# Influence maximization in random graphs*

Han Lin† Zhuangzhuang Jia† Ningyuan Tang†

May, 2019

## 1 Introduction

Our group project is mainly based on two papers, Akbarpour et al.(2018), Jackson and Storms (2018). In Akbarpour et al.(2018), the authors find that for a wide range of parameters, the random seeding strategy with $s + x$ seeds asymptotically performs as well as the omniscient strategy with $s$ seeds, where $x$ is vanishingly small relative to the size of the network. From this result, we can see when the target network is not too scattered or the communication probability between people is not too low, transmitting information through optimal seeds will have the same effect as transmitting such information through a few extra individuals.

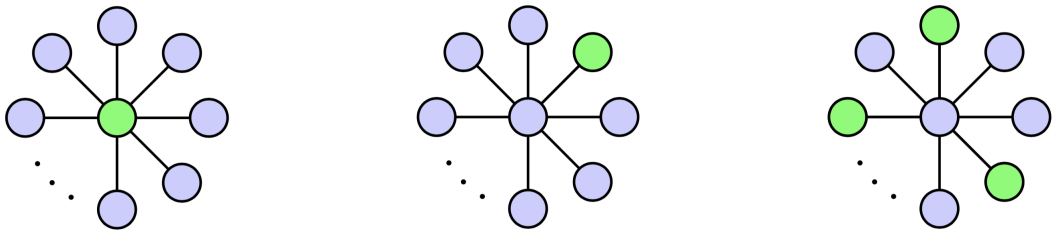
We followed the steps in Akbarpour et al.(2018) and test their results in both theoretical graphs and real-world networks, although Akbarpour et al.(2018) mention that when the diffusion process is such that neighbors are "complements", say when several of an agent's neighbors have to adopt a technology before he does the same, their results may fail to hold, they do not test this result under such circumstance in their paper. In order to get a better understanding of the results, we follow Jackson and Storms (2018) to test the performance of random seeding when a homogeneous threshold exists.

First, we test the result for 'sparse' Erdős-Rényi random graphs. From the definition of Erdős-Rényi Network, we can see for each node, they have the same expected degree, so the result may be just a coincidence. Therefore, we further test this result for a model of networks with power-law degree distributions and a generalized version of Erdős-Rényi graphs with clustering. Our simulation results are consistent with the theorems in Akbarpour

*Project in IEOR4408 Computational Discrete Optimization, Professor: Yuri Faenza

†Columbia University

et al.(2018). The intuition behind this is: as we can see from Figure 1, when we seed a few more seeds, the random seeding strategy will have a higher probability to select one of the neighbors of an agent that has many connections, and thus, central individuals will become informed through their neighbors.



(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 the probability that an informed node transfers information to its neighboring nodes is 0.5. 1(a): The optimal seeding strategy to choose one node is certainly to pick the central node and in expectation $\frac{n}{2}$ of nodes will be informed. 1(b): If we choose a node randomly, we will be very likely ($p \approx 1$) pick a non-central node. 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 $\frac{n}{4}$. 1(c): If we randomly choose $1 < x \ll n$ seeds. Now even if we choose $x$ neighboring nodes, the probability that a central seed is informed now becomes $1 - \left(\frac{1}{2}\right)^x$, so expected diffusion is nearly $\frac{n}{2}\left(1 - \left(\frac{1}{2}\right)^x\right)$ , 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. (Akbarpour et al.2018)

Then we test the result in some finite, real-world networks. Following the authors, we download the data of 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 from Leskovec and Krevl (2014). Our result shows that, compared with omniscient targeting strategies, random seeding performs well in both the extent and the speed of diffusion. From our simulation results, we also verify that the extra seeds required by random to beat network-guided heuristics in the speed of diffusion are smaller than the theoretical $o(\log(n))$ multiplicative bound.

After testing the results in Akbarpour et al.(2018), we extended our results to a threshold model follow Jackson and Storms (2018). We see that when the threshold is moderately high, random seeding (even with a few additional seeds) cannot perform well.

**Organization of our report.** We introduce three relevant seeding strategies in section 2. In section 3.1, we present our results in three theoretical networks. In section 3.2, we give our results in both undirected and directed real-world networks. We extend the objective

function to speed of diffusion in section 3.3. In section 4, we test the random seeding strategy under threshold model. Section 5 concludes.

# 2 Three Relevant Strategies

In the theoretical model, we assume we have $n$ nodes connected with each other through some kind of social network. In the beginning, only a small group of nodes acquire some information. The information percolates in the network according to a variant of the ubiquitous Susceptible-Infected-Recovered (SIR) diffusion model. The process goes as followed, in each round, the new informed nodes has one chance to transfer this information to its neighboring nodes. This success probability of such transfer is $c$ independently for each node. Therefore, the neighboring nodes may become informed in the next round. And the process will continue until no neighboring nodes will be informed.

In our project, we mainly focus on three relevant seeding strategies, the first one is optimal seeding strategy which is denoted by OPT. For a fixed network, this strategy picks the set of $s$ seeds that maximizes the expected diffusion, with an arbitrary selection when there are multiple optimal candidates:

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

It is known that computing this strategy is NP-hard (Kempe et al., 2003). In real world, instead, people often use heuristics such as seeding the $s$ most central individuals in the network, according to various measures of centrality instead. We will introduce them in the our test for real world networks.

The authors defined two seeding strategies as theoretical benchmarks. $\mathrm{RAND}(s)$ is the strategy which picks $s$ nodes uniformly at random in $G$. This strategy ignores all the information about the network structure.

Another one is *omniscient* seeding strategy, denoted by $\mathrm{OMN}(s)$, it assumes we already know the realization of each network and then pick $s$ initial seeds to maximize diffusion result. From this definition, we can see actually omniscient strategy is infeasible by construction because in real world, we can not know the exact network structure for each network. Obviously, for these three seeding strategies, we have the following relationship:

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

Here $\mathbf{H}$ is a function measures the performance of each strategy, the input is the strategy

and number of seeds, the output is the expected total number of informed nodes.

# 3 Basic Results

## 3.1 Theoretical Networks

A direct way to see how OPT strategy performs is to compare 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. From last section, however, we know that computing OPT is an NP-hard problem. Instead, we will compare the difference between the performances of OMN and RAND. Since for any realization of the diffusion process, OMN performs better than OPT, comparing RAND and OMN gives a generous upper bound on the value of network information and analysis.

Here is some explanation of asymptotical notations: We say that a function $f(n)$ asymptotically weakly dominates $g(n)$ if $\lim_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| \geq 1$. We also say $f$ is of $o(g), \omega(g)$, and $O(g)$ if and only if this limit is zero, infinity, and any finite constant respectively. For example, any divergent increasing function of $n$ is $\omega(1)$. We refer to $\omega(1)$ as a super-constant.

### 3.1.1 Erdős-Rényi Networks

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

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

*If $cd \leq 1$, then*

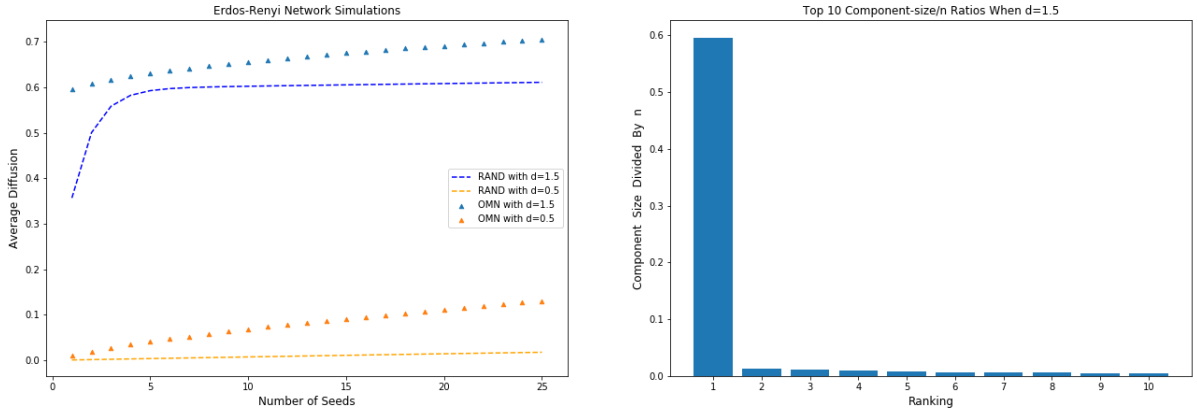
$$\lim_{n\to\infty} \frac{\mathbf{H}(\text{OMN}, s)}{n} = 0$$

From Theorem 1, we see that when $cd > 1$, which means for each node, the average node that it can transfer is larger than 1, than the random seeding strategy with super-constant extra seeds (asymptotically) performs better than omniscient seeding strategy. When $cd \leq 1$, even the omniscient strategy performs badly.

The whole proof can be found in Akbarpour et al.(2018), here we give a simple explanation, as in this theorem, our diffusion process is unbounded, from Lemma 1 in our appendix, we can see 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 is also our method for simulation as you can see from our code.

When $n$ is sufficiently large and $cd > 1$, we can prove that there exists a component in the communication network which contains a constant fraction of the total population. And all other remaining components are vanishingly small ($O(\log(n))$) in population size. So when we can choose a lot of seeds, we will have very high probability end up informing the nodes in the large component using random seeding strategy, which it can with high probability.

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

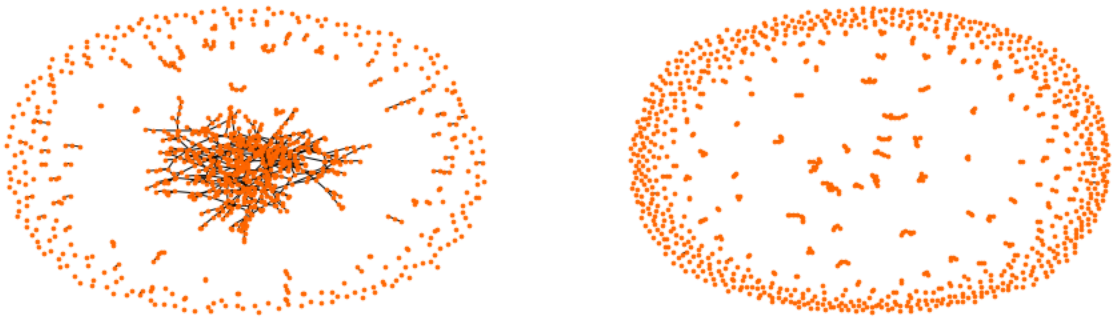



(a) Simulation results in Erdős-Rényi Network: $n = 1000, cd = 1.5$ and 0.5 respectively

(b) Size of Top 10 Connected Components in Erdős-Rényi Network

Figure 2: Simulation result for Erdős-Rényi Network

Figure 2 is our simulation result in ER network, from 2(a), we can see when $cd = 1.5$, randomly chooses 7 seeds performs as better as omniscient seeding strategy with 3 seeds, both have around 60% of average diffusion. When $cd = 0.5$, even the omniscient strategy can not have average diffusion larger than 20%. Figure 3 gives us a simple intuition behind Theorem 1.



(a) Erdős-Rényi Network: $n = 1000, cd = 1.5$ (b) Erdős-Rényi Network: $n = 1000, cd = 0.5$

Figure 3: Connected Components in Erdős-Rényi Network

### 3.1.2 Power-law Chung-Lu Networks

Lots of networks have highly central agents and have high 'clustering' coefficients, but either of these properties can be seen in Erdős-Rényi random graphs. Actually, from the definition of ER network, we can see every node have exactly the same expected degree, so maybe it's just a coincidence for such result to be true. Thus we test in a model of networks with power-law degree distributions, we first introduce the definition of such network.

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

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

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

From Figure 4 (b), we can see in such network, a small proportion of nodes have extremely high degrees and others have very small expected degree.

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

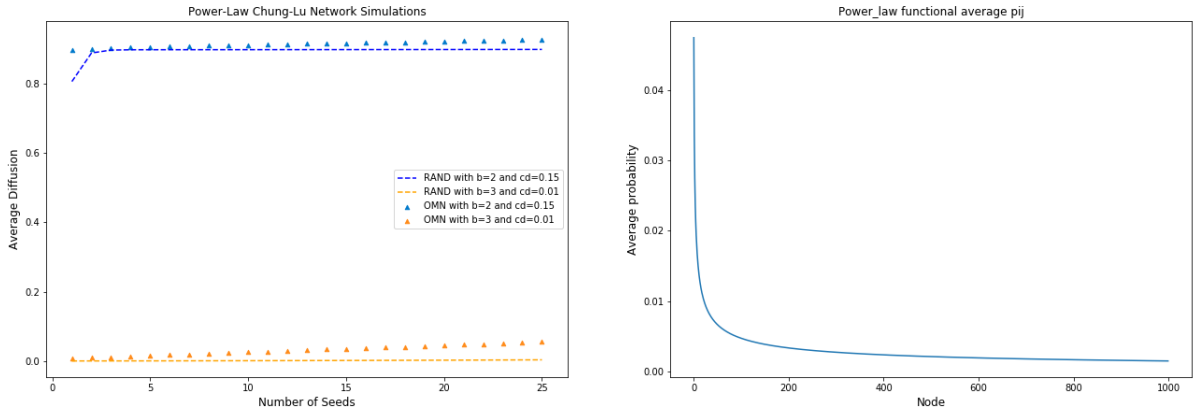
$$\lim_{n\to\infty} \frac{\mathbf{H}(\text{RAND}, s + x(n)}{\mathbf{H}(\text{OMN}, s)} \geq 1$$

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

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

The rigorous proof of Theorem 2 is very technical and can be found in Akbarpour et al.(2018), here like what we just give a simple intuition. The proof idea is quite similar to Theorem 1, here we need to extend the theorems over a range of parameters where standard techniques do not apply. In a power-law network, some nodes have very high degrees. The omniscient strategy will pick these nodes as seeds, but the random strategy will most likely never pick such nodes. So how can random strategy compete with omniscient strategy ? The intuition is just like what we see in Figure 1, when we choose asymptotically many nodes, the central nodes will be more likely to be informed.

Figure 4 is our simulation result in Power-Law network, from 2(a), we can see when $b = 2$, randomly chooses 15 seeds performs as better as omniscient seeding strategy with 10 seeds, both have around 90% of average diffusion. When $b = 3, cd = 0.01$, even the omniscient strategy can not have average diffusion larger than 20%. Figure 5 gives us a simple intuition behind Theorem 2.

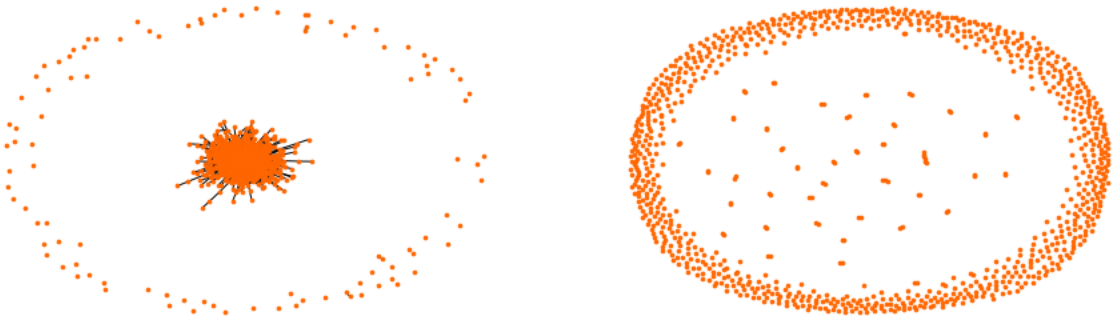



(a) Simulation results in Power-Law Network (b) Power-Law degree distribution

Figure 4: Simulation result in Power-Law Network

### 3.1.3 $k$-level Random Networks

In this part, we test the random seeding strategy in a high clustering networks. The $k$-level random network is based on Erdős-Rényi random graph with extra edges to neighbor of



(a) Power-Law Network: $n = 1000, cd = 1.5$ (b) Power-Law Network: $n = 1000, cd = 0.5$

Figure 5: Connected Components in Power-Law Network

neighbors, neighbor of neighbor of neighbors, and so on.

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

Our simulation result can be found in appendix B, from Figure 21 we can see first compared with Erdős-Rényi network, the average diffusion is higher, this is quite intuitive, because compared with Erdős-Rényi network, we now allow more links between different nodes. Also we can see that when $n$ is large, random seeding strategy performs asymptotically well as omniscient strategy. Our simulation result suggest that the presence of high clustering coefficient does not hinder the outcome of random seeding strategy with a few more seeds than omniscient strategy.

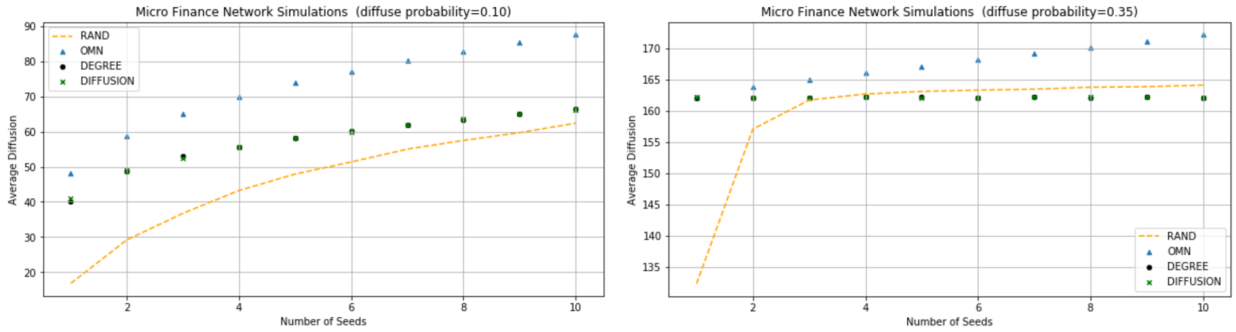
## 3.2 Real-world Networks

So far, we have we have considered random seeding strategy on some theoretical network models. But as we seen, the component size distribution under random graphs is either too dense or too sparse, which can rarely represent that in the real world. Therefore, we test the random seeding strategy on some real networks to see its real performance.

### 3.2.1 Undirected Networks

In this part, we study the diffusion model on two real world undirected networks: microfinance network data in Banerjee et al.(2013) as well as a network from Facebook.

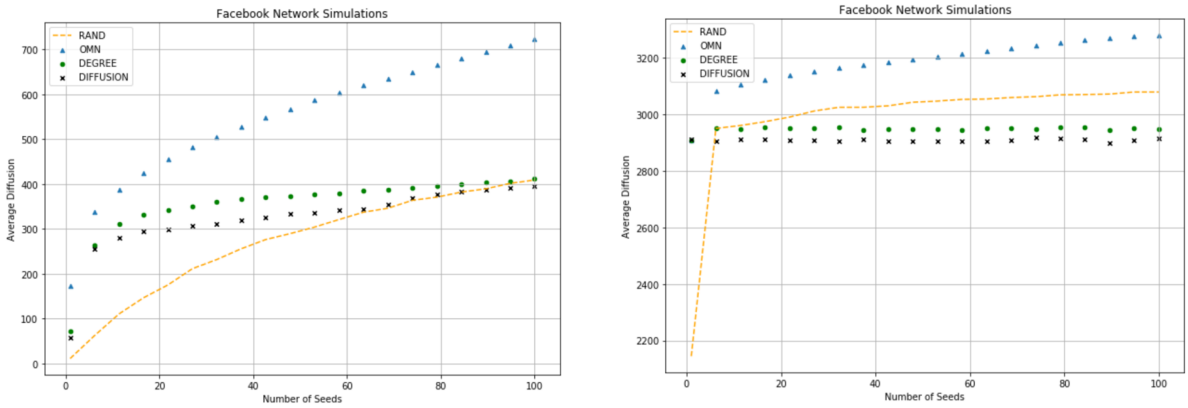
In the microfinance network example (See Figure 6), we take average of the number of households diffused over 77 Indian villages, with edges in each village represents any form of possible contacts between households. And the communication probability represents whether a households will let one of a given neighbor be involved in the same activity.




(a) 10% Communication Probability (b) 35% Communication Probability

Figure 6: Simulation results of various seeding strategies (omniscient, random, degree, diffusion) across in microfinance data

In the Facebook network example (See Figure 7), we analyze the social network with more than 4000 nodes, and average degree of 43 (which represents friends of a person).




(a) 1% Communication Probability (b) 10% Communication Probability

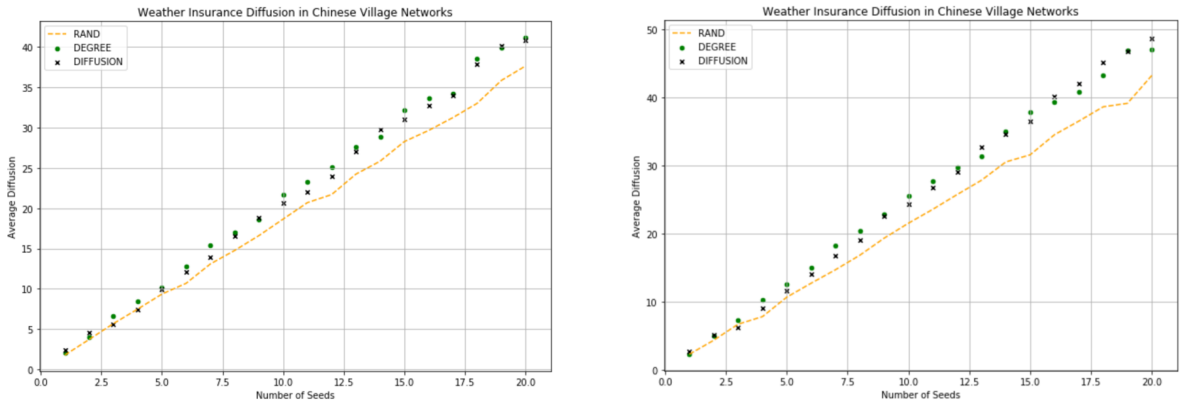
Figure 7: Simulation results of various seeding strategies (omniscient, random, degree, diffusion) across in facebook data

In both of these two real networks, the performance of random seeding strategy increases with higher communication probability, and it will outperform at least the degree centrality and diffuse centrality seeding strategy[1], and can achieve roughly the same performance as the omniscient one with some extra number of seeds.

[1]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 $T$

### 3.2.2 Directed Networks

In this part, we evaluated random seeding strategy on a directed real network. Our data is from Cai et al.(2015), where the authors studied diffusion of a new government offered weather insurance take-up across various villages in China. Every participants are asked to list at most 5 of their closest friends, with each friend listed on a person's list represents an directed edge to this friend. And if both people list each other on their lists, then we say that there's strong link between them, and there will be a weak link otherwise. As estimated in the paper, the probability of participation through a strong link is 21%, and 17% for a weak link. Besides, we assume that the communication only last for 2 periods. Our result on this real network with low average degree is consistent with Theorem 3, which means that random seeding can also perform well in directed graph.




(a) 35% Participation Probability

(b) 100% Participation Probability

Figure 8: Simulation results of various seeding strategies (omniscient, random, degree, diffusion) across in the directed network

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

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

*If $cp \leq 1$, then*

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

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).

The following figure compares the theoretical ratio of extra time of seeds we derived from Theorem 3 with the real lowest bound of this ratio that random seeding strategy can outperform omniscient seeding strategy. The simulation result indicates that we actually only need less than a half of the extra number of seeds that that provided in Theorem 3.

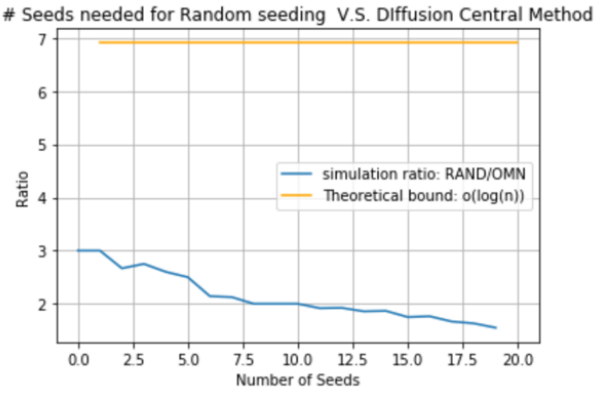



Figure 9: Simple Compare

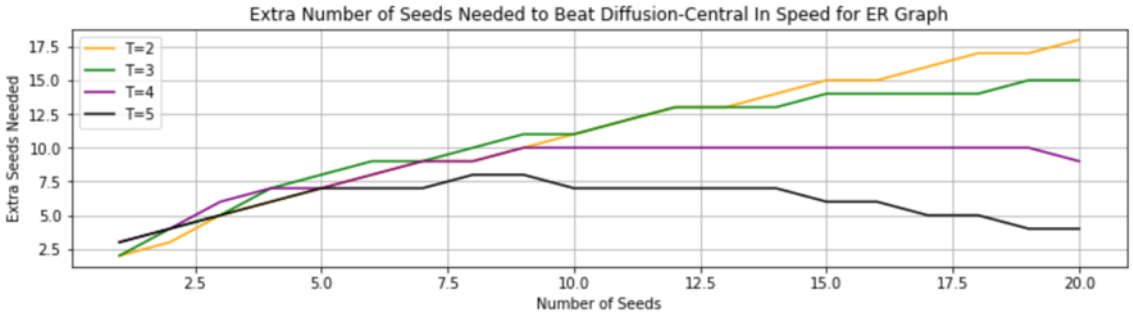
## 3.3 Speed of Diffusion

In this section, we test the result under bounded diffusion processes, where all communication ceases after a fixed number of rounds. Our simulation results, then, show that random seeding competes with omniscient seeding period by period, which can be interpreted as a statement on the relative speeds of diffusion in the unbounded case.

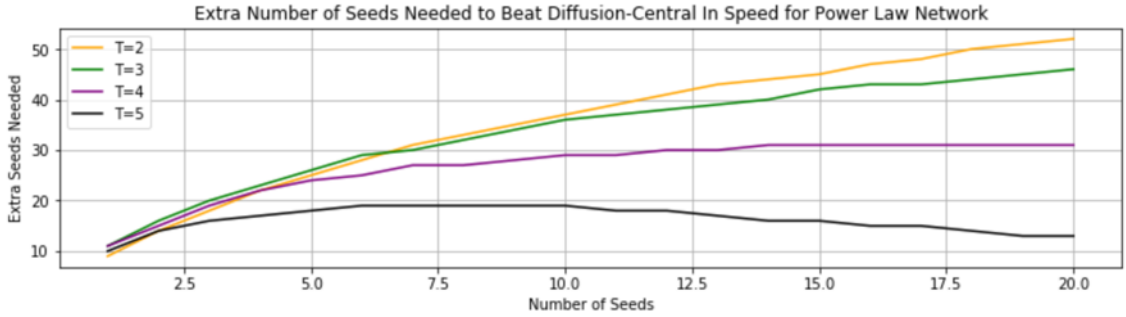
**Theorem 4.** *Consider an Erdős-Rényi network on $n$ nodes with average degree $d$ and a bounded diffusion process that ends in $T \geq 1$ periods and let $s$ be a non-negative integer. Then,* $\mathbf{H}(\text{RAND}, o(\log(n))s) \geq \mathbf{H}(\text{OMN}, s)$ *for $n$ sufficiently large.*

The rigorous proof of Theorem 4 can be found in Akbarpour et al.(2018), the idea here is that the extent of diffusion that happens in $T$ periods from any node must be $o(\log(n))$ for large $n$, therefore with $o(\log(n))$ times additional seeds, random seeding strategy competes with omniscient seeding strategy.

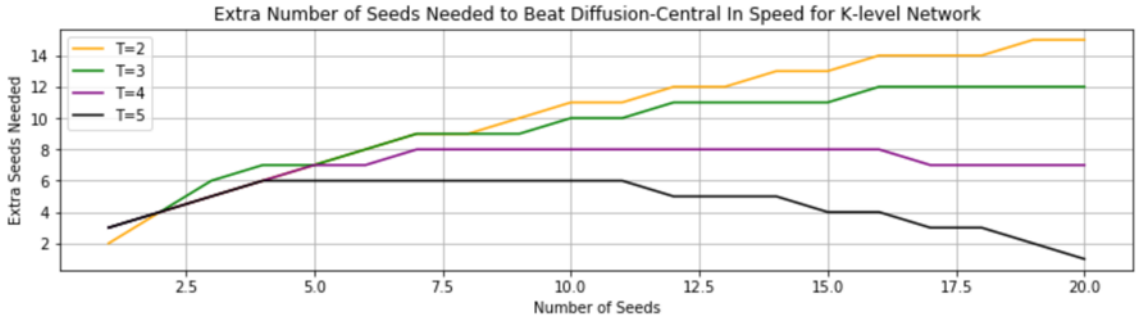
Figure 10 is our simulation results for three theoretical networks we have seen before, as is easy to see, when the number of periods become larger, fewer extra seeds are required .




(a) Erdős-Rényi Network




(b) Power-Law Chung-Lu Network




(c) $k$-level Network

Figure 10: Average number of extra seeds required by random to outperform diffusion-centrality seeding in three different networks.

# 4 Extension: Threshold Model

In the previous diffusion model, an individual can get informed by any single one of her informed neighbors with some probability in the communication network. If we consider the entire diffusion process as a set function, $f$, that maps the set of initial seeds to the number of informed agents finally, then $f$ exhibits diminishing returns, meaning that the number of newly informed agents decreases as the set of initial seeds increases in size. More rigorously, we give a formal definition of "diminishing returns", or "submodularity" (Krause and Golovin , 2014).

## 4.1 Submodularity and Random Seeding

Submodularity is a property of *set functions*, i.e., functions $f : 2^V \to \mathbb{R}$ that assign each subset $S \subseteq V$ a value $f(S)$. Hereby $V$ is a finite set, commonly called the *ground set*. In our example, $V$ may refer to the locations where sensors can be placed, and $f(S)$ the utility (e.g., detection performance) obtained when placing sensors at locations $S$. In the following, we will also assume that $f(\emptyset) = 0$, i.e., the empty set carries no value. Submodularity has two equivalent definitions, which we will now describe. The first definition relies on a notion of discrete derivative, often also called the marginal gain.

**Definition 3.** (Discrete derivative) For a set function $f : 2^V \to \mathbb{R}, S \subseteq V$ and $e \subseteq V$, let $\Delta_f(e \mid S) := f(S \cup \{e\}) - f(S)$ be the *discrete derivative* of $f$ at $S$ with respect to $e$.

Where the function $f$ is clear from the context, we drop the subscript and simply write $\Delta(e \mid S)$.

**Definition 4.** (Submodularity) A function $f : 2^V \to \mathbb{R}$ is *submodular* if for every $A \subseteq B \subseteq V$ and $e \in V \setminus B$ it holds that

$$\Delta(e \mid A) \geq \Delta(e \mid B)$$

Equivalently, a function $f : 2^V \to \mathbb{R}$ if for every $A, B \subseteq V$,

$$f(A \cap B) + f(A \cup B) \leq f(A) + f(B)$$

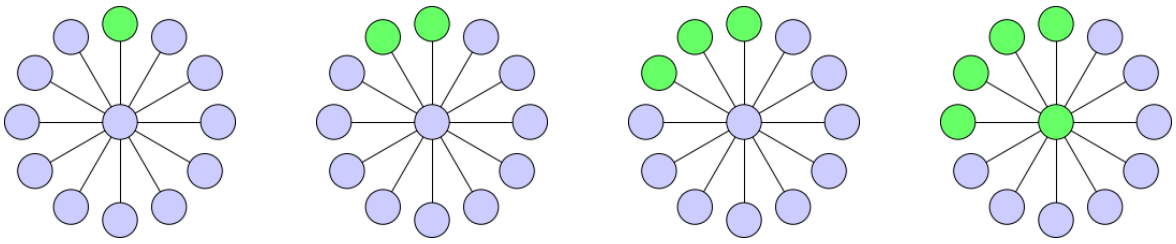
.

Random seeding works well in terms of influence maximization in the setting of submodular diffusion processes. However, we shall see that random seeding may not perform well in a setting that fails submodularity.

## 4.2 Thresholds

Consider a network where agents tend to follow their neighbors' behaviors. A behavior is characterized by a threshold $q \in (0, 1)$, and agents adopt that behavior if at least $q$ of neighbors do (Jackson and Storms , 2018). This type of behavior is a reasonable result of peer influence.

For example, Abby has 12 neighbors and she will buy a new iPhone only when 1/3 of her neighbors have done so. Compared with the previous diffusion model, this threshold model fails submodularity because the fourth neighbor's purchase has much more impact on Abby's decision compared with the previous three neighbors' purchases.



(a) One neighbor is informed

(b) Two neighbors are informed

(c) Three neighbors are informed

(d) The central node gets informed when the fourth neighbor is informed

Figure 11: A simple example to illustrate the intuition of threshold model. From the figures, we can imagine that random seeding may not perform well in terms of influence maximization when decisions are characterized by thresholds. Random seeding is unlikely to pick enough neighbors of one agent to make this agent informed if the threshold is high.

## 4.3 Conventions and Community Structures

As Jackson and Storms (2018) point out, optimal seeding in a diffusion model where threshold applies is an NP-hard problem, but since agents' behaviors depend on a fraction of their neighbors' behaviors, we can detect "communities" where agents within a community are closely related to each other and loosely related to those outside their block, and concentrate our initial seeds on such blocks so as to maximize influence while saving seeds. The greedy choice is seeding the communities with the largest size-to-cost ratio (number of agents informed finally/number of initial seeds in the community). Later we shall see that such communities are also called "atoms" in the formal definition. Intuitively, this algorithm can outperform random seeding when the threshold is not too low.

### 4.3.1 Notations

A finite set $N = 1, ..., n$ of people or nodes, with generic indices $i, j$, are connected in a network g. A *network* is a simple graph, $(N, g)$, consisting of a finite set of nodes (vertices...) $N$ together with a list of the undirected links (edges, ties...) that are present $g$. We let $ij \in g$ indicate that the undirected link between nodes $i$ and $j$ is present in the graph. We consider undirected networks (mutual friendships) and so $g$ is taken to be symmetric. We adopt the convention that $ii \notin g$ so that agents are not friends with themselves.

Agent $i$'s neighbors in g is the set $N_i(g) \subset N \setminus \{i\}$, $N_i(g) \equiv \{j | ij \in g\}$. Agent $i$'s degree is the size of $N_i(g)$, denoted $d_i(g) = |N_i(g)|$. Isolated nodes are not of much interest in our setting, so we ignore them. They can be incorporated by allowing them to each be their own convention (or by having them never be part of a convention). In what follows we presume

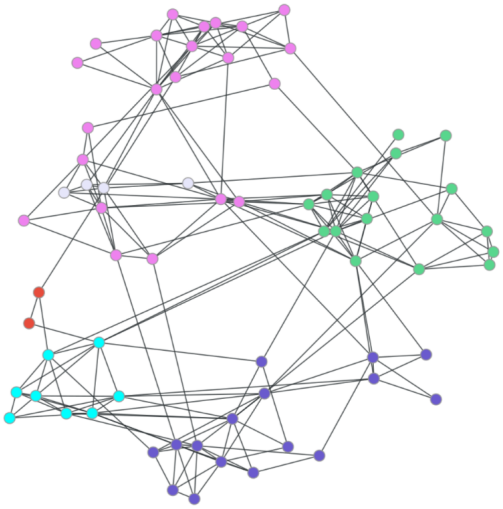


Figure 12: The figure is an example of communities that do not have overlaps with each other. We should allocate clusters of seeds to such communities that have the biggest size-to-cost ratios

that a network $g$ is such that each node has at least one neighbor ($d_i(g) \geq 1$ for all $i \in N$).

### 4.3.2 Conventions

A *convention* associated with some threshold $q$ on a network $g$ is a group $S \subset N$, for which all members of $S$ have a fraction of at least $q$ of their neighbors in $S$ ($|N_i(g) \cap S|/d_i(g)) \geq q$ for all $i \notin S$, and all agents not in $S$ have strictly less than $q$ of their neighbors in $S$ ($|N_i(g) \cap S|/d_i(g)) < q$ for all $i \notin S$. A couple of conventions are pictured in Figure 1 for an example of a behavior with a threshold of $q = 0.4$.

Figure 13, 14, and 15 show three possible conventions resulting from different choices of initial seeds with a threshold of $q = 0.4$. The green nodes represent informed nodes.

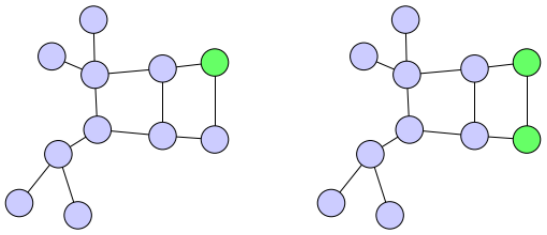


Figure 13: Convention 1

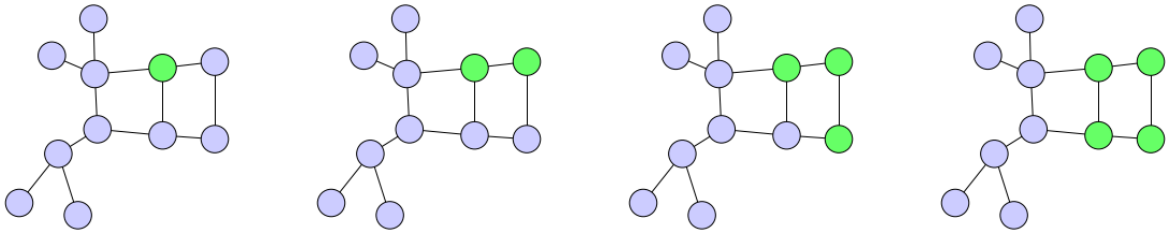


Figure 14: Convention 2

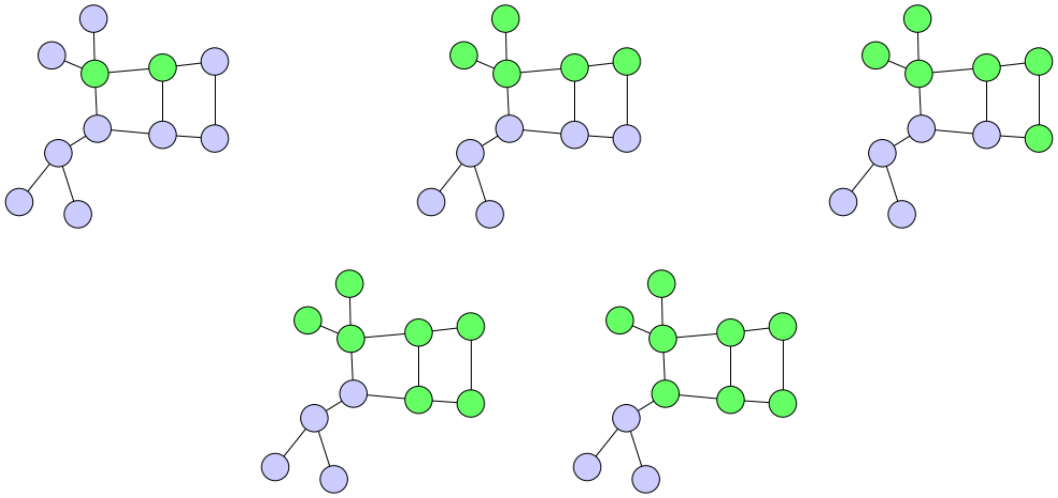


Figure 15: Convention 3

### 4.3.3 Cohesiveness and Closedness

A group of agents is said to be $q$-cohesive if every member of that group has at least proportion $q$ of her neighbors within the group (Morris ,2000). A group of agents is said to be $q$-closed if every agent outside the group has less than $q$ of her neighbors within the group (Jackson and Storms, 2018).

Intuitively, a set of agents is both $q$-cohesive and $q$-closed if the agents within the set are closely related to each other and they form a maximal set, meaning that they are loosely related to every agent outside the set.

Following Morris (2000), we define a group $S \subset N$ to be $q$-cohesive if each of its members have a fraction at least $q$ of their neighbors in the group ($|N_i(g) \cap S|/d_i(g) \geq q$ for all $i \in S$).

We say that a group $S \subset N$ is $q$-closed if every individual outside of $S$ (in $N \setminus S$) has a fraction of his or her friends in the group that is less than $q(|N_i(g) \cap S|/d_i(g) < q$ for all $i \notin S$).

### 4.3.4 Atoms

Nodes inside an atom behave the same way in every $q$-convention, and if two nodes are in different atoms then there is some convention under which behave differently (Jackson and Storms, 2018).

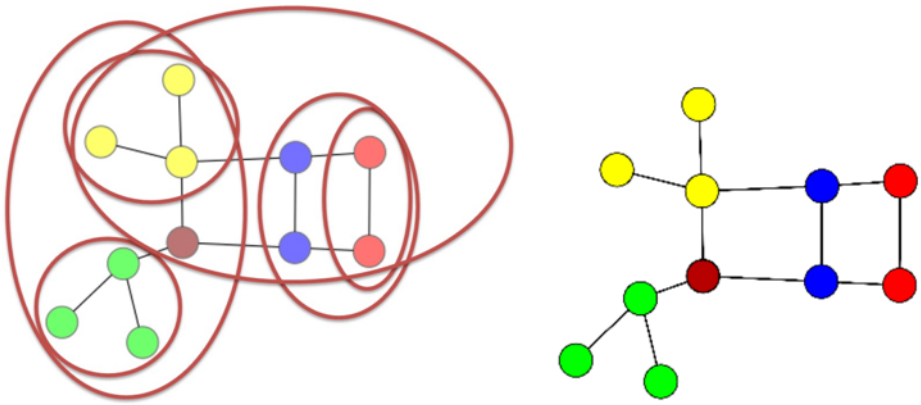


Figure 16: In this figure, the threshold is q=0.4, each red oval represents a convention, and nodes of the same color belong to the same atom. Two nodes belong to an atom if for every convention, they are either in or out of the convention together. Therefore, every convention can be partitioned into atoms, although an atom itself may not be a convention, and an atom may even consist of disconnected nodes (Jackson and Storms , 2018).

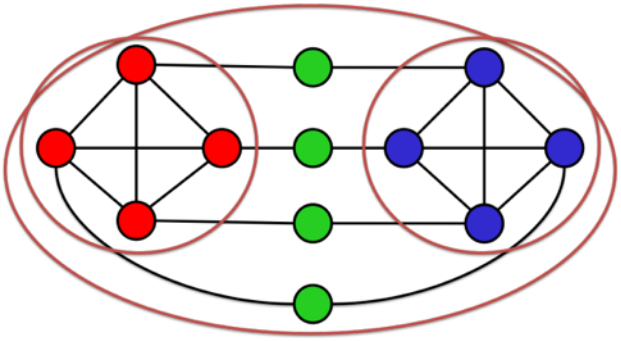


Figure 17: This figure illustrates the partition into atoms when the threshold is q=3/4.

### 4.3.5 Stochastic Block Model

Let $n$ index a sequence of random graph models, tracking the number of nodes in the society. The society is partitioned into different types of people or nodes indexed by $J(n)$, with generic indices of types $jj'$, and cardinalities $j(n)$. These might refer to demographic characteristics

like age, religion, gender, ethnicity, profession, etc. Let $\Pi(n)$ denote the associated partition of the nodes by types.

The probability that any node of type $j \in J(n)$ is linked to a node of type $j' \in J$ is given by some $p_{jj'}(n) = p_{j'j}(n)(j(n) - 1)$. Links are independent across all pairs of nodes. Let $d_{jj}(n) = p_{jj}(n)$ and $d_{jj'}(n) = p_{jj'}j'(n)$ denote the expected links of a type $j \in J(n)$ to types $j$ and $j'$, respectively, in society $n$, and $d_j(n) = \sum_{j' \in J(n)} d_{jj'}(n)$ be the overall expected degree of a type $j$ node. We use $g^n$ to denote a random network generated on the $n$ nodes.

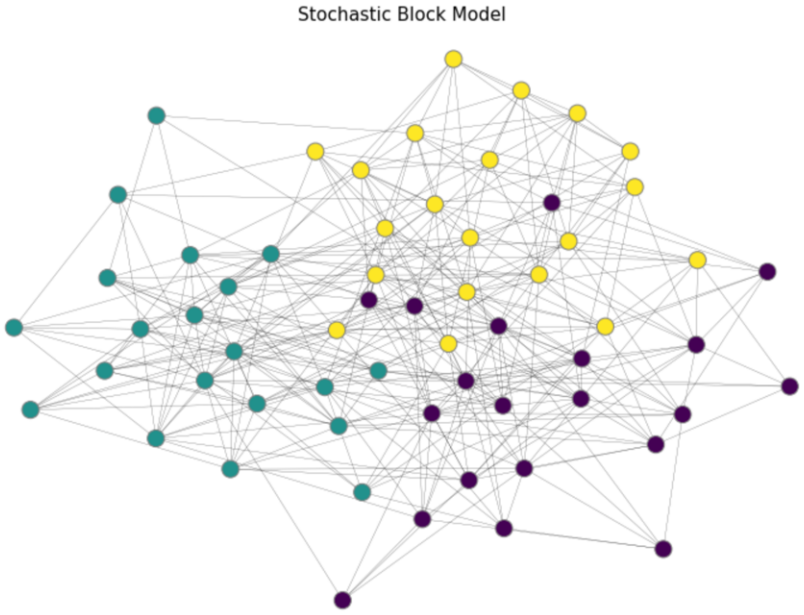



Figure 18: Stochastic model with 60 nodes, 3 blocks. Communication probability within block = 0.1, communication probability across block = 0.35.

## 4.4 Algorithm and Result

### 4.4.1 Atom-based Seeding Algorithm

**Algorithm 1.** (Atom-based Seeding Algorithm)(Jackson and Storms , 2018)

1. Find the $q$-atoms of $G$. Order them $A_1, A_2, ...A_m$.

2. For each atom $A_i$, find the minimal number of seeds needed to turn the entire atom on. Call this the cost of the atom, $c_i$.

3. Greedily seed the atoms in decreasing order of the size-to-cost ratio $|A_i|/c_i$ until we have used all $k$ seeds (skipping over any atoms that have seeding cost in excess of $k$) or there are not atoms left that can be seeded with the remaining number of seeds.

4. If there are seeds left over, select seeds uniformly at random from the set of nodes which are not in the $q$-closure of the set of seeds already selected.

In step 1 of the algorithm, we first find all conventions using the algorithm stated in the next section. Then, for each pair of nodes, we test if they are always together in/out of a convention for each convention, and then we put the nodes that behave the same in every convention into the same atom.

In step 2, for each atom, we can use brute force to determine the minimal set of nodes that, if they are informed, can make all the other nodes in the atom informed.

The greedy choice is seeding the atom with the largest size-to-cost ratio (atom size/ the minimal number of nodes to turn the entire atom on). If the seeds are not exhausted, we use the rest to seed the nodes that are not closely related to the nodes we have seeded randomly. Below is our own simulation result.

| # Blocks | Block Size | Density in Blocks | Density Across Blocks | # Seeds | Atom Based | Random |
|---|---|---|---|---|---|---|
| 2 | 20 | 0.2 | 0.1 | 5 | 1.000 000 | 0.803 625 |
| 2 | 20 | 0.5 | 0.1 | 5 | 1.000 000 | 0.359 250 |
| 2 | 20 | 0.2 | 0.1 | 8 | 1.000 000 | 0.984 375 |
| 2 | 20 | 0.5 | 0.1 | 8 | 1.000 000 | 0.953 750 |
| 2 | 30 | 0.2 | 0.1 | 5 | 0.116 667 | 0.243 500 |
| 2 | 30 | 0.5 | 0.1 | 5 | 0.500 000 | 0.081 750 |
| 2 | 30 | 0.2 | 0.1 | 8 | 1.000 000 | 0.801 917 |
| 2 | 30 | 0.5 | 0.1 | 8 | 0.500 000 | 0.249 833 |
| 3 | 20 | 0.2 | 0.1 | 5 | 0.183 333 | 0.445 000 |
| 3 | 20 | 0.5 | 0.1 | 5 | 1.000 000 | 0.110 583 |
| 3 | 20 | 0.2 | 0.1 | 8 | 1.000 000 | 0.874 583 |
| 3 | 20 | 0.5 | 0.1 | 8 | 1.000 000 | 0.444 167 |
| 3 | 30 | 0.2 | 0.1 | 5 | 0.055 556 | 0.057 500 |

Table 1: Simulation results for stochastic block model with various number of blocks, block size, number of seeds and connection probability within and across blocks. Atom-based seeding strategy can achieve high diffusion ratio and outperform random seeding strategy in most of the cases.

### 4.4.2 Approximation Method to Find Conventions

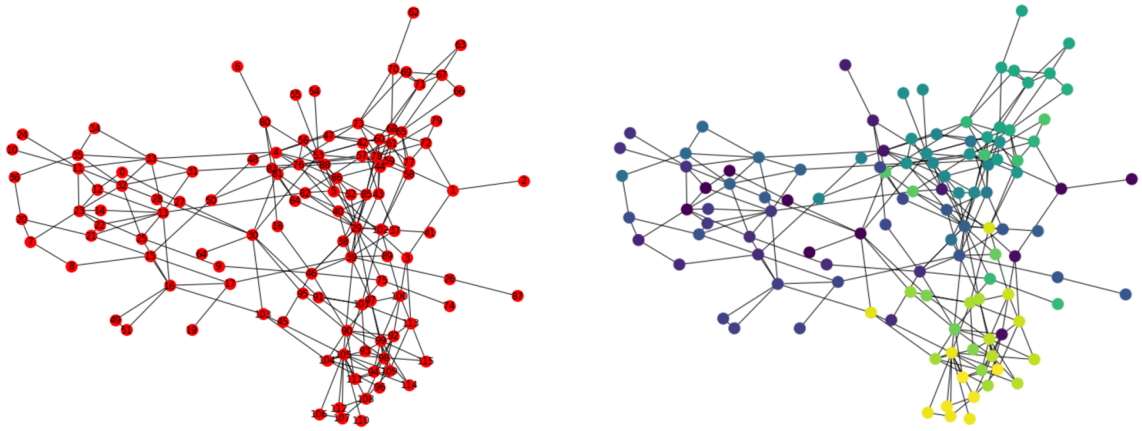
The algorithm starts by forming the collection of connected subsets of nodes of size less than some fixed $k$. The number of such subsets is at most $\binom{n}{k}$, and we will show later that $k$ can be fixed independent of $n$, so that the number of subsets formed is polynomial in $n$. The algorithm then selects a subset and generates its minimal $Q$-closed superset. If that superset also happens to be $Q$-cohesive, then it is a convention, and the algorithm stores the convention and moves on to the next subset in the working collection. If not, the algorithm adds to the superset the node whose addition most increases the node-wise minimum level of cohesion in the resulting subset, and then reiterates the preceding procedure with this new subset in place of the original subset. Since each such iteration adds at least one node to

the subset, the algorithm terminates in at most $\binom{n}{k} \times n$ steps.

Fix some $k$. The idea is to pick a subset $K$ of $k$ connected nodes and grow it through diffusion into a set as large as possible, $C$. The original set $K$ may contain some subset, $K'$, that have less than $q$ of their neighbors within $C$, so in this case, we add some of the neighbors of $K'$ to $C$ to make sure $K'$ have at least $q$ of their neighbors within $C$. Finally, $C$ is set where every node inside it has at least $q$ of its neighbors within $C$ and every node outside it has less than $q$ of its neighbors in $C$. We do this for every permutation of subsets of $k$ nodes and can terminate in polynomial time.

$k$ should not be too small, or the $k$ connected nodes are unlikely to form a large convention. $k$ should not be too large either, because the approximation algorithm can be too slow and can fail to form small conventions. In our application to the Indian village data, we choose $k = 3$.

### 4.5 Simulation Results on Real-world Networks



(a) Graph for connection activities among households in one of the Indian village real network.

(b) Atoms found in Indian village network, with threshold = 0.4. Nodes with the same color are in the same atom and tend to behave similarly when involved in activities.

Figure 19

## 5 Conclusion

Following Akbarpour et al.(2018) and Jackson and Storms (2018), we investigate the influence maximization problem in both theoretical and real-world networks. In our project, we

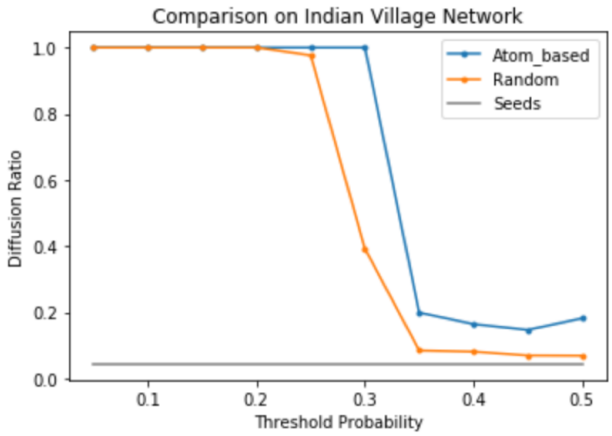



Figure 20: This figure illustrates Diffusion ratio comparison between atom-based seeding strategy and random strategy under different thresholds in the Indian village network. When the threshold becomes higher, the diffuse ability of random seeding strategy will merely be the same as the original number of seeds, whereas the atom-based strategy can achieve diffusion ratio around 20%.

mainly focus on three relevant seeding strategies, the optimal strategy, the omniscient strategy, and the random strategy. Our initial goal is to compare the performance of the optimal strategy and the random strategy, whereas even with full network information, it is still an NP-hard problem to execute the optimal seeding algorithm. Therefore, we modify our goal by comparing the omniscient strategy and random strategy to get an upper bound of the optimal result. For the omniscient strategy, we assume that we know everything about the network structure and the realized diffusion process, although this is infeasible in the real world.

Our tests in both theoretical networks and real-world networks prove that under some circumstances, the random seeding strategy with a few more seeds performs better than the omniscient strategy. Additionally, we test the random seeding strategy under the threshold model as described in Jackson and Storms (2018). We find that the random seeding strategy does not perform well when the homogeneous threshold becomes moderately high.

For future work, just as illustrated in Akbarpour et al.(2018), the analysis here shows that there are possible situations under which the position of seeds in the network is not a primary concern, although network targeting could be valuable under several circumstances. Whether those situations are satisfied in a specific context is an inherently empirical question. Much remains to be done to quantify the value of network information in other environments.

# References


[1] Akbarpour, M., Malladi, S., & Saberi, A. (2018). Just a few seeds more: value of network information for diffusion. *Available at SSRN* 3062830.

[2] Jackson, M. O., & Storms, E. (2018). Behavioral communities and the atomic structure of networks. *Available at SSRN* 3049748.

[3] Krause, A., & Golovin, D. (2014). Submodular function maximization.

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

[5] Erdős, P. (1959). A. R ényi, On random graphs .I *Publ. Math. Debrece*n, 6(290-297), 156.

[6] Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144), 1236498.

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

[8] Leskovec, J., & Krevl, A. (2014). SNAP Datasets : Stanford Large Network Dataset Collection.

[9] Morris, S. (2000). Contagion. *The Review of Economic Studies*, 67(1), 57-78.


# Appendices

## A Some Basic Proof

In this part, we give a lemma in proving Theorem 1. This lemma is focused on the performance of RAND and OMN on the communication graph $\mathcal{K}(G)$ for an arbitrary $G$.

**Lemma 1.** *Let $\mathcal{K}(G) = \mathcal{K}(G)$ 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,*

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

*and*

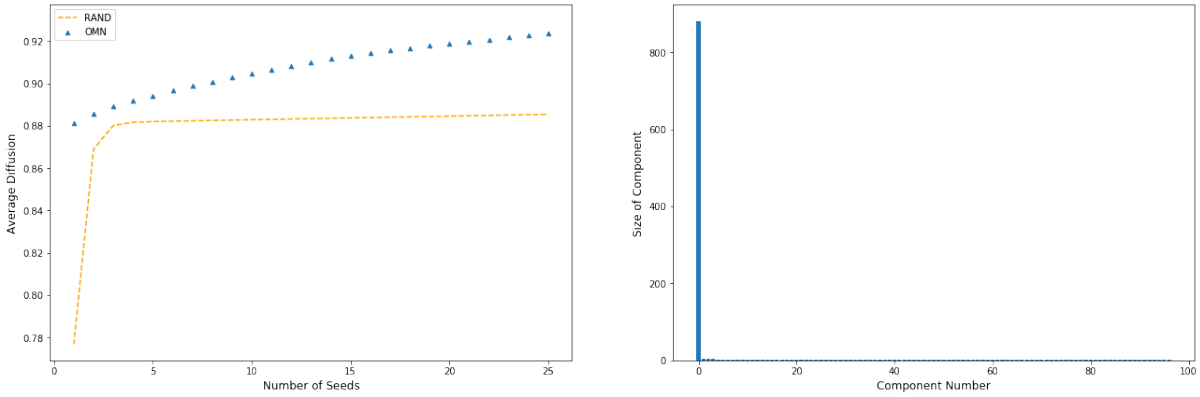
$$\boldsymbol{h}(G, s, RAND) = E[\sum_{i=1}^{CC} \mathcal{C}_i \left(1 - (1 - \frac{\mathcal{C}_i}{n})^s\right)] \tag{2}$$

*Proof.* 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 is just what equation (2) told us. □

## B $k$-level Network

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

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




(a) $k$-level Network Simulations: $n = 1000, d = 2.4, q_1 = 0.1, q_2 = 0.05$

(b) Size of Connected Components in $k$-level Network

Figure 21: Simulation result in $k$-level Network