W'}^2]}{E[W']} > 1$*

Proof. Note that when the CDF of W is given by F as defined above, its mean and variance exist when $b > 2$ and are given by $\frac{bd}{b-1}$ and $\frac{bd^2}{(b-1)^2(b-2)}$ respectively. Therefore, $E[{W'}^2]/E[W'] = \frac{d}{(b-1)(b-1)}$, so the parametric assumptions ensure that **C3** holds. Moreover, Van Der Hofstad (2016) show in Exercise 6.6 that conditions **C1** and **C2** hold for our choice of weights when $E[{W'}^2] < \infty$. □

When condition C1, C2, and C3 hold, Theorem 9.2 in Van Der Hofstad (2016) implies that there exists an $\alpha > 0$ such that with high probability $|\mathcal{C}(\mathcal{K}(CL_n(w)))|/n \to \alpha$ as $n \to \infty$. Therefore, this result along with the previous lemma gives us:

Corollary 1. *When $b > 2$ and $cd > (b-1)(b-2)$, there exists an $\alpha > 0$ such that with high probability $|\mathcal{C}_1(\mathcal{K}(CL_n(w)))|/n \to \alpha$ as $n \to \infty$.*

Note when $b \in (1, 2)$, $E[{W'}^2] = \infty$ (so condition C2 does not immediately follow) and for all large enough n there exist i such that $\frac{{w'_i}^2}{\sum_k w'_i} > c$, so the probability that i and j are linked is $\frac{{w'_i}^2}{\sum_k w'_i} \wedge c \neq \frac{{w'_i}^2}{\sum_k w'_i} \wedge 1$. Therefore, the aforementioned results of Van Der Hofstad (2016) are not applicable. We take an alternative and perhaps more illustrative route to show the existence of a linear sized giant component in this case.

Lemma 6. *Let $b \in (1, 2)$ and $d > 0$. Then there exists an $\alpha > 0$ such that with high probability, $|\mathcal{C}_1(\mathcal{K}(CL_n(w)))|/n \in (\alpha, 1]$ as $n \to \infty$.*

Proof. We will use a coupling argument. As before, let $w' = \{w'_i\}$ denote the sequence of weights with $F'(x) = 1 - (cd/x)^b$ as their cumulative distribution function. Choose ϵ small enough such that $cd > (1+\epsilon)(\epsilon)$ and let $\bar{b} = 2 + \epsilon$. Let $\bar{w} = \{\bar{w}_i\}$ denote the sequence of weights using $\bar{F}(x) = 1 - (cd/x)^{\bar{b}}$ as the cumulative distribution function.

The crucial observation is that w' dominates $\bar{w}$, i.e., for any n and for any i, $w'_i > \bar{w}_i$. Now for every n, we can couple the random graph $\bar{G}_n$ generated when weights are given by $\{\bar{w}_i\}$ with the random graph G'_n generated when weights are given by $\{w'_i\}$ by coupling the edges one by one, so that $\bar{G}_n$ is a subgraph of G'_n.

We know by the earlier case that there is an $\alpha > 0$ such that with high probability, the largest connected component in G'_n is of linear size. Therefore, G'_n also has a connected component of size linear in n. □

Now it remains to show that the second largest component of $\mathcal{K}(CL_n(w))$ is of size $O(log(n))$. Again, the case $b > 2$ and $cd > (b-1)(b-2)$ follows directly from Van Der Hofstad (2016) (see Exercise 9.40).

For case $b \in (1, 2)$ we take advantage of the notion of kernel of a random graph family. We refer the reader to the section 9.5 of Van Der Hofstad (2016) for the relevant definitions. For our purpose, we define our kernel function $\kappa(x, y) = \frac{[1-F']^{-1}(x)[1-F']^{-1}(y)}{\frac{1}{n}\sum_k w'_k} \wedge c$. It follows from the definition of kernel that such a kernel function is *graphical* and *irreducible*. Now, we only need to show that

(C4) $\frac{c|\{ij: \frac{w'_i w'_j}{\sum_k w'_k} > c\}|}{\sum_{i>j} \frac{w'_i w'_j}{\sum_k w'_k} \wedge c} \to 0$ as $n \to \infty$

(C5) $\inf_{x,y,n} \frac{[1-F']^{-1}(x)[1-F']^{-1}(y)}{\frac{1}{n}\sum_k w'_k} > 0$

C4 ensures that a vanishing fraction of potential edges in $CL_n(w')$ are ensured to exist with probability 1 (along such edges, the probability of the edge existing in the communication graph is c) [12].

Lemma 7. *When $b \in (1, 2)$, C4 and C5 are satisfied.*

Proof. We will start by showing C4. Recall W'_n is the weight of a randomly selected node among the first n nodes in $CL_n(w')$. The number of nodes i (among the first n) for which $\frac{{w'_i}^2}{\sum_k w_k} > c$ is

$$n\Pr({W'_n}^2 > cnE[W'_n]) \le \frac{E[{W'_n}^2]}{cE[W'_n]} \le ([1-F']^{-1}(1/n))^2 C_1 = (cdn^b)^2 C_1,$$

where the first inequality follows from Markov's inequality, and C_1 is a constant independent of n.[13] That is to say, the number of such nodes is $O(n^{2b})$ and therefore $o(n)$. Denote S as the set of the remaining $\Theta(n)$ nodes. By construction, for all nodes i and j in S, $C_2 \equiv \frac{(cd)^2}{nE[W']} < \frac{w'_i w'_j}{\sum_k w'_k} < c$, for all large enough n. Now note that all edges for which $\frac{{w'_i}^2}{\sum_k w'_k} > c$ must either be among nodes in $[n] - S$ or between nodes in $[n] - S$

[12] In particular, this result combined with the fact that a sequence of Chung-Lu graphs with edge probabilities given by $\frac{w'_i w'_j}{\sum_{k=1}^n w'_k} \wedge 1$ is *graphical* and *irreducible* implies the same for a sequence of graphs with edge probabilities $\frac{w'_i w'_j}{\sum_{k=1}^n w'_k} \wedge c$ (see chapter 9 of Van Der Hofstad (2016) for definitions).

[13] Here, we used the fact that the sample average of weights converge to something finite. To see this, note that $\sum_{i=0}^{n-1}(1-i/n)^{-1/b}\frac{1}{n} < \int_0^{\frac{n-1}{n}}(1-x/n)^{-1/b}dx = \frac{b}{1-b}(1-x)^{1-1/b} + C_2|_{x=\frac{n-1}{n}} \to C_2$ as $n \to \infty$, since $b > 1$

and S. But since $|[n] - S| = o(n)$, there are at most $o(n^2)$ such edges. Therefore, $\frac{c|\{ij: \frac{w'_i w'_j}{\sum_k w'_k} > c\}|}{\sum_{i>j} \frac{w'_i w'_j}{\sum_k w'_k} \wedge c} = o(n^2)/\Theta(n^2) \to 0$ as $n \to \infty$.

For any x, y, note $\frac{[1-F']^{-1}(x)[1-F']^{-1}(y)}{\frac{1}{n}\sum_k w'_k} \leq \frac{(cd)^2}{\frac{1}{n}\sum_k w'_k}$. Now $b \in (1,2)$ means $\frac{1}{n}\sum_k w'_k \to E[W'] < \infty$. C5 immediately follows. □

With this, Theorem 9.33 of Van Der Hofstad (2016) ensures that when $b \in (1,2)$, any non-giant component is $O(log(n))$ in size, completing the proof of Theorem 2.

D Directed Networks and Communication: Proof of Theorem 3

Consider a model of directed networks similar to Erdős-Renyi: $D(n,p)$ is a random directed network on n nodes in which directed edge (i,j) is drawn with probability $\frac{p}{n}$. In this setting, OMN observes a realization of the directed communication network and chooses the best nodes to seed using this information. A *strongly connected component* is a subgraph for which there exists a directed path between any two member nodes. A relevant concept for directed graphs is that of a strongly connected *giant component*, which is a strongly connected component containing a linear fraction of the nodes, asymptotically. We will follow the arguments of Karp (1990) to show Theorem 3.

Proof of Theorem 3. First we note three facts from Karp (1990).

1. Under the condition $cp > 1$, there exists a strongly connected giant component (s.g.c.) with high probability.

2. If asymptotically, the s.g.c. contains Θn nodes, then $pc(1-\Theta) < 1$.

3. Let $f(n)$ be any superconstant that is also $o(\sqrt{n})$, and let $R(v)$ be the vertices reachable from any node v through some path. Then there exists a $B > 0$ such that with high probability, $|R(v)| \in [0, B\log(n)] \cup [\Theta n - f(n)\sqrt{n}, \Theta n + f(n)\sqrt{n}]$.

From fact 3, we know that $H(\text{RAND}, s + x(n))$ gets at least $\Theta n - f(n)\sqrt{n}$ nodes, whereas a single omnisciently chosen seed may reach up to $\Theta n + f(n)\sqrt{n}$ nodes. The difference of $2f(n)\sqrt{n}$ is irrelevant for our result on the convergence of the performance ratio to 1. However, with $s = \sqrt{n}$ initial seeds, it is theoretically possible that OMN collects sufficiently many $\sqrt{n}$ sized clusters of nodes that have paths leading to the s.g.c. but are not reachable (due to the directed nature of communication) from nodes in the s.g.c. This raises the possibility that OMN reaches as many as $(\Theta + \mu)n$ nodes where $\mu \in (0, 1-\Theta)$. This would overturn the ratio result, and so it remains to show that this will not happen,

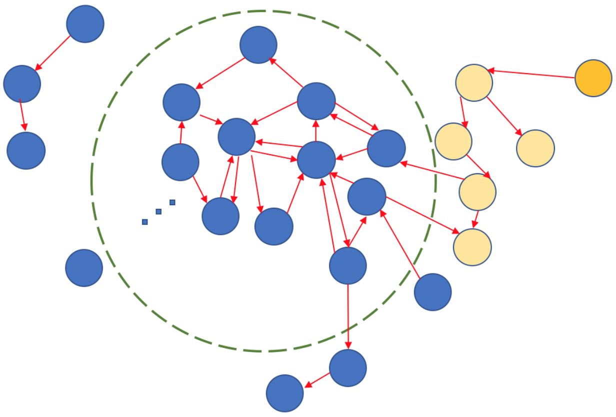


Figure 7: Above is an example communication network when communication is directed. The outgoing edges represent the nodes that a given node would inform if given information. The nodes within the dotted dashed circle represent the strongly connected giant component. If any node is informed within the s.g.c., all nodes in the s.g.c. become informed. Random seeding with enough seeds will land a seed in the s.g.c. with sufficiently high probability. The orange nodes, if informed, also disseminate information to the s.g.c. In particular, OMN might choose to seed the dark orange node, given a single seed (and there could be many such useful entry points, though only one set of orange nodes is pictured above). In the proof of 3, we want to show that the size of the set of any cluster of orange nodes is $o(\log(n))$ so that OMN cannot significantly outperform RAND.

with high probability, and that a propitiously chosen seed can reach at most $O(log(n))$ more nodes than a randomly chosen seed which lands in the s.g.c (see D for clarification).

More precisely, let C be the set of nodes reachable from any vertex in the strongly connected component. We want to show that with high probability, for every vertex v, $|R(v) - C| = \mathrm{O(log\ n)}$. If v is in C, we are done, so suppose $v \notin C$. If V is the set of nodes in the graph, it suffices to show that there are at most $O(log(n))$ nodes in $V - C$ for which there exists a path from v entirely consisting of nodes not in C.

To see this, consider the subgraph consisting of only nodes in $V - C$. The probability of communication between any two nodes is at most pc, and $|V - C|$ is at most $(1 - \Theta)n + f(n)\sqrt{n}$ by fact 3. By fact 2, there exists an $\epsilon > 0$ such that $pc(1 - \Theta + \epsilon) \equiv pc\Theta' < 1$. Therefore, the number of neighbors of a given node (within the subgraph in consideration) is asymptotically dominated by $Bin(\Theta' n, pc)$. Using the Poisson approximation to the binomial distribution, a standard result on bounding the population of a Galton-Watson branching process, and the Chernoff bound, we get:

$$Pr(|R(v)| > k) \leq e^{-k(t - pc\Theta'(e^t - 1))}$$

for t of our choice. Since $pc\Theta' < 1$, t can be chosen small enough such that $-k(t - pc\Theta'(e^t - 1))$ is strictly negative. When $k = B\log(n)$, for large enough B, we can apply the union bound and show that $Pr(|R_1| > k)$ is vanishing, where $R_1 = max_{v \in V-C}|R(v)|$. □

An alternate model is one in which the original graph is undirected, but communication is directed. This is not altogether a superficial change from the $D(n,p)$ model. In particular, the probability that i communicate with j is correlated to the probability j communicates with i, since communication is only possible if an edge existed between the two nodes in the first place (in $D(n,p)$, the directed edges exist with independent probabilities, so there is no such correlation). In such a model, it can be shown that a result analogous to 3 holds by symmetric arguments.

E Simulations of microfinance diffusion model

Banerjee et al. (2013) study the following diffusion model: There is a piece of information being spread about a program. Agents are in one of three states with respect to knowledge of and participation into the program: uninformed, informed non-participants, and informed participants. Each agent is a node in the network. Each period, every informed, non-participating agent communicates information about the program with each of his direct neighbors with an independent probability q_N. Similarly, each informed participant communicates information about the program with each of his direct neighbors with an independent probability $q_P \geq q_N$. The interpretation is that participants are more likely to talk about the program than non-participants. All communication ceases after T periods. For small T, this can be thought of as a crude way of imposing the fact that people eventually stop talking about the program (although a model in which each informed individual stops talking about the program T periods from the date she was first informed better suits this interpretation). Upon becoming informed about the program, a node makes an irrevocable decision to adopt with probability p. In the case where $q_N = q_P$ and $T = \infty$, the previous model becomes an instance of the SIR model with $k = \infty$. In the case where $k = 1$, this is the independent cascade model Kempe et al. (2003). The objective function for this diffusion process can be defined to be either the expected number of nodes which are informed or the number of nodes which participate–the authors of the microfinance paper use the latter measure.

To keep the focus on the model of diffusion , we simply model acceptance probabilities as being constant across all nodes without taking into consideration demographics. This gives the cleanest comparison between the seeding strategies based on two notions of centrality. In the simulations, we use the probability of adoption of 0.24, which is the observed in sample probability of adoption among initial seeds when this study was carried out. In two different estimates, the authors of the microfinance study estimated that participants spread information with probability 0.35 while non-participants spread information with probability 0.05. In another specification, these parameters were found to be 0.45 and 0.1 respectively. Appendix E shows the results of simulations for both sets of parameter estimates. We include simulations for the sparser kerosene and rice

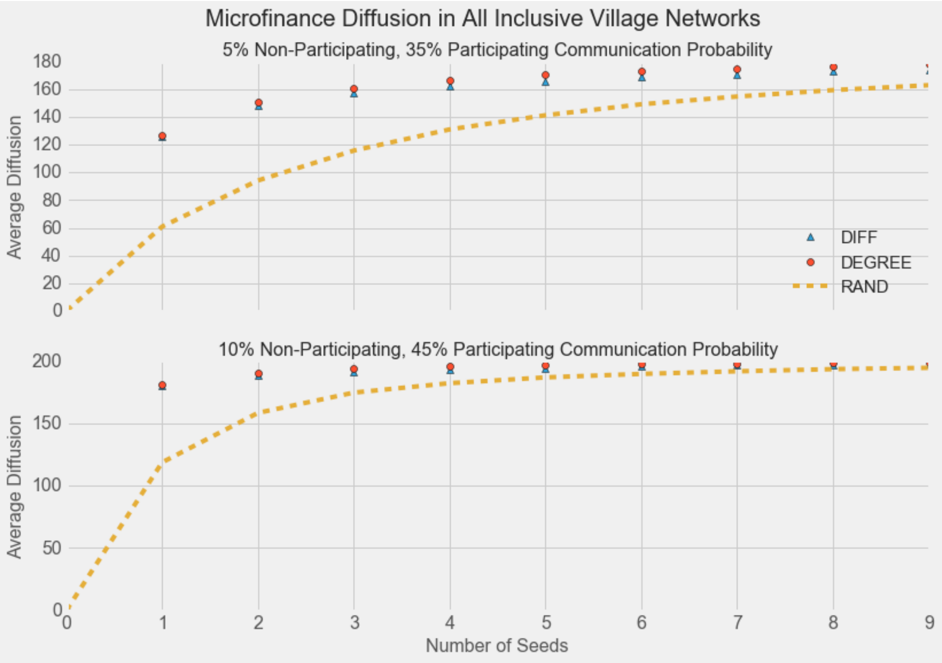


Figure 8: This is an analogue of Figure 2 with the diffusion process specified in Banerjee et al. (2013) rather than the model studied in this paper. As the number of seeds increases, random seeding performs as well as the centrality-guided seedings.

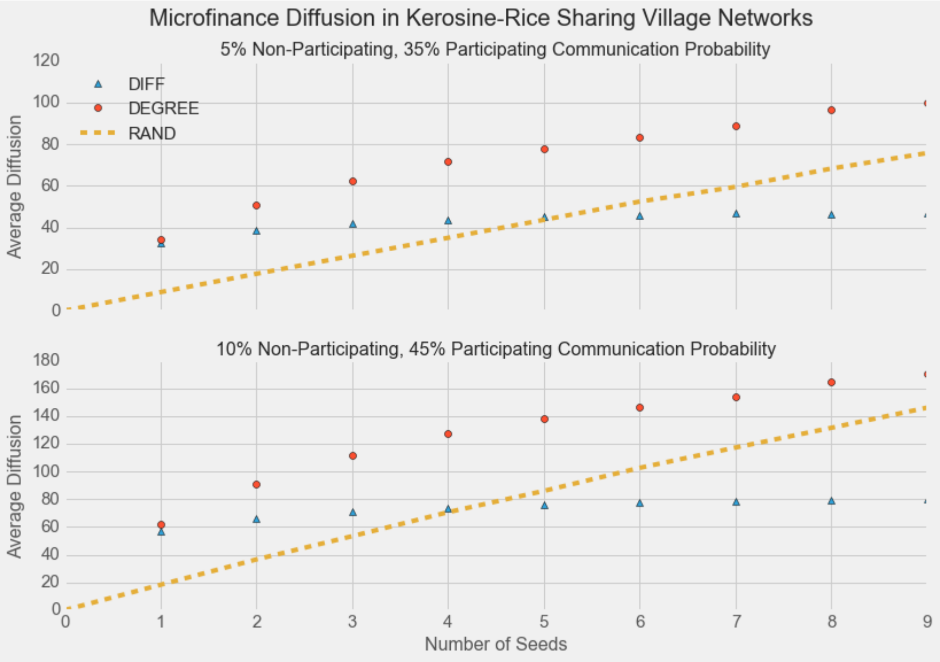


Figure 9: Random seeding performs well relative to the other seeding strategies. Moreover, it performs better than the seeding guided by the diffusion centrality when the number of seeds is more than 5.

borrowing network in Figure E.

For simulations of section 6, we use the same model and data, and vary T between 1 to 4. We conduct simulations on all village networks and take average among them to calculate the extra number of seeds needed.

F Simulations of weather insurance diffusion model

In this section, we will evaluate the benefit of targeting in the setting studied by Cai et al. (2015). The authors study diffusion of a new government offered weather insurance take-up by rice farmers across various villages in China. To understand spill-over effects in information and take-up decisions, the authors randomly choose injection points for simple and intensive information sessions about the program. A social network survey ask participants to list their 5 closest friends, yielding networks in which nodes have close to identical out-degree, barring some instances of under reporting [14]. They find that an important channel through which take-up happens is by learning about the program from friends. On the other hand, the purchase decisions of neighbors is not so relevant to a farmer's own decision, conditional on learning about the program. Finally, intensive sessions are more effective than simple sessions in generating uptake.

The authors show these effects in reduced form regressions and without explicitly laying out a model of diffusion. They find that if a strongly-linked [15] neighbor of an untreated node learns about the program, this increases the chance of adoption for the untreated node by 7.5%. If a weakly linked neighbor learns the same, the probability of adoption goes up by 6%.

Since the authors do not explicitly describe a model of diffusion, we make some assumptions about the process to interpret their results in back-of-the-envelope simulations. We assume that the probability of adoption for untreated nodes who hear about the program from their friends is 35%, the same as the treatment effect of the simple program. This along with the coefficient of the regressions of fraction of informed friends on uptake give us a 17% probability of communication occurring along a weak link and a 21% probability of communication occurring along a strong link in any given period. Since the channel of diffusion is information, we assume communication occurs each period with the aforementioned probabilities (unlike our model in which communication ceases for a node after a single period). Finally, we assume communication happens only two periods, since only two rounds were studied in Cai et al. (2015). Note these are conservative assumptions in that they stack the performance of careful seeding algorithms

[14] The authors find that even without an explicit constraint on the number of reported friends, most survey participants list 5 friends anyway.

[15] Two nodes i and j in a directed network are strongly linked if edges (i, j) and (j, i) are present in the network. In the present setting, this means both farmers listed each other as friends in the survey.

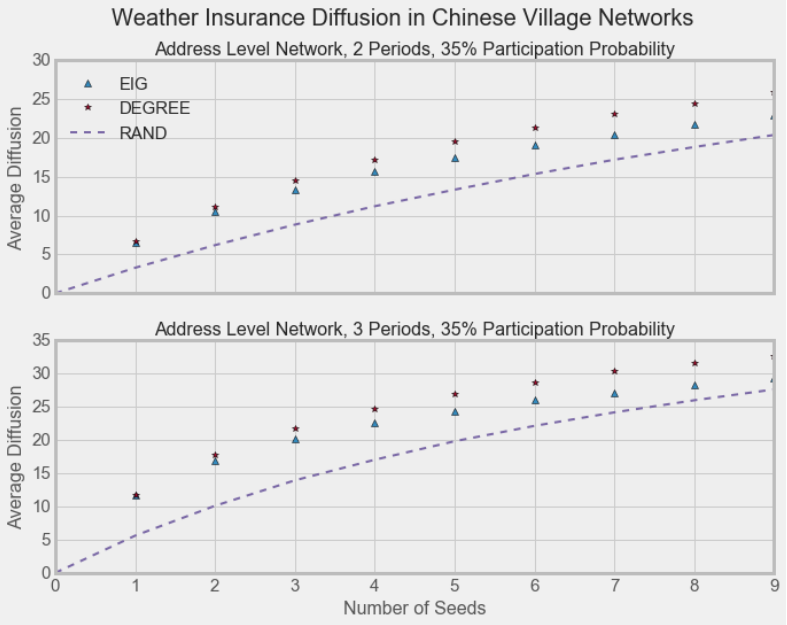


Figure 10: DEGREE seeding refers to seeding those with the highest degree, considering the undirected version of the village network. RAND seeding only chooses out of those villagers who participated in the social network survey, though they may name individuals who have not been surveyed as neighbors. Finally EIG refers to eigenvector centrality seeding. Note the average network size is 50 farmers.

against RAND—the latter, for example, does better when the assumed diffusion process is unbounded.

We compare random seeding to degree seeding and seeding based on eigenvector centrality[16], two measures of centrality the authors suggest for targeting. Since all nodes more or less report the same number of friends, variation in degree mostly arises from variation in the number of friends that named the node in question as a friends. The authors find that under a permissive specification, central nodes do not wield additional influence over a given neighbor than less central counterparts. Therefore, in our simulations, the benefit of seeding central nodes arises purely from their connection to more immediate neighbors and paths to other nodes. The results of our simulations show again in a different network and setting that the presence of network effects and positive association between centrality and diffusion does not immediately imply that carefully targeting nodes will make a large difference. Indeed one of the striking findings in Cai et al. (2015) is that social learning is a powerful vehicle of information transmission–strong enough that a policymaker may safely ignore minutiae of network structure.

[16] This is defined by the eigenvector of the largest eigenvalue of the adjacency matrix, ignoring direction of edges.

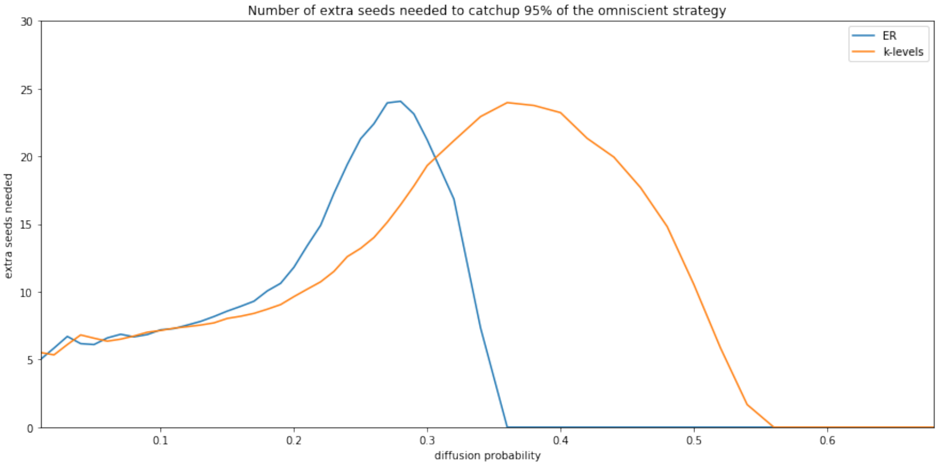


Figure 11: Extra number of seeds needed by ER and k-level graphs. x-axis is the diffusion probability and y-axis is the extra seeds needed.

G Extra Seeds Needed in Random Finite Networks

In this section we show that the number of extra seeds needed for the random seeding strategy to get close to the omniscient seeding can be comparable for k-level and ER random graphs.

We run simulations on two kinds of networks on $n = 1000$ nodes: An Erdős-Rényi random graphs with parameter $d = 5$, and a k-level random graph with $d = 2.4$, and $q_1 = 0.1$ and $q_2 = 0.05$. The parameters for the k-level graph are chosen so that the average degree of a node is roughly 5, so comparisons with ER graphs are on more even footing. We assume the omniscient has access to $s = 10$ seeds.

Figure 11 shows the simulation results. When the diffusion probability is large enough (so that the condition of the theorem is satisfied), the extra number of seeds required by random to get to 95% of the omniscient goes down quickly. When $c > 0.38$, random without *any* additional seeds performs as well as 95% of omniscient.

When the diffusion probability is small, our theoretical results are silent about the performance of random relative to the omniscient, though we noted that the total diffusion is vanishingly small in either case. Still, Figure 11 indicates that except for the interval around the 'phase transition' diffusion probability, the additional seeds required for random seeding is small (note also that the maximum number of additional seeds required would fall as average degree rises). Qualitatively, similar results go through for k-level random networks, though the fall in the number of seeds required to catch up with omniscient seeding seems slower. The latter fact can be explained by the fact that the base ER random graph is sparser, so the giant component of the k-level graph is smaller and more difficult for random seeding to target. If we raise the degree of its base ER graph but adjust the remaining parameters in a way that average degree is still 5, the latter graph will better resemble that of the ER catchup plot.

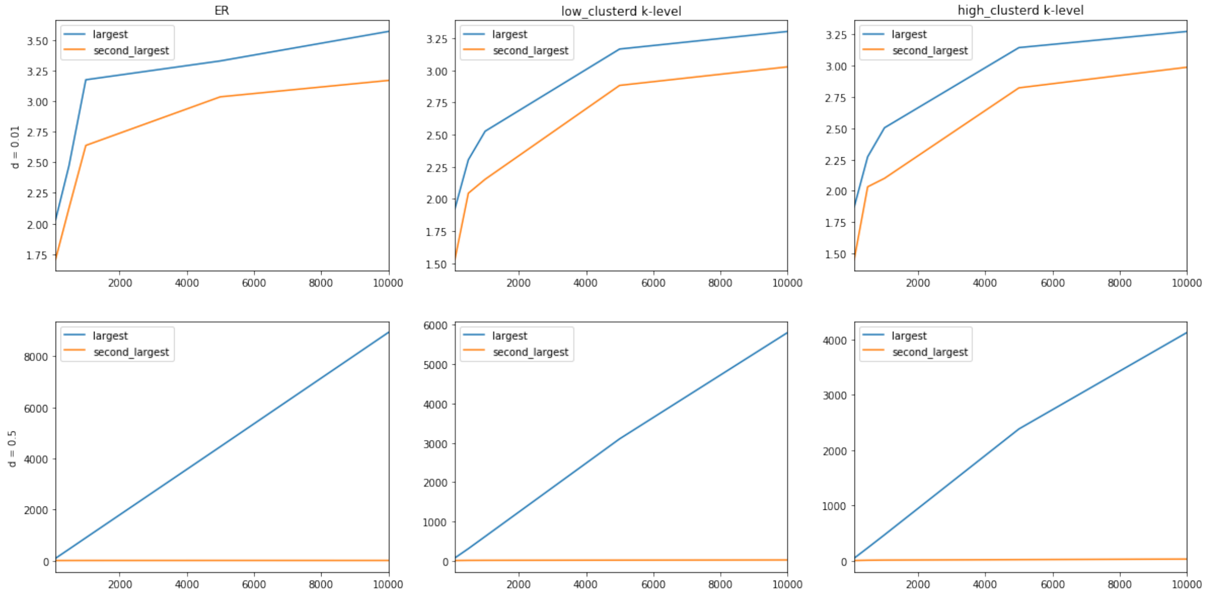


Figure 12: Size of largest and second largest components in ER and k-level graphs in both regimes. The x- axis is the size of the network and y-axis is the sizes of components.

H Component Sizes in ER and k-Level Graphs

The top row of Figure 12 shows that for both ER and k-level random graphs, when we are in the regime that the communication network is very sparse (hence the diffusion will be unsuccessful), the sizes of the largest and second largest components of the networks are very small essentially for all network sizes. For percolated k-level graphs, proving that component sizes are order $log(n)$ is analytically challenging. Simulations, however, indicate that a similar result is true for such graphs

Figure 12 also shows that in the regime where ER and k-level graphs have a giant component, the smaller component are $O(\log(n))$ in size. While Theorem 5 keeps s fixed, these simulations suggest that using similar arguments as in the proof of Theorem 1, one can perhaps let s belong the class $o(\frac{n}{log(n)})$.