Just a Few Seeds More: The Inflated Value of Network Data for Diffusion[*]

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

First Draft: September 2017
This Draft: September 2023

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 diffusion of information, spread of microfinance programs, and adoption of 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 the classic independent cascade model of diffusion, either picking a few more seeds at random can prompt a larger diffusion than optimal seeding, or even optimal seeding does not prompt a significant diffusion. We prove this result for large classes of random networks, and verify them for several real-world networks.

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

JEL classification codes: D85, D83, O12, Z13

[*] We thank Daron Acemoglu, Nava Ashraf, Francis Bloch, Stephane Bonhomme, Arun Chandrasekhar, Raj Chetty, Darrell Duffie, Pascaline Dupas, Matthew Elliot, Drew Fudenberg, Andrea Galeotti, Ben Golub, Sanjeev Goyal, Zoe Hitzig, Matthew Jackson, Emir Kamenica, Scott Kominers, Shengwu Li, Yucheng Liang, Greg MacNamara, Erik Madsen, Mihai Manea, Ilya Morozov, Michael Ostrovsky, David Pearce, Debraj Ray, Peter Reiss, Phil Reny, Evan Sadler, Andy Skrzypacz, Alex Teytelboym, Christopher Tonetti, Carlos Varjao, as well as several other seminar participants, for valuable comments. We also thank several anonymous referees for invaluable feedback. We thank Jing Cai for helping us navigate their dataset. We also thank Maxime Rivet and Mahdi Shahrabi for coding some of the simulations reported here.

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

[‡] Managerial Economics and Decision Sciences Department, Kellogg Business School. suraj.malladi@kellogg.northwestern.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 contexts like brand awareness (Richardson and Domingos, 2002), microfinance (Banerjee et al., 2013), and agricultural technologies (Beaman et al., 2021). Since this problem is known to be computationally intractable (Kempe et al., 2003), a large body of theoretical and empirical studies introduced various heuristics for ranking candidate individuals to target. But implementing such heuristics requires knowledge of the network structure, which may not be readily available or may be difficult to attain.[1] Studies such as Banerjee et al. (2019) or Breza et al. (2020) develop methods for identifying central nodes or approximating the network structure without conducting a thorough census.

By contrast, our goal here is *not* to identify the central individuals or propose methods for acquiring network information but to quantify the value of doing so. Our main contribution is to recast the benefit of following a network-guided seeding heuristic in terms of the extra seeds required for a seeding strategy that ignores network information to perform just as well. We address our question theoretically and investigate how our results are borne out in simulations on real-world network data.

We consider a population of n individuals (or nodes) who are connected to each other through a random graph.[2] Individuals are either informed or uninformed about some technology. Information spreads according to a variant of the Susceptible-Infected-Recovered (SIR) diffusion model. At time $t = 0$, all individuals (nodes) outside a small group (seeds) selected by the policymaker are initially uninformed. Once informed at time t, a node has one chance to speak to 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 process continues until no new individual becomes informed.

To quantify the value of targeted seeding, we consider the following thought experiment: Suppose in one setting, the policymaker has access to full network data and unlimited computational power to optimally pick s individuals as initial seeds. In the second setting, the policymaker observes nothing about the network and picks $s + x$ initial seeds uniformly at random. For what value of x will random seeding inform as many individuals, in expectation, as the optimal seeding?[3]

In fact, we compare random seeding to a *better than optimal* strategy. Suppose, in ad-

[1] Breza et al. (2020) estimate that conducting network surveys in 120 Indian villages would cost approximately \$190,000 and take over eight months.

[2] The model is general and subsumes Erdős-Rényi graphs, but for other parameters, also produces graphs exhibiting homophily or power-law degree distributions.

[3] This thought experiment mirrors the famous comparison in Bulow and Klemperer (1996), who ask 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.

dition to network information, the policymaker has a perfect forecast of who would share information with whom and seeds the best s individuals given this information. Comparing such ‘omniscient’ seeding with random seeding gives a generous upper bound for the value of network-guided seeding, because the omniscient strategy performs better than the optimal one, which itself performs better than computationally feasible heuristics.

Our main theorem shows that under one set of conditions (the *viral* regime), the difference in the expected fraction of informed individuals between the random seeding strategy with $s+x$ seeds and the omniscient strategy with s seeds vanishes exponentially in x. Precisely when those conditions fail, even omniscient seeding produces an expected diffusion that reaches only a vanishing fraction of individuals.

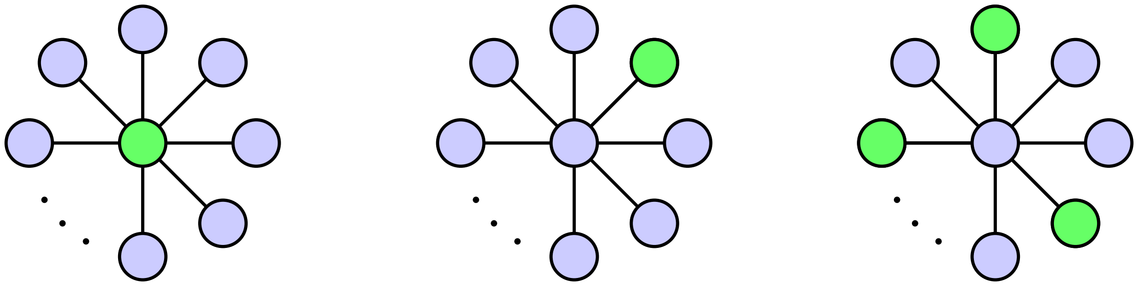
This theorem holds for the general *Inhomogeneous Random Networks* (IRN) model, which subsumes several well-known random network formation models as its special cases. This means that our result readily applies to simple Erdős-Rényi graphs (where any pair of nodes is connected with the same probability), networks with *homophily* (where nodes are more intensely connected to nodes with “similar” types), and networks with *power-law degree distribution* (where some individuals are connected to a large fraction of the population).

In particular, it may be surprising random seeding can work well even on networks with highly unequal degree distributions, where it may seem that informing central nodes is important. The explanation is that random seeding is likely to seed connections of those highly central nodes, precisely because they are highly connected. Therefore, central individuals will become informed through their neighbors and broadcast information throughout the network. Figure 1 provides a simple example for this intuition.

After presenting our asymptotic results, we turn to the question of whether similar findings hold for small, real-world networks. To this end, we simulate heuristic strategies and random seeding for the basic SIR model on network data, such as the Indian village household networks of Banerjee et al. (2013). Here, for instance, we find that if nodes pass information to 10% of their neighbors on average, random seeding with $s+3$ seeds outperforms a host of centrality-based seeding strategies with s seeds, for any s. As expected, the number of additional seeds required for random seeding to catch up decreases in the communication probability.

We next explore whether analogues of our main theorem hold when we consider variance in diffusion size as the objective function. This is an important consideration for a risk-averse planner. A careful targeting strategy may guarantee some baseline level of diffusion, while random outreach strategies risk fizzling out. Still, we show that the variance of diffusion from random seeding goes to zero at an exponential rate in the number of seeds.

On the other hand, if the policymaker cares about the diffusion achieved in the first few periods, the differences between random and omniscient strategies can be large. To



(a) Optimal seeding

(b) Random seeding

(c) Random seeding with additional seeds

Figure 1: 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.

see why, consider again the example of Figure 1. If the objective is to maximize diffusion only in the first period, the optimal strategy seeds the center and reaches in expectation $\frac{n}{2}$ nodes. Random with $s + x$ seeds can only reach $s + x + 1$ nodes in one period (seeds and possibly the center). One period is not enough time for the random to outperform the optimum. More generally, informing highly central nodes through their neighbors requires more time than diffusing by informing them directly. Therefore a concern for fast diffusion tilts the balance in support of targeted seeding.

Finally, we investigate the robustness of our results to some variations to the model of diffusion. We simulate the more complex models estimated in the development economics literature, which depart from the SIR model in different ways. For example, for a model of diffusion estimated in Banerjee et al. (2013), random seeding with 11 seeds performs nearly as well as central seeding with 10. With the estimated diffusion model and the farmer social networks in Cai et al. (2015), random seeding with 6 seeds performs nearly as well as central seeding with 5.

Still, there are many diffusion models for which targeting central nodes can be valuable. Suppose central individuals communicate to their connections with a higher probability than others or listen to fewer friends than they speak to. Then, it can help to target such individuals. This might explain why some policymakers and firms are willing to pay to target central individuals. Alternatively, if the product diffusion follows a threshold model, where an agent is informed if sufficiently many of his neighbors are informed, random seeding, even with many seeds, may perform poorly.

Regardless of what is the true underlying diffusion model or network structure, our framework suggests that the "extra number of seeds required for random seeding strategy to reach $(1-\delta)\%$ of of a prescribed network-based heuristic" is a useful statistic for diffusion and centrality studies to report. This number can be interpreted as the economic value of careful targeting in a given setting. As an example, for $\delta = 0.05$ and for Banerjee et al. (2013) and Cai et al. (2015), this statistic is smaller than 3.

Related literature. A large, cross-disciplinary literature focuses on developing algorithms for influence maximization over networks. Watts and Dodds (2007) are one of the first to dispute the value of targeting 'influencers' in networks. They compare through simulations, and for a variety of diffusion models, the size and dynamics of diffusion achieved by seeding high degree nodes to low degree nodes and find that seeding highly central nodes may not be as important. We complement this work by asking a different question, quantifying the value of targeted seeding in terms of the extra seeds needed, and theoretically identifying conditions under which this number is small. In work that models a monopolist facing a network with word-of-mouth communication about her product, Campbell (2013) also notes the high value of randomized awareness for a small seed set. Complementing this work, we bound value of network information in terms of the extra seeds required for random seeding to outperform optimal seeding, noting that the catch-up is slow precisely when the achievable diffusion is small.

Influence maximization in networks has more recently received attention in development economics, following Banerjee et al. (2013).[4] They find that centrality of initial seeds is strongly correlated with total eventual participation into a microfinance program, and this correlation is not explained simply by the degree or demographic characteristics of these nodes. Cai et al. (2015) also conduct a randomized experiment in which they seed certain individuals in Chinese villages with information about a weather insurance program and observe how take-up rates among neighbors vary with centrality of the seeds. Beaman et al. (2021) study technological adoption by farmers as they vary seeding rules over 200 independent village-networks in Malawi in an experimental setting. These papers find that network theoretic seeding improves diffusion, whereas we are interested in studying how quickly additional outreach makes up for network-agnostic seeding.

Two particularly relevant papers conduct multiple RCTs to compare random seeding to other targeting methods. Kim et al. (2015) compares random seeding with seeding a nominated friend of a random individual and an individual with the most number of ties. They find that targeting nominated friends increased adoption of the nutritional intervention by 12.2% compared to random targeting. Banerjee et al. (2019) similarly

[4] For an older theoretical literature, see Galeotti and Goyal (2009); Young (2009); Goyal et al. (2014); Lim et al. (2015); Mobius et al. (2015); Sadler (2020); Bloch (2016); Galeotti et al. (2020); Bloch et al. (2018) .

compare seeding based on identifying "gossips" and seeding "trusted" individuals with seeding randomly, all with six seeds. In their first experiment, they received on average 8.1 phone calls in villages with random seeding, and 11.7 in villages with gossip seeding. In a second experiment studying vaccination rates, they find that with random seeds, 18.11 children attended and received at least one shot, compared to 23 in villages with gossip-based seeding.

Our results indeed suggest that, *fixing the number of seeds*, random seeding may perform worse than other heuristics. Neither Kim et al. (2015) nor Banerjee et al. (2019) test how many extra seeds random needs to compete with other heuristics, although the gap between random and gossip seeding (3.6 individuals in the first RCT and 4.89 in second) suggests that a few more random seeds can indeed help random catch up.[5]

Throughout the paper, we mainly focus on random seeding—which requires *no* network data—and omniscient seeding—which requires *full* network data. Of course, in practice, policy makers will always know something about the network. For instance, the policymaker may know the pattern of segregation by geographic proximity. In such cases, it seems reasonable that using this data is better than ignoring it, and thus there may be simple heuristics that are more cost-efficient than random with a few extra seeds. In subsequent work and complement to our results, Sadler (2022) considers a policymaker who can collect information about types (e.g., men vs. women) and, equipped with this information, identifies the best seeds.

Organization of the paper. We introduce our diffusion and network models in Section 2. In Section 3, we present our main theorem and simulations on real-world networks. In Section 4, we study the robustness and limitations of the results, considering alternative objectives and diffusion models. In Section 5, we discuss our results and conclude.

2 Model

The set of *agents* or *nodes* are denoted by $N = \{1, 2, \cdots, n\}$. Agents are connected in a *network* represented by a simple graph $G = (N, E)$, where E is the set of unordered pairs of agents. If $\{i, j\} \in E$, agent i and agent j are *neighbors*. A node's *degree* in G is the number of its neighbors and $|E(G)|$ denotes the number of edges in network G.

[5] Banerjee et al. (2019) consider another experiment in which SMS blast reminders are sent to 33% and 66% of village households selected at random. They find that the SMS blasts do worse than targeted seeding. However, individuals in gossip-guided seeding were contacted by phone and given regular personalized reminders rather than SMS blasts. Given the incomparable modes of communication and the magnitude of coverage, this do not cleanly test whether random seeding with a few extra seeds may perform well. Indeed, as Banerjee et al. (2023) find, broadcasting information to a large group of people may drastically change the dynamics of information acquisition.

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. 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. We highlight that c is a key parameter of the model, as we will be interested in characterizing how diffusion patterns change when c is low versus high.

An alternative description of the diffusion process is each link of the social network is maintained probability c and dropped otherwise. The graph with the set of maintained links is the *communication network*, denoted as $\mathcal{K}(G) \subseteq G$. The communication network represents which pairs of agents 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. In Section 4.2, we discuss how results extend to settings with directed communication and discuss alternative diffusion models.[6]

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 $f : \mathcal{U}_n \times [n] \to \Delta 2^N$, with the property that $f(G, s)$ is supported on sets with cardinality s.

We say seeding strategy f is *feasible* if for all networks $G = (N, E) \in \mathcal{U}_n$ and $s \leq |N| = n$, $f(G, s)$ and $\mathcal{K}(G)$ are independent. That is, the event that a particular set of nodes are seeds and the event that a particular communication graph is realized are independent. A seeding strategy that does not satisfy this property uses information on who would speak to whom in determining the choice of seeds. A policymaker with no knowledge beyond the network structure can only use feasible seeding strategies. Let $\mathcal{F}$ be the space of feasible seeding strategies for graphs on n nodes.

Performance of seeding strategies. Let $A_t(G, s, f) \subseteq N$ denote the (random) set of informed nodes at time $1 \leq t \leq T$ for a given network G, number of seeds s and seeding strategy f. Let $\mathbf{h}(G, s, f) = \mathbb{E}[|A_T(G, s, f)|]$ be the expected number of informed agents at the end of the process.

We denote the optimal seeding strategy by OPT. For a given network G, this strategy

[6] In addition, simulations of Appendix C and Appendix D consider models of D(G) communication.

picks a set of s seeds that maximizes the expected diffusion:

$$\text{OPT}(G, s) \in \operatorname*{argmax}_{f \in \mathcal{F}} \mathbf{h}(G, s, f).$$

In our main results, we study how other seeding strategies compare to OPT on average for a 'typical network'. To that end, we define a network formation process as a probability distribution over all possible networks of size n. Fixing some network formation process, $\mathbb{P}_n$, let $\mathbf{H}(f, s) = \frac{1}{n}\mathbb{E}_{G \sim \mathbb{P}_n}[\mathbf{h}(G, s, f)]$. $\mathbf{H}$ returns the expected fraction of informed agents for a given seeding strategy and number of seeds, when the network G is drawn according to $\mathbb{P}_n$.[7] $\mathbf{H}$ serves as our measure of the performance of a seeding strategy.

Bounding the performance of OPT. Computing OPT requires knowing the precise realization of the network. Even with that information, it is known that computing this strategy is NP-hard (Kempe et al., 2003). Therefore, it is generally intractable to characterize the performance of OPT. Here we introduce two seeding strategies whose performances bound that of OPT.

Let RAND denote the *random* seeding strategy which picks s nodes uniformly at random in G. Let OMN denote the *omniscient* seeding strategy which for every realization of $\mathcal{K}(G)$ picks s initial seeds to maximize diffusion. OMN is not a feasible strategy.

Clearly, for any s we have:

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

2.1 Network Model: Inhomogeneous Random Networks

We now introduce the *inhomogeneous random networks* (IRN) model (Bollobás et al., 2007). The IRN model is a general network model that subsumes several random network models as special cases. In this model, there is a set of potential "types" and each agent has a specific type. Any two individuals are connected with some exogenously given probability that is a function of their types. We state our main theorem for a general version of this network model with a finite type space. Then we explore the consequences of this theorem by specializing the result to certain familiar instances of the IRN model.

Fix some $\mathcal{T} = \{1, 2, \cdots, \tau\}$ as the set of different *types* of agents. Let n_i denote the number of agents of type $i \in \mathcal{T}$. Define a *kernel* as any arbitrary symmetric function $\kappa : \mathcal{T}^2 \to (0, n]$, and let

$$p_{ij}(\kappa) = \frac{1}{n}\kappa(i, j).$$

[7] $\mathbf{H}$ depends on $\mathbb{P}_n$, though we suppress this dependence in the notation, as it will be clear from the context which network formation process we are assuming.

Let $\boldsymbol{p}(\kappa)$ be a matrix with p_{ij} as its elements. Then, $\text{IRN}_n(\boldsymbol{p}(\kappa))$ is a random network on n nodes, where an agent of type i is linked to an agent of type j with probability p_{ij}. Let κ_{ij} be the expected number of type j neighbors of an agent of type i. Let $\mathbf{T}_\kappa = [\kappa_{ij}]_{i,j\in[n]}$ be the *types matrix*. Since $\kappa(i,j) > 0$ for all i and j, $\mathbf{T}_\kappa$ is a positive matrix. Therefore by the Perron-Frobenius theorem, and letting $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{\mathcal{T}} x_i^2}$, the largest eigenvalue $||\mathbf{T}_\kappa||$ can be computed as

$$||\mathbf{T}_\kappa|| = \sup_{\mathbf{x}:||\mathbf{x}||_2\leq 1} ||\mathbf{T}_\kappa \mathbf{x}||_2.$$

The IRN model admits some classic network models as special cases:

Erdős-Rényi networks. In an Erdős-Rényi random network on n nodes, there is independently a link between a pair of agents (i,j) with probability d/n. The average degree for large N is approximately d.

Erdős-Rényi model is a special case of the IRN model we just described. $\mathcal{T}$ is a singleton type space, and $\kappa = d$, so $\mathbf{T}_\kappa = [d]$, and $||\mathbf{T}_\kappa|| = d$.

The Islands-connections networks and homophily. The islands-connections model of network formation starkly captures the feature that people are more likely to be connected to others of the same type—a feature that is referred to as *homophily* (see Jackson (2010), Chapter 6). The islands model is another special case of the IRN model. An $\text{IRN}_n(\boldsymbol{p}(\kappa))$ is an islands network with parameters (m, d_{in}, d_{out}), all positive, if (1) there are m types of agents and their sizes are n/m for all types, (2) for two agents i and j with the same type $p_{ij} = d_{in}/n$, and (3) for two agents i and j with different types $p_{ij} = d_{out}/n$. The matrix $\mathbf{T}_\kappa$ will have d_{in}/m in the diagonal entries and d_{out}/m elsewhere. Simple calculations show that $||\mathbf{T}_\kappa|| = \frac{1}{m}(d_{in} + (m-1)d_{out})$.

Chung-Lu networks and highly central nodes. Fix a sequence $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$ such that $\max_k(w_k^2) \leq \sum_k w_k$. A *Chung-Lu* network on n nodes, $CL(n, \mathbf{w})$, includes each edge $\{i,j\}$ independently with probability $p_{ij} = \frac{w_i w_j}{\sum_k w_k}$.[8] The Chung-Lu model fits into the framework of the IRNs but with a continuum of types (see Appendix A.2.1).

Simple calculation shows that the sequence of weights $\mathbf{w} = (w_1, \ldots, w_n)$ is the same as the sequence of expected node degrees. Thus, a Chung-Lu network can capture, for example, a power-law degree distribution by using a parametric power-law functional form for the weights. In particular, suppose that for all i,

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

[8] More generally, we could let $p_{ij} = \min(\frac{w_i w_j}{\sum_k w_k}, 1)$ and allow $\max_k(w_k^2) > \sum_k w_k$. But this variation is known to be equivalent asymptotically to the original (see Hofstad (2016), Section 6.6).

This generates a network on n nodes with minimal expected degree d. The scale parameter b determines the thickness of the right tail of the distribution F. As b grows, the tail becomes thinner. We say a distribution has a power-law tail if the mass of the cumulative distribution function lying to the right of some large enough k is proportional to $k^{-\tau}$. The degree distribution a Chung-Lu graph follows a power law for $\tau = b - 1$.

Although the IRN model is capable of capturing several classic network models as special cases, it falls short in representing many features of real-world networks. One such example is the inability to capture clustering, which refers to the tendency of nodes to be connected if they share a common neighbor. We address this issue by proving more general results (that include networks with clustering) in the Online Appendix, as well as a series of simulations on a variety of real-world social networks.

3 Main Theorem

To quantify the value of network information and optimization, we pose the following question: Fixing the number of seeds available to the omniscient seeding strategy, how many *additional seeds* are required in order for random seeding to perform as well as the omniscient? Let $\alpha = \lim_{n\to\infty} \mathbf{H}(\mathrm{OMN}, 1)$ be the fraction of nodes informed by the omniscient seeding with one seed.

Theorem 1. *Consider a sequence of random graphs,* $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$*. Let s be the number of seeds.*

Then, if $||\mathbf{T}_\kappa|| > 1/c$, random seeding catches up to the omniscient seeding at an exponential rate in the number of extra seeds, i.e., $\alpha > 0$ and for any x,

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

If $||\mathbf{T}_\kappa|| < 1/c$, then any seeding strategy diffuses to only a vanishing fraction of the population:

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

This theorem tells us that there are generically two states of the world. In one state, where $||\mathbf{T}_\kappa|| > 1/c$, random seeding catches up with OMN exponentially fast in the number of extra seeds. In the other, where $||\mathbf{T}_\kappa|| < 1/c$, even OMN cannot inform a non-negligible fraction of the population, at least in large networks.

A policymaker might be interested in knowing how many extra seeds are needed for RAND to achieve at least a fraction $1-\epsilon$ of the diffusion of some targeted seeding strategy. Our result gives an upper bound on the number of additional seeds required. By letting x be large enough so that $1 - (1-\alpha)^{s+x} > 1 - \epsilon$, in a sufficiently large network, having

$\max(0, \frac{\log(\epsilon)}{\log(1-\alpha)} - s)$ extra seeds guarantees that RAND's expected performance is withing $1 - \epsilon$ of OMN.

We now sketch the ideas of the proof. The formal proof can be found in Appendix A.1.

Proof overview of Theorem 1. Recall that the realization of the communication network $\mathcal{K}(G) \subseteq G$ determines the pairs of agents who will speak to each other upon becoming informed. The connected components of this communication network reveal the performance of RAND and OMN. 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 a method of computing the expected diffusion for each of the strategies for any distribution of component sizes for a communication network.

When n is sufficiently large and $||\mathbf{T}_\kappa|| > 1/c$, by the phase transition results of the IRN model (Bollobás et al., 2007), there exists a component in the communication network which contains a constant α fraction of the total population. Informing one node in that component is enough to inform a constant fraction of the population. The remaining components of the communication network, on the other hand, are vanishingly small (*i.e.*, $o(n)$) in population size.[9]

Therefore, the omniscient seeding strategy with only one seed informs the largest component. With additional seeds, it picks the largest of the small components, but the total fraction of additionally informed nodes is $o(n)(s-1)/n$, which is asymptotically a negligible fraction. Thus, $\lim_{n\to\infty} \mathbf{H}(\text{OMN}, s) = \lim_{n\to\infty} \mathbf{H}(\text{OMN}, 1) = \alpha$.

Similarly, as $n \to \infty$, the expected diffusion of random seeding with $s + x$ seeds is $\alpha(1 - (1-\alpha)^{s+x})$. This is because it is enough for one of the random seeds to hit the constant size component, while the other components are irrelevant. Therefore, the limit ratio of random with $s + x$ seeds and omniscient with s seeds is $1 - (1-\alpha)^{s+x}$, which goes to 0 as n grows.

When $||\mathbf{T}_\kappa|| < 1/c$, then size of even the largest component is $o(n)$, meaning that any informed agent only informs (directly or through a chain of messages passing) $o(n)$ other agents. Hence, the omniscient seeding strategy with s seeds can at most inform $o(n)s/n$ fraction of the population. □

We make a few remarks about the theorem and its interpretation.

[9] We use standard "big o" and "little o" notation: Let $f, g : \mathbb{N} \to \mathbb{R}$, with g being positive valued. We say $f \in O(g(n))$ if there exists some $B > 0$ such that $|f(n)| \leq Mg(n)$ for all sufficiently large n. We say $f \in o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

On asymptotic results. We investigates the applicability of asymptotic results to small networks and addresses this question through simulations. For example, we show that for an Erdős-Rényi network with 100 nodes and when $cd = 1.5$, random with 3 extra seeds performs better than omniscient with one seed. In Section 3.4, we show that similar results hold in simulations on several real-world networks, including the Indian village networks of Banerjee et al. (2013).

The growth rate of s. Theorem 1 holds for any constant number of seeds $s \leq n$ but can be extended to allow s to grow in n in the first part of the theorem. When $||\mathbf{T}_\kappa|| > 1/c$, the size of the second-largest component of any IRN is $O(\log(n))$, and so even if $s = o(\frac{n}{\log(n)})$, the first part of the theorem holds. This is because $o(\frac{n}{\log(n)}) \times O(\log(n))$ is $o(n)$, and thus the fraction of extra nodes reached by OMN (after reaching the giant component by the first seed) is asymptotically zero.

For the second part of the theorem too, we can relax the limitation on s for specific classes of IRNs. For instance, for Erdős-Rényi random networks, we know that even if $||\mathbf{T}_\kappa|| < 1/c$, all components are $O(\log(n))$-sized, and thus if $s = o(\frac{n}{\log(n)})$, the second part of the theorem holds.

When s grows even more quickly in n, Theorem 1 need not hold: OMN may diffuse to a larger fraction of nodes than RAND, both when $||\mathbf{T}_\kappa|| < 1/c$ and $||\mathbf{T}_\kappa|| > 1/c$. It is unclear that the same would hold true for OPT or any other feasible heuristic. This depends on whether such strategies can reliably discover the largest of the small components as OMN can.

Is one seed enough? One observation arising from our analysis is that, for a given network, OMN with one seed can diffuse to the same proportion of the network (i.e., the giant component) as OMN with multiple seeds. This finding may seem counterintuitive given the common practice of seeding multiple nodes in a network. However, we caution against interpreting this observation too literally for two reasons.

First, in reality, a policymaker cannot know in advance which nodes belong to the giant component. As such, there is a risk that a single-seed policy could fail to reach the largest component of the communication network. Consequently, policymakers may choose to hedge against this risk by seeding multiple nodes.

Secondly, there is a possibility that the chosen seeds may not adopt and spread the information, product or idea being propagated, even though our stylistic model assumes that they do so with probability 1. In the case of the Indian village networks studied in Banerjee et al. (2013), for instance, some of the seeds did not adopt microfinance. Incorporating non-adoption into our model would potentially overturn the conclusion that a single seed suffices for OMN, while leaving Theorem 1 unaffected.

3.1 Erdős-Rényi Networks

The next result follows immediately from Theorem 1 and the fact that $||\mathbf{T}_\kappa|| = d$ for Erdős-Rényi networks, as discussed in Section 2.1.

Corollary 1. *Consider an Erdős-Rényi network on n nodes with average degree d. If $dc > 1$, then for any s and x,*

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

If $dc \leq 1$, then $\alpha = 0$. Furthermore, for every $s > 0$,

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

For Erdős-Rényi networks, the $||\mathbf{T}_\kappa|| > 1/c$ condition translates into $dc > 1$. Intuitively, if this condition holds, then each informed individual (on average) talks to at least one of their friends. Under this condition, random seeding catches up with the omniscient seeding at an exponential rate. On the other hand, when $dc \leq 1$, the fraction of informed nodes even under the omniscient seeding strategy goes to zero as $n \to \infty$.

In addition, a known feature of Erdős-Rényi networks is that the (asymptotic) size of their giant component (when $cd > 1$) can be implicitly calculated by solving for $1 - \alpha = e^{-cd\alpha}$. Thus, we can easily calculate the rate by which the gap between random and OMN closes. For instance, if $cd = 1.5$, then $\alpha \simeq 0.58$. Thus, the performance of random with 5 seeds is roughly 99% of the omniscient with one (or more) seeds. If $cd = 2$, then $\alpha \simeq 0.8$. Thus, the performance of random with 3 seeds is roughly 99% of the omniscient with one (or more) seeds.

On the other hand, computing OPT is NP-hard. With only one seed, however, we can calculate the exact OPT numerically, since there are only n possible options to consider. We compute OPT for an ER network of size 100 with $cd = 1.5$ and find that random seeding with only two extra seeds improves over optimal seeding with one seed. To overtake OMN, random needs 4 extra seeds. Similar numbers are enough for random to catch up when $cd = 2$: random needs 3 extra seeds to overtake OMN and 2 extra seeds to overtake OPT. These findings suggest that the limit results pertain to small networks.

3.2 Power-law Chung-Lu Networks

Several real-world networks are characterized by degree distributions with fat tails, in the sense that few nodes that 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. Erdős-Rényi networks fail to capture this feature. Since Chung-Lu power-law networks are special cases of the IRN model, Theorem 1 implies the following corollary.

Corollary 2. *Consider a power-law Chung-Lu network on n nodes with scale parameter b and minimal expected degree d. If either* (1) $b \in (1, 2]$ *or* (2) $b > 2$ *and* $cd > \frac{b-2}{b-1}$, *then for any s and x,*

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

If $b > 2$ *and* $cd \leq \frac{b-2}{b-1}$, *then for every s,*

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

The corollary is proved in Appendix A.2.

Barabasi and Albert (1999) estimate the scale parameter for the tails of different real-world network degree distributions and find this lies in the $(1, 2]$ interval for most of their examples. Corollary 2, therefore, implies that precisely in the regime where the network admits highly central agents, no further assumptions on communication probability are needed to ensure that random with a few more seeds can beat the omniscient. The intuition, as depicted in Figure 1, is that random seeding is likely to pick neighbors of the highly connected nodes, precisely because they are highly connected. Highly connected nodes, then, are informed through their randomly seeded friends.

3.3 Networks with Homophily: The Island Model

The relationship between homophily and the conditions for the comparability between random and optimal seeding is easiest to see in the context of the islands model of networks. The next result follows immediately from the calculation of $||\mathbf{T}_\kappa||$ for the Island model in Section 2.1 and Theorem 1.

Corollary 3. *Consider an islands network model on n nodes with parameters* (m, d_{in}, d_{out}), *all positive. If* $\frac{1}{m}(d_{in} + (m-1)d_{out}) > 1/c$, *then for any x and s,*

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

If $\frac{1}{m}(d_{in} + (m-1)d_{out}) \leq 1/c$, *then for every s,*

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

Note that the term $\frac{1}{m}(d_{in} + (m-1)d_{out})$ is simply the average degree of a node in the islands model. Thus, the condition $\frac{1}{m}(d_{in} + (m-1)d_{out}) > 1/c$ translates into the requirement that on average informed agents speak to at least one of their friends.

It is instructive to investigate the relationship between homophily and the performance of random seeding. The homophily measure for the islands network, as defined by Golub and Jackson (2012), is given by

$$\frac{d_{in} - d_{out}}{d_{in} + (m-1)d_{out}}.$$

We can see that the existence of significant homophily is neither necessary nor sufficient for the conditions of Corollary 3 to be met. For instance, when $d_{out} \approx 0$, so there are few cross-group links and the network exhibits extreme homophily, the average degree is close to d_{in}/m. If $c\frac{d_{in}}{m} > 1$, the first part of the theorem still holds. On the other hand, when $d_{in} = d_{out}$ and thus the network exhibits no homophily, spectral homophily is 0 and the average degree is d_{in}. Thus, the condition $\frac{1}{m}(d_{in} + (m-1)d_{out}) > 1/c$ translates to $d_{in} > 1/c$.

3.4 Real-world Networks

None of the existing network formation models are perfect representations of the real-world networks. They can match degree distributions, or even incorporate clustering, but they cannot match all moments of the data. It is therefore important to simulate the diffusion model studied here on real-world networks to see whether our results continue to hold.

We will conduct simulations on the microfinance network data in Banerjee et al. (2013) and compare the performance of various seeding strategies.[10] The networks in Banerjee et al. (2013) have households as nodes, with edges representing the existence of some relationship. For example, edges may indicate that incident households go to temple, mosque or church together. Alternatively, edges may represent the fact that members of one household have borrowed or loaned money to those in the other, give or take advice, and so on. While some of these relationships are directed, we ignore the direction. An interpretation is that any sort of contact may create an opportunity to share some information being diffused.

Simulations in Figure 2 compare the average performance of random, degree-central, diffusion-central[11], eigenvector-central, and omniscient seeding strategies on village networks, which includes an edge between two households whenever either party indicated

[10] We do a similar exercise on a subnetwork of Facebook as a robustness check and observe similar findings. See Appendix B for details.

[11] 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).

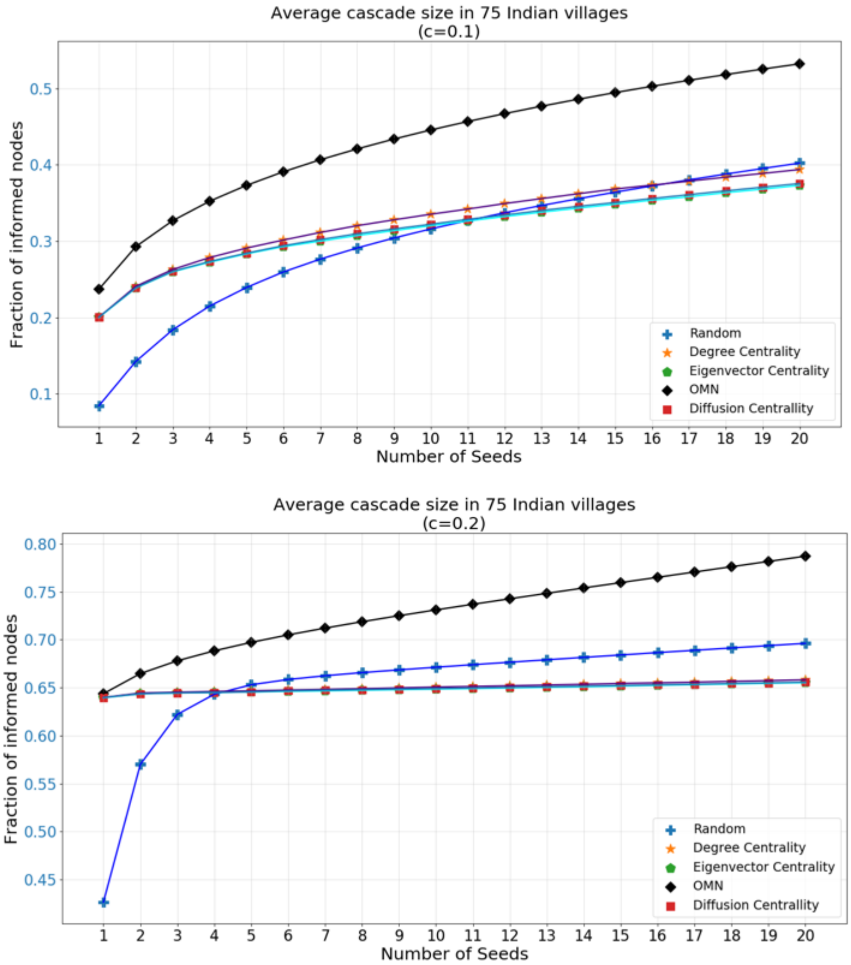


Figure 2: A comparison of average diffusion for various seeding strategies (omniscient, random, degree-, diffusion-, and eigenvector-central seeding) across 'all inclusive networks' in the village network data, for two different levels of communication probabilities.

some contact with the other group of any form. Results are included for two different values of communication probability c. For instance, when $c = 0.1$, random with 5 seeds performs as well as degree- and diffusion-central seeding with two seeds, and better than omniscient with one seed. When $c = 0.2$, random with 5 seeds performs better than all heuristics *with an equal number of seeds*, and better than omniscient with one seed.

Comparison to the OPT. With only one seed, we can solve for OPT numerically. Simulations a sample Indian village network show that when $cd = 1.5$, random seeding with 3 extra seeds is better than both OPT and OMN. When $cd = 2$, meanwhile, random with 2 and 3 extra seeds beat OPT and OMN, respectively.

A virtue of random seeding. Our simulations show that as the number of seeds increase, random seeding is better than diffusion-central and eigenvector-central seeding with an equal number of seeds. Centrality-guided seeding heuristics pick redundant agents, who are likely to be part of the connected core of the network. Seeding those individuals has decreasing marginal value. Random seeding performs better because it is more likely to seed individuals in the small components as well.

4 Robustness and Limitations

Here, we extend the comparison between seeding strategies beyond the expected value of eventual diffusion. In particular, we see how random seeding compares to optimal seeding when taking into account variance and speed of diffusion. We then investigate the robustness of our results to alternative diffusion models.

4.1 Alternative Objective Functions

4.1.1 Variance of Random Seeding

In some settings, maximizing the expected diffusion might not be the only objective. One reason for using network information and optimal seeding might be to guarantee some baseline level of diffusion. In this sense, the variance of the performance of a seeding strategy is an important measure. Here we note the rate with which the variance in diffusion produced by a random seeding converges to zero. Recall that $\alpha = \lim_{n\to\infty} \mathbf{H}(\mathrm{OMN}, 1)$ is the limit fraction of informed agents under the omniscient seeding strategy.

Proposition 1. *Consider a sequence of* $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$. *Let s be the number of seeds. Then*

$$\lim_{n\to\infty} \mathrm{Var}(\mathbf{H}(\mathrm{RAND}, s)) \le \alpha^2(1-\alpha)^s(1-(1-\alpha)^s).$$

We prove this proposition in Appendix A.4.

Proposition 1 shows that the variance of random seeding decreases exponentially in the number of extra seeds used. For instance, in diffusion in a large Erdős-Rényi network with $dc = 2$, $\alpha \simeq 0.8$, and thus the variance of random seeding is less than 0.0014. This convergence rate is faster when α is larger, i.e., there is more diffusion to be achieved.

4.1.2 Speed of Diffusion

Can random seeding compete with network-guided heuristics in *speeds* of diffusion? This question addresses the economically salient concern that even if both seeding strategies eventually reach the same diffusion level, network information might allow policymakers to accelerate the rate of adoption.

To address this question, we consider a bounded diffusion process where diffusion stops after T periods. We ask whether random seeding with extra seeds improve over OPT for any T? Figure 1 already shows that in general, the answer to this question is *no*. In order for random to beat OPT in the star network example, the process should

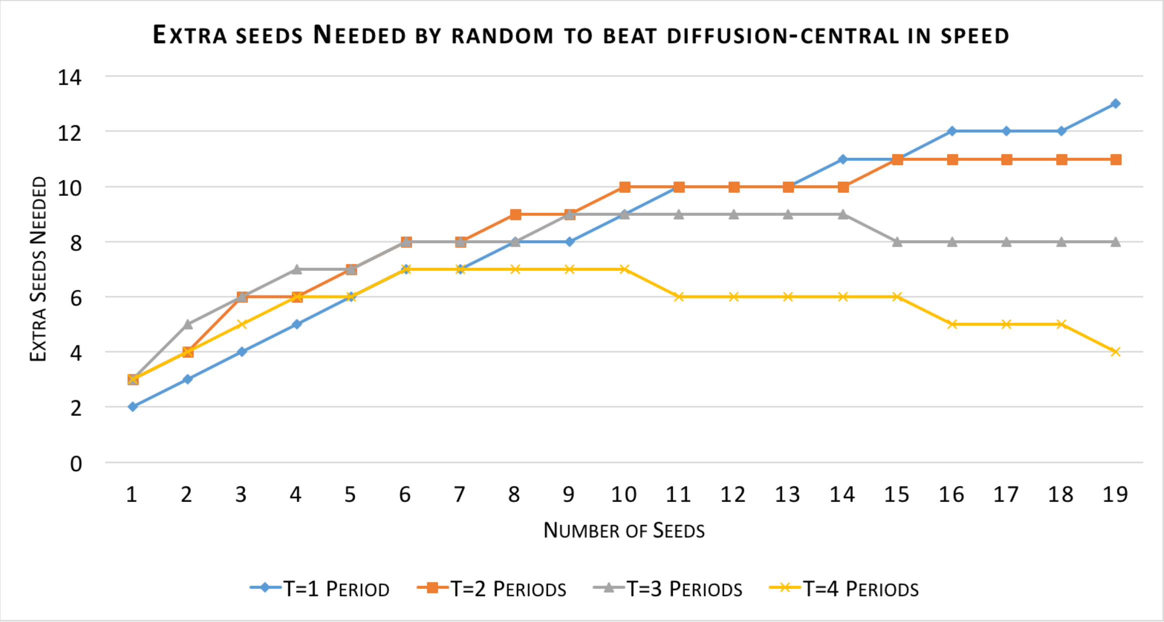


Figure 3: Average number of extra seeds required by random to outperform diffusion-centrality seeding in Indian village networks (in of *speed of diffusion*). If objective is diffusion in the first T=1 or T=2 periods, then extra seeds required is relatively high (still less than 15), but once total outreach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

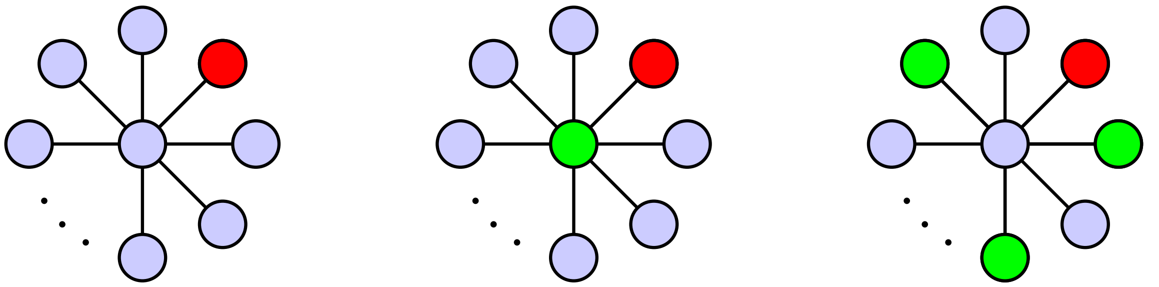
continue for at least two periods. Thus, if a policymaker's objective is to maximize the extend of diffusion in one period (i.e., only those who are directly informed by seeds), then random seeding has a hard time catching up. The result fails for power law networks for similar reasons.

Speed of diffusion in Microfinance. Figure 3 depicts the extra number of seeds needed for random to beat diffusion-central seeding, simulating the microfinance diffusion model on Indian village networks. We consider a bounded diffusion process, where the diffusion stops after 1, 2, 3 or 4 periods. When the diffusion ends in $T = 1$ or $T = 2$ periods, the extra number of seeds required for random to beat diffusion centrality is between 3 to 13, depending on the number of seeds. Note, in particular, that for $T = 1$ the number of extra seeds is increasing in s.

4.1.3 Diffusion Minimization by Vaccination

Network information can be highly valuable when a policymaker wishes to *minimize* the spread of some diffusion, as in the cases of fake news or infections. In such cases, a policymaker may try to inform agents about such news or to vaccinate them so as to lessen the spread.

To fix ideas, suppose some random individual is infected with a disease, and the diffusion process is captured by our baseline model. A policymaker seeks to 'vaccinate'



(a) No vaccination

(b) Optimal vaccination of 1 agent

(c) Random vaccination of several agents

Figure 4: 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. 4(a): Without vaccination, the central node will be infected with probability 0.5, and thus $\frac{n}{4}$ of agents get infected in expectation. 4(b): Vaccinating the central node is optimal, as it stops the diffusion completely. 4(c): Randomly vaccinating $x = o(n)$ individuals picks the central node with vanishing probability. The chance that the central node gets infected is around 50%, so nearly $\frac{(n-x)}{4}$ of agents get infected in expectation.

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, the number of additional individuals needed for random vaccination to outperform the optimum can be as large as a constant fraction of all agents. Figure 4 shows one such example. An individual is randomly infected, and the goal is to vaccinate one individual to minimize the size of diffusion. Optimal vaccination will choose the central node and the diffusion stops. Random vaccination, even with a few extra seeds, is not going to pick the central node, and thus performs poorly.

4.2 Alternative Diffusion Models

So far, our theoretical results focused on a simple variant of the undirected SIR model of diffusion, which is used to study processes such as diffusion of information and ideas, rumors, or infectious diseases. Here we discuss how our results hold up under alternative diffusion models.

4.2.1 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. What if the communication is directed?

It is clear that if a node's propensity to have in-links versus out-links is correlated

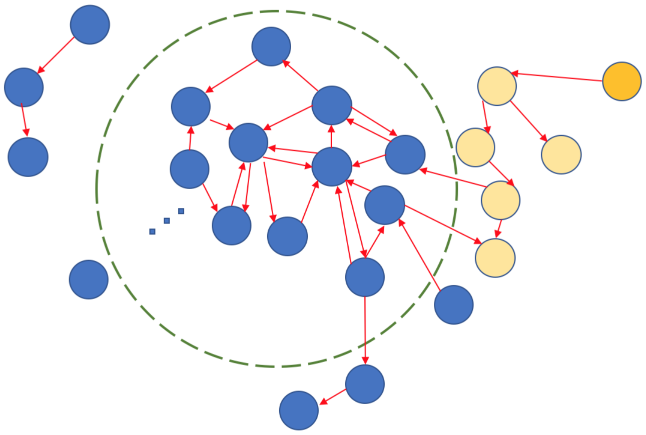


Figure 5: Above is an example communication network when communication is directed. The outgoing edges represent the nodes that a given node would inform if given information. The nodes within the dotted dashed circle represent the strongly connected giant component. The orange nodes, if informed, disseminate information to the SGC. In this example, OMN will choose to seed the upper right node, given a single seed. In the proof of Proposition 2, we show that the size of the set of any cluster of orange nodes (paths to the SGC) is $o(n)$ and the OMN cannot significantly outperform RAND.

with degree, diffusion may be impossible without targeting central nodes (e.g., consider a star network where the center speaks to the leaves but not vice-versa). We investigate whether an analogue of our main result holds when any direction of a relationship is equally likely.

In the Online Appendix, we show that our main theorem in fact goes through for a very general class of random networks. Here, for simplicity, we consider a model of directed networks similar to Erdős-Renyi. Let $D(n, d)$ be a random directed network on n nodes in which directed edge (i, j) is drawn with probability $\frac{d}{n}$. In this setting, OMN observes a realization of the directed communication network and chooses the best nodes to seed using this information.

Proposition 2. *Consider a random directed network, $D(n, d)$. If $cd > 1$, then for any x and s,*

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

If $cd < 1$, then for every s,

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

Proof overview. 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 that are reachable via a *directed path* 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 i and the number of vertices that can be reached from i that are outside the SGC. In Appendix A.3, we establish that the number of such vertices is a vanishing fraction of the total number of vertices, and therefore their impact is minimal.

4.2.2 Models from development economics

The diffusion models used in Banerjee et al. (2013) and Cai et al. (2015) are more complex. 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) consider a linear probability model, where the chance that an agent gets informed is proportional to the number of its informed neighbors. Still, these processes share the feature of the SIR model that it is possible to be informed even with only one informed neighbor.

We 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 C (for the Microfinance model) and Appendix D (for the weather insurance model) show that the number of additional seeds required for random to perform no worse than 95% of centrality-guided heuristics is typically less than 3.

When the diffusion process is such that several of an agent's neighbors have to adopt a technology before he does the same, our results may not 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. This seems to be the case in Beaman et al. (2021), who study technological adoption by farmers as they vary seeding rules in village-networks in Malawi in an experimental setting. Since random seeding is unlikely to inform multiple neighbors of the same node, it will fail to prompt any diffusion if thresholds are uniformly high across all agents unless a sufficiently large fraction of nodes are seeded.[12]

4.2.3 Heterogeneous communication probabilities

In the IRN model, each pair of agents i and j are connected and communicate with probability $c\kappa(i,j)$, where c is the communication probability regardless of types. In principle, we could instead consider a model where each type i agent communicates

[12] Jackson and Storms (2023) calculate how many random seeds are needed to compare to optimal strategies in a threshold model and find that this number is typically high. On the other hand, Watts and Dodds (2007) consider a threshold model where individuals may have different thresholds and show that a large diffusion is possible only when a sufficiently large fraction of nodes would adopt whenever even one of their neighbor adopts. In such settings, it is plausible that random seeding strategies would perform well.

with a type j agent with probability c_{ij}. Then, a type i individual is connected to and speaks to a type j individual with probability $c_{ij}\kappa(i,j)$. Note that in the diffusion model we consider, connection probability $(\kappa(i,j))$ and communication probability (c_{ij}) play a similar role. Therefore, as long as $c_{ij} = c_{ji}$, we can simply define a new symmetric kernel function $\kappa'(i,j) = c_{ij}\kappa(i,j)$ and assume $c = 1$ is the communication probability for all types. This then becomes a special case of our analysis.

4.3 Alternative Network Structures

We have shown that our findings apply to various types of random networks, including those with power-law patterns and homophily. A remaining question is how would results hold up in networks that have features not captured by the IRN model (e.g., networks with clustering, like "small-world" networks (Watts and Strogatz, 1998), or networks with power-law *and* homophily). Moreover, for what sort of networks would our results fail?

To address this, we will first state a sufficient condition for *any* sequence of graphs. This will clarify that our results continue to hold in any network model that admits a unique giant component and a phase transition with high probability.

To fix ideas, let $\{G_n\}_{n\in N}$ be a sequence of (possibly random) graphs in which the number of vertices of G_n is equal to n. Let us build the communication network $\mathcal{K}(G_n)$ by keeping each link in the graph independently and with probability c. Let C_i to be the set of vertices in the i'th largest connected component of $\mathcal{K}(G_n)$. Let us use $\xrightarrow{\mathbb{P}}$ to denote convergence in probability with respect to both the random graph distribution and the process of forming the communication network.

Definition 1. *We say a sequence of graphs has the* ***solar system structure*** *if and only if there exists a threshold probability $0 < c_p < 1$ such that*

1. *If $c < c_p$ then,*
$$\frac{|C_1|}{n} \xrightarrow{\mathbb{P}} 0.$$
2. *If $c > c_p$, then there exists a function $\zeta : [0,1] \to (0,1]$*
$$\frac{|C_1|}{n} \xrightarrow{\mathbb{P}} \zeta(c),$$
3. *Moreover, for all $c \in [0,1]$, $\frac{|C_2|}{n} \xrightarrow{\mathbb{P}} 0$.*

In words, in a graph with solar system structure, either all connected components are of size $o(n)$ or there is a unique giant connected component that is linear size in n. The same argument as in the proof of Theorem 1 shows that our results hold for all such networks.

Theorem 2. *Let $\{G_n\}_{n\in\mathbb{N}}$ be a sequence of (possibly random) graphs with solar system structure and s be the number of seeds.*

- *If $c > c_p$, random seeding catches up to the omniscient seeding at an exponential rate in the number of extra seeds, i.e., for any x,*

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

- *If $c < c_p$, then any seeding strategy diffuses to only a vanishing fraction of the population:*

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

The proof of this result is the workhorse behind the proof of Theorem 1.

In the Online Appendix, we consider even more general classes of random graphs than Inhomogeneous Random Networks exhibit the solar system property. We build on recent results related to graphs with "local convergence" (Alimohammadi et al., 2023, 2022) to generalize our theoretical findings. In particular, our results hold for small-world networks (Watts and Strogatz, 1998) and a model of networks with power-law degree distributions and homophily.

Still, there are network structures for which random seeding performs poorly, i.e., catch-up to optimal seeding might be very slow relative to the amount of diffusion that is achievable. For example, consider a set of disjoint components, each with n/k nodes. Omniscient seeding with k seeds achieves a diffusion of 1. While random seeding eventually catches up, it does so an arbitrarily slow rate for large k. A random strategy would likely seed the same components redundantly without reaching other unseeded components. This cannot happen for the networks considered in Theorem 1, where the convergence rate of the random strategy grows with the fraction of achievable diffusion.

Such networks may arise when agents form into distinct groups, and the policymaker has no information about the group identity of nodes. In practice policymakers may know at least some members of large groups, and use these as seeds. For example, the fact that one village is distinct form another is readily observable, making it easy for the policymaker to split seeds across villages. Castes of some members within a village where links across castes are sparse may also be known, again allowing the policymaker to split seeds between castes. This highlights the value of information about group identity. We show in a concrete example how such information can improve the convergence rates of random strategies in Appendix E.

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).

5 Concluding Discussions

Our main result does not provide specific prescription on how policymakers should approach seeding decisions. Random seeding gives a lower-bound for what a policymaker can do without any network information, and omniscient seeding gives a generous upper-bound for what she can do with full information about the network and the diffusion process, together with computational power. Neither strategies are policy prescriptions. However, our results may provide a helpful perspective for researchers and policymakers.

For researchers reporting on the effectiveness of different seeding strategies, we suggest measuring and reporting: *how many extra seeds would a random seeding strategy need to be within z% of the proposed seeding strategy, for a small z?* Regardless of the diffusion model or class of networks in consideration, this measure provides an economically meaningful quantity for the value of careful targeting in terms of the extra seeds. For example, for the diffusion model of Banerjee et al. (2013) and with $s = 10$ initial seeds, random seeding with 1 extra seed performs within 95% of 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).[13]

Policymakers, on the other hand, may ask: *how, then, should we seed a network?* The answer depends on several factors that we do not model, such as the policymaker's beliefs about the network structure and the diffusion process, the costs of acquiring network data, and the cost of seeding. For instance, if the policymaker knows a lot about the structure of the network *a priori*, using this information is clearly better than throwing it away. In the setting of informing homeless individuals about HIV prevention (Yadav et al., 2016; Rice, 2010), as shown in Appendix E, dividing seeds between distant homeless populations is better than random seeding that ignores geography, given the

[13] 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. Additional numbers are reported in Table 1.

low likelihood of communication across locations. Moreover, it may be very cheap, and therefore advisable, to obtain partial network information (e.g., who lives close to whom) but much more expensive to acquire more detailed data (e.g., who speaks to whom). Lastly, careful targeting can be more valuable in settings where seeding is very costly (e.g., providing an extra free version of an expensive product).

While our main theorem alone does not directly prescribe seeding decisions, it can sometimes be useful—combined with the policymaker's decision theoretic primitives—to assess whether her intended heuristic seeding strategy is dominated by a random strategy with additional outreach. To see this, consider a simple setting where the policymaker's payoff depends on whether or not the fraction of adopting nodes exceeds or falls below a certain threshold. For example, a referendum might go up for a vote only if at least half the town hears about it and shows up to the town hall meeting. Or a political party using a word-of-mouth campaign may be rewarded only when a sufficient fraction of the community registers to vote.

In such settings, the policymaker's beliefs about the parameters of the network structure and the diffusion process are both complex, high dimensional objects. However, the key idea of our analysis is that when the network is not too small, this prior maps to some marginal distribution G over the size of the giant component, $\alpha \in [0, 1]$.

To fix ideas, suppose the policymaker gets a payoff normalized to 1 if diffusion reaches 50% of nodes and zero otherwise. Let c_s be the cost of seeding, and c_I be the cost of learning the network structure. In this case, if an omniscient policymaker pays a cost c_I to acquire network data and learns that $\alpha < 0.5$, she avoids paying the cost of seeding and gets a payoff of $-c_I$. On the other hand, if $\alpha \geq 0.5$, she will seed the giant component and her payoff will be $1 - c_s - c_I$. Hence, her payoff will be $(1 - G(0.5))(1 - c_s) - c_I$.

Meanwhile, the payoff to implementing the random seeding strategy with k seeds is given by $\int_{0.5}^{1}(1 - (1 - \alpha)^k)dG(\alpha) - k \cdot c_s$. This is the probability of seeding a node in the giant component minus the cost of seeding k nodes. Let k^* denote the number of seeds that maximizes the value of the random seeding strategy.

Putting these together, the policymaker prefers random seeding to acquiring full network information whenever the following condition holds:

$$\Big((1 - G(0.5)) - \int_{0.5}^{1}(1 - (1 - \alpha)^{k^*})dG(\alpha)\Big) + (k^* - 1 + G(0.5)) \cdot c_s \leq c_I$$

The first term on the left hand side of this inequality is an upper bound on the value of acquiring full network information—the value of OMN minus the value of RAND with k^* seeds. The second term on the left of the inequality is the cost of additional seeding. If these terms sum to less than or equal to c_I, then a policymaker can conclude that random seeding beats any network-guided heuristic, including the infeasible omniscient strategy.

This example suggests a way to include costs, benefits and prior beliefs into a tractable, decision-theoretic framework to compare network targeting and expanded outreach. Carrying out this exercise more generally, we suspect, is a fruitful avenue for future research.

References

Alimohammadi, Y., Borgs, C., and Saberi, A. (2022). Algorithms using local graph features to predict epidemics. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3430–3451. SIAM.

Alimohammadi, Y., Borgs, C., and Saberi, A. (2023). Locality of random digraphs on expanders. *The Annals of Probability*, 51(4):1249–1297.

Banerjee, A., Breza, E., Chandrasekhar, A. G., and Golub, B. (2023). When less is more: Experimental evidence on information delivery during india's demonetization. *Review of Economic Studies*.

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. (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 86(6):2453–2490.

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. (2021). Can network theory-based targeting increase technology adoption? *American Economic Review*, 111(6):1918–1943.

Bloch, F. (2016). Targeting and pricing in social networks. In *The Oxford Handbook of the Economics of Networks*.

Bloch, F., Demange, G., and Kranton, R. (2018). Rumors and social networks. *International Economic Review*, 59(2):421–448.

Bollobás, B., Janson, S., and Riordan, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122.

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. (2020). Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–2484.

Bulow, J. and Klemperer, P. (1996). Auctions versus negotiations. *The American Economic Review*, pages 180–194.

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.

Campbell, A. (2013). Word-of-mouth communication and percolation in social networks. *American Economic Review*, 103(6):2466–2498.

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.

Galeotti, A., Golub, B., and Goyal, S. (2020). Targeting interventions in networks. *Econometrica*, 88(6):2445–2471.

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

Golub, B. and Jackson, M. O. (2012). How Homophily Affects the Speed of Learning and Best-Response Dynamics. *Quarterly Journal of Economics*, 127(3):1287–1338.

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

Hofstad, R. v. d. (2016). *Random graphs and complex networks*, volume 43. Cambridge University Press.

Hofstad, R. v. d. (2024). Random graphs and complex networks. Vol. 2. In preparation, see `http://www.win.tue.nl/∼rhofstad/NotesRGCNII.pdf`.

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

Jackson, M. O. and Storms, E. C. (2023). 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.

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

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

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.

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. (2020). Diffusion games. *American Economic Review*, 110(1):225–70.

Sadler, E. (2022). Seeding a simple contagion. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 247–248.

Watts, D. and Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

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

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 Proofs

We start with a lemma on the performance of RAND and OMN on the communication graph $\mathcal{K}(G)$ for an arbitrary G:

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. □

A.1 Proof of Theorem 1

Proof of Theorem 1. When $||\mathbf{T}_\kappa|| > 1/c$, as $n \to \infty$, by Theorems 3.1 and 3.12 (Bollobás et al., 2007) on the sizes of connected components of a random graph in the IRN model, there exists an $\alpha \in (0, 1]$ such that with high probability (i.e., with probability approaching 1, as n approaches infinity) for graph $\kappa(G)$, $\mathcal{C}_1 = \alpha n + o(n)$ and $\mathcal{C}_i \in O(\log(n))$ for all $2 \leq i \leq CC$. Let G_n be a randomly realized network from $\text{IRN}_n(\boldsymbol{p}(\kappa))$.

The combination of the above result with Lemma 1 implies that for G_n, $\mathbf{h}(G_n, s, \text{OMN}) \leq \mathcal{C}_1 + (s-1)\mathcal{C}_2 = \alpha n + O(\log(n))s$ with high probability.

With $s + x$ seeds, the probability that a node in the largest component is randomly seeded converges in probability to $(1-(1-\alpha)^{s+x})$. Again, using Lemma 1, $\mathbf{h}(G_n, s + x, \text{RAND}) \geq \mathcal{C}_1(1-(1-\frac{\mathcal{C}_1}{n})^s) \geq \alpha n(1-(1-\alpha)^{s+x}) + o(n)$, with high probability. Taking expectations over the realizations of G_n:

$$\frac{\mathbf{H}(\text{RAND}, s+x)}{\mathbf{H}(\text{OMN}, s)} \geq \frac{\alpha n(1-(1-\alpha)^{s+x}) + o(n)}{\alpha n + o(n)},$$

which approaches $(1-(1-\alpha)^{s+x})$ as $n \to \infty$.

When $||\mathbf{T}_\kappa|| < 1/c$, then even $C_1 = o(n)$, so $\mathbf{H}(\text{OMN}, s) = o(n)$, which shows the second part of the theorem. □

A.2 Proof of Corollary 2

Unlike the definition that appears in the main body of the text, the Chung-Lu graph considered here has infinitely many types. To extend phase transition results similar to those used in Theorem 1, we need to modify the stated model of inhomogeneous random graphs appropriately. The next subsection closely follows Hofstad (2016) to extend the IRN model to a setting with potentially infinite type-space.

A.2.1 Infinite Type IRNs

A *ground space* is a pair $(\mathcal{T}, \mu)$, where $\mathcal{T}$ is a separable metric space and μ is a Borel probability measure on $\mathcal{T}$. The set $\mathcal{T}$ is the set of agent *types* and it can include finite or infinite types of agents. The measure $\mu(A)$ denotes the proportion of agents having a type in A, for $A \in \mathcal{T}$ in the limit as n grows, in a manner to be formalized now. A *node space* $\mathcal{V}$ is a triple $(\mathcal{T}, \mu, (\mathbf{x_n})_{n\geq 1})$ where $(\mathcal{T}, \mu)$ is a ground space and, for each $n \geq 1$, $\mathbf{x_n}$ is a random sequence $(x_1, x_2, ..., x_n)$ of n points of $\mathcal{T}$, such that:

$$\mu_n(A) = \#\{i : x_i \in A\}/n \to \mu(A),$$

for every μ-continuity set $A \in \mathcal{T}$.

A *kernel* $\kappa : \mathcal{T}^2 \to [0, \infty)$ is a symmetric (Borel) measurable function. For a fixed kernel κ and $n \in \mathbb{N}$, $\text{IRN}_n(\boldsymbol{p}(\kappa))$ is the random network on $[n] = \{1, 2, \cdots, n\}$, where each possible link ij, $i, j \in [n]$ is present with probability

$$p_{ij}(\kappa) = p_{ij} = \left(\frac{1}{n}\kappa(x_i, x_j)\right) \wedge 1,$$

and links are present independently of each other. Note that this model allows for type-specific correlations among agents. While the choice of a kernel is arbitrary, for typical applications we want the graph to be "connected". This motivates the following definition. We say a kernel κ is *reducible* if there exists some $A \subseteq \mathcal{T}$ with $0 < \mu(A) < 1$ such that $\kappa = 0$ on $A \times (\mathcal{T} \backslash A)$ almost everywhere. The kernel is *irreducible* if it is not reducible. Irreducibility means that the graph $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$ cannot be split into two graphs so that the probability of a link from one part to the other is zero. This is a natural restriction, since if it fails, then the graph is split into two independent random graphs, so we could have considered each of them separately.

We now define the notion of a *regular* kernel.

Definition 2 (Regular Kernels)**.** *A kernel κ is* regular *if it is irreducible and the following conditions are satisfied:*

1. *κ is continuous on $\mathcal{T}^2$ almost everywhere.*
2. *$\iint_{\mathcal{T}^2} \kappa(x,y)\mu(dx)\mu(dy) < \infty$*
3. *$\frac{1}{n}\mathbb{E}[|E(\mathrm{IRN}_n(\boldsymbol{p}(\kappa))|] = \frac{1}{2}\iint_{\mathcal{T}^2} \kappa(x,y)\mu(dx)\mu(dy)$.*

Similarly, a sequence (k_n) of kernels is called *regular with limit κ* when $x_n \to x$ and $y_n \to y$ imply that $\kappa_n(x_n, y_n) \to \kappa(x,y)$, where κ is regular and:

$$\frac{1}{n}\mathbb{E}[|E(\mathrm{IRN}_n(\boldsymbol{p}(\kappa_n))|] \to \frac{1}{2}\iint_{\mathcal{T}^2} \kappa(x,y)\mu(dx)\mu(dy)$$

Conditions (1), (2), and (3) imply that the expected number of edges in the graph is proportional to n, with the proportionality constant being equal to $\iint_{\mathcal{T}^2} \kappa(x,y)\mu(dx)\mu(dy)$. This ensures that the average degree per node "converges".

Finally, let:

$$(\mathbf{T}_\kappa f)(x) = \int_{\mathcal{T}} \kappa(x,y)f(y)\mu(dy),$$

for any measurable function f such that this integral is defined for (almost every) $x \in \mathcal{T}$. We can now define the key mathematical object:

$$||\mathbf{T}_\kappa|| = \sup\{||\mathbf{T}_\kappa f||_2 : f \geq 0, ||f||_2 \leq 1\}.$$

We are now ready to state a result from Hofstad (2016) that is useful in extending Theorem 1 to infinite type spaces:

Theorem 3 (Hofstad (2016))**.** *Let (κ_n) be a sequence of regular kernels with limit κ, and let $\mathcal{C}_1$ denote the largest connected component of $IRG_n(\mathbf{p}(\kappa_n))$. Then $|\mathcal{C}_1|/n \to \alpha$ for some $\alpha \in [0,1]$. Moreover, $\alpha > 0$ if and only if $||\mathbf{T}_\kappa|| > 1$.*

Proof of Corollary 2. Let $CL_n(n, \mathbf{w}^n)$ be the power-law Chung-Lu network with scale parameter b and minimum expected degree d i.e., $w_i^n = [1-F]^{-1}(i/n)$, where $F(x) = 1 - (\frac{d}{x})^b$.

The first observation is that the probability that nodes i and j are connected in $\mathcal{K}(CL_n(w))$ is

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

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

Therefore, $\mathcal{K}(CL_n(w))$ is also a power-law network with scale parameter b and minimum expected degree cd. Equivalently, ${w'_i}^n = [1-F']^{-1}(i/n)$, where $F'(x) = 1 - (\frac{cd}{x})^b$.

Let W' be a random variable with cumulative distribution function F' on $[cd, \infty)$. W' follows a Pareto distribution with scale parameter b and the minimum support cd, therefore $E[W'] = \frac{bcd}{b-1}$ when $b > 1$ and infinity when $b \leq 1$. Similarly $E[W'^2] = \frac{b(cd)^2}{b-2}$ when $b > 2$ and infinity for $b \leq 2$.

The rest of the proof follows the same lines as the analysis of the connected components of Chung-Lu graphs (e.g., see Section 3.5.2 of Hofstad (2024)).

When $b > 2$, $E[W'] < \infty$, so Conditions (1)-(3) of Definition 2 hold. Therefore, kernels $\kappa_n(i/n, j/n) = np_{ij} = \frac{w'_i w'_j}{\frac{1}{n}\sum_k w'_k}$ are regular and have a limit $\kappa(x,y) = [1-F']^{-1}(x)[1-F']^{-1}(y)/E[W']$. Furthermore, $||\mathbf{T}_\kappa|| = \frac{E[W'^2]}{E[W']} = \frac{cd(b-1)}{(b-2)}$. So, $||\mathbf{T}_\kappa|| > 1$ if and only if $cd > \frac{b-2}{b-1}$. Applying Theorem 3, we obtain that the largest connected component of $\mathcal{K}(CL_n(w))$ is of linear size in n if and only if $cd > \frac{b-2}{b-1}$. Theorem 3.17 of **?** implies that the rest of the connected components are all of size $o(n)$. Now, applying Lemma 1, we can prove Corollary 2 in the case $b > 2$.

When $b \in (1,2)$, $E[W'^2] = \infty$. So, $||\mathbf{T}_\kappa|| = \infty$ or all non-negative values of c. Therefore, the graph has a unique giant connected component with size linear in n. Lemma 1 implies Corollary 2 for this case similarly.

□

A.3 Proof of Proposition 2

We will follow the arguments of Karp (1990) to show Proposition 2. But first we need to define a few concepts. A *strongly connected component* is a subgraph with vertex set C, such that for every $u, v \in C$, there is a directed path from u to v and a directed path from v to u. A relevant concept for our study is that of a *strongly-connected giant component*, which is a strongly connected component containing a linear fraction of the nodes, asymptotically.

Proof of Proposition 2. Consider $D(n,d)$, a random directed network on n nodes in which each directed edge (i,j) is drawn independently with probability $\frac{d}{n}$. First, we note a few facts from Karp (1990) about these networks. Let $R(i)$ be the vertices reachable from a i through some directed path.

1. If $cd < 1$, then with probability tending to 1 as n tends to infinity, for all v, $|R(i)| < \frac{8\log n}{1-cd}$.

2. If $cd > 1$, with probability tending to 1 as n tends to infinity, there exists a unique Strongly-connected Giant Component (SGC) which contains about $\theta^2 n$ vertices, where θ is the unique root in $[0, 1]$ of the equation $1 - x - e^{cdx} = 0$. More precisely, with probability tending to 1 as n tends to infinity, and for any non-decreasing unbounded function $w(n)$, the size of the SGC lies in the interval $[\theta^2 n - w(n)\sqrt{n \log n}, \theta^2 n + w(n)\sqrt{n \log n}]$.

3. Suppose $cd > 1$, $w(n)$ is a non-decreasing unbounded function, and C denotes the set of vertices in the the SGC. Let $R^+(i)$ denote the set of vertices that are reachable from i, and let $R^-(i)$ denote the set of vertices that can reach i. Similarly, define C^+ to be the set of all vertices that can be reached from some vertex in the SGC. And define C^- to be the set of all vertices that can reach some vertex in the SGC. Note that SGC $= C^+ \cap C^-$ and every vertex in C^- can each every vertex in C^+.

   Consider a vertex i. If $i \in C^-$, then with probability tending to 1 as $n \to \infty$, the size of the reachable set $|R^+(i)|$ lies within the interval $[\theta n - w(n)\sqrt{n}, \theta n + w(n)\sqrt{n}]$. On the other hand, if $i \notin C^-$, then $|R^+(i)|$ is at most $O(\log n)$ with probability tending to 1.

   Moreover, for any vertex i, the difference $|R^+(i) - C^+|$ is less than $w(n)\sqrt{n}$ with probability going to 1 as $n \to \infty$.

Fact 1 shows that if $cd < 1$, the fraction of nodes informed by OMN is at most $O(\frac{\log n}{n})$, and therefore $\lim_{n\to\infty} \mathbf{H}(\text{OMN}, s) = 0$.

Let $w(n)$ be a function growing sublinearly in n, like $\log n$. From fact 3, we know that each trial of RAND gets at least $\theta n - w(n)\sqrt{n}$ nodes with probability θ. On the other hand, the first seed chosen by the OMN can get at most $\theta n + w(n)\sqrt{n}$, and the remaining $s - 1$ can add at most $w(n)\sqrt{n}$ each. That proves the first part of the proposition for $\alpha = \theta$. □

A.4 Proofs of Proposition 1

Proof of Proposition 1. Note that with s seeds, the probability that at least one belongs to the giant component is $p_n = 1 - (1 - \alpha)^s + o(1)$. This follows since a giant component exists with high probability. Since the remaining components are $o(\log(n))$ with high probability, random seeding reaches a fraction α nodes with probability p_n and a fraction $o(1)$ with probability $1 - p_n$. The variance in the fraction of nodes reached is therefore $\alpha^2(1 - p_n)p_n + o(1) \to \alpha^2((1 - \alpha)^s)(1 - (1 - \alpha)^s)$. □

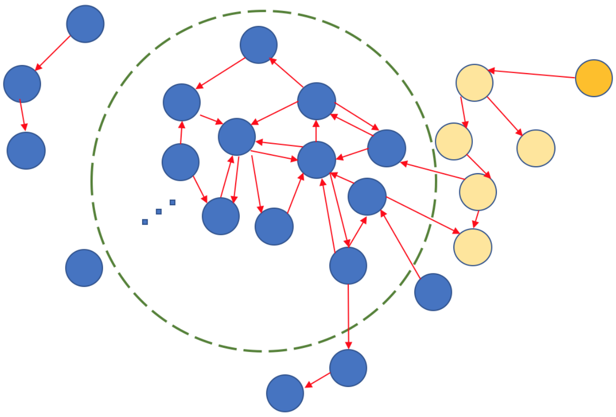


Figure 6: Above is an example communication network when communication is directed. The outgoing edges represent the nodes that a given node would inform if given information. The nodes within the dotted dashed circle represent the strongly connected giant component. If any node is informed within the SGC, all nodes reachable from the SGC become informed. Random seeding with enough seeds will land a seed in the SGC with sufficiently high probability. The orange nodes, if informed, also disseminate information to the SGC. On the other hand, OMN might choose to seed the dark orange node, and gain an advantage over RAND. The proof of Proposition 2 shows that the size of cluster of orange nodes reachable from any node is $o(n)$ and therefore OMN cannot significantly outperform RAND.

B Simulation of SIR model on Facebook network

We replicate the comparison between diffusion strategies on a Facebook subnetwork in Figure 7 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. As it can be seen, for both $c = 0.02$ or $c = 0.05$, random seeding quickly catches up with network-guided seeding heuristics. For instance, when $c = 0.05$, random seeding with 5 seeds beats omniscient seeding with one seed.

C 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

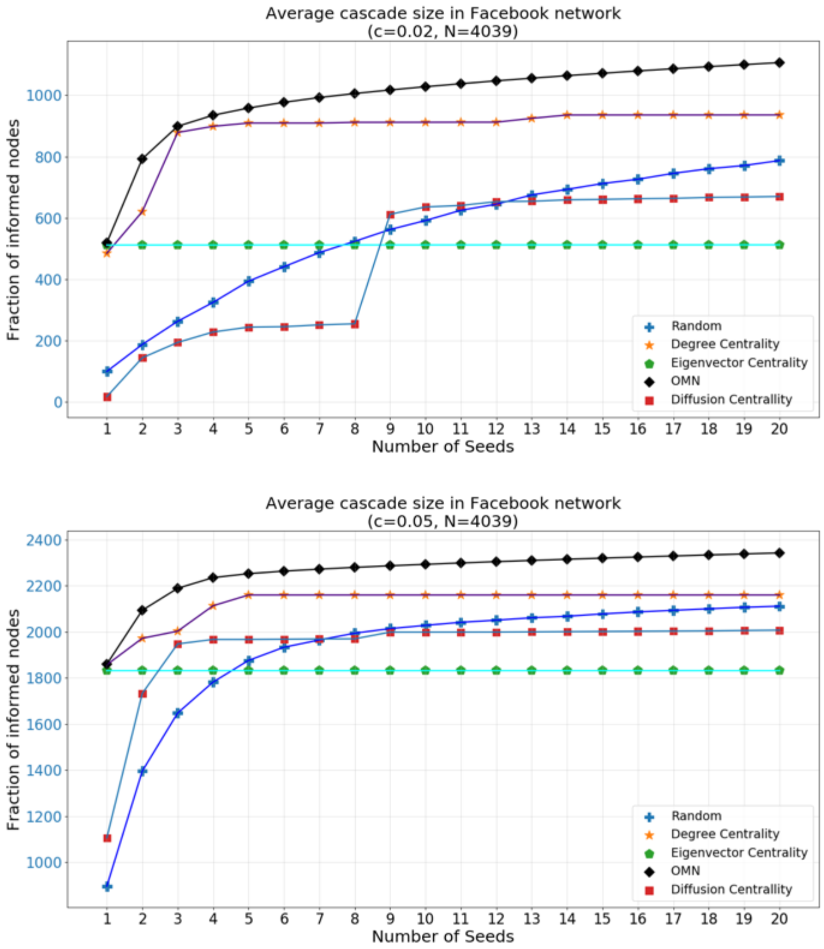


Figure 7: A comparison of average diffusion for various seeding strategies (omniscient, random, degree-central, diffusion-central, and eigenvector-central seeding) in a subnetwork of Facebook, for two different levels of communication probabilities..

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 that are informed or the number of nodes that 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

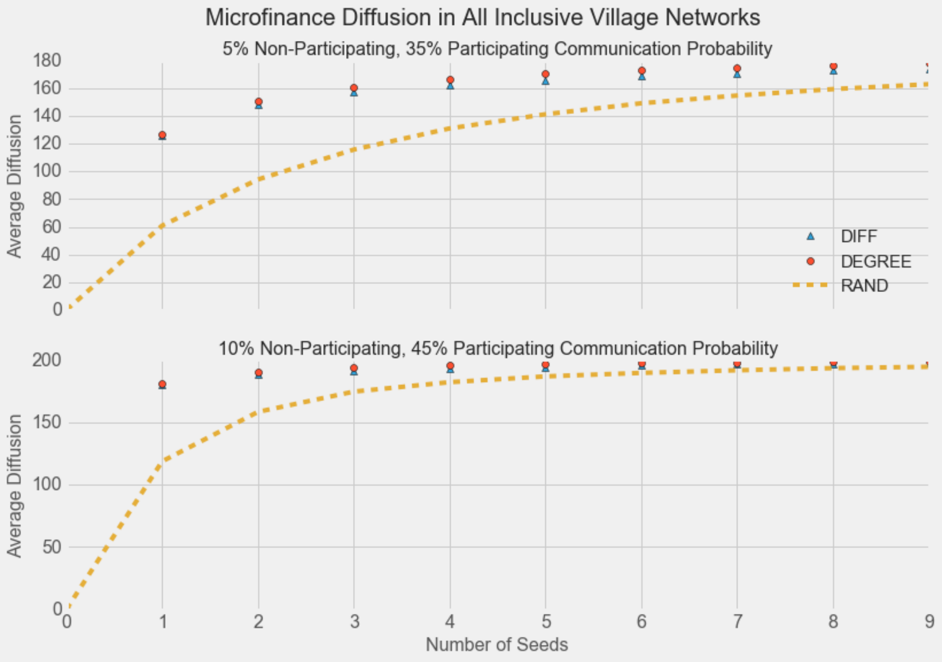


Figure 8: This is an analogue of Figure 2 with the diffusion process specified in Banerjee et al. (2013) rather than the model studied in this paper. As the number of seeds increases, random seeding performs as well as the centrality-guided seedings.

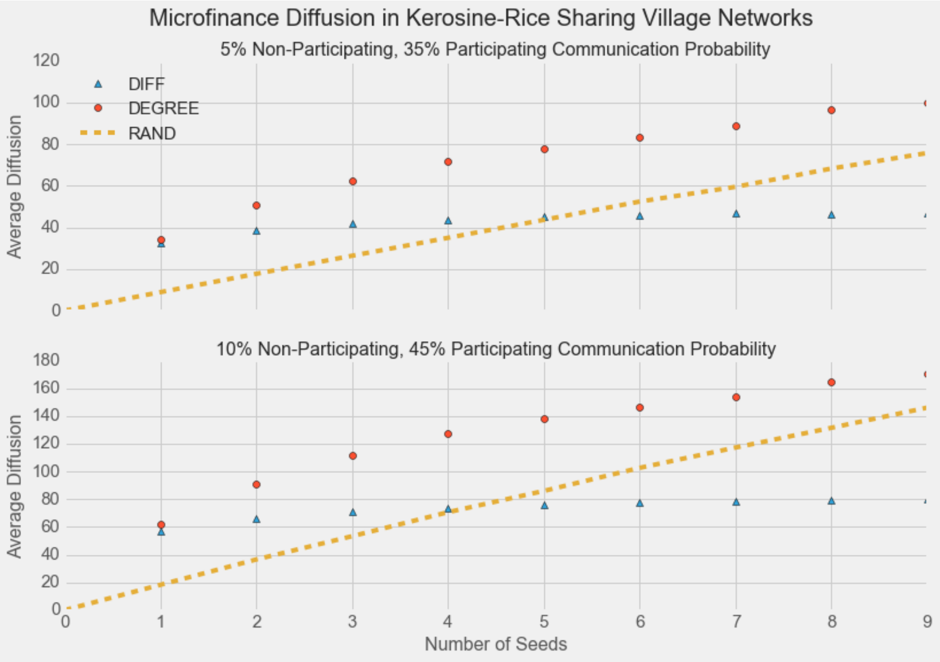


Figure 9: Random seeding performs well relative to the other seeding strategies. Moreover, it performs better than the seeding guided by the diffusion centrality when the number of seeds is more than 5.

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 C shows the results of simulations for both sets of parameter estimates. We include simulations for the sparser kerosene and rice borrowing network in Figure C.

For simulations of Section 4.1.2, 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.

D 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

[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.

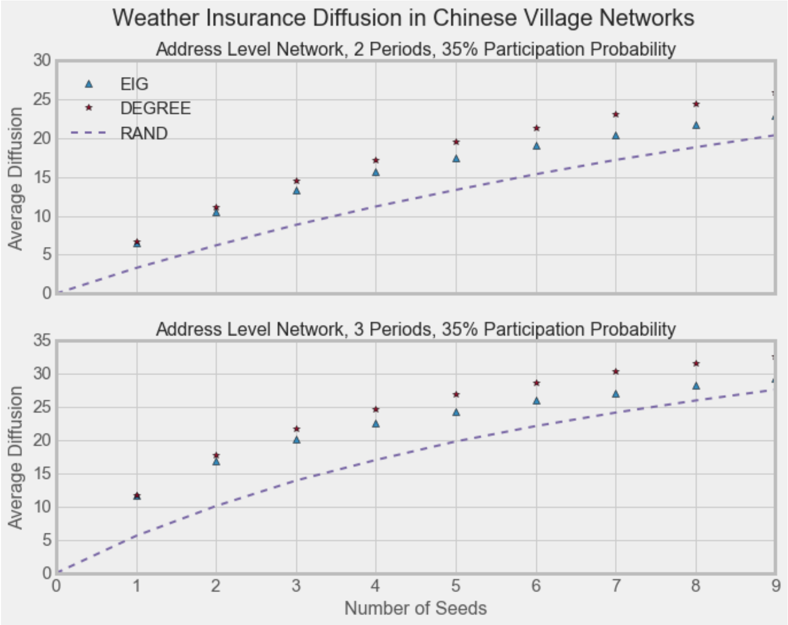


Figure 10: DEGREE seeding refers to seeding those with the highest degree, considering the undirected version of the village network. RAND seeding only chooses out of those villagers who participated in the social network survey, though they may name individuals who have not been surveyed as neighbors.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

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 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–

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

strong enough that a policymaker may safely ignore minutiae of network structure.

E Value of Knowing Homophily Patterns

In many settings, policymaker can identify distinct community membership without observing exact connections. For instance, a policymaker can distinguish between people who are in different villages or may readily observe the cohort to which college students belong. In such cases, a simple strategy that performs better than seeding agents uniformly at random is to first split seeds between communities and then seed randomly. The goal, however, is not to necessarily prescribe this as a policy, but to show that such partial network data can improve the rate at which random seeding converges to omniscient.

Consider a communication network G' with a single connected component containing a fraction 0.5 of all nodes, while the remaining nodes are isolated. Consider a different network G' with two disjoint components, each with a fraction 0.25 of all nodes, while the remaining nodes are isolated.

Omniscient strategies with $k > 2$ seeds diffuse to half the nodes in G' and G''. Random seeding strategies with k seeds also achieve an expected diffusion of approximately 0.5 in both networks when k is sufficiently large. However, expected diffusion converges to 0.5 more quickly for graph G' than for graph G''.

To see this, note that in G',

$$\mathbf{H}(OMN, k) - \mathbf{H}(RAND, k) = 0.5 - 0.5(1 - (1 - 0.5)^k) = 0.5^{k+1}.$$

Meanwhile, in G'', letting $p(k) = \sum_{m=1}^{k} \binom{k}{m} 2(0.25)^m (0.5)^{k-m}$,

$$\begin{aligned}\mathbf{H}(OMN, k) - \mathbf{H}(RAND, k) &= 0.5 - (0 \times 0.5^k + 0.25 \times p(k) + 0.5 \times (1 - 0.5^k - p(k))) \\ &= 0.5 - 0.5 \times (1 - 0.5^k - 0.5p(k)) \\ &= 0.5 - 0.5 \times (1 - \sum_{m=0}^{k} \binom{k}{m} 2(0.25)^m (0.5)^{k-m}) \\ &= 0.5 - 0.5(1 - (0.5 + 0.25)^k) \\ &= 0.5(0.75)^k\end{aligned}$$

Therefore, while random seeding eventually catches up to omniscient seeding in both cases, it converges at a slower rate when the giant component is split into two.

Suppose that each giant component in G'' and half of the isolated nodes belonged to two identical and independent villages. Suppose, moreover, that the policymaker knows this and splits the k seeds randomly among the two villages. Then,

$$\begin{aligned}\mathbf{H}(OMN, k) - \mathbf{H}(RAND, k) &= 0.5 - 2 \times 0.25 \times (1 - (1 - 0.5)^{k/2}) \\ &= 0.5 - 0.5 \times (1 - (\sqrt{0.5})^{k}) \\ &\approx 0.5 \times (0.71)^{k}\end{aligned}$$

Note that the convergence rate of random seeding is faster when the policymaker is equipped with partial network information, which in this case the identity of the villages. This suggests that homophily patterns are worthwhile knowing when they are particularly stark.