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

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

First Draft: September 2017
This Draft: August 2020

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 SIR model of diffusion and some of its generalizations, randomly seeding $s + x$ individuals can prompt a larger diffusion than optimally targeting the best s individuals, for a small x. We prove our results for large classes of random networks, and verify them in several small, real-world networks. Our results identify practically relevant settings under which collecting and analyzing network data to boost diffusion is not cost-effective.

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, 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

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

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

Contents

1 Introduction

How to identify individuals who are the best 'seeds' for maximizing the spread of information in a social network is a widely studied policy question in settings such as the diffusion of brand awareness for products (Richardson and Domingos, 2002), the propagation of microfinance programs (Banerjee et al., 2013), and the adoption of agricultural technologies in developing economies (Beaman et al., 2019). Since this problem is known to be computationally intractable (Kempe et al., 2003), a large body of theoretical and empirical studies introduce heuristics such as 'degree centrality,' 'eigenvector-centrality,' 'diffusion-centrality,' and 'k-shell index' as proxies for ranking candidate individuals to target. While such heuristic approximations are computationally feasible, implementing them requires knowledge of the network structure, which can be extremely costly to acquire in field settings.[1] This is part of the motivation for studies such as Banerjee et al. (2019a) or Breza et al. (2020), which develop methods for identifying central nodes or approximating the network structure without conducting a thorough census. Here, our goal is *not* to identify the central individuals, but instead to quantify the value of doing so. We are interested in questions such as: when is it important to target central individuals? What is the value of having access to the network information? And how does this value compare with the cost of seeding?

The main contribution of this paper is to recast the benefit of following a network-guided seeding heuristic in terms of the extra seeds required for a heuristic that ignores the network structure to perform just as well. We show that for the widely studied 'SIR' model of diffusion and many of its extensions, seeding a slightly larger number of individuals randomly may be more economical than network-guided targeting. In particular, we prove that there are only two scenarios that the diffusion process can follow. In one scenario, the diffusion reaches a non-negligible fraction of the population, and the difference in the expected diffusion of the optimal seeding strategy and random seeding is exponentially decaying in the number of extra seeds that the random seeding strategy uses. Consequently, seeding a slightly larger number of individuals randomly can prompt a larger diffusion than seeding by optimizing over the network structure. In the second scenario, even the optimal seeding strategy diffuses to only a vanishing fraction of the population, which means there is not much value in network targeting regardless of the seeding strategy. Such results hold in simulations on real-world network data and some diffusion models studied in the development economics literature.

In our model, we consider a population of n individuals (or nodes) who are connected to each other through a social network. Individuals are either informed or uninformed about some product. The information percolates in the network according to a variant

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

of the ubiquitous Susceptible-Infected-Recovered (SIR) diffusion model. In this model, individuals behave in a “mechanical” fashion. At time $t = 0$, all individuals (nodes) other than a small group (seeds) selected by the policymaker are initially uninformed. Once informed at time t, a node has one chance to speak to each of its uninformed neighbors. This information sharing is successful with probability c independently for each neighbor, in which case the corresponding neighbors become informed by time $t + 1$. This cascade of information continues until no new individual has the opportunity to become informed.

To quantify the value of network information in a policy-relevant way, we consider the following thought experiment: Suppose in one setting, we have access to full network data and unlimited computational power to optimally pick s individuals as initial seeds. In the second setting, we ignore the network and simply pick $s + x$ initial seeds uniformly at random. For what value of x will random seeding inform as many individuals, in expectation, as the optimal seeding?[2]

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

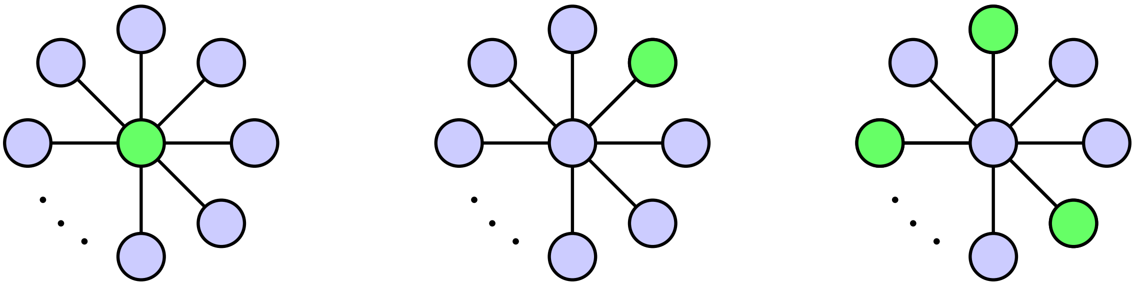
Our main theorem shows that under one set of conditions, the difference in expected fraction of informed individuals between the random seeding strategy with $s + x$ seeds and the omniscient strategy with s seeds vanishes exponentially in x. Thus, the random seeding strategy with $s + x$ seeds asymptotically performs as well as the omniscient strategy with s seeds, for a small x. Precisely when those conditions fail, even the omniscient seeding produces an expected diffusion that reaches only a vanishing fraction of individuals.

This suggests learning the network is not valuable if seeding costs are small relative to the costs of collecting and analyzing network data. A careful seeding strategy has the potential of outstripping a random strategy with additional seeds only in the state of the world where even the best strategy fails to reach a significant fraction of the population.

This theorem holds for the general *Inhomogeneous Random Networks* (IRN) model[3], which subsumes several well-known random network formation models as its special cases.

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

[3] In this model, there is an arbitrary set of *types* for nodes and an agent of type i is connected to an agent of type j with some probability p_{ij}. We do not impose any restrictions on these probabilities. We explain this model in detail in Section 2.1.



(a) Optimal seeding

(b) Random seeding

(c) Random seeding with additional seeds

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

Thus, this 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).

How does random seeding work well even on networks with highly unequal degree distributions, where it appears that informing one of the few highly central nodes can be very 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. We show through simulations that these asymptotic theoretical results materialize in such networks. For the Indian village household networks of Banerjee et al. (2013), the Chinese village rice farmer networks of Cai et al. (2015), and a small subnetwork of Facebook, we verify that random seeding competes well with both typically proposed and omniscient targeting strategies. For instance, in the Facebook subnetwork with 4039 nodes, if each node speaks to her neighbors with probability 5%, random seeding with 5 seeds prompts a larger diffusion than all seeding strategies—including the omniscient seeding—with one seed. A similar result holds even for smaller networks: In an Indian village network with only 99 nodes, when each informed node on average speaks to two of her neighbors, random seeding with 3 extra seeds beats omniscient seeding with one seed.

We next explore the robustness of our main theorem to alternative objectives, namely, the variance and the speed of diffusion.

Variance in diffusion size 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. Our next result proves that this is not the case. Variance of random seeding goes to zero at an exponential rate in the number of seeds. In fact, our simulations on Indian village networks show that random seeding with four extra seeds first-order stochastically dominates *any* seeding strategy with one seed.

Speed of diffusion is another important consideration when policymakers are concerned with the rate of adoption, rather than just the eventual reach of a new product. Indeed, insofar as imitation of neighbors' technology is a driving force of local economic growth, the speed of diffusion may be a primary concern.[4] We consider a variant of the diffusion model in which the diffusion process ends after $T \geq 1$ periods and find the differences between random and omniscient strategies can be larger than in an unbounded diffusion model. Even in relatively homogeneous networks, random seeding strategy needs an order $\log(n)$ times as many seeds as an omniscient strategy to perform just as well. In networks with highly central nodes, the comparison is even less favorable. 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 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. Random seeding works well when there are multiple rounds of communication.

Finally, we investigate the robustness of our results with respect to the model of diffusion. Our simulations show that similar results hold for some of the more complex models estimated in the development economics literature, which generalize the SIR model in different ways. For example, for a version of the model of diffusion studied and estimated in Banerjee et al. (2013), random seeding with 11 seeds performs nearly as well as central seeding with 10. For the 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. Our main result naturally goes through in the "game-theoretic" model of Sadler (2020). In addition, the model we studied exhibits undirected relationships and communication. We prove that our main result holds in a model of random directed relationships and communication, so long as communication probabilities are symmetric.

[4] This has been theoretically studied in the growth literature, see (Alvarez et al., 2013; Perla and Tonetti, 2014)

Still, there are many diffusion models for which targeting central nodes can be valuable. Importantly, suppose central individuals communicate to their connections with a higher probability (or more convincingly) than other people or listen to fewer friends than they speak to. Then, it may be especially important to target such individuals. Scenarios like these could be why some policymakers and firms are willing to pay a lot to target "influentials." In addition, 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 a few extra 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 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. Our paper is connected to the large body of literature around influence maximization. This problem was originally motivated by product diffusion and viral marketing. Domingos and Richardson (2001) is one of the firsts to introduce the influence maximization problem in the context of viral marketing, and other papers followed by documenting contagion in consumer networks.[5] Kempe et al. (2003) formally introduce the influence maximization problem. They consider two common diffusion models and asks how difficult it is to generally solve for the optimal size k set of initial targets when the objective is maximum contagion. They show that computing the optimal set is NP-hard. Leskovec et al. (2007) take this problem to data and discover patterns of influence by studying person-to-person recommendations for books and videos.

The computational complexity of this problem, on the one hand, and its practical importance, on the other hand, led to developing algorithms for influence maximization over networks in a wide range of disciplines. For examples in computer science and operations research, see Chen et al. (2009); Goyal et al. (2011); Chen et al. (2016); Wilder et al. (2017); in health-care, see Rice (2010); Rice et al. (2012); Kim et al. (2015); Yadav et al. (2016); and in physics, see Kitsak et al. (2010); Chen et al. (2012). Watts and Dodds (2007), like us, dispute the idea of targeting influentials, but in a conceptually and methodologically different way. They use simulations, and consider a single seed to compare different heuristics. In contrast, we theoretically quantify the value of network information as the extra seeds required by random seeing to beat the exact optimum, and identify conditions under which careful seeding may or may not matter.

Mechanisms for social learning and diffusion have long been recognized in the de-

[5] For a review of network studies in marketing and empirical work disambiguating network effects from other confounds see Hill et al. (2006) and Iyengar et al. (2011).

velopment economics (e.g., see Duflo and Saez (2003); Conley and Udry (2010); Dupas (2014)).

The topic of the current paper is most related to a newer literature that studies how network theoretic characteristics of learners and spreaders determine the extent of these processes. Banerjee et al. (2013) pioneer this approach. 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. (2019) study technological adoption by farmers as they vary seeding rules over 200 independent village-networks in Malawi in an experimental setting. Our theoretical model of diffusion pertains more to the diffusion processes observed in Banerjee et al. (2013) and Cai et al. (2015) rather than Beaman et al. (2019), which finds data more consistent with a threshold type diffusion model. While these papers find that network theoretic seeding improves diffusion, we are interested in studying how quickly additional outreach makes up for network-agnostic seeding.

Two more papers in this literature are of direct relevance: Kim et al. (2015) and Banerjee et al. (2019b) conduct multiple RCTs to directly compare random seeding and 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 with random targeting. We note that our results confirm that *fixing the number of seeds* random seeding may perform worse than other heuristics, and thus this experiment is consistent with our theoretical findings.

Banerjee et al. (2019a) conduct two randomized control trials to compare random seeding with "deliberately seeding" the network. In particular, they 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 the second experiment, they looked at the rate of vaccination as the outcome. They find that with random seeds, 18.11 children attended and received at least one shot. In villages with gossip-based seeding, this number was 23.

These findings are consistent with our results. As we argue and simulations confirm, *fixing the number of seeds*, random seeding typically performs worse that network-guided heuristics. Banerjee et al. (2019b) do not test how many extra seeds random needs to compete with other heuristics.[6] However, Banerjee et al. (2019b) consider one more

[6] We could try to guess the extra number of seeds needed for random seeding to beat other heuristics in their second experiment. With six initial network-guided seeds, total diffusion is 23. With six random

experiment in which SMS "blast" reminders are sent to 33% and 66% of village households selected at random. They find that the SMS blasts did not lead to greater adoption than targeted seeding. However, individuals in gossip-guided seeding were contacted by phone and given regular personalized reminders, whereas random seeds were contacted through SMS blasts. These incomparable modes of communication do not constitute a clean test of the hypothesis that random seeding with a few extra seeds may perform well. Indeed, as Banerjee et al. (2018) find, broadcasting information to a large group of people may change the dynamics of information acquisition.

Of particular relevance is a game-theoretic diffusion model studied in Sadler (2020). Here, agents hear information, update their beliefs about their network position in a Bayesian fashion, and subsequently choose whether or not to adopt a product. Sadler (2020) exploits percolation results to show if the diffusion process reaches a positive fraction of the population, there will be a giant component of informed individuals. The resulting diffusion process is similar to what we study here, and our results on seeding would go through in this context. Indeed, our object of study is the comparison of seeding strategies for any model that produces the patterns of transmission studied here, whether these arise mechanically or from a game.

Finally, a theoretical literature in economics studies the optimal seeding problem under various diffusion processes and competition in diffusion (Morris (2000); Galeotti and Goyal (2009); Young (2009); Goyal et al. (2014); Bloch et al. (2014); Lim et al. (2015); Mobius et al. (2015); Sadler (2020); Galeotti et al. (2017); Banerjee et al. (2018)). Meanwhile, other papers describe game-theoretic foundations for the traditional measures of centrality (e.g., Ballester et al. (2006); Bloch et al. (2016); Bloch (2016)) or role of influential nodes (e.g., Galeotti and Goyal (2010)).

Organization of the paper. We introduce our diffusion and network models in Section 2. In Section 3, we present our main theorem, as well as its many special cases. We also present simulations on real-world networks. In Section 4, we study robustness and limitations of the results. We consider alternative objectives (variance and speed of diffusion) as well as alternative diffusion models. In Section 5, we discuss our results and conclude.

seeds, total diffusion is 18.11. In the experiment, each random seed is inducing more than 3 households to participate. Thus, if diffusion of random seeding was linearly increasing, 2 extra seeds would be enough to perform better than the network-guided seeding. Conservatively, if each random seed only induces 2 households to participate (the seeded household plus one more household), then 3 extra seeds are enough for random to surpass the network-guided strategy.

2 Model

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

Diffusion process. Time passes in discrete periods $t = \{0, 1, 2, \ldots\}$. An agent is either *informed* or *uninformed*. Once an agent becomes informed, it remains informed forever after. Initially, a subset $A_0 \in N$ of individuals are informed. Once informed at time t, an agent has one chance to speak to each of its uninformed neighbors. We focus on the case that an informed individual has only one chance to speak to her neighbors, but this can be easily generalized to multiple chances. This information sharing is successful with probability c independently for each neighbor, in which case the corresponding neighbors become informed by time $t + 1$. Diffusion continues until no new individual has the opportunity to become informed. Our main theorem is stated for this diffusion model. In Section 4.1.2, we consider cases where all communication ceases after some $T \geq 1$ periods. The case where T is finite is called a *bounded* diffusion process. Otherwise the diffusion process is called *unbounded*.

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

The diffusion process considered here is one in which communication is undirected. In particular, the event that node i talks to j if informed is coupled with the event that j talks to i if informed. Moving from undirected communication settings to directed communication requires addressing some technical issues. We postpone the discussion of this case to the Section 4.2.1, where we discuss under what conditions our results can be extended to settings with directed communication.[7]

Importantly, the diffusion model studied here is a "mechanical" model in which individuals' incentives are not explicitly considered. We focus on this model simply because it is a workhorse model of the diffusion literature. We will discuss how our results alter by considering other models of diffusion in Section 4.2.

Seeding strategies. A seeding strategy takes as input a network and a constant number of initial seeds $s \leq n$ and outputs a (random) set of s initial seeds to be informed at

[7] In addition, simulations of Appendix F and Appendix G consider models of directed communication.

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

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

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

Let $\mathbf{h}(G, s, f) = \mathbb{E}[|A_T(G, s, f)|]$ be the expected number of informed agents at the end of the process. Here the expectation is taken over the diffusion process. Let $\mathbf{H}(f, s) = \frac{1}{n}\mathbb{E}_{G\sim\mathbb{P}_n}[\mathbf{h}(G, s, f)]$, where $\mathbb{P}_n$ is a network formation process—a probability distribution over all possible networks of size n. The function $\mathbf{H}$ measures the performance of a seeding strategy by taking the strategy and number of seeds as inputs and producing the *expected fraction of informed agents* as output, for a given network formation process. The goal of the planner is to choose a seeding strategy f to maximize $\mathbf{H}(f, s)$.

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

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

For a given network formation process, this strategy depends on $\mathbb{P}_n$. For a given $\mathbb{P}_n$, OPT solves:

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

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

Let RAND(s) be the strategy which picks s nodes uniformly at random in G. Implementing this strategy does not require any information about the network structure. To quantify the value of learning the network and identifying the optimal seeds, we would ideally like to compare the performances of OPT and RAND. Recall that OPT exploits the full knowledge of the structure of the network and solves a computationally hard optimization problem, while RAND ignores any information about the network. Therefore, the difference between these two can be interpreted as the value of network information and analysis.

As noted earlier, however, computing OPT is an NP-complete problem. Instead, we measure the difference between the performances of the *omniscient* seeding strategy and RAND. The omniscient seeding strategy, denoted by OMN(s), is a strategy that for every realization of the communication network picks s initial seeds to maximize diffusion. Notice that this strategy is infeasible by construction because it knows who is going to speak to whom, and it performs better than any feasible strategy for any realization of the diffusion process. In particular, for any initial number of seeds s:

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

Since for any realization of the diffusion process, OMN performs better than OPT, comparing RAND and OMN provides an upper bound for the value of network information and analysis.

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. This is a rich framework, as the set of types can be arbitrarily general. For instance, in a college network, types can represent major, cohort, gender, and race. Types can also represent the ages of individuals in the population, their genders, occupations, or combinations thereof. 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

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

Let $\boldsymbol{p}(\kappa)$ be a matrix with p_{ij} as its elements. Then, $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$ 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, the *spectral radius* of $\mathbf{T}_\kappa$ is an eigenvalue of $\mathbf{T}_\kappa$ and all other eigenvalues of $\mathbf{T}_\kappa$ have a strictly smaller absolute value. Therefore, the largest eigenvalue $||\mathbf{T}_\kappa||$ can be computed as

$$||\mathbf{T}_\kappa|| = \sup_{\mathbf{x}:||\mathbf{x}||_2\le 1} ||\mathbf{T}_\kappa \mathbf{x}||_2,$$

where $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{\mathcal{T}} x_i^2}$.

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

Erdős-Rényi networks. In an Erdős-Rényi random network on a set N of nodes with $|N| = n$, there is a link between a pair of agents $(i,j) \in N^2$ with probability d/n, independently of other agents and links. This structure is perhaps the simplest and most widely used random network model. The average degree of nodes in this model is 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 captures the idea that people are more likely to be connected to their own "type" of people—a feature that is referred to as *homophily* (see Jackson (2010), Chapter 6). For instance, Stanford college students are more likely to be connected to each other than to UC Berkeley college students. The islands model is another special case of the IRN model. In this model, each type has the same number of agents and an agent only distinguishes between agents of one's own type and agents of a different type. More precisely, an $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$ is an islands network with parameters (m, d_{in}, d_{out}) 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. We also consider the class of networks introduced in Chung and Lu (2002) which generalizes the Erdős-Rényi model by supporting any degree distributions.

Fix a sequence $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$. A *Chung-Lu* (undirected) network on n nodes, $CL(n, \mathbf{w})$, is generated by including each edge $\{i,j\}$ independently with probability $p_{ij} = \min(\frac{w_i w_j}{\sum_k w_k}, 1)$. For the convenience of notation, we will assume that $\max_k(w_k^2) \le \sum_k w_k$, so we don't have to take the maximum with 1. These two variations are known to be equivalent asymptotically (see Van Der Hofstad (2016), Section 6.6).

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,

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

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 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 $\tau = b - 1$, see (Van Der Hofstad, 2016) for a more detailed discussion.

Chung-Lu model fits into the framework of the IRNs. However, since the support of the degree distribution can be a continuous variable, the set of types is not necessarily finite. Incorporating a continuum of types into the IRN model is straightforward, but requires developing a measure-theoretic setup, which we do in the Appendix B.1.

3 Main Theorem

The resource constraint in the optimal seeding problem is the *number of seeds*. Therefore, to quantify the value of network information in a policy-relevant way, 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?

We emphasize that our goal here is not to recommend random seeding as a seeding strategy to be used in practice—in fact, there are cheap ways to perform better than random. We study random seeding as a tractable theoretical benchmark that does not require information about the social network. Of course, even random seeding requires *some* information to be feasible; for instance, in the context of villages in developing countries, the policymaker at least needs access to a list of all households in the village.

Let $\alpha = \lim_{n\to\infty} \mathbf{H}(\text{OMN}, 1)$ be the fraction of nodes informed by the omniscient seeding with one seed. This constant is important in our analysis, as shown below in the statement of the main theorem.

Theorem 1. *Consider a sequence of* $\text{IRN}_n(\boldsymbol{p}(\kappa))$*. 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,

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

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

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

There is a lot to discuss about this theorem. It is clear that the essence of the theorem is in the $||\mathbf{T}_\kappa|| > 1/c$ condition. Also, the parameter α call for some discussions. It is easier to discuss these once we clarify the role of each condition in the proof. Thus, we first sketch the ideas of the proof. The formal proof can be found in Appendix A.

Proof overview of Theorem 1. Recall that the communication network $\mathcal{K}(G) \subseteq G$ is a way to think about the set of all pairs of agents who will speak to each other, once one of them becomes informed. We can consider the connected components of this communication network to better understand the behavior of random and omniscient seeding strategies. Note that in the SIR model, a node becomes informed if and only if one of the nodes in its connected components in $\mathcal{K}$ is seeded. This implies that an omniscient seeding strategy with s seeds would simply seed one node in each of the s largest connected components of $\mathcal{K}$. On the other hand, for each seed, the probability that the random strategy informs a given component is proportional to the component's size. This gives us a method of computing the expected diffusion for each of the strategies, once we are given the 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, meaning that informing one node in that component is enough to inform a constant fraction of the population through information cascade. The remaining components of the communication network, on the other hand, are vanishingly small*i.e.*, $o(n)$) in population size.

Therefore, the omniscient seeding strategy informs the constant size component with only one seed. With the additional seeds, it picks the largest of the small components, but with s seeds the total fraction of informed nodes cannot be more than $o(n)s/n$, which is asymptotically 0. Thus, $\lim_{n\to\infty} \mathbf{H}(\mathrm{OMN}, s) = \lim_{n\to\infty} \mathbf{H}(\mathrm{OMN}, 1) = \alpha$.

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, and the other components are irrelevant. Therefore, the limit ratio of random with $s + x$ seeds and omniscient with s seeds is $1 - (1 - \alpha)^{s+x}$, where α is the size of the constant size component.

On the other hand, 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 cascade) $o(n)$ other agents. Hence, the omniscient seeding strategy with s seeds can at most inform $o(n)s/n$ fraction of the population. □

The proof overview shows that the $||\mathbf{T}_\kappa|| > 1/c$ condition is equivalent to having a *giant component* with size α in the communication network, while its failure means all components are (very) small. Clearly, as communication probability increases, the condition is more likely to be satisfied. What the condition $||\mathbf{T}_\kappa|| > 1/c$ implies about network structure becomes clearer for the special cases of IRN networks considered in Section 3.1 to Section 3.4.

A few remarks about the theorem worth pointing out:

On asymptotic results. Stating clean mathematical results for diffusion on random networks is typically feasible only in the limit as $n \to \infty$. That is why our main theorem is a limit result. One may question the relevance of this asymptotic result for small networks, such as those in development economics studies. This is a valid concern, and we address it in our simulations. We conduct a series of simulations in small networks. The findings confirm that our theoretical result holds far from the limit. For example, we show that for an Erdős-Rényi network with 100 nodes and when $cd = 1.5$ (where d is the average degree), random with 3 extra seeds performs better than omniscient with one seed.

On $||\mathbf{T}_\kappa|| < 1/c$ ***regime.*** Our results show that in the regime where the limit fraction of informed individuals is zero, random seeding with a few extra seeds cannot beat omniscient. As discussed in Section 3.5, in applications such as the microfinance diffusion in Indian villages or weather insurance diffusion in Chinese villages, the diffusion is indeed in the regime where a giant component emerges, so random seeding performs well.

Nonetheless, for most marketing campaigns on Twitter or Instagram, the fraction of informed individuals is negligible relative to the size of these networks. Does it mean that the $||\mathbf{T}_\kappa|| < 1/c$ regime is the relevant regime for these networks? It depends. When measuring the fraction of informed individuals, it is important to identify the 'relevant' network. An economics paper may go viral in Twitter in economists' subnetwork, but even if all economists are informed of the paper, the total diffusion is still a negligible fraction of the Twitter network. Therefore, the virality should be evaluated with respect to the agents for whom the information has relevance.

Having said that, we emphasize that $\lim_{n\to\infty} \mathbf{H}(\mathrm{OMN}, s) = 0$ should not be read literally when dealing with small networks. The omniscient strategy can indeed reach up to an $s \times O(log(n))$ many nodes, which may amount to a sizeable fraction of nodes in small networks.

However, it is important to note that the omniscient seeding strategy is too strong of a benchmark for the $||\mathbf{T}_\kappa|| < 1/c$ regime, as it is able to precisely pick the largest (of small) components. In practice, the fact that random cannot compete well with omniscient does not necessarily mean that it cannot compete with typically used network-guided heuristics. In Appendix H, we conduct simulations on Erdős-Rényi networks and Indian village networks and show that random seeding, while being far from omniscient, competes well with typical seeding strategies even in this regime.

How many extra seeds? Suppose our goal is to ensure that the limit ratio between RAND and OMN is at least $1-\epsilon$. Then, it must be that $1-(1-\alpha)^{s+x} > 1-\epsilon$, and thus, having $\max(0, \frac{\log(\epsilon)}{\log(1-\alpha)} - s)$ extra seeds is enough to guarantee that RAND performance is at least $1-\epsilon$ of the OMN.

The limitation on the number of seeds. In stating Theorem 1, we considered any constant number of seeds $s \leq n$. In particular, we did not allow the number of seeds to grow with n. We can relax this requirement for the first part of the theorem, when $||\mathbf{T}_\kappa|| > 1/c$. In this setting, the size of the second-largest component of any IRN is $O(\log(n))$, and thus 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 is asymptotically zero. For the second part of the theorem, however, we cannot let s grow faster than a constant in general since the exact sizes of the small components depend on parameters. That said, 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.

We emphasize that the limitation on s is an artifact of using OMN as a bound, which is at the end too strong of a benchmark. In fact, for commonly used seeding strategies and the OPT strategy, our simulations suggest that flooding many individuals with information makes a careful selection of initial seeds *less* valuable, and random has an easier time catching up with network-guided strategies when s is large. As such, having a constant number of seeds (and, as an extreme case, if $s = 1$) is a 'worst-case' scenario for random seeding.

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 $s > 0$,

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

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.

3.1.1 Small Erdős-Rényi networks and the exact optimum

Computing OPT is NP-hard in general. 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 with only two extra seeds beats the optimum with one seed. To beat omniscient, random needs 4 extra seeds. Similar numbers are enough for random to catch up when $cd = 2$: random needs 3 extra seeds to beat OMN and 2 extra seeds to beat OPT. These findings show that the theoretical limit results quickly kick in.

3.2 Power-law Chung-Lu Networks

Several real-world networks are characterized by degree distributions with fat tails, in the sense that they exhibit few nodes that have significantly greater degrees than others. For example, Barabasi and Albert (1999) describe a variety of social networks, such as the network of linked web pages or collaborating actors, exhibiting a power-law like degree distribution on its right tail. 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 B.

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, means 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. This raises the question of how random seeding—which is going to miss highly central nodes with high probability—can compete with omniscient seeding. 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, which is another special case of the IRNs. 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}). 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 does not necessarily imply that the condition of Corollary 3 is (or is not) satisfied. For instance, when $d_{out} = 0$, so there are no cross-group links and the network exhibits extremely homophily, the average degree becomes d_{in}/m. If $c\frac{d_{in}}{m} > 1$, the result of our 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 $d_{in} > 1/c$. This illustrates that there is no immediate relationship between homophily and the conditions required for random seeding to perform close to optimal.

3.4 Erdős-Rényi Networks with Clustering

In all of the network models considered so far, the probability that a node i is connected to some node j is independent of whether they are both friends with k. In many real-world networks, on the other hand, having a common friend increases the probability of connection. We now present a simple network model that admits clustering to prove that existence of this property—known as *clustering*—will not affect our results. Our simulations of the next section show that even in real-world networks with significant clustering, our results hold.

Here, we consider a network formation model that allows for higher clustering. Incorporating clustering into IRNs is technically challenging, and thus we state this last result for a new model of network formation that includes clustering into a generalized version of the Erdős-Rényi model.[8] As in the model of Jackson and Rogers (2007), nodes meet each other randomly at first and then make a few random friendships with the neighbors of their initial neighbors, which can be thought of as a natural model of how clustered relationships arise. To that end, we define a k-level random network in the following way.

Definition 1 (k-Level Random Network). *Let $\phi = (\lambda, q_1, \ldots, q_k) \in [0,1]^{k+1}$. A k-level network on n nodes, denoted $L_n(\phi)$, is constructed in two steps: first, sample a random graph G_n from the family of Erdős-Rényi networks with n vertices and average degree d. Second, include for every node, a link with one of its neighbors of neighbors with probability $1 - \sqrt{1-q_1}$, a link with one of its neighbors of a neighbor of a neighbor with probability $1 - \sqrt{1-q_2}$ and so on up to k.*

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

[8] Even then, we cannot theoretically study this model for the case when the total diffusion size is vanishingly small, and rely on simulations in that regime. See Appendix I for simulations.

Corollary 4. *Consider a k-level random network with base-level average degree d. If $cd > 1$, then for any x and s,*

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

We prove this Corollary in Appendix C.1.

This result suggests that even in networks with clustering, the gap between random with x extra seeds and omniscient seeding is exponentially small in x.

3.5 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. A curious reader may wonder whether our results are robust with respect to the network formation models we considered here. To address this concern, we now offer a (network formation) model-free perspective on the main theorem. We simulate the diffusion model studied here on the microfinance network data in Banerjee et al. (2013) as well as a subnetwork of Facebook, and compare the performance of various seeding strategies.

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

Simulations in Figure 2 compare the average performance of random, degree-central, diffusion-central[9], eigenvector-central, and omniscient seeding strategies on village networks, which includes an edge between two households whenever either party indicated some contact with the other group of any form. Results are included for two different values of communication probability c. In both cases, random with a few extra seeds can compete well with network-guided seeding heuristics. For instance, when $c = 0.1$, random with 5 seeds performs as well as degree- and diffusion-central seeding with two seeds, and

[9] 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).

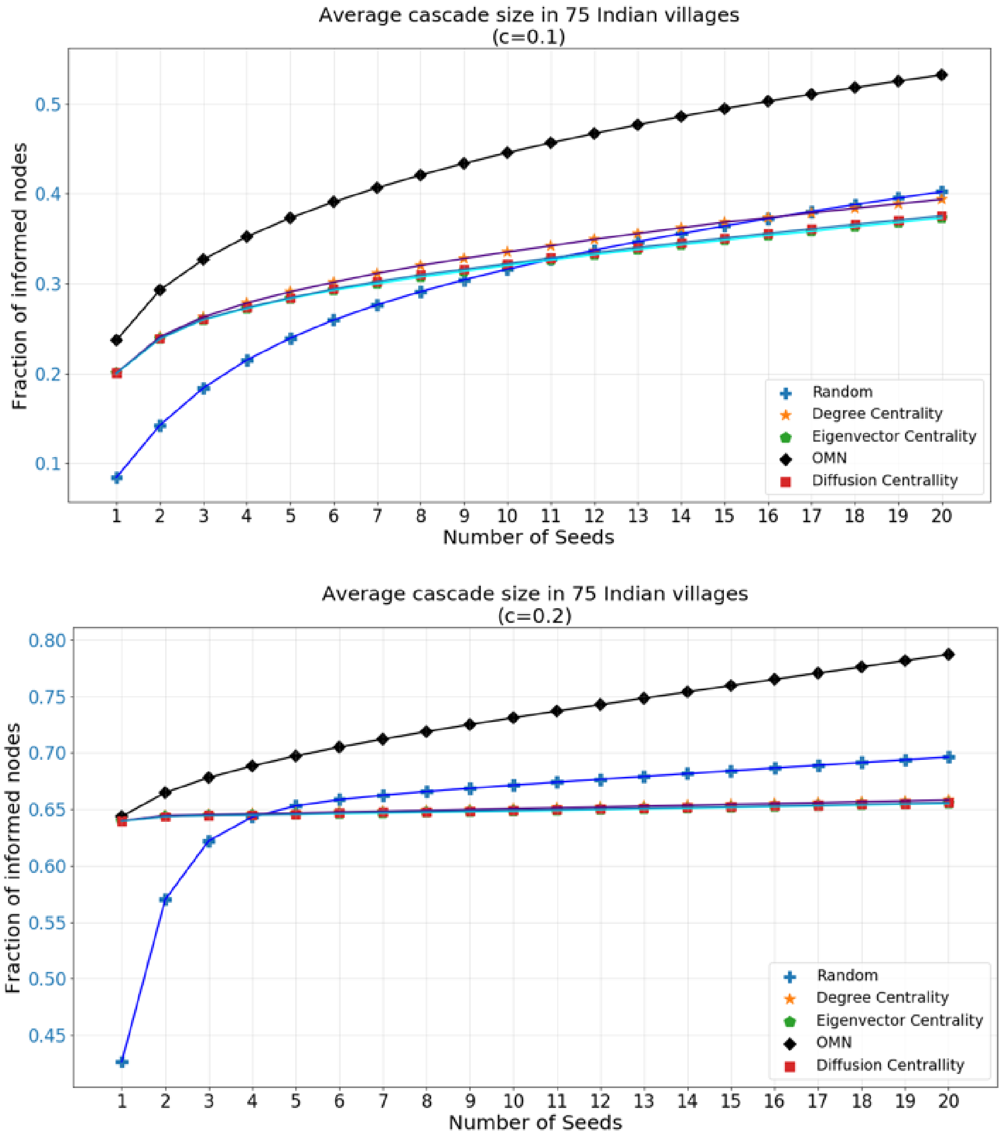


Figure 2: A comparison of average diffusion for various seeding strategies (omniscient, random, degree-, diffusion-, and eigenvector-central seeding) across 'all inclusive networks' in the village network data, for two different levels of communication probabilities.

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. We discuss the observation that, fixing the number of seeds, random beats network-guided heuristics in Section 5.

Comparison to the OPT. With only one seed, we can calculate the exact OPT numerically. We compute OPT for a sample Indian village network. Our simulations show that when $cd = 1.5$, random with 3 extra seeds beats both OPT and OMN. When $cd = 2$, meanwhile, random with 2 and 3 extra seeds beat OPT and OMN, respectively. In our discussion of variance in Section 4.1.1, we show that random with 4 extra seeds empirically first-order stochastically dominates both OPT and OMN with one seed.

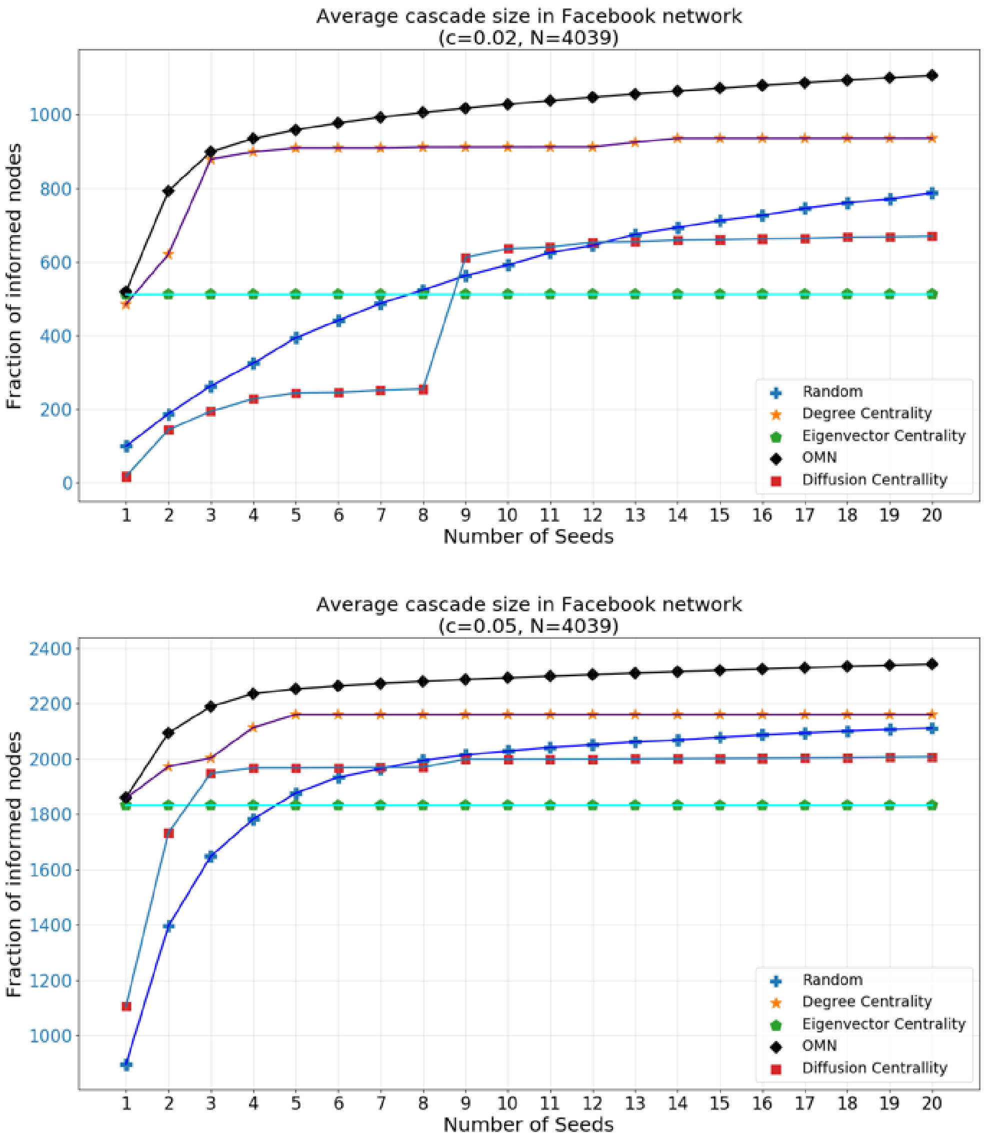


Figure 3: A comparison of average diffusion for various seeding strategies (omniscient, random, degree-central, diffusion-central, and eigenvector-central seeding) in a subnetwork of Facebook, for two different levels of communication probabilities..

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

A virtue of random seeding. Simulations on Facebook network show that for 12 or more seeds, random beats diffusion- and eigenvector-central seeding with an equal number of seeds. These simulations highlight a more general point that when the number of available seeds is not too small, random seeding can perform *better* than centrality-guided seeding heuristics. 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. As the number of seeds increases, seeding an additional individual in the giant component is less valuable than seeding individuals in small components. 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. In this sense, the variance of the performance of a seeding strategy is an important measure. Here we show that random seeding can compete with network-guided heuristics not only in expected value, but also in variance. 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*

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

We prove this proposition in Appendix E.

Proposition 1 shows that the variance of random seeding is exponentially small 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.

Small networks and stochastic dominance. Proposition 1 proves that random seeding as a small variance in large networks. We verify that the same thing holds in small, real-world networks by conducting simulations on the Indian village networks. In fact, our simulations show something even more powerful: random seeding with just a few seeds more beats (first-order) stochastically dominates OPT and OMN, as shown in Figure 4. Here, we simulate an SIR diffusion model with $c = 0.15$ on an Indian

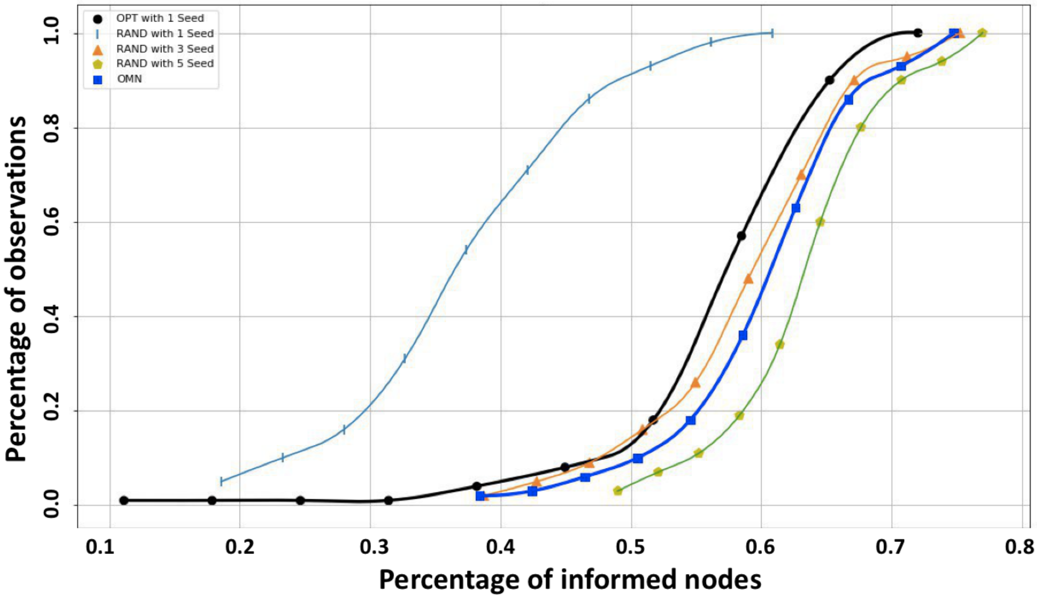


Figure 4: The empirical CDF of the percentage of informed nodes for various seeding strategies in a sample Indian village network. Random with 3 seeds stochastically dominates OPT with 1 seed and random with 5 seeds stochastically dominates OMN with 1 seed.

village network with 99 nodes and depict the CDF of the percentage of informed nodes in different simulations. Random with only one seed performs poorly, but random with 2 additional seeds stochastically dominates OPT with one seed, and random with 4 extra seeds stochastically dominates OMN with one seed.

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 allows policymakers to significantly accelerate the speed with which the information spreads. As an example, policymakers may be concerned with how quickly farmers adopt a new technology, so that the developing economies may grow at faster rates.

To address this question, we now consider a bounded diffusion process, where the diffusion stops after T periods. We then ask: can random seeding with extra seeds beat 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 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.

Nonetheless, our theoretical result in this section shows that at least for Erdős-Rényi networks (potentially with clustering), with $O(\log(n))$ times additional seeds, random

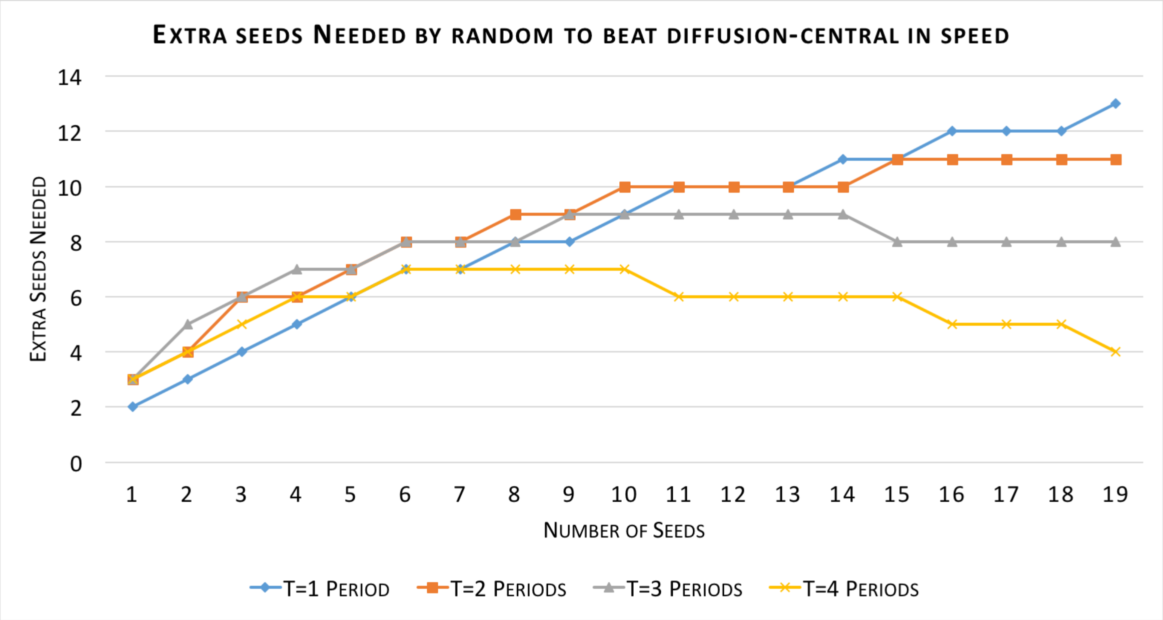


Figure 5: Average number of extra seeds required by random to outperform diffusion-centrality seeding in Indian village networks (in of *speed of diffusion*). If objective is diffusion in the first T=1 or T=2 periods, then extra seeds required is relatively high (still less than 15), but once total outreach in the first T=3, T=4 or more periods is the objective, less than 9 extra seeds is enough.

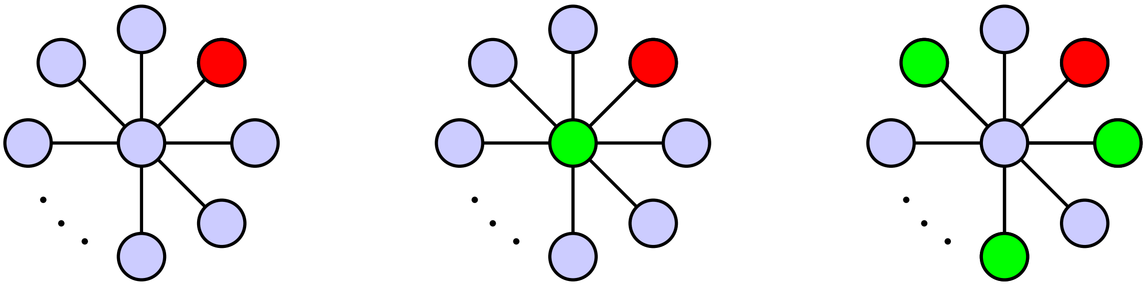
seeding competes with the omniscient seeding even in the speed of diffusion. For the proof see Appendix C.2.

Proposition 2. *Consider an Erdős-Rényi network on n nodes with constant average degree d and a diffusion process with communication probability c that ends in $T \geq 1$ periods. Then, with high probability, for every s,*

$$\mathbf{H}(\text{RAND}, 2T^3 \log(n)s) \geq \mathbf{H}(\text{OMN}, s).$$

As described in the appendix, this bound straightforwardly extends to a model of graphs with clustering inspired by Jackson and Rogers (2007), where friend of friends and friend of friend of friends etc. are likely to be one's direct neighbors as well. In this sense, clustering does not change the performance comparison between omniscient and random seeding in speed of diffusion.

For power-law networks, this result fails, since for $T = 1$, when only the first period diffusion matters), it is clearly important to identify highly central nodes. Thus, extra seeds required by random seeding to compete with network-guided heuristics in the speed of diffusion depends crucially on the underlying network structure. We now investigate this question in the context of Microfinance diffusion of Banerjee et al. (2013).



(a) No vaccination

(b) Optimal vaccination of 1 agent

(c) Random vaccination of several agents

Figure 6: Random strategy with a few additional individuals can perform poorly when the goal is to 'vaccinate' individuals to halt the diffusion. Consider a star network with n leaves, for some large n. Suppose some random individual gets infected with some disease (the red node), and any infected node infects its neighbors with probability $c = 0.5$. The goal is to vaccinate a single individual to minimize diffusion. 6(a): Without vaccination, the central node will be infected with probability 0.5, and thus $\frac{n}{4}$ of agents get infected in expectation. 6(b): Vaccinating the central node is optimal, as it stops the diffusion completely. 6(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.

Speed of diffusion in Microfinance. Figure 5 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. As noted above, the linearly increasing bound proved in Proposition 2 is for the worst-case scenario, when $T = 1$.

When $T = 3$ and $T = 4$, the extra number of seeds needed for random is always less than 9 and 7, respectively.

4.1.3 Diffusion Minimization by Vaccination

Network information can be highly valuable when a policymaker wishes to *minimize* the spread of some diffusion. This is a relevant point for the diffusion of fake news (or an infections), where a policymaker wants to inform individuals that the news is fake (or to vaccinate them) so that they stop spreading it.

To fix ideas, suppose some random individual is infected with a disease, and the diffusion process is the diffusion model studied in this paper. A policymaker seeks to 'vaccinate' a group of individuals to *minimize* the extent of the diffusion. It is known that it is important to pick the optimal individuals for vaccination (Bollobás and Riordan, 2004; Drakopoulos et al., 2016). In fact, we conjecture that the number of additional individuals that we need in order for random vaccination to beat the optimum can be as large as a constant fraction of all agents. Figure 6 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 the undirected SIR model of diffusion, which is used to study processes such as diffusion of information and ideas, rumors, or infectious diseases. We focused on the SIR model since this is a workhorse model, studied and estimated in several economic environments. We will now discuss some alternative diffusion models under which our results will (or will not) hold.

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. The assumption of undirected relationships and communication may both be called into question. Indeed, it frequently happens in surveys that one individual names another as a close friend, without the other declaring in kind. In addition, even if relationships are undirected, it is not a foregone conclusion that just because one agent would have informed a friend of some information, that the reverse would have occurred had the latter party learned of the information first.

We will now consider a model of directed networks similar to Erdős-Renyi. $D(n, d)$ is 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 3. *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 for which a *directed path* exists from a seed. The analog of

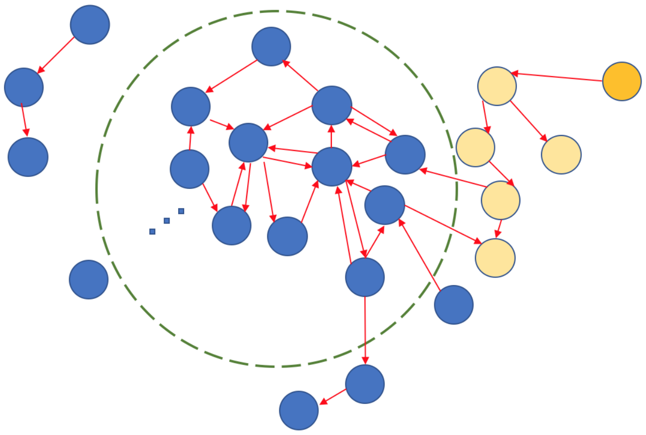


Figure 7: Above is an example communication network when communication is directed. The outgoing edges represent the nodes that a given node would inform if given information. The nodes within the dotted dashed circle represent the strongly connected giant component. The orange nodes, if informed, disseminate 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 3, we show that the size of the set of any cluster of orange nodes (paths to the SGC) is $o(\log(n))$ so that OMN cannot significantly outperform RAND.

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 7). Consider such a node and the length of the shortest path leading from this node into the SGC. If the path length is long, an omniscient strategy would go choose this node as an entry point into informing the nodes in the SGC. But a random seeding strategy with any number of seeds would not hit such a node, other than through sheer luck. Results in Karp (1990) indicate that such paths are $o(\sqrt{n})$ in length, which is too generous of an upper bound for our results to hold. In Appendix D, we establish that these paths are in fact $O(\log(n))$ in length, and can therefore be safely ignored.

4.2.2 Models from development economics

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

Our basic insight goes through for all diffusion models discussed above. To show this, we will consider the diffusion models and the social network data of Banerjee et al. (2013) and Cai et al. (2015) and compare centrality-guided and random seeding strategies. Simulations reported in Appendix F (for the Microfinance model) and Appendix G (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 small, typically less than 3.

When the diffusion process is such that neighbors are "complements", say when several of an agent's neighbors have to adopt a technology before he does the same, our results may fail to hold. For instance, in the *threshold* type models of diffusion, agents will only adopt a behavior if at least a certain number (or fraction) of their neighbors adopt, so there are complementaries in the inputs of propagation. Beaman et al. (2019) study technological adoption by farmers as they vary seeding rules in village-networks in Malawi in an experimental setting. Their result suggests a threshold-type diffusion process, although they observe little diffusion. Since random seeding is unlikely to inform multiple neighbors of the same node, random seeding will fail to prompt any diffusion if thresholds are uniformly high across all agents. This intuition has been subsequently formalized in Jackson and Storms (2017). Typically, these models assume a uniform threshold across agents. However, if thresholds are heterogeneous and sufficiently many agents have a threshold of 1, then results similar to our main theorems may continue to hold.

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 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 models such as the SIR, connection probability ($\kappa(i,j)$) and communication probability (c_{ij}) play a similar. 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.

Results will change, however, if we drop the symmetry assumption; that is, if $c_{ij} \neq c_{ji}$. In some real-world settings, an individual might be valuable to target not because she is central in the network, but because she is more "diffusive" or "persuasive." Our model abstracts from this asymmetry in persuasiveness. Consider the example of Figure 1. Suppose the central agent talks to her neighbors with probability 0.5, but her neighbors talk to her with probability 0.05, so the central node is more persuasive. Then, in order for random to beat the strategy of seeding the central node, it needs 55 extra seeds.

Whether the symmetry assumption is plausible is context-dependent. It is likely a more plausible one in village settings, where two people communicate when they meet or call, and degree distributions are more concentrated than online social networks, where highly connected individuals are typically more likely to communicate to their friends than the reverse. It may also be a more plausible assumption for pure information diffusion than settings where individuals are 'learning' from others.

5 Concluding Remarks

Our formulation for the value of network data can be generalized to settings different from ours. In particular, consider a general network setting, where a research study aims to identify optimal nodes of a network for maximizing diffusion for a given diffusion model. The diffusion model could be the model studied in this paper or generalizations thereof, or any other diffusion model of interest. Suppose the researchers identify a specific seeding heuristic to perform well. These researchers can report the following statistic as a policy-relevant quantity: *How many extra seeds does the random seeding strategy need to be within z% of their proposed strategy, for a small z?*

For example, for the diffusion model of Banerjee et al. (2013) and with $s = 10$ initial seeds, random seeding with 1 extra seed 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).[10] Additional numbers are reported in Table 1.

Whether seeding a few extra individuals is cheaper than collecting and analyzing network data is, of course, context-specific. We quantified the value of network data by comparing it to the number of extra seeds needed, precisely because policymakers are better positioned to compare the costs of the two. However, it appears to us that policymakers would benefit a lot comparing these two methods—expanded outreach versus network targeting—before spending resources to identify the optimal seeds.

[10] 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.

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).

References

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

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

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

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2019a). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 86(6):2453–2490.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2019b). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 86(6):2453–2490.

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

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

Beaman, L., BenYishay, A., Magruder, J., and Mobarak, A. M. (2019). Can network theory based targeting increase technology adoption. *Unpublished Manuscript*.

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. (2014). Rumors and social networks.

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

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. *arXiv preprint arXiv:1703.04157*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hill, S., Provost, F., Volinsky, C., et al. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276.

Iyengar, R., Van den Bulte, C., and Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212.

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

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

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

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

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

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

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *arXiv preprint arXiv:1001.5285*.

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

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.

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

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

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

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

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

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

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

Van Der Hofstad, R. (2020). *Random graphs and complex networks*, volume 2. Cambridge university press.

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

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

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

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

A Proof of Theorem 1

In this appendix, we prove Theorem 1.

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

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

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

and

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

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

Proof of Theorem 1. 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 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 $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$.

The combination of the above result with Lemma 1 implies that for G_n, $\mathbf{h}(G_n, s, \mathrm{OMN}) \leq \mathcal{C}_1 + (s-1)\mathcal{C}_2 = \alpha n + O(\log(n))s$ 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, $\mathbf{h}(G_n, s + x, \mathrm{RAND}) \geq \mathcal{C}_1(1 - (1 - \frac{\mathcal{C}_1}{n})^s) \geq \alpha n(1 - (1 - \alpha)^{s+x}) + o(n)$, with high probability. Taking expectations over the realizations of G_n:

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

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

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

B A General IRN model and 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 introduces the appropriate model.

B.1 Infinite Type IRNs

This section closely follows Van Der Hofstad (2016) to extend the IRN model to a setting with potentially infinite type-space.

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 $n \in \mathbb{N}$, $\mathrm{IRN}_n(\boldsymbol{p}(\kappa))$ is the random network on $[n] = \{1, 2, \cdots, n\}$, where each possible link ij, $i, j \in [n]$ is present with probability

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

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 $A \subsetneq \mathcal{T}$ 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 Van Der Hofstad (2016) that is useful in extending Theorem 1 to infinite type spaces:

Theorem 2 (Van Der 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 $IRG_n(\boldsymbol{p}(\kappa_n))$. Then $|\mathcal{C}_1|/n \to \alpha$ for some $\alpha \in [0,1]$. Moreover, $\alpha > 0$ if and only if $||\mathbf{T}_\kappa|| > 1$.*

B.2 Proof of Corollary 2

In this section, we prove 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 easily along the same lines as the analysis of the connected components of Chung-Lu graphs. The reader can also see Section 3.5.2 of Van Der Hofstad (2020) for more details.

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 2, 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 Van Der Hofstad (2020) 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 connected component with size linear in n. Lemma 1 implies Corollary 2 for this case similarly.

C Clustering and Speed of Diffusion

This appendix shows that the main insight goes through for a model of networks with clustering as stated in Corollary 4. We will subsequently prove Proposition 2.

C.1 Proof of Corollary 4

Let G be the k-level network and G_1 its base graph. The base graph G_1 is an Erdős-Rényi network with n vertices and average degree d. Since $cd > 1$, with high probability, the communication network $\mathcal{K}(G_1)$ has a unique connected component $\mathcal{C}_1$ of size $C_1 n + o(n)$ and the rest of its connected components $\mathcal{C}_2, \mathcal{C}_3, \cdots \mathcal{C}_k$ are of size at most $O(\log(n))$.

Now, note that by definition, $\mathcal{K}(G_1) \subset \mathcal{K}(G)$. Furthermore, the connected components of $\mathcal{K}(G)$ are formed by merging connected components of $\mathcal{K}(G_1)$ using edges of level $2, 3, \cdots k$. By the symmetry of Erdős-Rényi networks, conditioned on components $\mathcal{C}_1, \mathcal{C}_2, \cdots \mathcal{C}_k$, the probability that there is an edge in G_1 between two components $\mathcal{C}_i$ and $\mathcal{C}_j$, for $1 \leq i < j \leq k$ depends only on the cardinality of these two subsets. Similarly, the probability that two connected components of $\mathcal{K}(G_1)$ have an edge of level 2 to k between them depends only on their sizes too.

The above observation, along with straightforward calculations show that with high probability $\mathcal{K}(G)$ has a unique giant component of size $C_1^{'} n + o(n)$, where $C_1^{'} \geq C_1$ and the rest of the components are of size $o(n)$. The rest of the proof follows from the proof of Theorem 1.

C.2 Proof of Proposition 2

Suppose $cd > 1$. Then classic results in random graph theory (e.g. see Theorem 2.11 of ?) establish that Erdos-Renyi graphs with average degree cd converge in the local weak sense to a Poisson branching process with mean offspring cd. Specifically, $E[N_t(v)]$ is at least $(cd)^T$, where $N_T(v)$ is the number of vertices within distance T from v. That shows that RAND can reach $s(cd)^T$ vertices in expectation.

For bounding the performance of OMN, we need to bound $\max_{v \in \mathcal{K}(G)}(N_t(v))$. For that, we use Lemma 1 from ?, which in our setting implies that with high probability, $\mathcal{K}(G)$ is such that, for every vertex v, $|N_T(v)| \leq 2T^3 \log(n)(cd)^T$. Therefore, OMN can reach at most $2T^3 \log(n)(cd)^T s$ with s seeds.

For $cd < 1$, we can plug in the above bound for $cd = 1$ to get an upper bound of $2T^3 \log(n)$ on $|N_T(v)|$. This observation implies that OMN cannot reach more than $2T^3 \log(n)s$ vertices. Obviously RAND reaches at least s vertices with s seeds.

Finally, note that for a k-level random network with a Erdos-Renyi random graph with average degree d as a base, the maximum size of a t-neighborhood is no more than the maximum size of tk-neighborhoods in the base random graph. Therefore, the largest $t-$neighborhood in $L_n(\phi)$ is of size $o(\log(n))$ as well.

D Directed Networks and Communication: Proof of Proposition 3

Consider a model of directed networks similar to Erdős-Renyi: $D(n, d)$ is 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. A *strongly connected component* is a subgraph for which there exists a directed path between any two member nodes. A relevant concept for directed graphs is that of a strongly connected *giant component*, which is a strongly connected component containing a linear fraction of the nodes, asymptotically. We will follow the arguments of Karp (1990) to show Proposition 3.

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

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

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

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

From fact 3, we know that each trial of RAND gets at least Θn nodes with probability Θ, whereas a single omnisciently chosen seed may reach up to $\Theta n + f(n)\sqrt{n}$ nodes. The difference of $f(n)\sqrt{n}$ is irrelevant for our result when s is a constant. So the combination of the above three observations already proves Proposition 3.

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

To see this, consider the subgraph consisting of nodes in $V - C$. The probability of communication between any two nodes is at most pc, and $|V - C|$ is at most $(1 - \Theta)n + f(n)\sqrt{n}$ by fact 3. By fact 2, there exists an $\epsilon > 0$ such that $dc(1 - \Theta + \epsilon) = dc\Theta' < 1$.

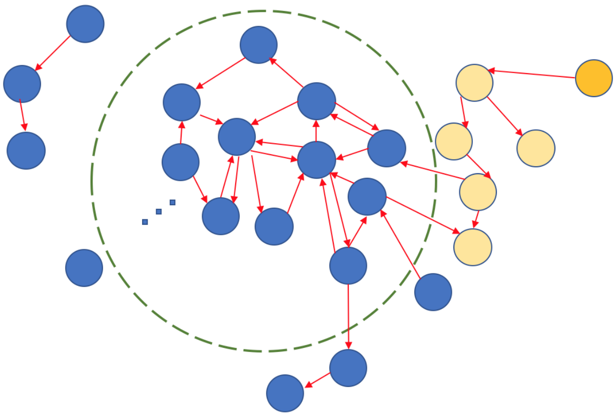


Figure 8: Above is an example communication network when communication is directed. The outgoing edges represent the nodes that a given node would inform if given information. The nodes within the dotted dashed circle represent the strongly connected giant component. If any node is informed within the s.g.c., all nodes in the s.g.c. become informed. Random seeding with enough seeds will land a seed in the s.g.c. with sufficiently high probability. The orange nodes, if informed, also disseminate information to the s.g.c. In particular, OMN might choose to seed the dark orange node, given a single seed (and there could be many such useful entry points, though only one set of orange nodes is pictured above). In the proof of Proposition 3, we want to show that the size of the set of any cluster of orange nodes is $o(\log(n))$ so that OMN cannot significantly outperform RAND.

Therefore, the number of neighbors of a given node (within the subgraph in consideration) is asymptotically dominated by $Bin(\Theta' n, dc)$. Using the Poisson approximation to the binomial distribution, a standard result on bounding the population of a Galton-Watson branching process, and the Chernoff bound, we get:

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

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

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

E Proofs missing from Section 4.1.1

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 $\alpha^2(1-p_n)p_n + o(1) \to \alpha^2((1-\alpha)^s)(1-(1-\alpha)^s)$. □

F Simulations of microfinance diffusion model

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

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

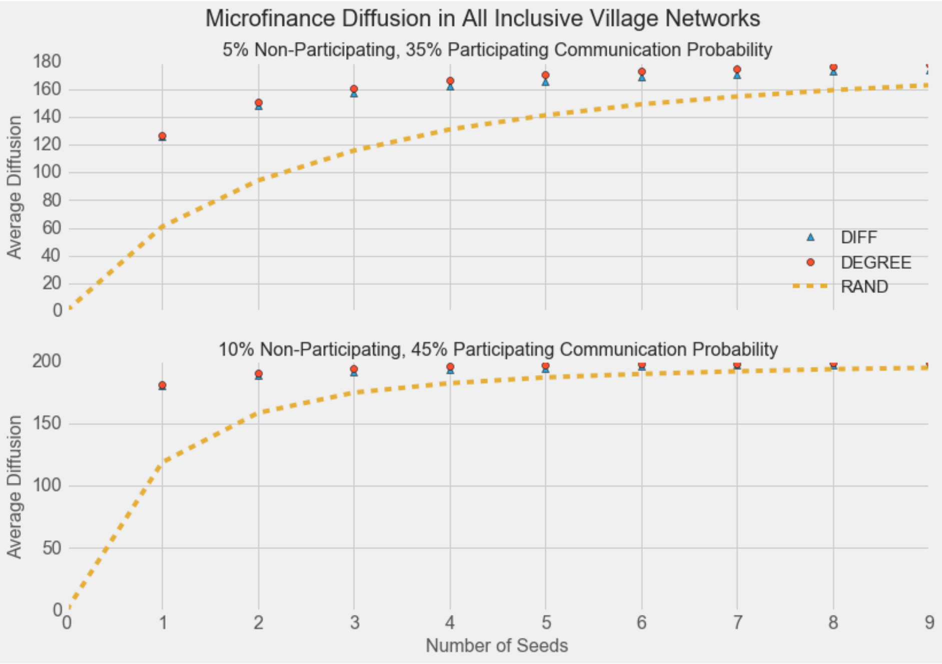


Figure 9: This is an analogue of Figure 2 with the diffusion process specified in Banerjee et al. (2013) rather than the model studied in this paper. As the number of seeds increases, random seeding performs as well as the centrality-guided seedings.

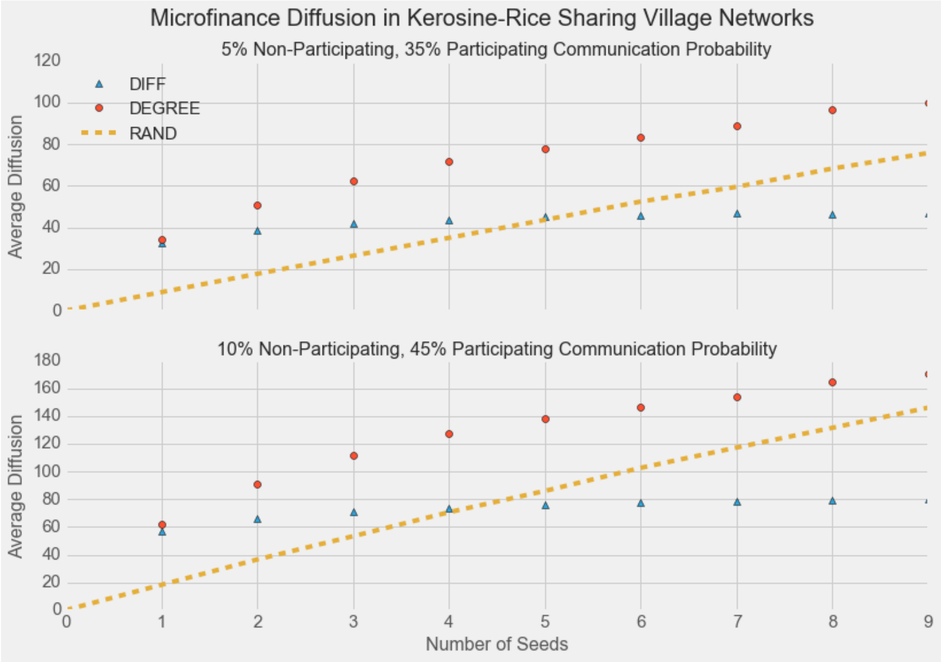


Figure 10: Random seeding performs well relative to the other seeding strategies. Moreover, it performs better than the seeding guided by the diffusion centrality when the number of seeds is more than 5.

sets of parameter estimates. We include simulations for the sparser kerosene and rice borrowing network in Figure F.

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.

G 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 [11]. 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 [12] neighbor of an untreated node learns about the program, this increases the chance of adoption for the untreated node by 7.5%. If a weakly linked neighbor learns the same, the probability of adoption goes up by 6%.

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

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

[12] 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.

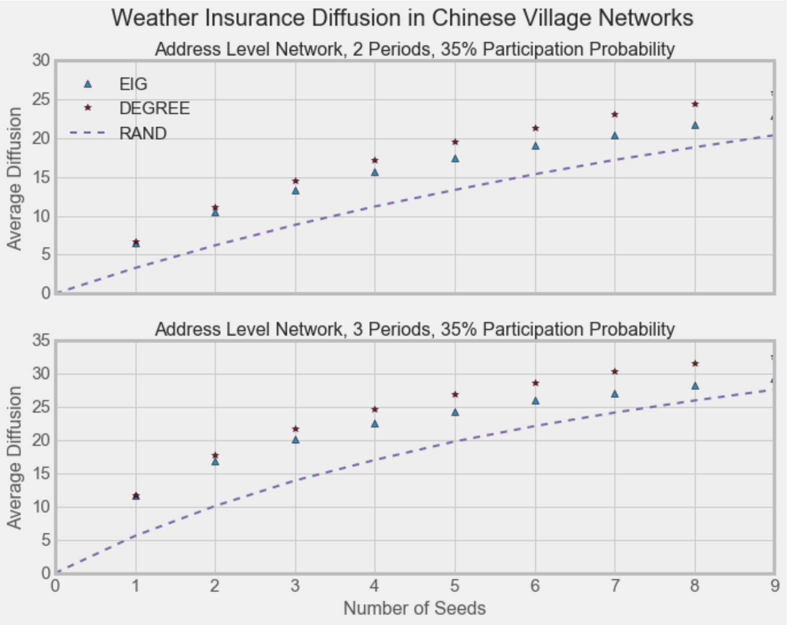


Figure 11: DEGREE seeding refers to seeding those with the highest degree, considering the undirected version of the village network. RAND seeding only chooses out of those villagers who participated in the social network survey, though they may name individuals who have not been surveyed as neighbors. Finally EIG refers to eigenvector centrality seeding. Note the average network size is 50 farmers.

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

We compare random seeding to degree seeding and seeding based on eigenvector centrality[13], two measures of centrality the authors suggest for targeting. Since all nodes more or less report the same number of friends, variation in degree mostly arises from variation in the number of friends that named the node in question as a friends. The authors find that under a permissive specification, central nodes do not wield additional influence over a given neighbor than less central counterparts. Therefore, in our simulations, the benefit of seeding central nodes arises purely from their connection to more immediate neighbors and paths to other nodes. The results of our simulations show again in a different network and setting that the presence of network effects and positive association between centrality and diffusion does not immediately imply that carefully targeting nodes will make a large difference. Indeed one of the striking findings in Cai et al. (2015) is that social learning is a powerful vehicle of information transmission–strong enough that a policymaker may safely ignore minutiae of network structure.

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

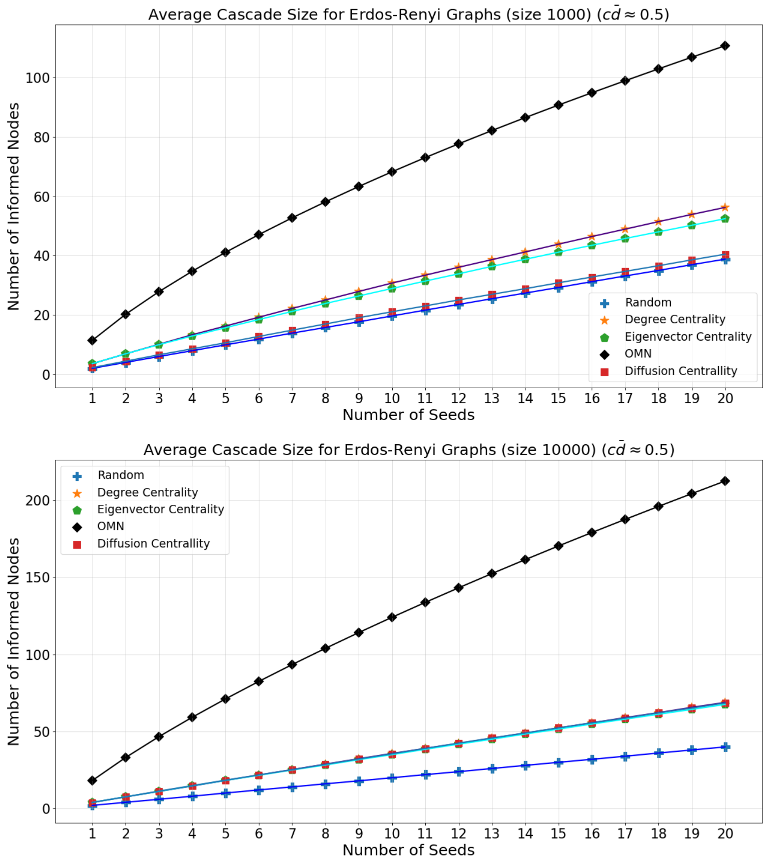


Figure 12: Performance of different algorithms in Erdős-Rényi graphs, when $cd = 0.5$.

H Simulations of $||\mathbf{T}_\kappa|| < 1/c$ regime in small networks

To check the behavior of OMN, random, and other seeding strategies when $pd < 1$, we conducted some simulations on both Erdős-Rényi graphs (as your comment suggested) as well as Indian village networks.

As Figure 12 shows, with 1000 nodes, random is overall very close to typically used network-guided strategies. The OMN algorithm, however, performs unusually good. Even then, starting from one seed, it can not reach more than 1.3% of the nodes. With 20 seeds, random can reach 4% of the nodes, typically used seeding strategies reach less than 6% of the nodes, while OMN reaches around 11% of the nodes.

When there are 10000 nodes, our limit results are seen much more cleanly. OMN with 1 seed can only reach 0.2% of the nodes. Even with 20 seeds, OMN cannot reach more than 2.2% of the nodes. Again, random and typically used seeding heuristics cannot reach more than 0.7% of the nodes, even with 20 seeds.

Figure 13 shows our simulation results for Indian village networks (where the average size is just below 200). Because these networks are relatively small, OMN can indeed reach a sizeable fraction of the network when it is given a lot of seeds. Thus, our limit

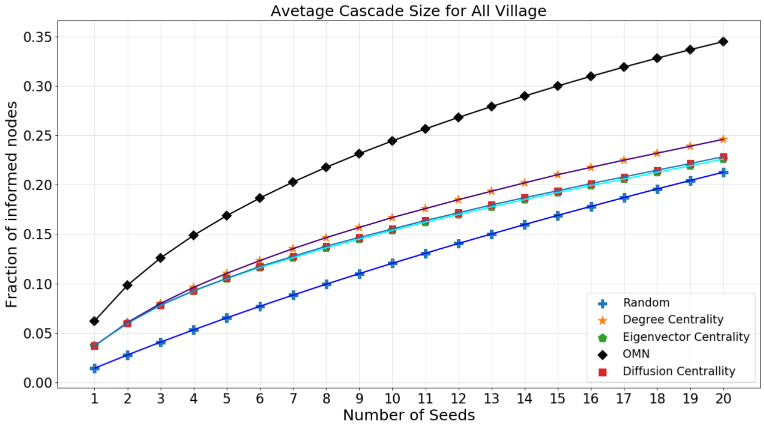


Figure 13: Performance of different algorithms averaged over all Indian village networks, when $cd = 0.5$.

result for this regime (that $\lim_{n\to\infty} \mathbf{H}(OMN, s) = 0$) clearly fails to hold. Even then, random competes well with typically used heuristic algorithms.

These simulations illustrate two points about the $||\mathbf{T}_\kappa|| < 1/c$ regime: (1) OMN is too good of a benchmark for this setting, as it performs much better than not only random seeding, but also than network-guided heuristics. (2) Random seeding (with a few extra seeds) competes well with typically used heuristics even in this regime.

I Component Sizes in ER and k-Level Graphs

The top row of Figure 14 shows that for both ER and k-level random graphs, when we are in the regime that the communication network is very sparse (hence the diffusion will be unsuccessful), the sizes of the largest and second largest components of the networks are very small essentially for all network sizes. For percolated k-level graphs, proving that component sizes are order $log(n)$ is analytically challenging. Simulations, however, indicate that a similar result is true for such graphs

Figure 14 also shows that in the regime where ER and k-level graphs have a giant component, the smaller component are $O(\log(n))$ in size. While Corollary 4 keeps s fixed, these simulations suggest that using similar arguments as in the proof of Corollary 1, one can perhaps let s belong the class $o(\frac{n}{log(n)})$.

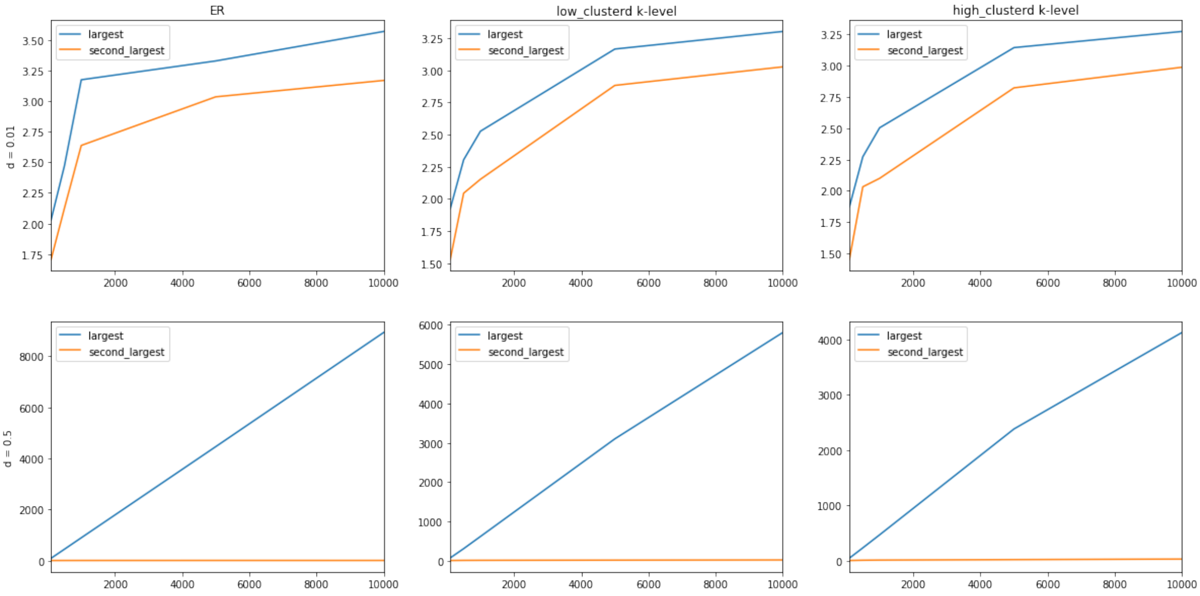


Figure 14: Size of largest and second largest components in ER and k-level graphs in both regimes. The x- axis is the size of the network and y-axis is the sizes of components.