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

Virtual Market Design Seminar

Mohammad Akbarpour, *Stanford*
Suraj Malladi, *Stanford*
Amin Saberi, *Stanford*

June 2020

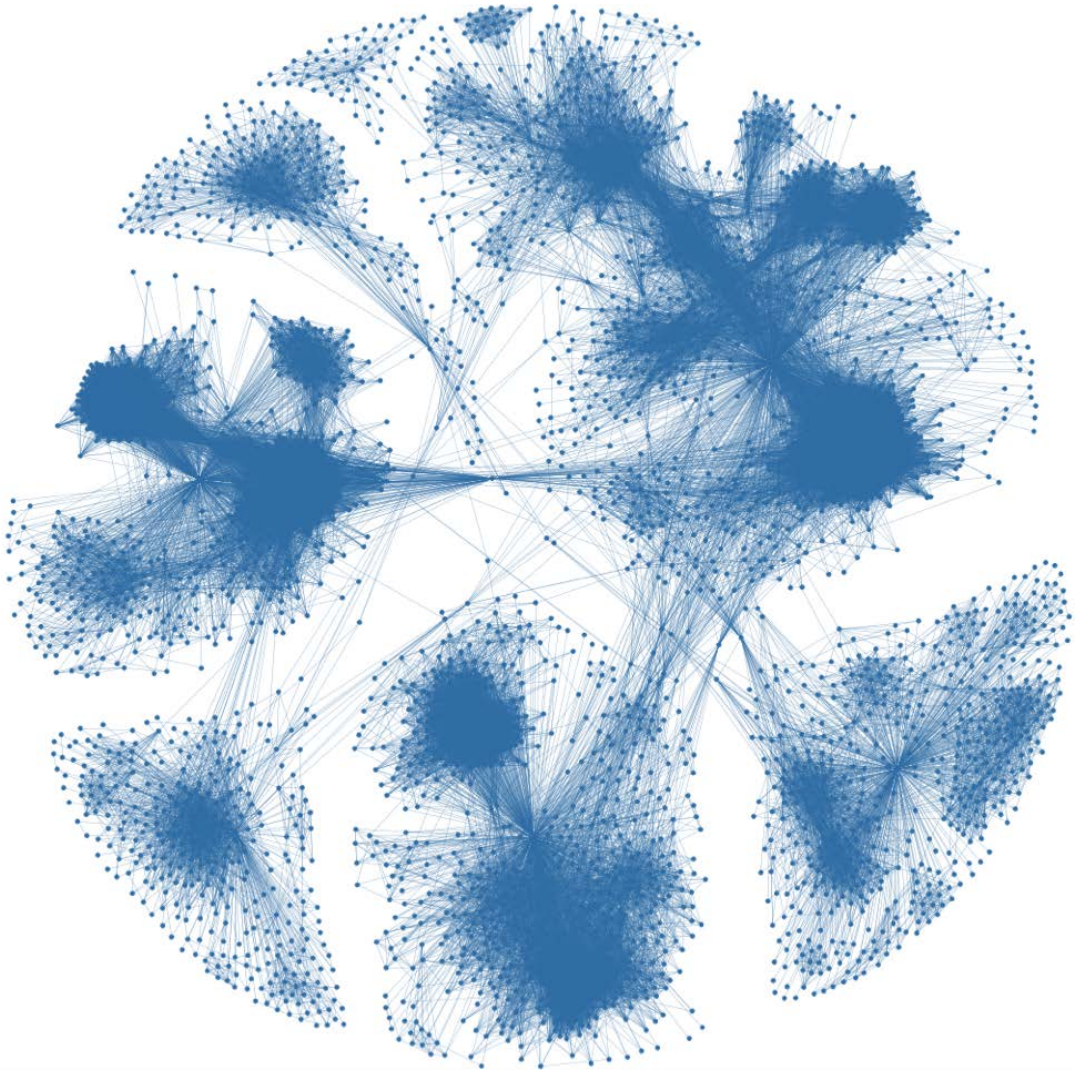
Optimal Seeding

- Fix:
  - Social network with *n* people
  - Diffusion model
    - For example: Once informed, you inform your friends with some probability.
  - Budget for informing S “seeds” initially

- Who are the optimal S individuals to seeds? (from $\binom{n}{S}$ options)

Sometimes Easy, Often (Very) Hard

- With 3 seeds:



Thanks to Ozan Candogan for this network visualization of their paper

Optimal Seeding

- Fix:
  - Social network with *n* people
  - Diffusion model
    - For example, SIR: Once informed, you inform your friends with probability ***c***.
  - Budget for informing S “seeds” initially

- Who are the optimal S individuals to seeds? (from $\binom{n}{S}$ options)

- **The problem is NP-Complete (Kempe-Kleinberg-Tardos (2003))**

Network Seeding: Many Applications

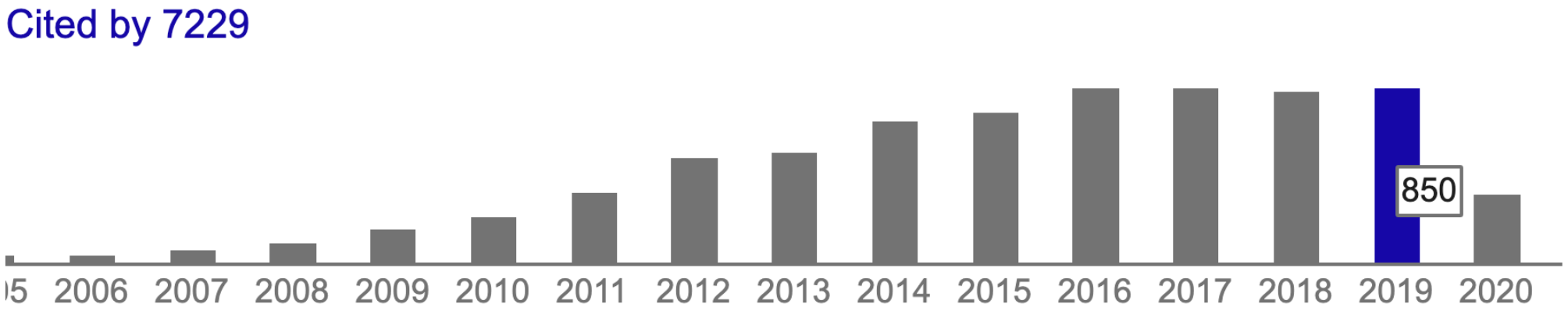
- **Diffusion of microfinance** [Banerjee *et al*, 2013, …]
- **Spread of new technologies** [Beaman *et al*, 2018, …]
- **HIV prevention information** [Wilder *et al*, 2017, …]
- **Word-of-mouth marketing** [Domingos, 2005, …]
- **Spread of political ideas** [Lazarsfield *et al,* 1944, …]
- **…**

Heuristic Solutions for Optimal Seeding

- **Eigenvector centrality** (Cai *et al* 2015)
- **Diffusion centrality** (Banerjee *et al* 2013)
- **K-shell index** (Kitsak et al 2010)
- **Discounted Degree** (Chen et al 2009)
- **Diffusion-Based Detection** (Wang et al 2010, Jackson-Storms 2017)
- **Others** (Narayanam et al 2010)**,** (Leskovec et al 2007)**,** (Jiang et al 2011)**,** (Zhou et al 2014)...

Heuristic Solutions for Optimal Seeding

Kempe-Kleinberg-Tardos (2003)



What Have We Learned?

[Network targeting] has important implications for policy makers to pick the right people to inform in order to ensure that a new idea or product or piece of information reaches the maximum number of people.

Banerjee-Chandrasekhar-Duflo-Jackson, *Diffusion of Microfinance*, (2013)

What Have We Learned?

Targeting individuals who are more central in the village network for this intervention can make a significant difference in the size of the multipliers achieved.

Cai-Janvry-Sadoulet, *Social Networks and the Decision to Insure* (2015)

What Have We Learned?

Theory-driven targeting using detailed social network data can increase technology adoption relative to the status quo approach to agricultural extension services.

Beaman-BenYishay-Magruder-Mobarak, *Can Network Theory-based Targeting Increase Technology Adoption?* (2020)

Computationally Feasible, But Require…

- **Eigenvector centrality** (Cai *et al* 2015)
- **Diffusion centrality** (Banarjee *et al* 2013)
- **K-shell index** (Kitsak et al 2010)
- **Discounted Degree** (Chen et al 2009)
- **Diffusion-Based Detection** (Wang et al 2010, Jackson-Storms 2017)
- **Others** (Narayanam et al 2010)**,** (Leskovec et al 2007)**,** (Jiang et al 2011)**,** (Zhou et al 2014)...

Computationally Feasible, But Require…

- **Eigenvector centrality** (Cai *et al* 2015)
- **Diffusion centrality** (Banarjee *et al* 2013)
- **K-shell index** (Kitsak et al
- **Discounted Degree** (Che
- **Diffusion-Based Detection** (Wang et al 2010, Jackson-Storms 2017)
- **Others** (Narayanam et al 2010), (Leskovec et al 2007), (Jiang et al 2011), (Zhou et al 201

Network Data

Expensive

Imperfect

Which network we're talking about?

Banerjee-Chandrasekhar-Duflo-Jackson (*ReStud*, 2020)

Breza-Chandrasekhar-McCormick-Pan (Forthcoming, *AER*)

Quantifying the Value of Network Data: A Simulation

Diffusion of Microfinance [Banerjee-Chandrasekhan-Duflo-Jackson, 2013]:

- Network data
- Model of diffusion:
  - Once informed: **participate** or **not participate** with prob. **p**
  - Participating agents communicate with prob. c_p
  - Non-participating agents communicate with prob. c_n
  - The whole diffusion process stops after **T** periods
- Structural estimates of parameters

Quantifying the Value of Network Data: A Simulation

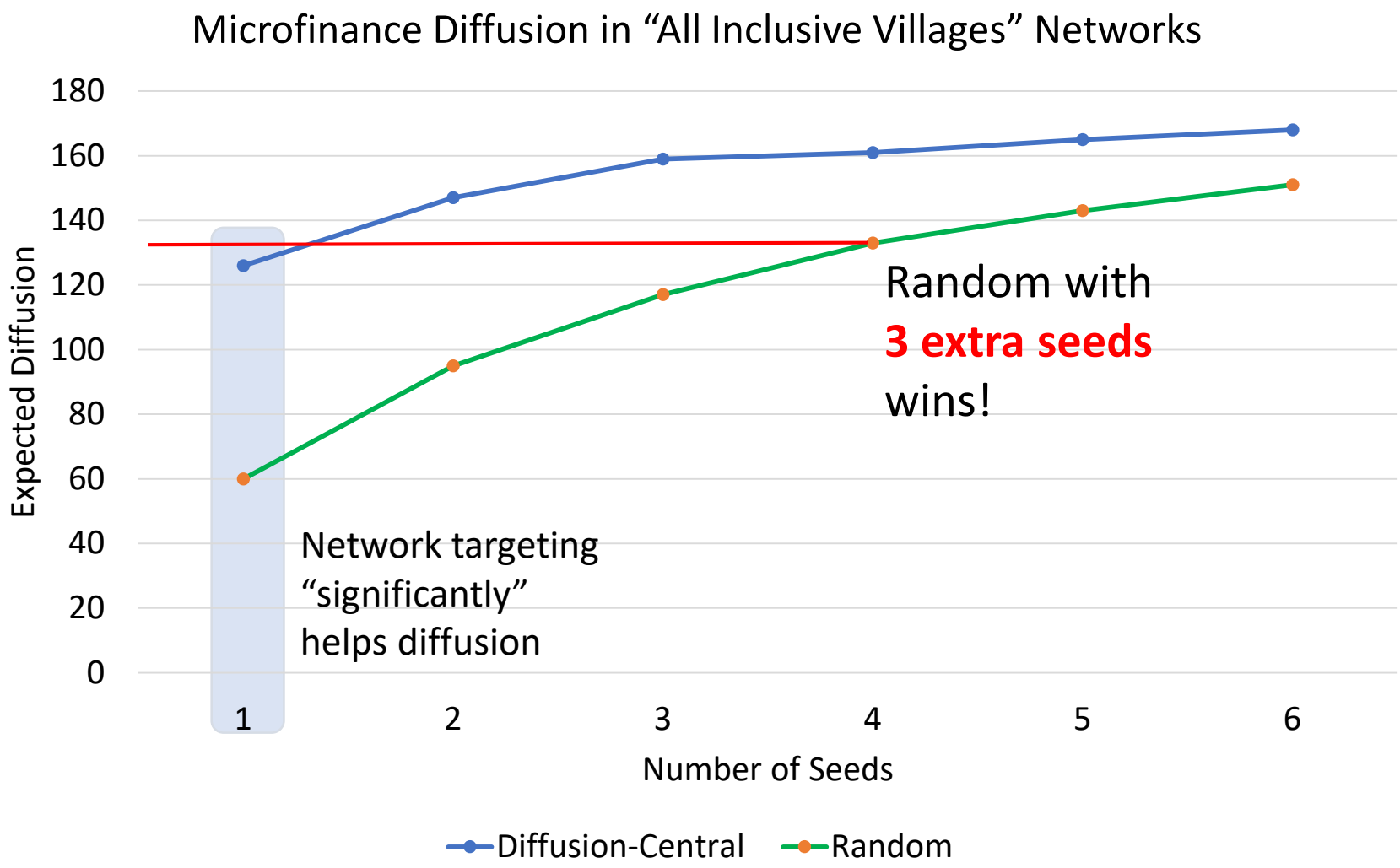
Diffusion of Microfinance [Banerjee-Chandrasekhan-Duflo-Jackson, *Science*, 2013]:

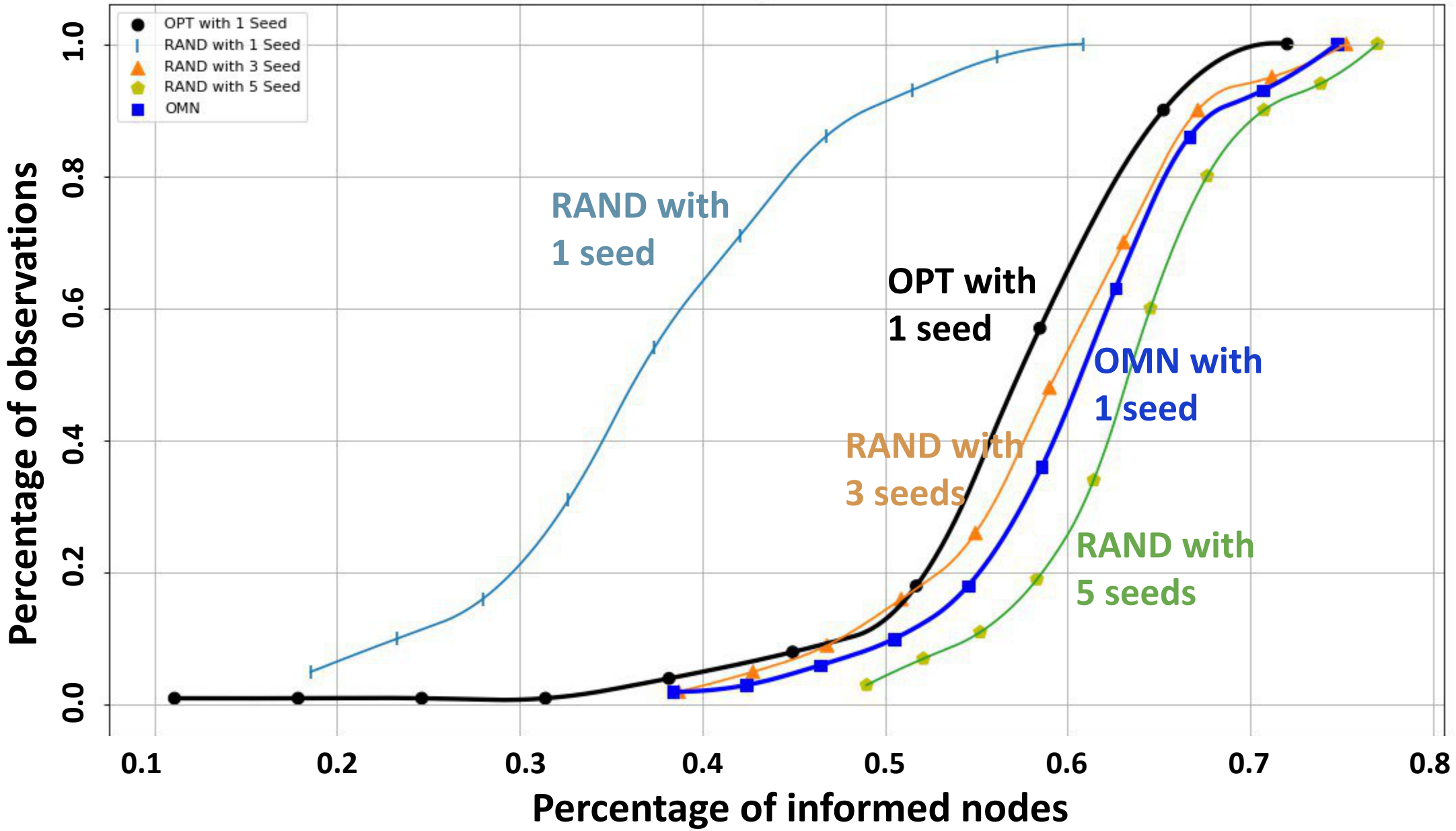
- Network data
- Model of diffusion:
  - Once informed: **participate** or **not participate** with prob. **0.24**
  - Participating agents communicate with prob. **0.55**
  - Non-participating agents communicate with prob. **0.05**
  - The whole diffusion process stops after **7** periods
- Structural estimates of parameters

The heuristic suggested based on experiment is “diffusion-centrality”

Network Targeting vs. “Naïve” Expanded Outreach

asd




OPT with 1 Seed
RAND with 1 Seed
RAND with 3 Seed
RAND with 5 Seed
OMN
RAND with
1 seed
OPT with
1 seed
OMN with
1 seed
RAND with
3 seeds
RAND with
5 seeds
Percentage of observations
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Percentage of informed nodes

This Paper: Value of Network Data and Analysis

Exploit the network:
Inform s agents
optimally

VS.

Ignore the network:
Inform $s + x$ agents
randomly

For what values of x does random seeding beat optimal?

Quantifies the value of network data + computational power using a policy-relevant measure.

Result: *In viral diffusion, for a wide class of diffusion models, random with "a few" extra seeds outperforms the optimum.* (and yes, even for power-law networks)

Outline

1. **Model**

2. **Main Theorem & Proof ideas**

3. **Power-Law and Real-world Networks**

4. **Limitations: Towards Guiding Empirical Research**
    1. Diffusion model
    2. Speed of diffusion
    3. Diffusion minimization

5. **Concluding Remarks**

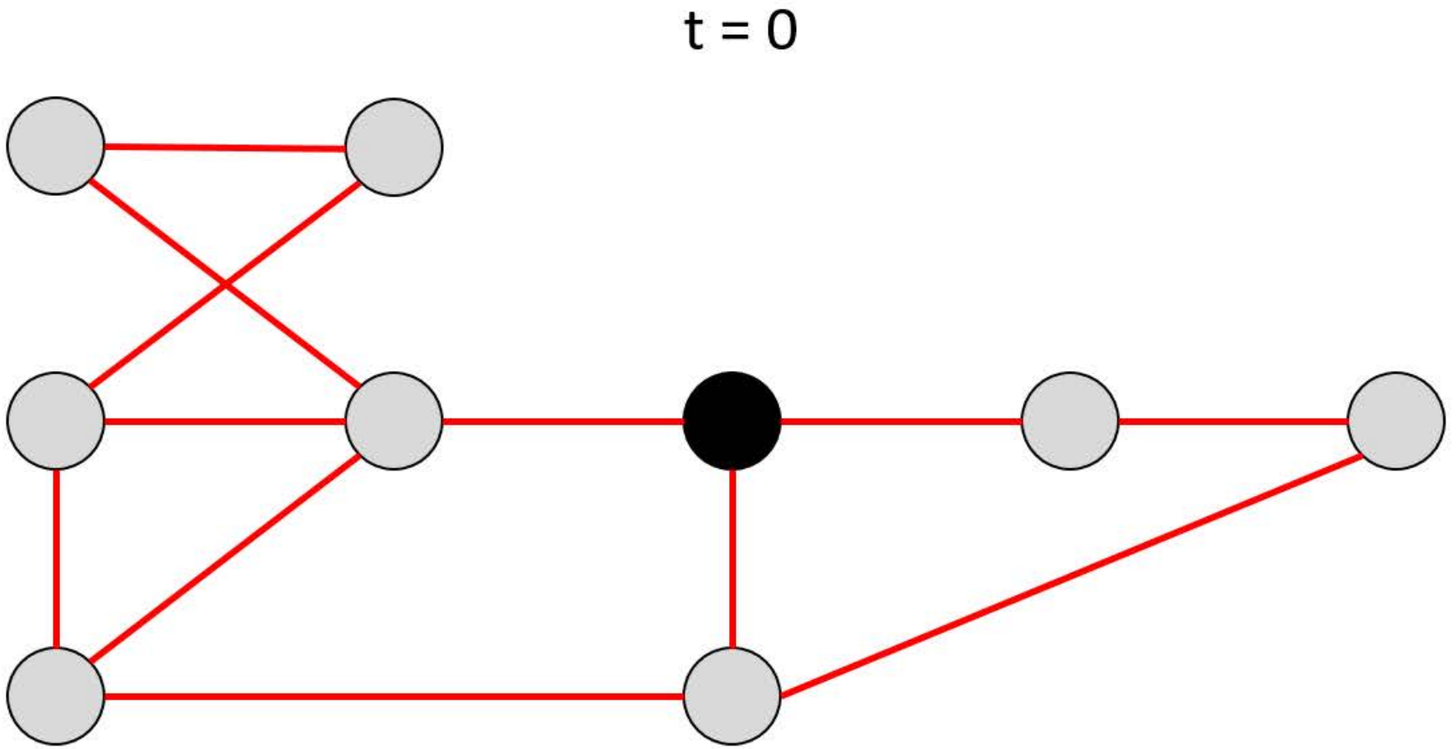
Model

Model: Basics

- $N = \{1, \ldots, n\}$ is the set of **agents** (or nodes)

- $G = (N, E)$ is the **social network**
  - $E \subseteq N^2$
  - $ij \in E$ if agents i and j are **linked** (or friends, neighbors)
  - Consider **undirected networks:** if $ij \in E \leftrightarrow ji \in E$ (study directed in paper)

Seeding and Diffusion

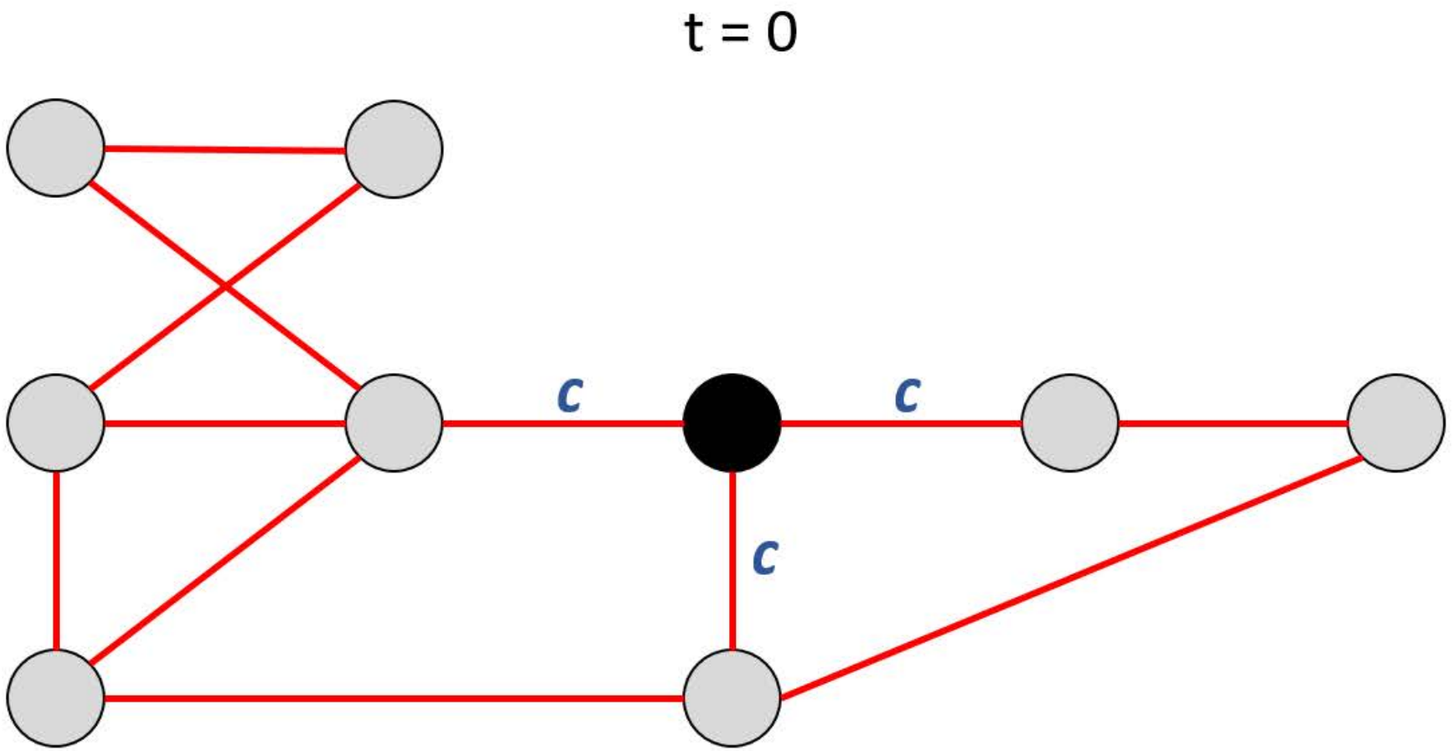


c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion

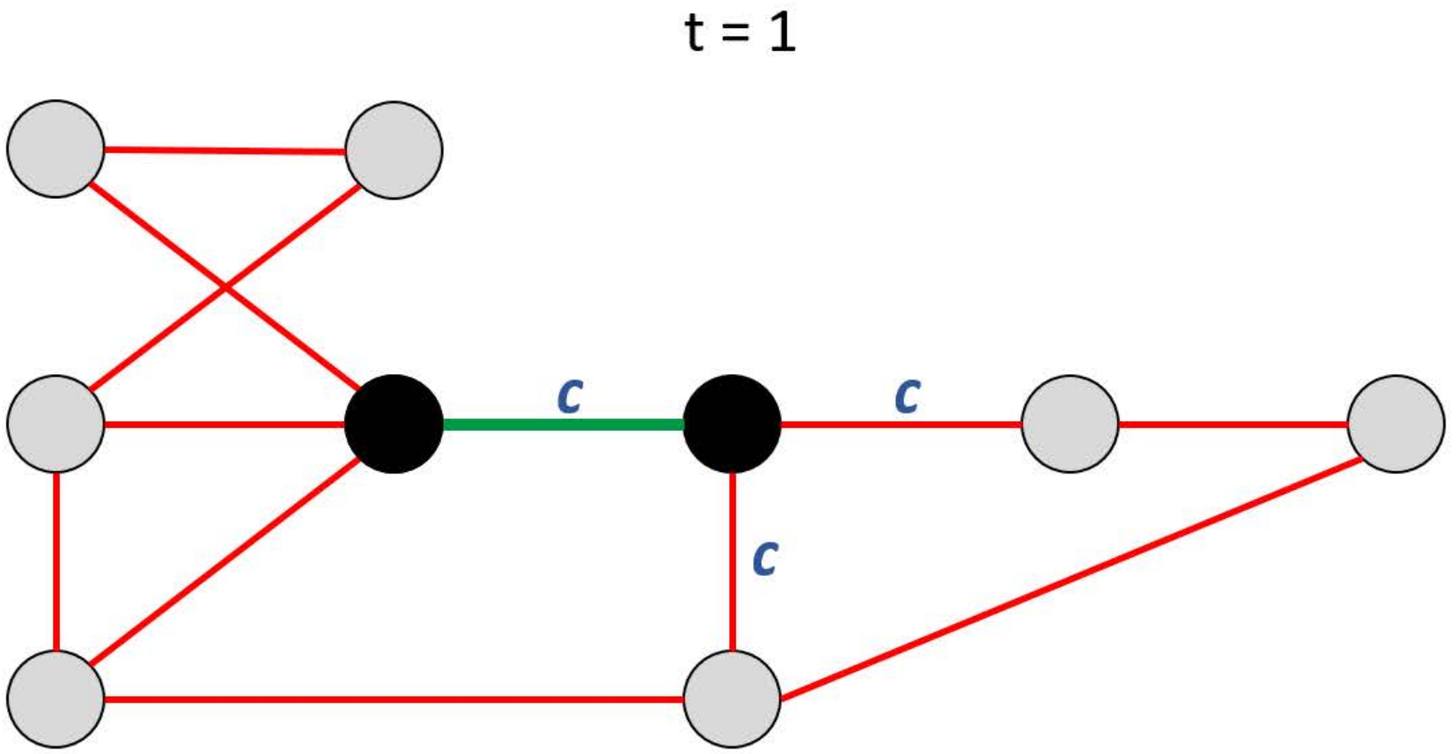


c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion

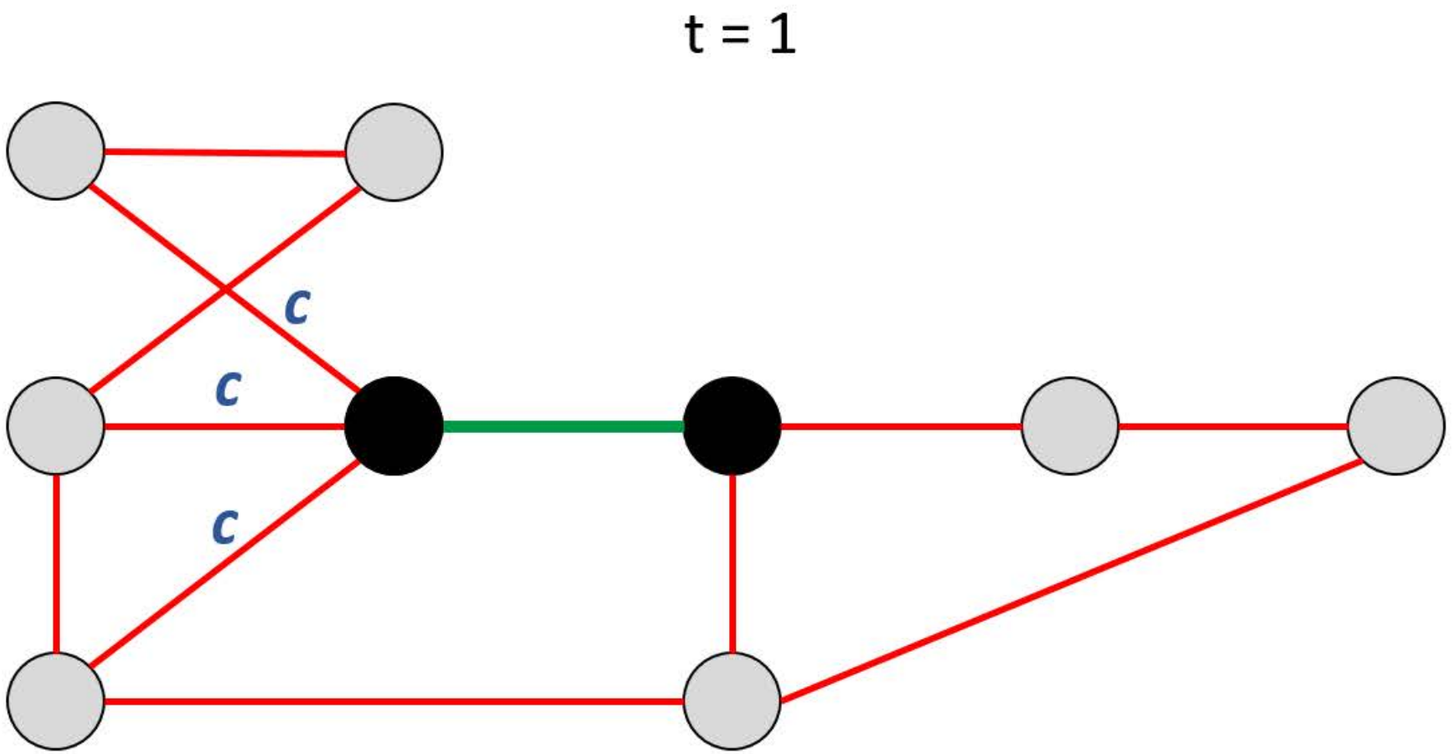


c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion



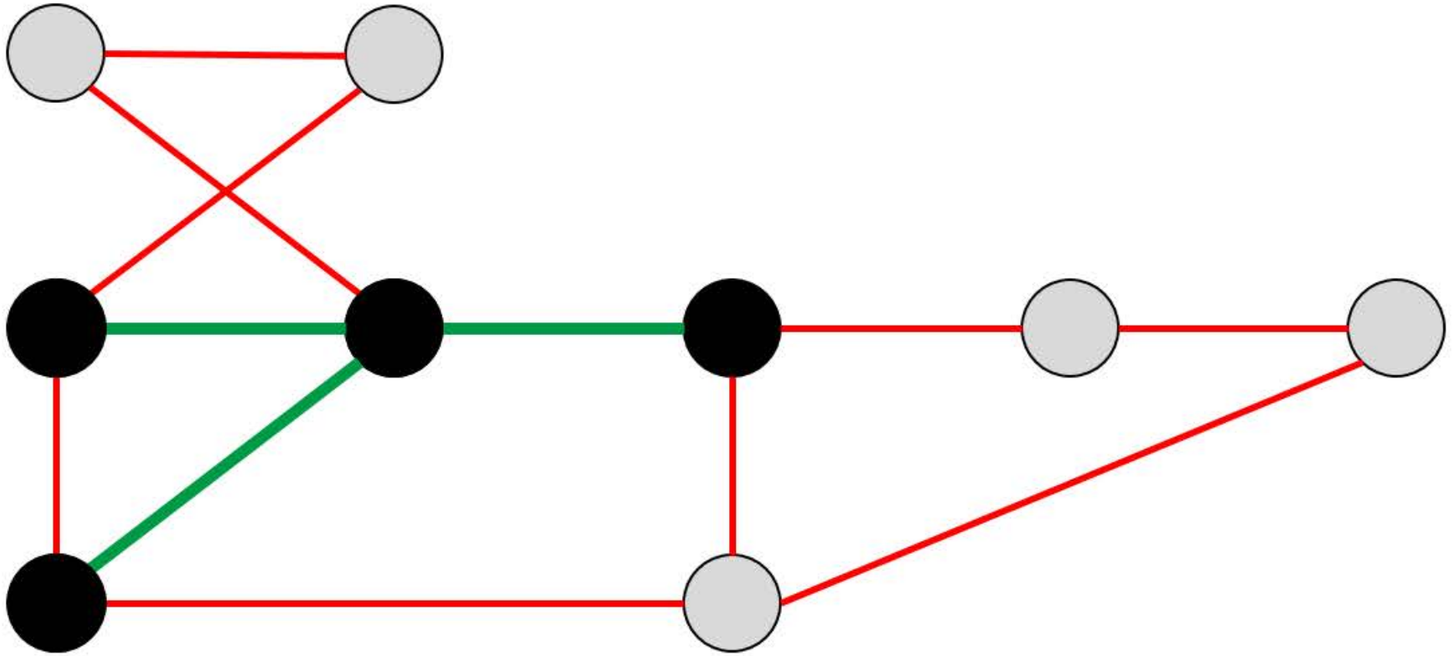
c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion

t = 2

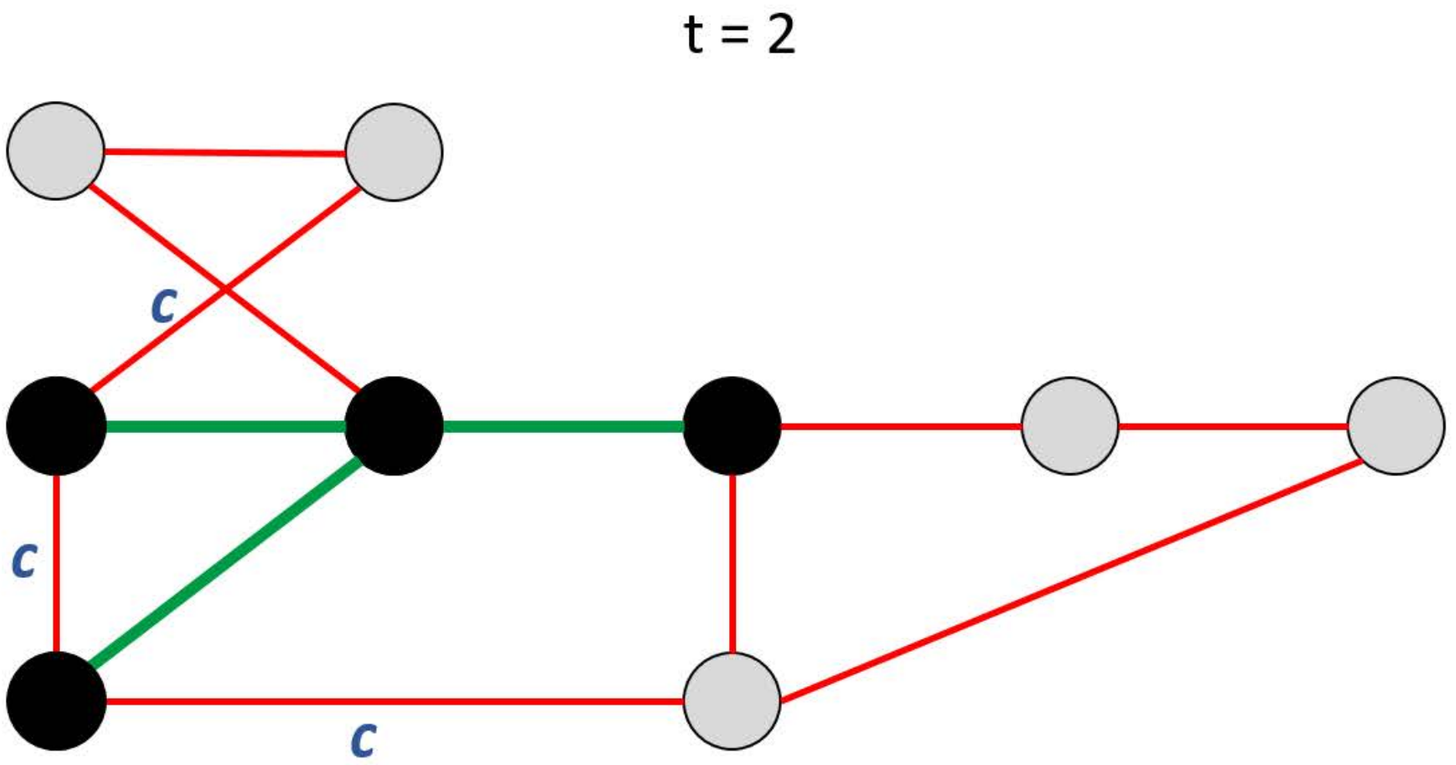


c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion

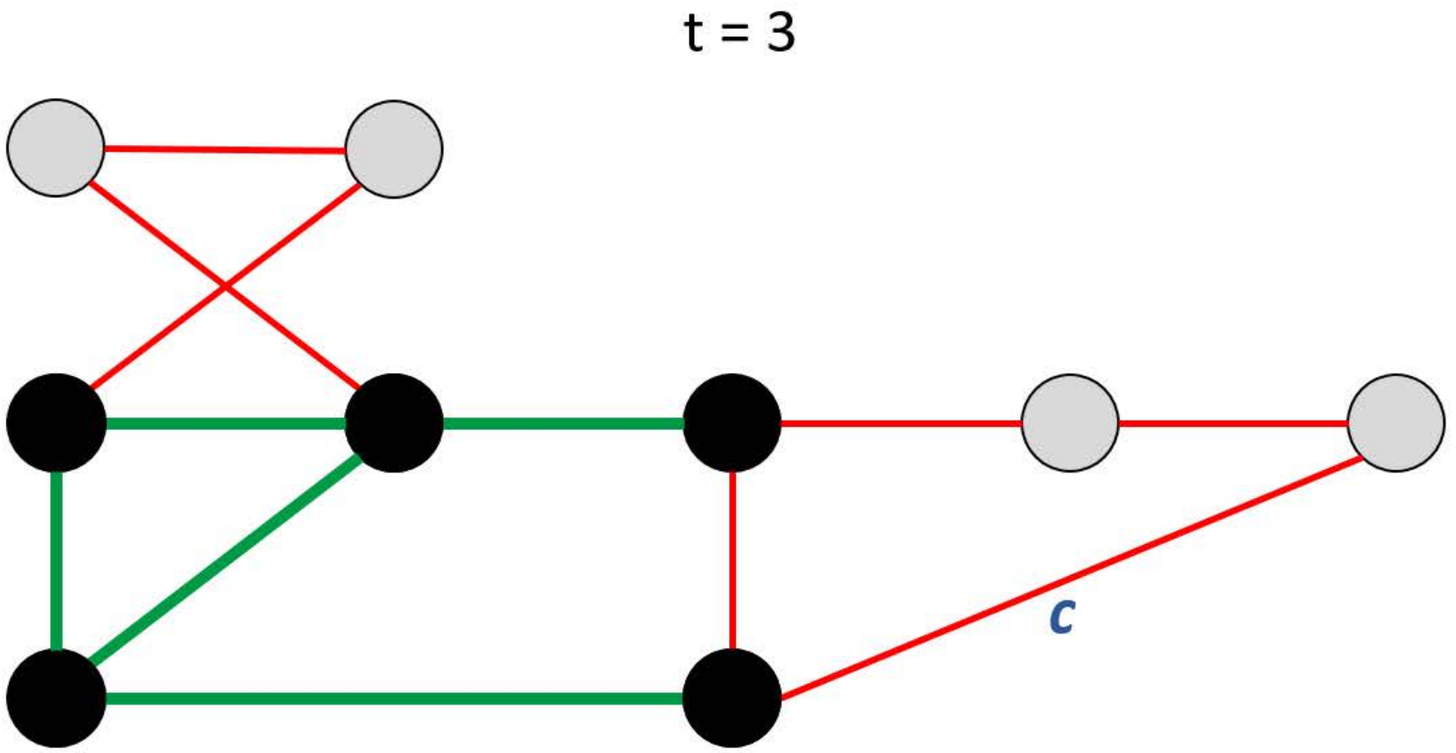


c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion

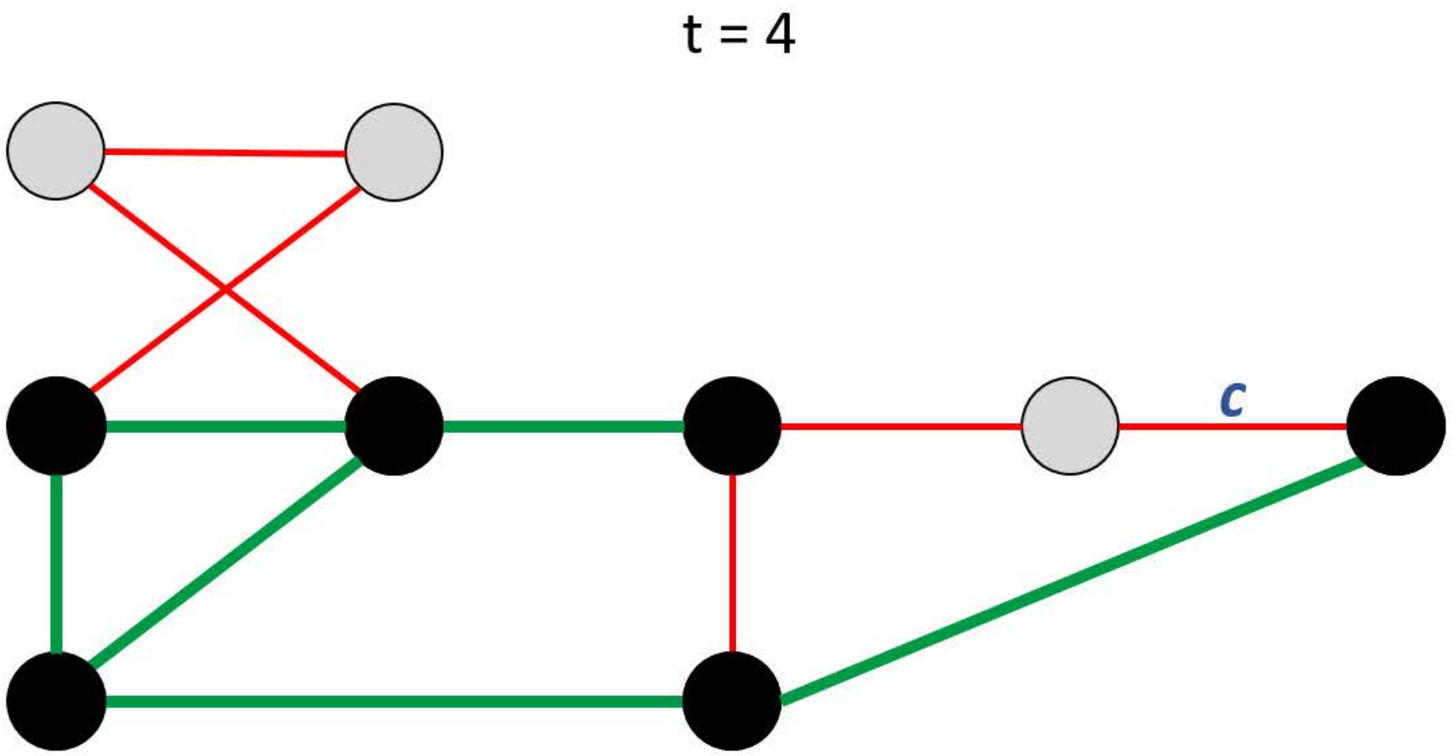


c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion



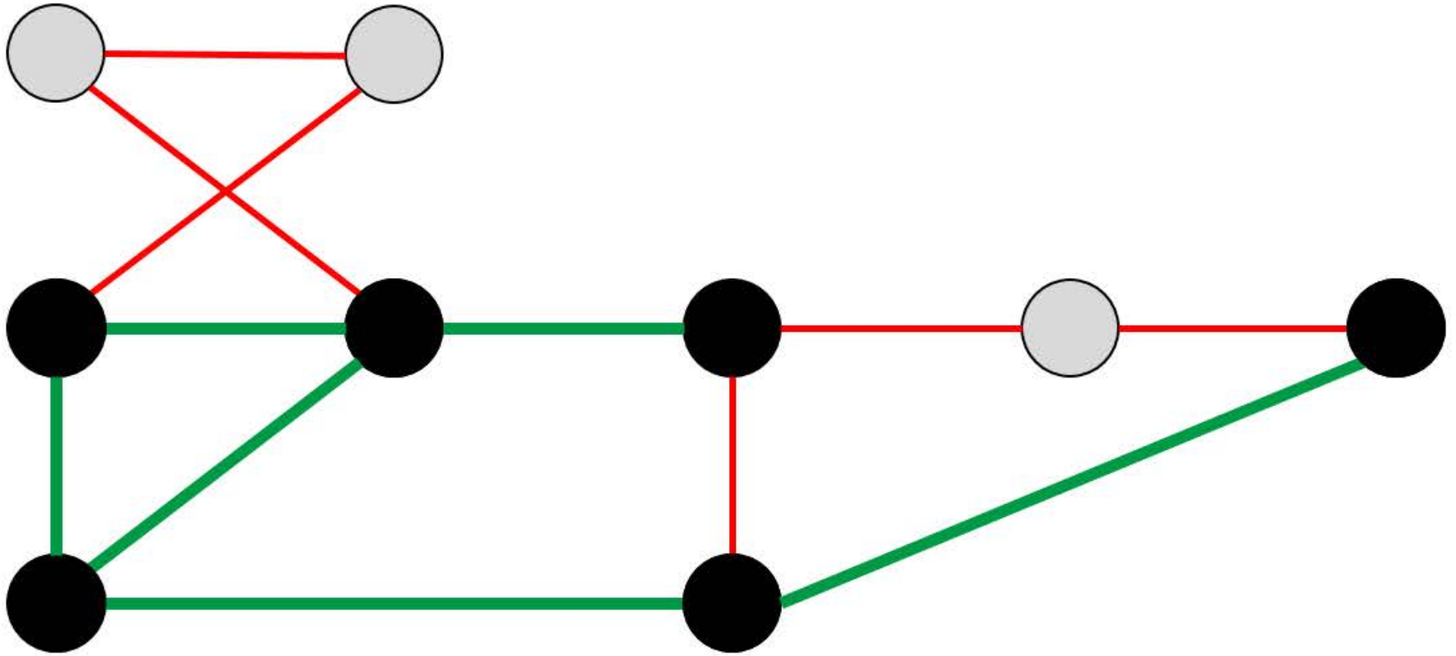
c is the *communication probability*

Informed

Uninformed

Seeding and Diffusion

t = 4



c is the *communication probability*

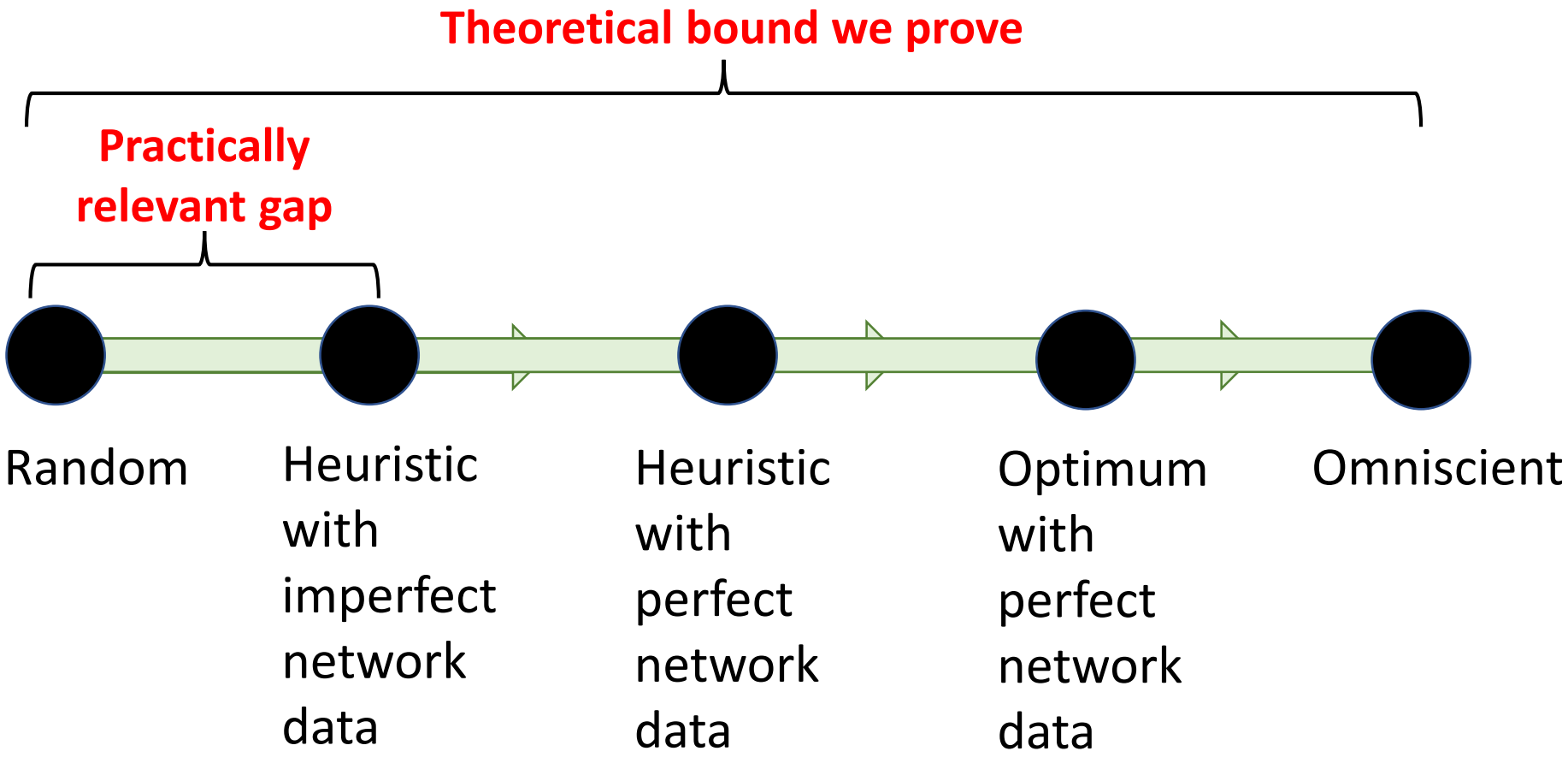
Informed

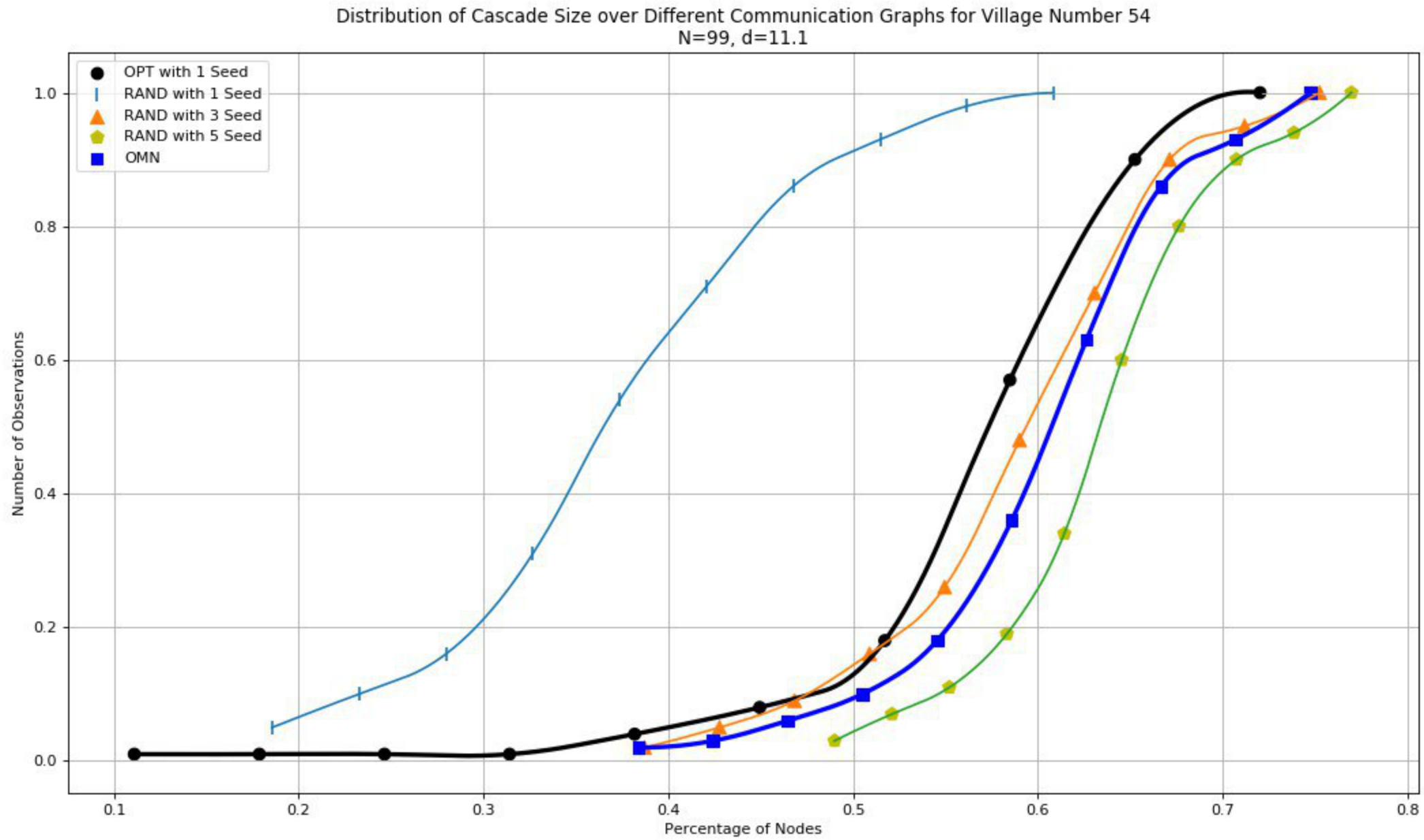
Uninformed

Seeding Strategies

- $\boldsymbol{f}(G, s)$*:* seeding strategy
  - DEGREE($\boldsymbol{G}, \boldsymbol{s}$) returns the $\boldsymbol{s}$ nodes with the highest degree in network $\boldsymbol{G}$
  - RAND($\boldsymbol{G}, \boldsymbol{s}$) returns a set of $\boldsymbol{s}$ nodes in the network $\boldsymbol{G}$, selected uniformly at random
- $\mathbf{A}(G, s, f)$: the (stochastic) set of agents eventually informed
- $\mathbf{h}(G, s, f) \stackrel{\text{def}}{=} \mathbb{E}[|\mathbf{A}(G, s, f)|]/n$, expected fraction of informed agents
- $\mathrm{OPT}(G, s) \in \operatorname{argmax}_f \mathbf{h}(G, s, f)$
- $\mathrm{OMN}(G, s)$: selects the best s seeds, given the realization of who would speak with whom

The Intellectual Exercise




Distribution of Cascade Size over Different Communication Graphs for Village Number 54
N=99, d=11.1
OPT with 1 Seed
RAND with 1 Seed
RAND with 3 Seed
RAND with 5 Seed
OMN
Number of Observations
Percentage of Nodes
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Network Model: Inhomogeneous Random Networks (IRN)

- Each node i has some *type* $\theta_i \in \tau = \{1, 2, \ldots, r\}$
- *Kernel function* is a symmetric function k: $\tau^2 \to [0, n]$
- Each type θ_i and θ_j are linked with probability $0 \leq k(\theta_i, \theta_j)/n \leq 1$

- Let k_{ij} be the expected number of type θ_j friends of a θ_i node
- Let $\mathbf{T}_k = [k_{ij}]_{i,j \in \tau}$ and consider its *largest eigenvalue*:

$$\|\mathbf{T}_\kappa\| = \sup_{\mathbf{x}:\ \|\mathbf{x}\|_2 \leq 1} \|\mathbf{T}_\kappa \mathbf{x}\|_2, \quad \text{where} \quad \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{r} x_i^2}.$$

Main Theorem

Define: $\mathbf{H}(f, s) \stackrel{\text{def}}{=} \mathbb{E}_{\mathsf{G}}[\mathbf{h}(G, s, f)]$

- Expected fraction of infected nodes given seeding strategy with ***s*** seeds, drawing an IRN

Let $\alpha = \lim_{n \to \infty} \mathbf{H}(\mathrm{OMN}, 1)$

Theorem. Let $s = o(\frac{n}{\log(n)})$. If $||\mathbf{T}_k|| > \frac{1}{c}$, then $\alpha > 0$ and for any x:

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

A Corollary: Single-type IRN

Erdős–Rényi random graph is a special case of IRN with $\boldsymbol{k(\theta_i, \theta_j) = d}$ for all types.

Definition. In an **Erdős–Rényi random graph** on $\boldsymbol{n}$ nodes and parameter $\boldsymbol{d}$, link exists between any two nodes independently with probability $\boldsymbol{d/n}$

- $\boldsymbol{d}$ is the expected number of friends (**degree)** of a node

Value of Network Data: Erdős–Rényi $(n, d/n)$

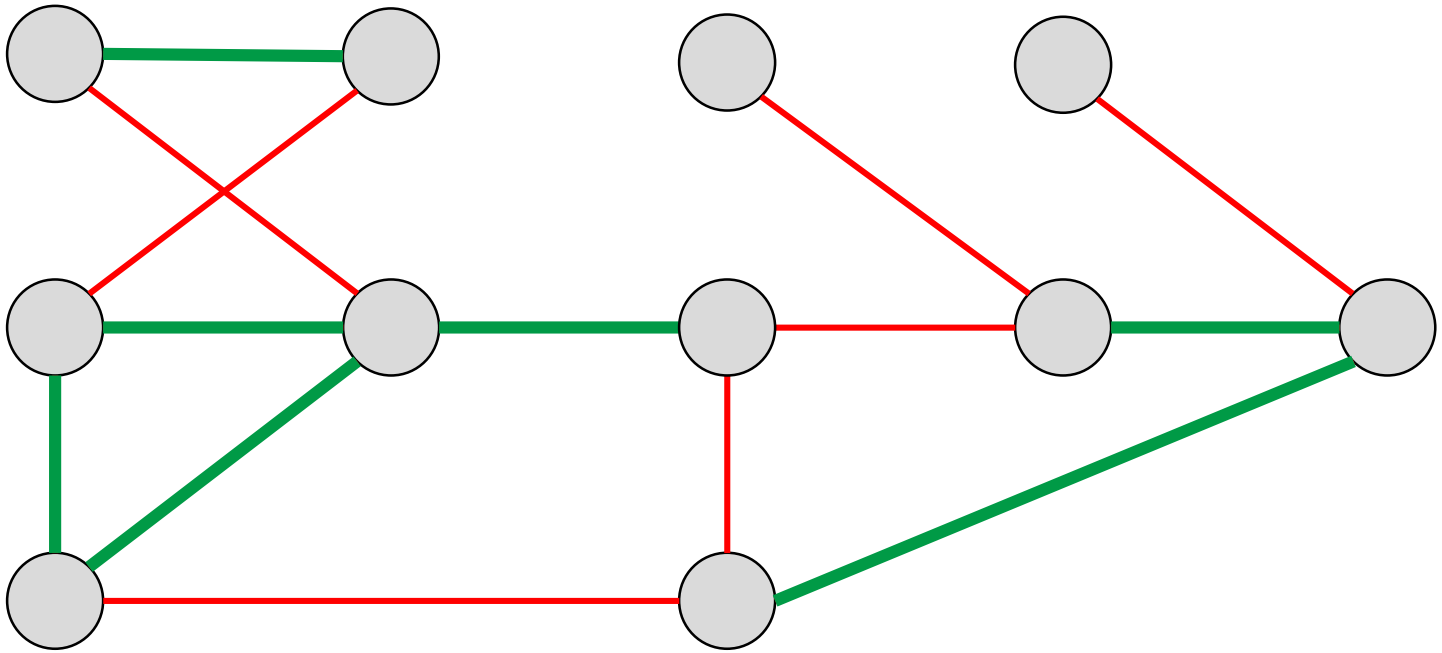
Theorem 1. Let $s = o(\frac{n}{\log(n)})$. If $dc > 1$, then for any x:

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

Proof Ideas

(it’s not about friendship, it’s about *communication*)

Proof Ideas: Communication Network

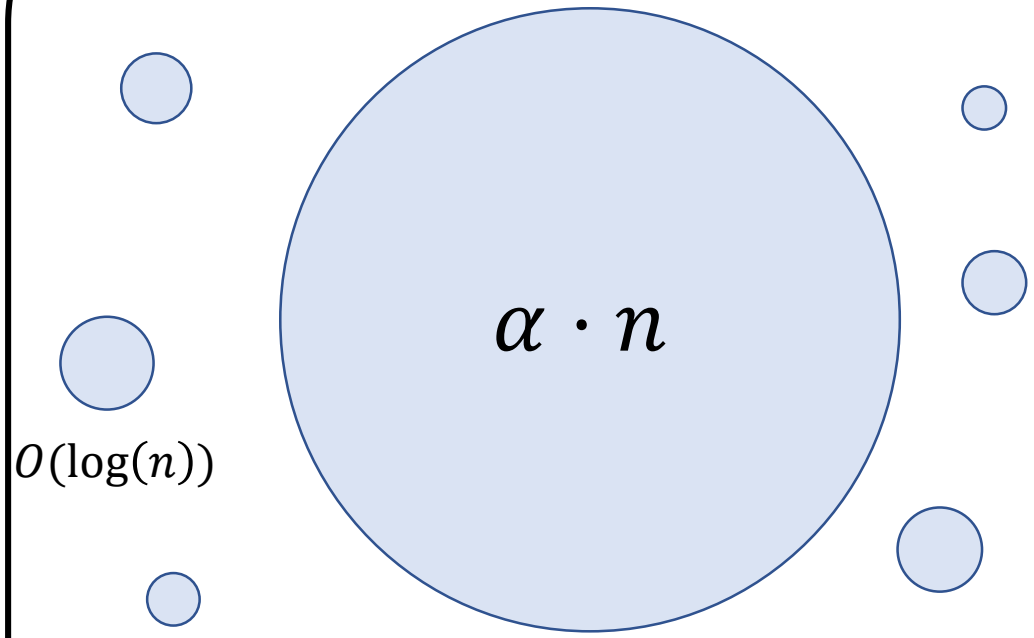


If any agent in a connected component is informed, all others are.

- OMNICIENT seeding strategy picks **top *s*** component sizes.
- RAND picks them with prob. **proportional to their size**

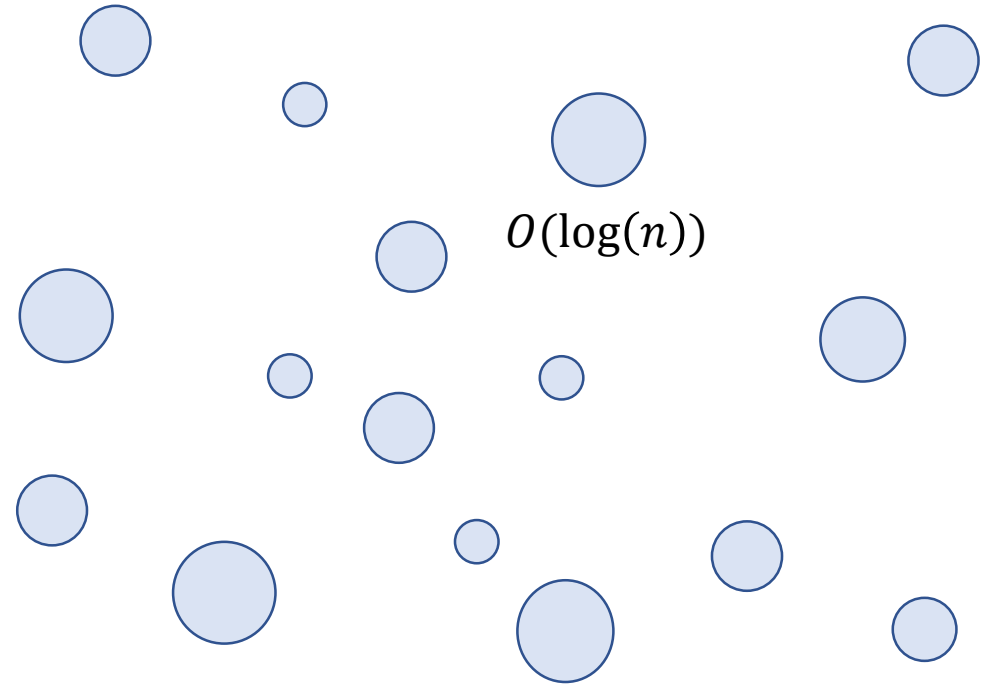
Communication Network Components' Sizes

Regime 1: **if** $||\mathbf{T}_k|| > \frac{1}{c}$



Chance random hits the giant component: $1 - (1 - \alpha)^{s+x}$

Regime 2: **if** $||\mathbf{T}_k|| < \frac{1}{c}$

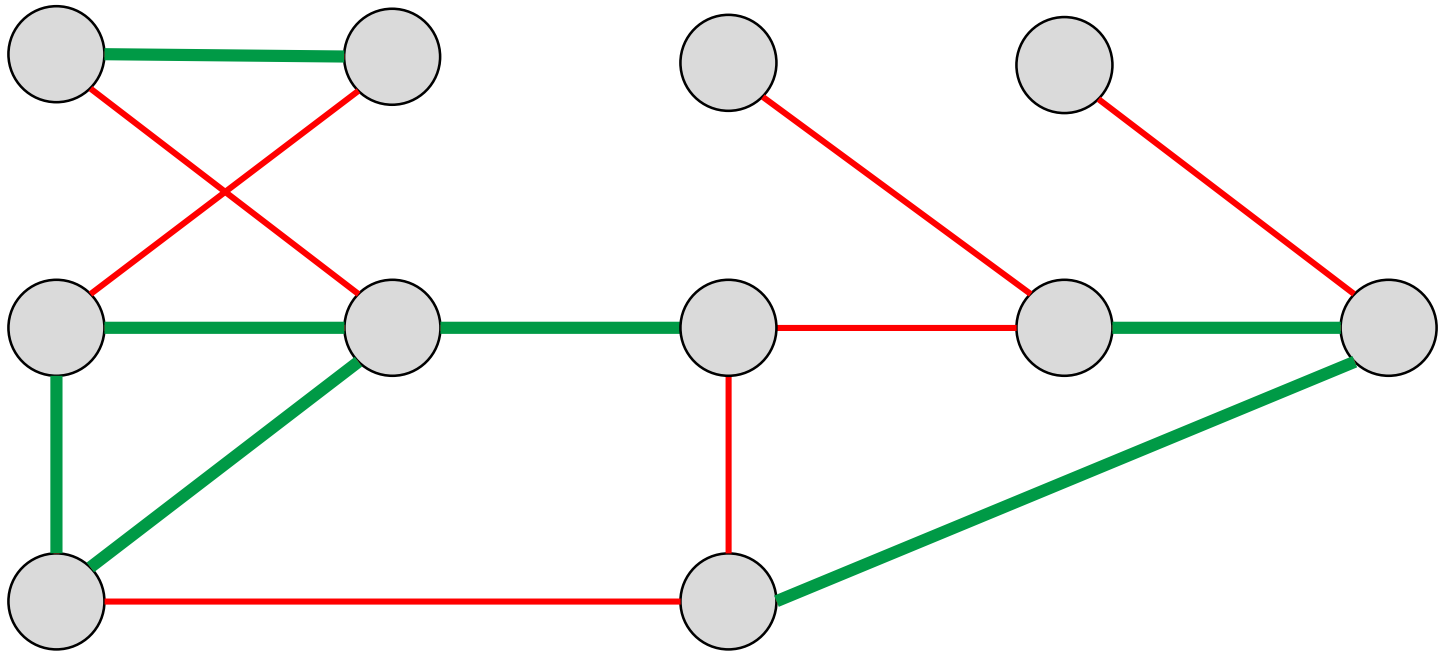


Even omniscient cannot do much

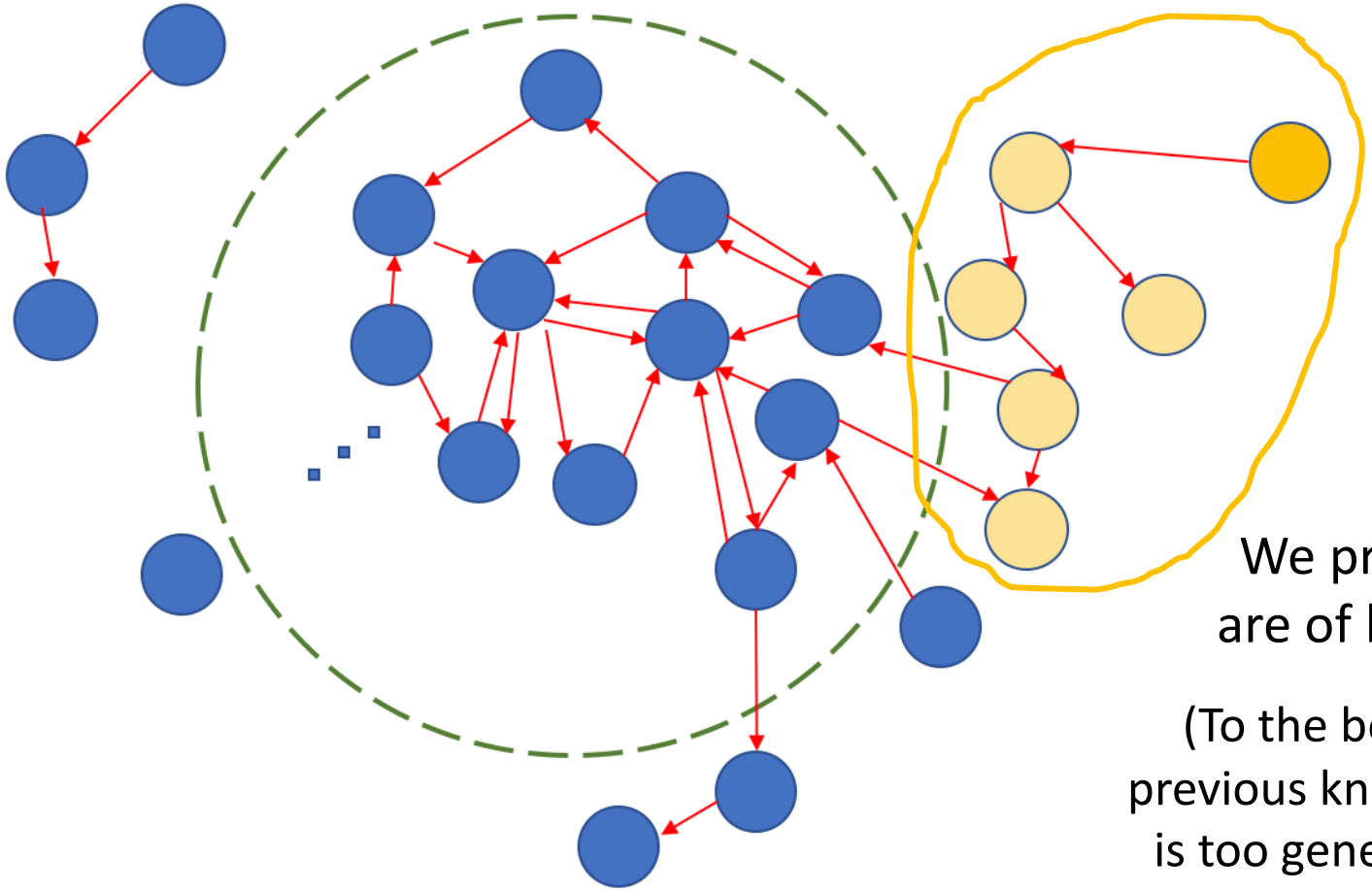
Some Lessons for Corona (March 4th, 2020)

- In ER, if $cd = 2$, then giant component is 79% of the network!
- $||\mathbf{T}_k|| > \frac{1}{c}$ is the condition for the pandemic to go viral.
  - LHS is only a function of social network structure
  - RHS is only a function of the virus, hand-washing, etc
- If the condition holds, just a few (random) seeds are enough!
- Power-law networks makes it even easier for the condition to be satisfied!

(Un)Directed Communication



Directed Communication



OMN will pick these *in addition* to those in the giant component.

We prove: These paths are of length $\boldsymbol{O(log(n))}$

(To the best of our knowledge, previous known bound is $\sqrt{n}$, which is too generous for our purposes)

Power-law Networks

(What about @TaylorSwift or @LeoMessi?)

Chung-Lu (2002) Networks

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

- Node i expected degree is w_i
- A special case of IRN !

Random vs. OMN: Power-law Networks

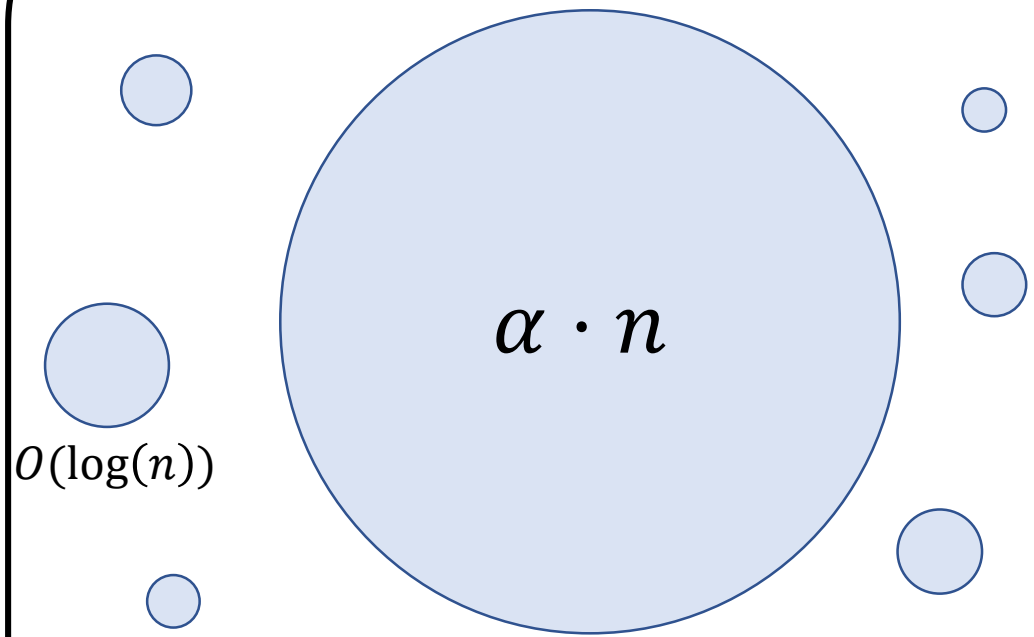
- Let $F(x) = 1 - (d/x)^b$ be weight distribution in a Chung-Lu graph, for $b > 1$.

Theorem. Let $s = o(\frac{n}{\log(n)})$. If either $b \in (1, 2]$, or if $b > 2$ and $dc > (b - 1)(b - 2)$, then for any x :

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

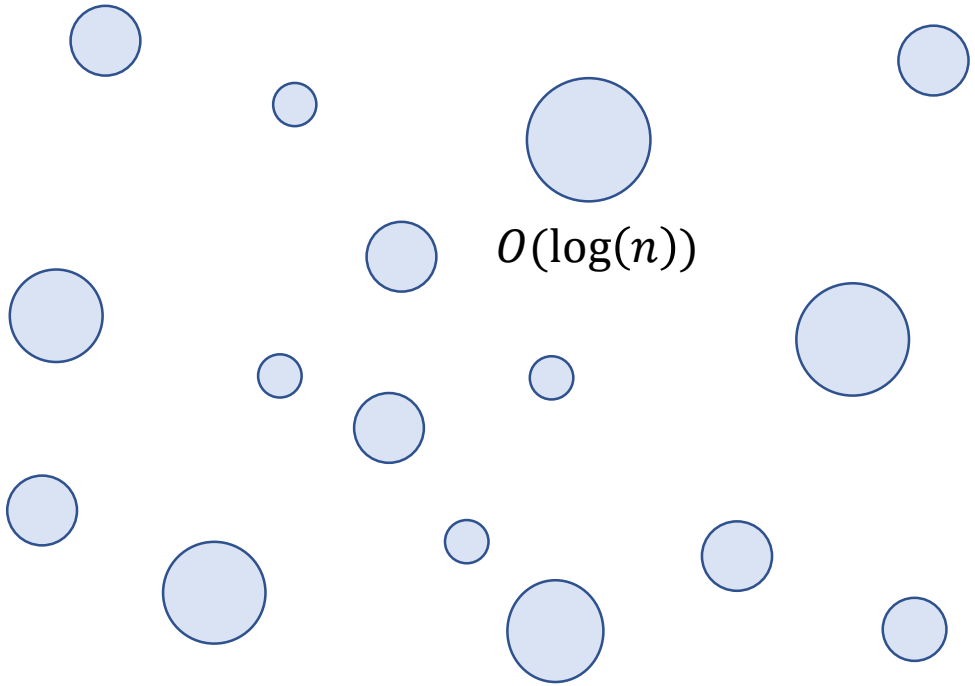
Communication Network Components' Sizes

$b \in (1, 2]$, **or** if $b > 2$ and $dc > (b-1)(b-2)$



Random with superconstant extra seeds hits the giant component

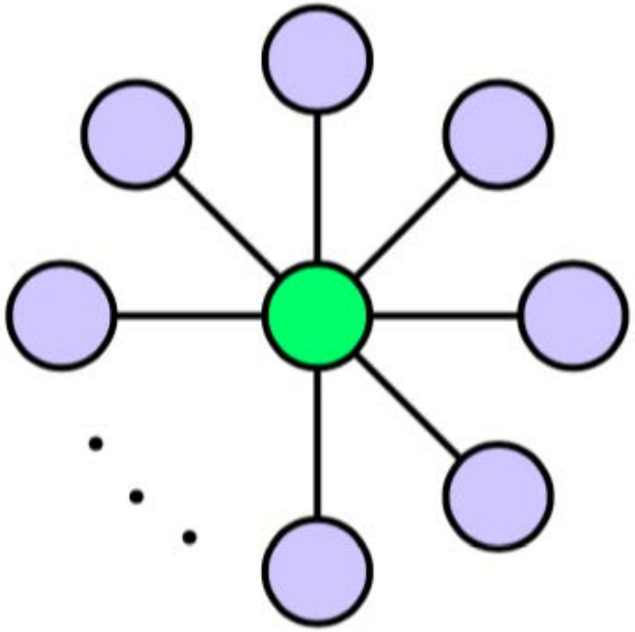
If $b > 2$ **and** $dc < (b-1)(b-2)$



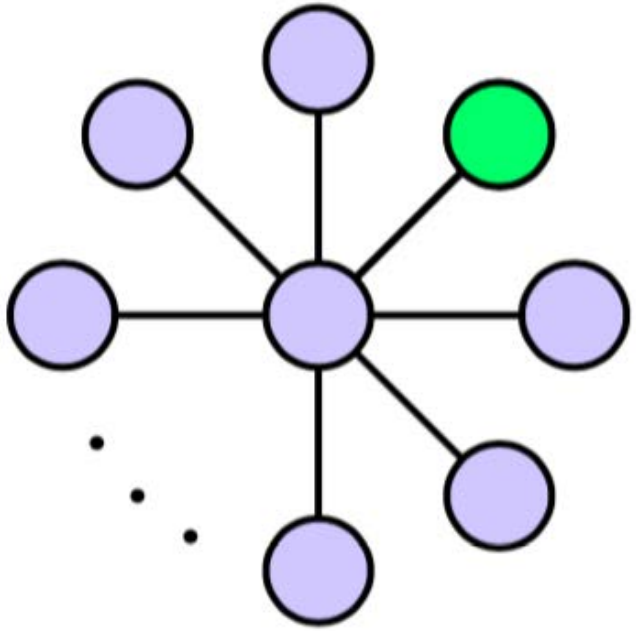
Even omniscient cannot do much

With Equal Seeds, Random Performs Poorly

Communication probability = 0.5
1 central node, $n = 1000$ leaves, 1 seed



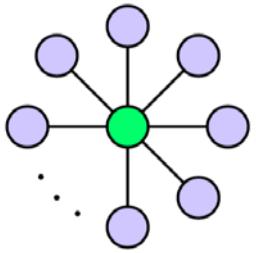
Optimal seeding
Diffusion $\simeq$ 500



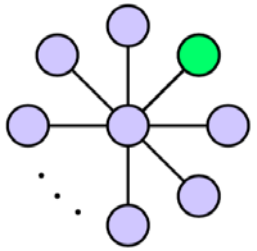
Random seeding
Diffusion $\simeq$ 250

With Additional Seeds, Random Catches up

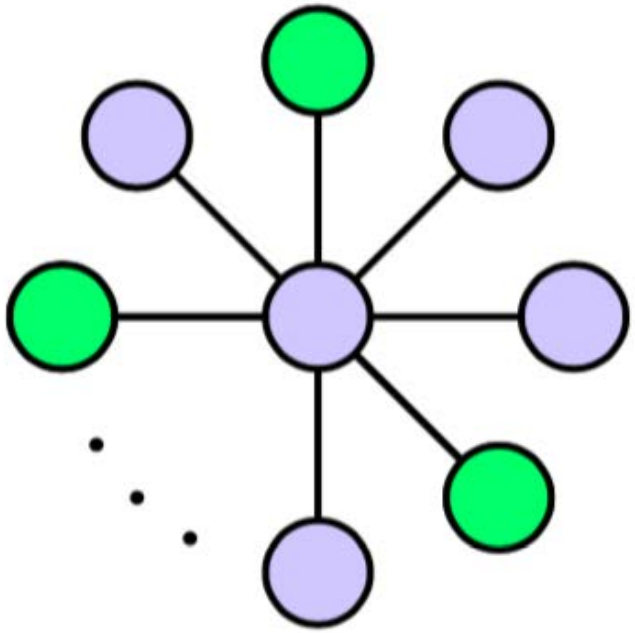
Communication probability = 0.5
1 central node, $n = 1000$ leaves, 1 seed



Optimal seeding
Diffusion $\simeq$ 500



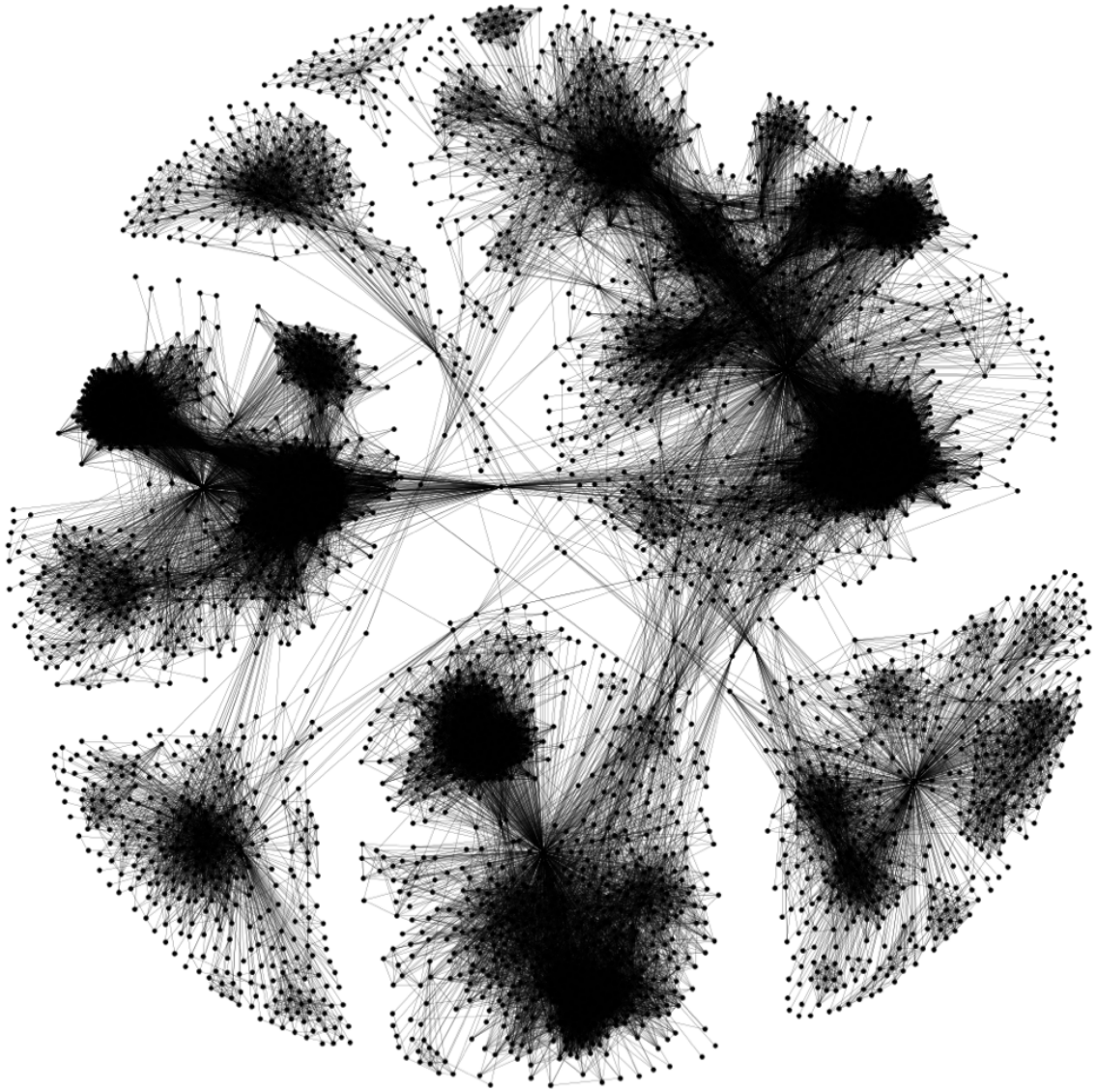
Random seeding
Diffusion $\simeq 250$

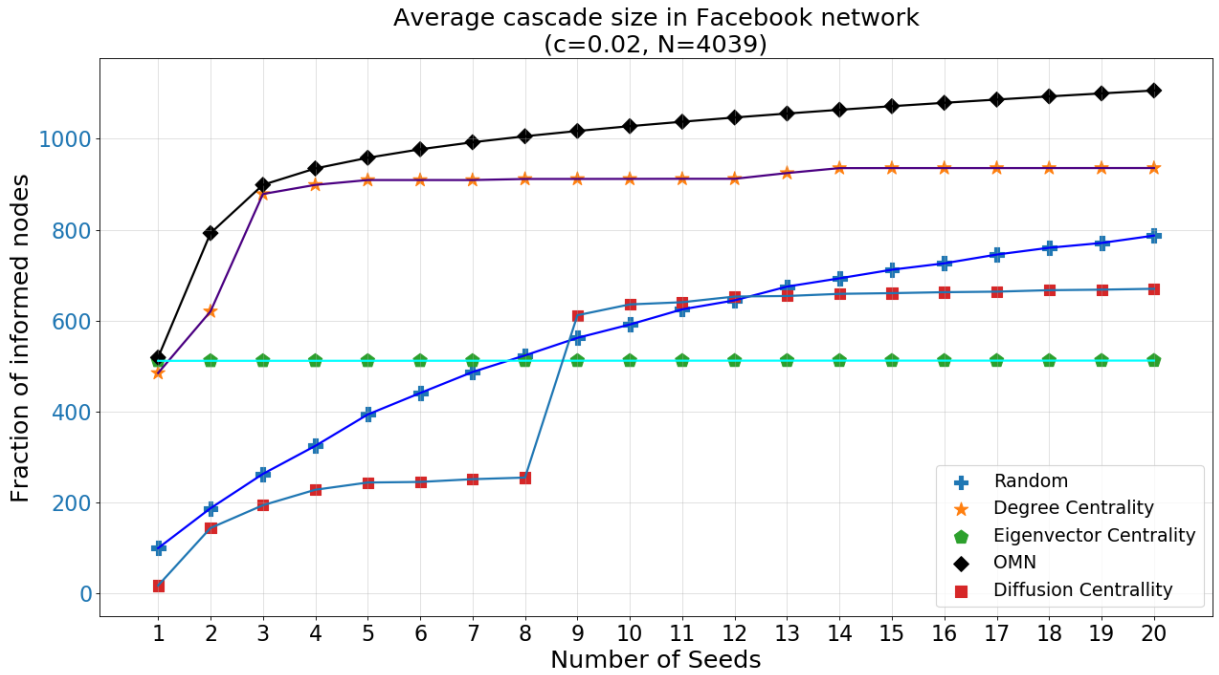


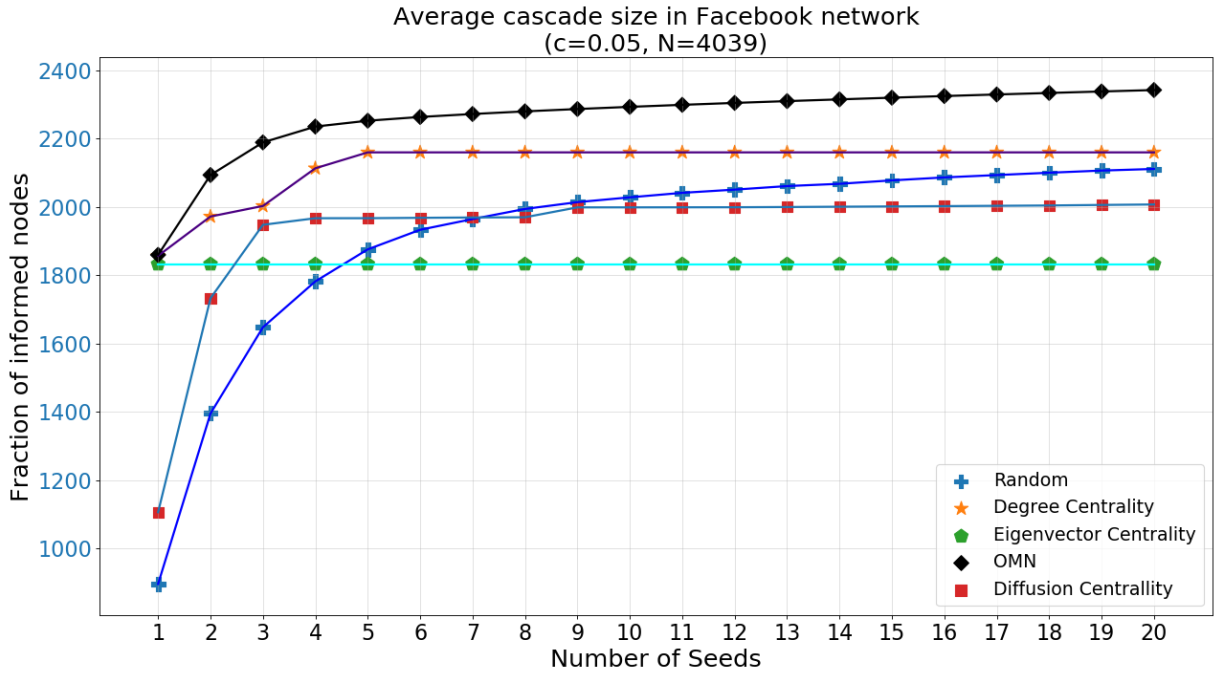
Random seeding with x extra seeds

Diffusion $\simeq \frac{n}{2}\left(1-\left(\frac{1}{2}\right)^{x}\right) \rightarrow \frac{n}{2}$ as x grows very quickly!

Real-world Networks: Facebook Subnetwork




Average cascade size in Facebook network
(c=0.02, N=4039)
Fraction of informed nodes
Number of Seeds
Random
Degree Centrality
Eigenvector Centrality
OMN
Diffusion Centrallity


Average cascade size in Facebook network
(c=0.05, N=4039)
Fraction of informed nodes
Number of Seeds
Random
Degree Centrality
Eigenvector Centrality
OMN
Diffusion Centrallity

Limitations

Diffusion Model
Speed of diffusion
Diffusion minimization

Beyond Simple SIR Diffusion

- Main insight holds for some more complex models:

  - ✓ **Directed communication** (theoretical results)
  - ✓ **Microfinance** [Banerjee *et al,* 2013]: Participants vs. Non-participants
  - ✓ **Weather insurance** [Cai *et al*, 2015]: Linear probability model
  - ✓ **Diffusion games** [Sadler, 2018]: Agents "decide" to adopt or not

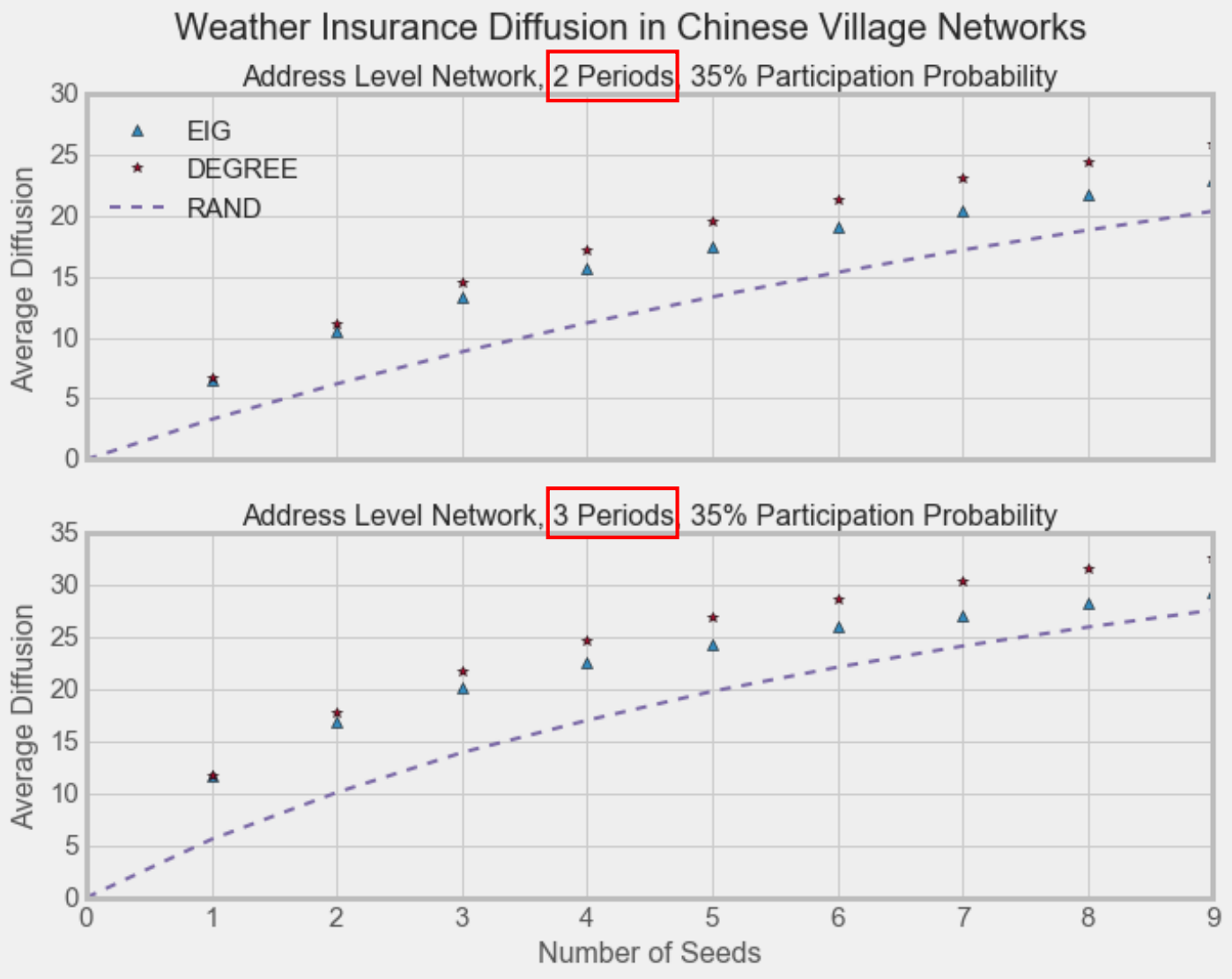
- But not for all:

  - Threshold model of diffusion
  - Limited capacity to listen
  - …

Alternative Diffusion Model: Weather Insurance

- **Cai-Janvry-Sadoulet (2015)**
  - Diffusion of information in weather insurance programs in Chinese villages
- Your chance of adoption increases *linearly* by the # of your friends
- We repeat a similar exercise on their model and network data.

Simulations: Weather insurance



Limitation: “Threshold” Model

- Suppose agents adopt only if a certain fraction of their friends do
- Then it is important to pick agents in “clusters”, so random performs poorly
- *Jackson & Storms (2018)* formalized this

Limitation: Limited Capacity Model

- Suppose agents (regardless of how many links they have) don’t listen to more than a certain number of them.
  - @Taylor_Swift and @Leo_Messi do not listen to all 100M links they have!

- Then results may fail.

Speed of Diffusion

- Clearly, results will not go through if you care about first period diffusion

- But that's not really "viral" diffusion…

Speed of Diffusion

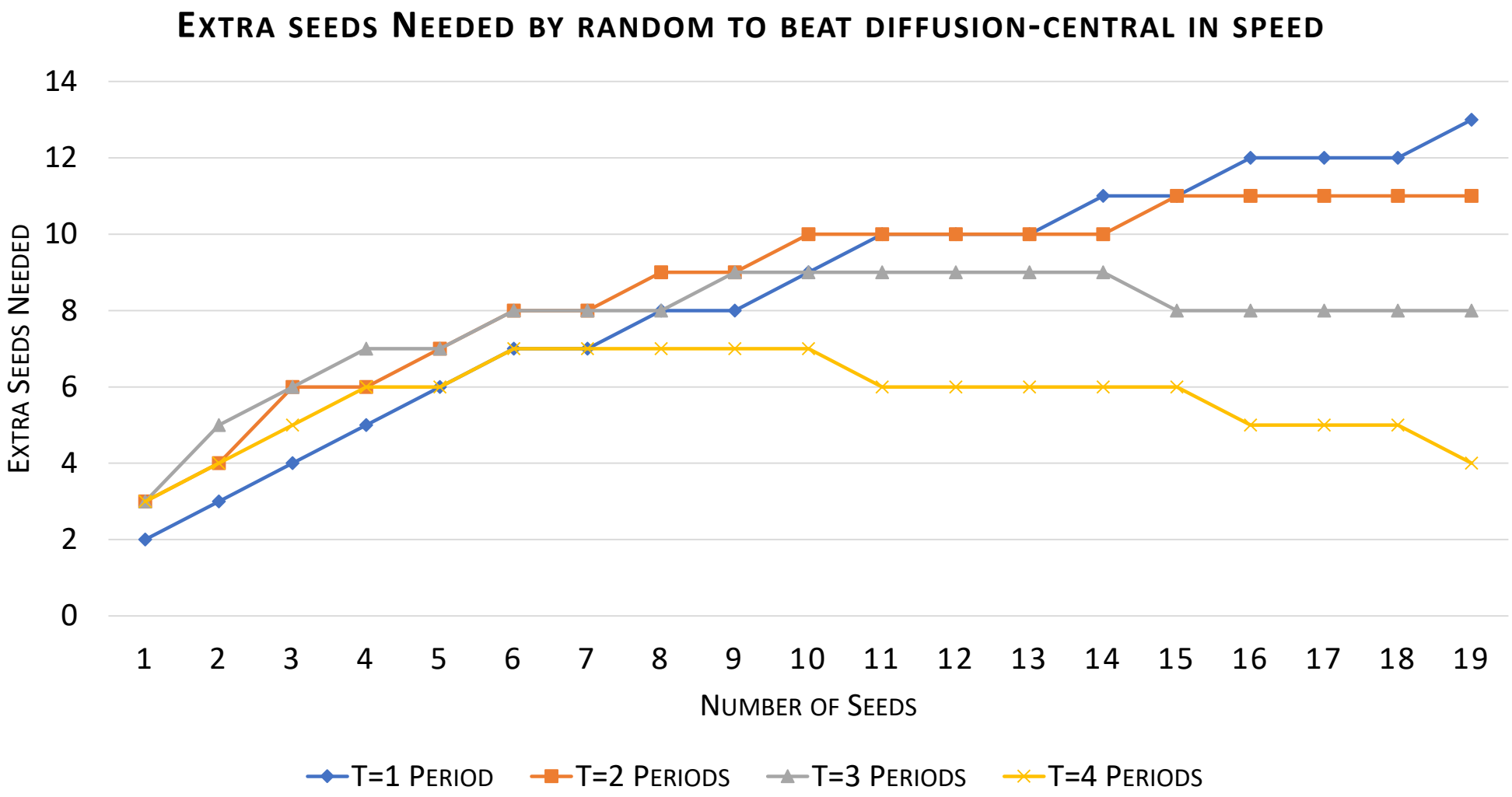
- A *k*-level random network and a bounded diffusion process that ends in **T**$\geq$1

Theorem. Let s be a non-negative integer.

$$\mathbf{H}(\mathrm{RAND}, s + log(n))/\mathbf{H}(\mathrm{OMN}, s) \to 1 \text{ as } n \to \infty.$$

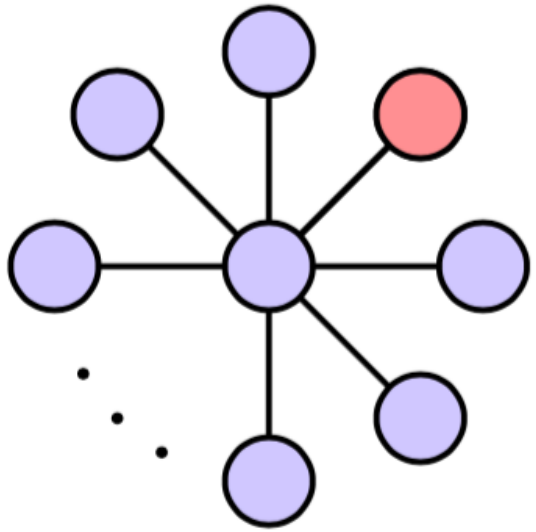
- Similar result will **not** hold true for power law graphs
  - Remember the star example and T=1

Speed of Diffusion in Microfinance Data

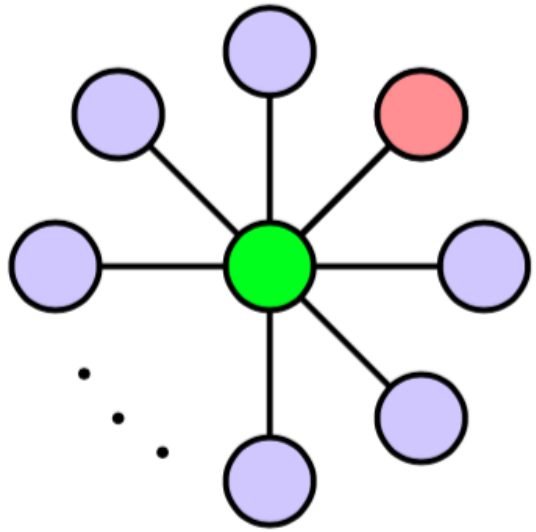


“Vaccination”: Network Can Matter *a lot*

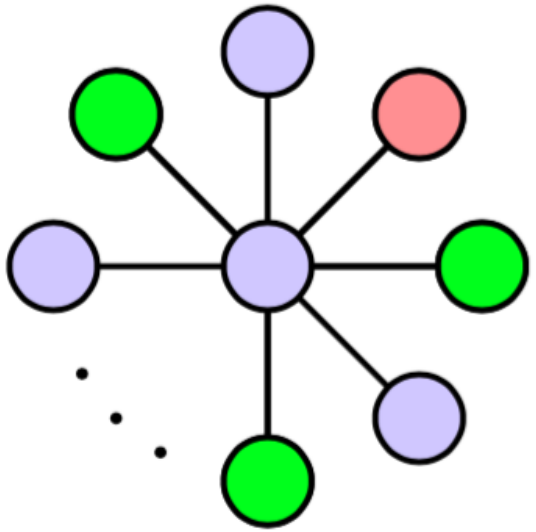
Consider the “minimization” problem: Some agent initially infected, and goal is to curb spread of infection through vaccinations



No vaccination



Optimal vaccination



Random vaccination of several agents

Outline

1. **Model**
   - Diffusion model, Network model

2. **Value of network information**
   - Random vs. Optimum

3. **Proof ideas**

4. **Generalizations**
   1. Network model: Power-law, Clustering, Real-world data
   2. Objective: Speed of diffusion
   3. Alternative diffusion models & Limitations

5. **Concluding remarks**

Networks *Can* Matter

"Our results identify conditions under which network optimization is not very important.

... If an analyst believes (or finds out) that employing a complicated algorithm that accounts for the network structure will yield large gains, then their environment must depart materially from the setting studied here."

A'-Li-Oveisgharan, *JPE* (2020)

Some Questions to Ask Before Seeding

- What is the underlying diffusion process?
- Do you hope the diffusion is going to become viral?
- Do you particularly care about what happens in the first 1-2 periods?
- Do you want to maximize or minimize diffusion?
- And more!

Statistical vs. Economic Significance

In addition to **statistical significance**, also report:

Extra seeds required by the a network-agnostic seeding strategy to get to the $(1-\delta)\%$ *of the network-guided heuristic*

is a useful and easily interpretable information about the **economic significance** of the results.

A Statistic to Report

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

Thank you!

Illustration by
Harriet Lee-Merrion

A “Micro-view” at Diffusion Functions