Diversity and Inequality in

Information Diffusion on Social Networks

Ana-Andreea Stoica (Simons Fellow / MPI-IS Tübingen)

Epidemics and Information Diffusion Workshop

October, 2022

Information diffusion

Information propagated through a social network:

- The news we read
- The technologies we hear about
- Running marketing promotions
- Public health issues

Information diffusion

Information propagated through a social network:

- The news we read
- The technologies we hear about
- Running marketing promotions
- Public health issues

Information diffusion

Information propagated through a social network:

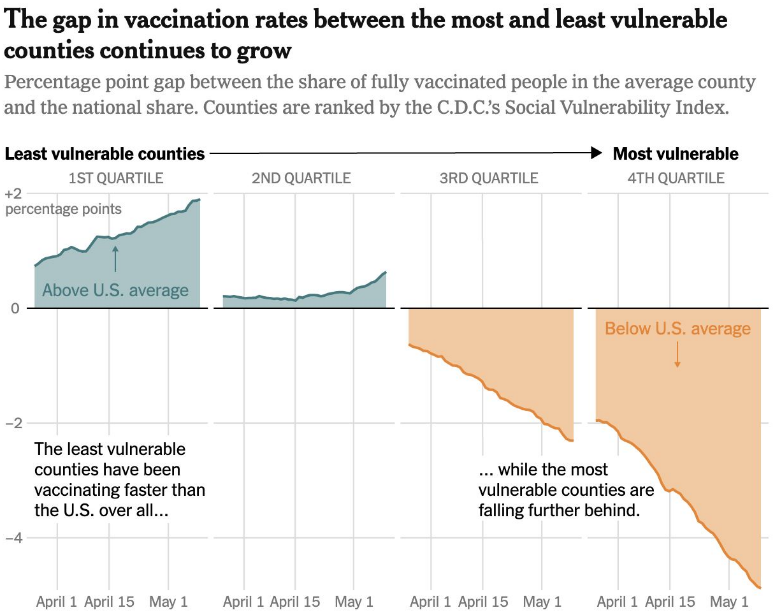
- The news we read
- The technologies we hear about
- Running marketing promotions ← Small targeted campaign
- Public health issues ← Large campaign

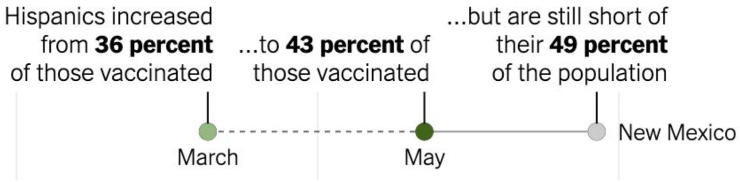
Information diffusion

Information propagated through a social network:

- The news we read
- The technologies we hear about
- Running marketing promotions ← Small targeted campaign
- **Public health issues** ← Large campaign

Social influence & public health





Latino adults in the United States have the lowest rates of Covid-19 vaccination, but among the unvaccinated they are the demographic group most willing to receive the Covid shots as soon as possible, a new survey shows.

The findings suggest that their depressed vaccination rate reflects in large measure misinformation about cost and access, as well as concerns about employment and immigration issues, according to the latest edition of the Kaiser Family Foundation Covid-19 Vaccine Monitor.

https://www.nytimes.com/2021/05/12/us/covid-vaccines-vulnerable.html
https://www.nytimes.com/interactive/2021/05/14/us/vaccine-race-gap.html
https://www.nytimes.com/2021/05/13/health/covid-vaccine-latino-hispanic.html

Social influence & opportunities

The Diffusion of Microfinance

Abhijit Banerjee,* Arun G. Chandrasekhar,* Esther Duflo,* Matthew O. Jackson*

Introduction: How do the network positions of the first individuals in a society to receive information about a new product affect its eventual diffusion? To answer this question, we develop a model of information diffusion through a social network that discriminates between information passing (individuals must be aware of the product before they can adopt it, and they can learn from their friends) and endorsement (the decisions of informed individuals to adopt the product might be influenced by their friends' decisions). We apply it to the diffusion of microfinance loans, in a setting where the set of potentially first-informed individuals is known. We then propose two new measures of how "central" individuals are in their social network with regard to spreading information; the centrality of the first-informed individuals in a village helps significantly in predicting eventual adoption.

Access to information is access to opportunity/healthcare

Information diffusion
(Social influence maximization problem)

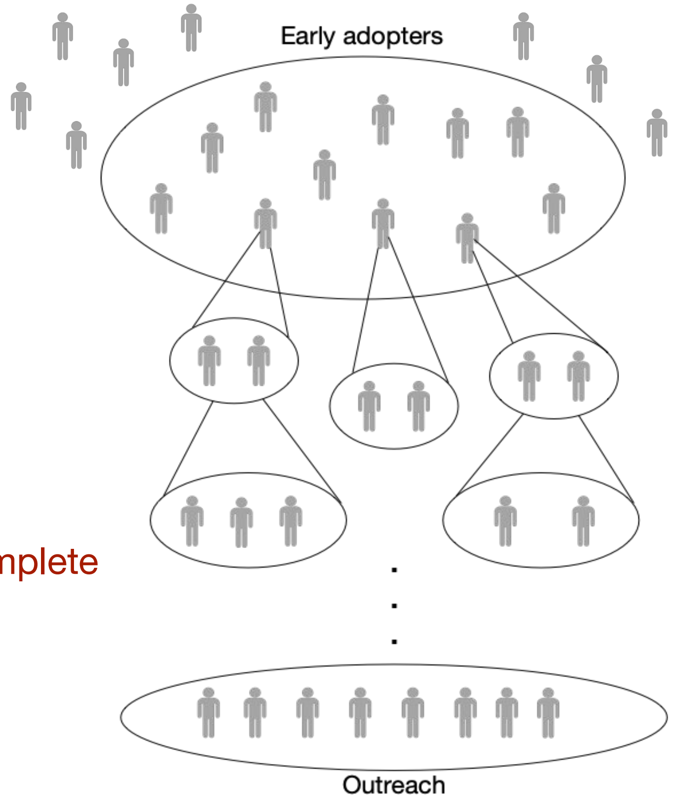
- Given a network G, with diffusion model as independent cascade with probability p, pick the best k early-adopters ('seeds') that maximize outreach:[1]

$$S^* = \operatorname{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|, \quad \text{s.t. } |S| \leq k$$

NP-complete

- Algorithms that choose based on:
  - Centrality: degree, distance centrality, …
  - Iteratively: greedy

Agnostic to communities



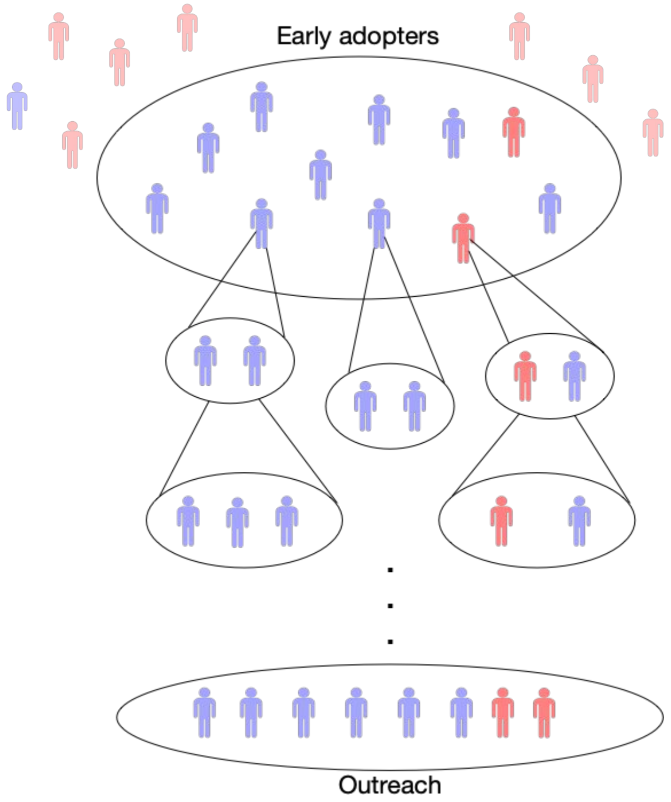
[1] Kempe, David, Jon Kleinberg, and Éva Tardos. "*Maximizing the spread of influence through a social network.*" In Proceedings of the ninth ACM SIGKDD Conference, pp. 137-146. 2003.

Information diffusion
(Social influence maximization problem)

- Given a network G, with diffusion model as independent cascade with probability p, pick the best k early-adopters ('seeds') that maximize outreach:

$$S^* = \text{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S, p)|, \quad \text{s.t. } |S| \leq k$$

- Algorithms that choose based on:
  - Centrality: degree, distance centrality, …
  - Iteratively: greedy



⇒ Bias in centrality measures and social structure gets reproduced[2]

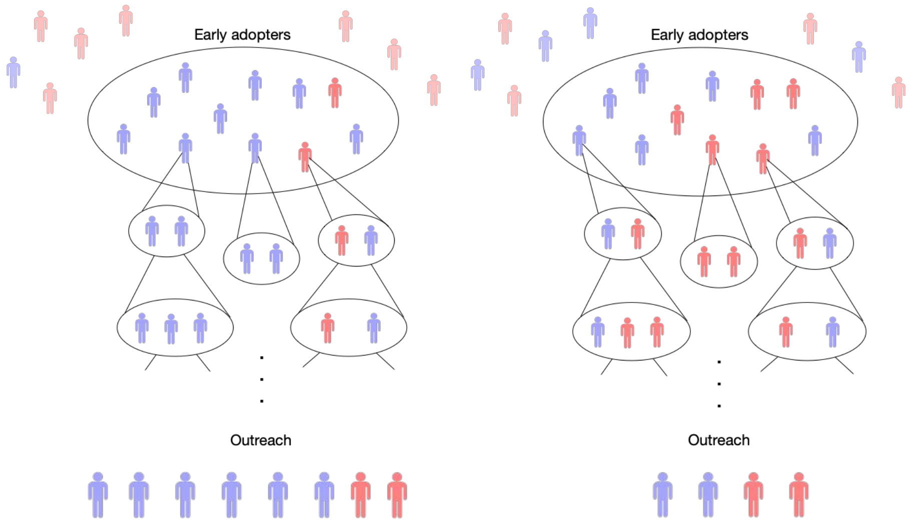
[2] Fish, Benjamin, et al. "Gaps in information access in social networks". *The World Wide Web Conference.* ACM, 2019.

Information diffusion

- Parity constraint in an optimization function based on greedy algorithms:

$$S^* = \text{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|,$$
$$\text{s.t. } |S| \leq k \text{ and } \frac{\mathbb{E}(|\phi_G(S,p) \cap R|)}{\mathbb{E}(|\phi_G(S,p) \cap B|)} \simeq \frac{|R|}{|B|}$$

➡ Fairness-efficiency trade-off



no constraint *parity constraint*

Our approach:

- Partially known networks ⇒ centrality measures (# of connections etc)
- Model of network growth & tap into inactive communities
- Theoretical conditions for when **equity increases efficiency (outreach)**

[2] Fish, Benjamin, et al. "Gaps in information access in social networks". *The World Wide Web Conference.* ACM, 2019.

Information diffusion

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

Mohammad Akbarpour†
Suraj Malladi‡
Amin Saberi§

Random seeding with extra x nodes is comparable to optimal seeding (for small x)

Our approach:

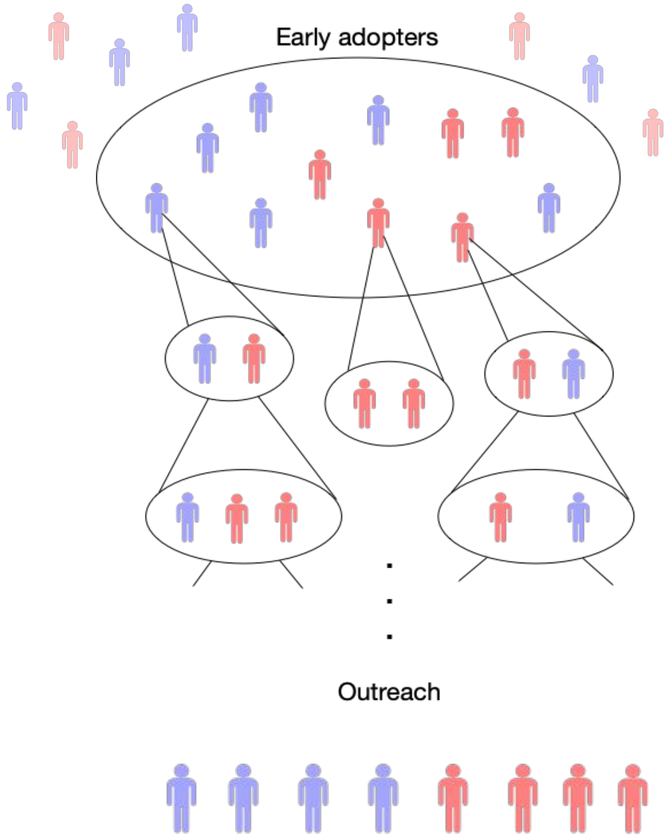
- Partially known networks ⇒ centrality measures (# of connections etc)
- Model of network growth & tap into inactive communities
- Theoretical conditions for when **equity increases efficiency (outreach)**

Information diffusion

- **Our vision:** bias as a sign of inefficiency
  - Diversity: tap into inactivated communities in the *early adopters* set

$$S^* = \text{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|,$$
$$\text{s.t. } |S| \leq k \text{ and } \frac{\mathbb{E}(|S \cap R|)}{\mathbb{E}(|S \cap B|)} \simeq \frac{|R|}{|B|}$$

- Seeding can be done with awareness of labels: **statistical parity** in your campaign (even if choosing less connected people)
  - Parity seeding (strict)
  - Diversity seeding (relaxed)

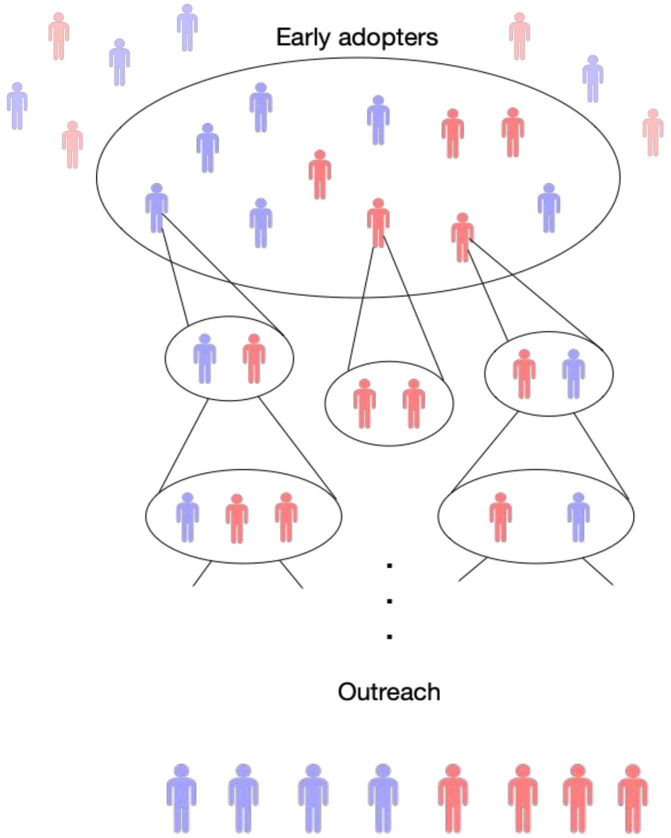


Information diffusion

- **Our vision:** bias as a sign of inefficiency
  - Diversity: tap into inactivated communities in the *early adopters* set

$$S^* = \mathrm{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|,$$
$$\text{s.t. } |S| \leq k \text{ and } \frac{\mathbb{E}(|S \cap R|)}{\mathbb{E}(|S \cap B|)} \simeq \frac{|R|}{|B|}$$

- Seeding can be done with awareness of labels: **statistical parity** in your campaign (even if choosing less connected people)
  - Parity seeding (strict)
  - Diversity seeding (relaxed)

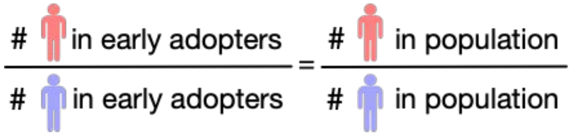


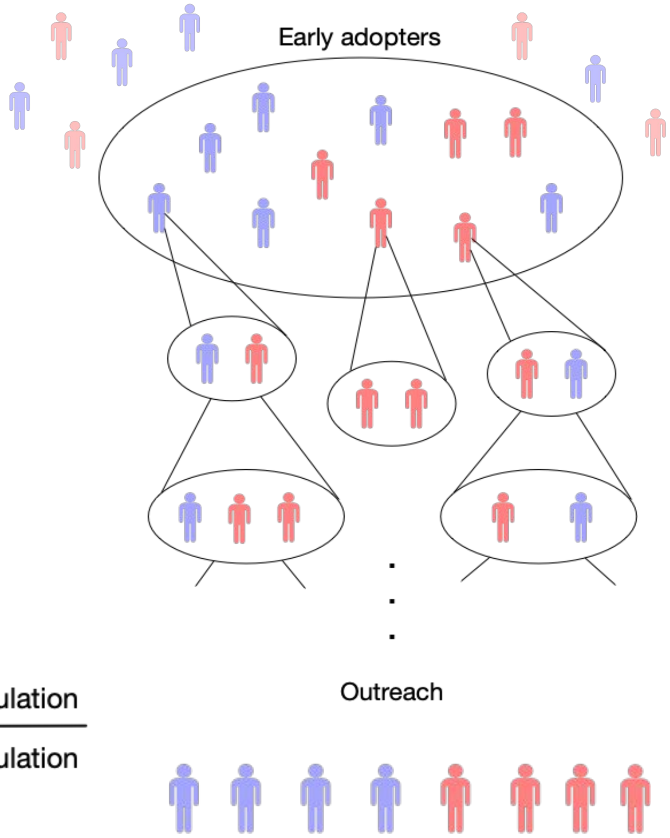
Information diffusion

- **Our vision:** bias as a sign of inefficiency
  - Diversity: tap into inactivated communities in the *early adopters* set

$$S^* = \text{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|,$$
$$\text{s.t. } |S| \le k \text{ and } \frac{\mathbb{E}(|S \cap R|)}{\mathbb{E}(|S \cap B|)} \simeq \frac{|R|}{|B|}$$

- Seeding can be done with awareness of labels: **statistical parity** in your campaign (even if choosing less connected people)
  - **Parity seeding (strict)**
  - Diversity seeding (relaxed)





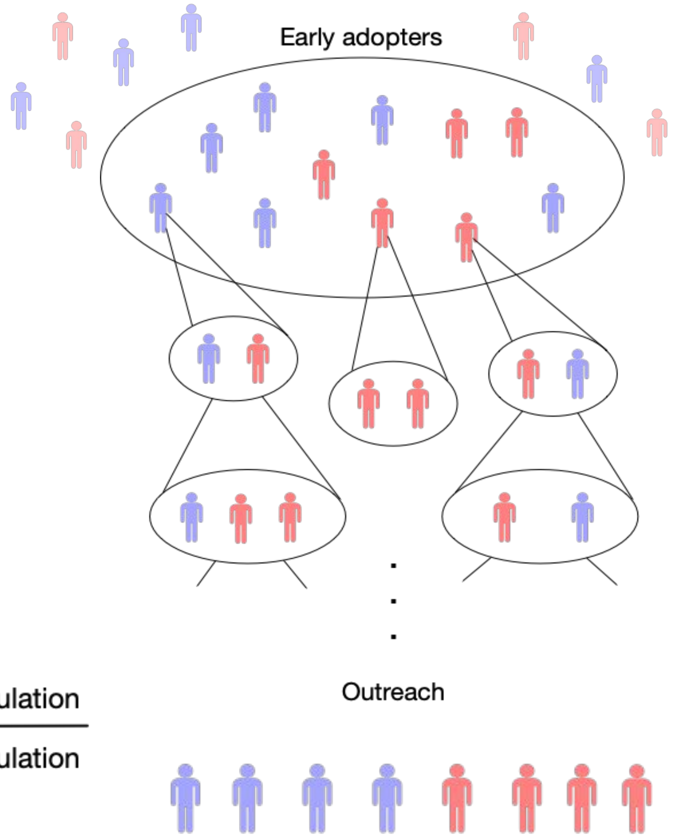
Information diffusion

- **Our vision:** bias as a sign of inefficiency
  - Diversity: tap into inactivated communities in the *early adopters* set

$$S^* = \text{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|),$$
$$\text{s.t. } |S| \leq k \text{ and } \frac{\mathbb{E}(|S \cap R|)}{\mathbb{E}(|S \cap B|)} \simeq \frac{|R|}{|B|}$$

- Seeding can be done with awareness of labels: **statistical parity** in your campaign (even if choosing less connected people)
  - Parity seeding (strict)
  - **Diversity seeding (relaxed)**

$$\frac{\#\ \text{in early adopters}}{\#\ \text{in early adopters}} \cong \frac{\#\ \text{in population}}{\#\ \text{in population}}$$

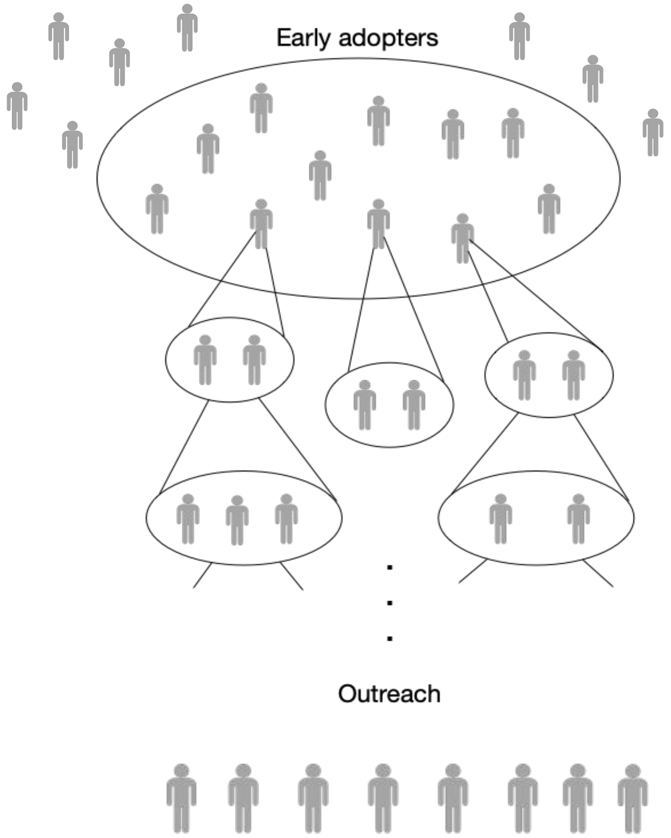


Information diffusion

- **Our vision:** bias as a sign of inefficiency
  - Diversity: tap into inactivated communities in the *early adopters* set

$$S^* = \text{argmax}_{S \subseteq V(G)} \mathbb{E}(|\phi_G(S,p)|,$$
$$\text{s.t. } |S| \leq k \text{ and } \frac{\mathbb{E}(|S \cap R|)}{\mathbb{E}(|S \cap B|)} \simeq \frac{|R|}{|B|}$$

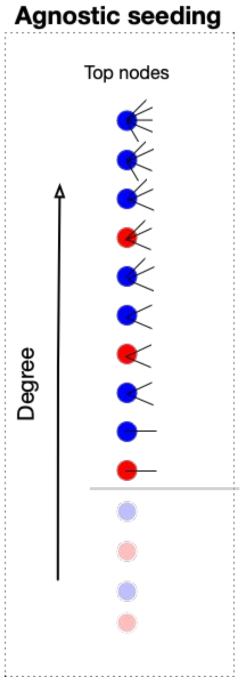
- Seeding can be done with awareness of labels: **statistical parity** in your campaign (even if choosing less connected people)
  - Parity seeding (strict)
  - Diversity seeding (relaxed)
- Baseline: Seeding can be done **agnostically**: ignore labels, already takes into account network structure



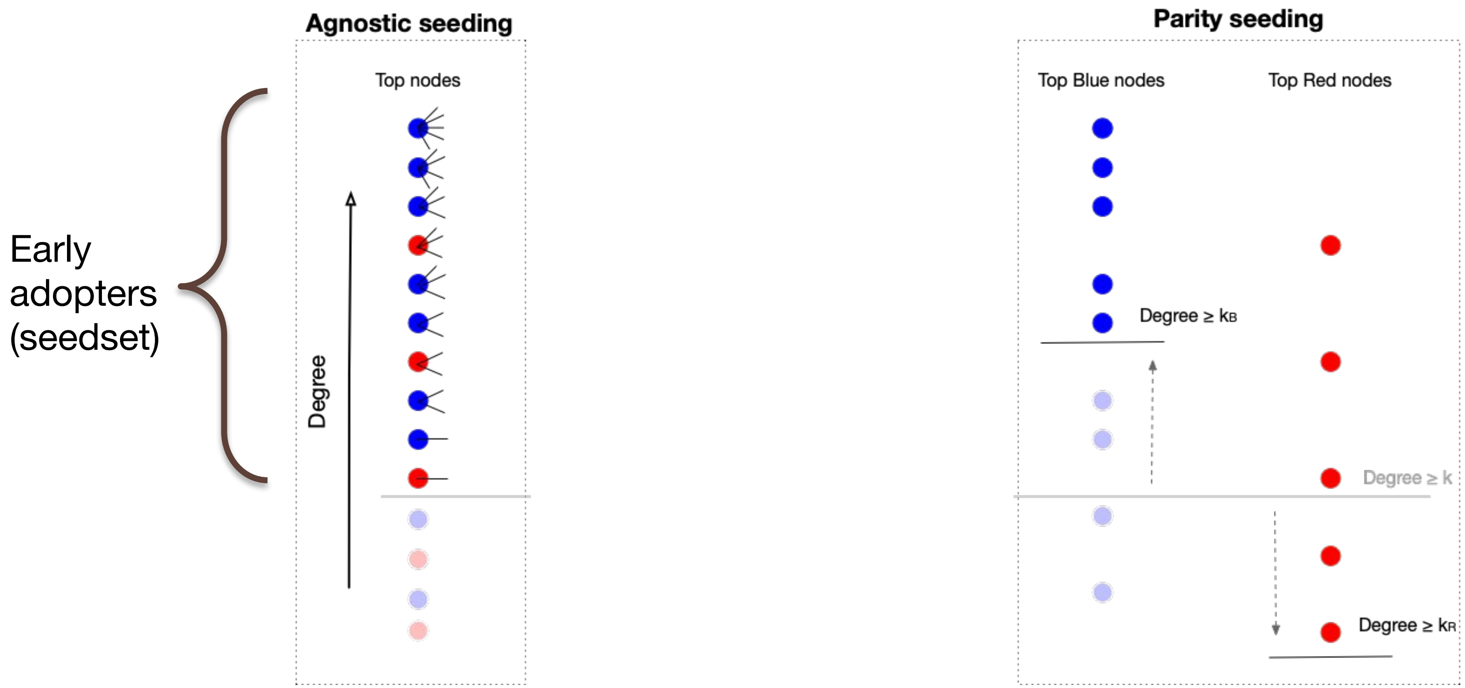
[3] Stoica, Ana-Andreea, Jessy Xinyi Han, and Augustin Chaintreau. "Seeding Network Influence in Biased Networks and the Benefits of Diversity." WWW. 2020.

Color-agnostic v. Diversity Seeding

Early
adopters
(seedset)



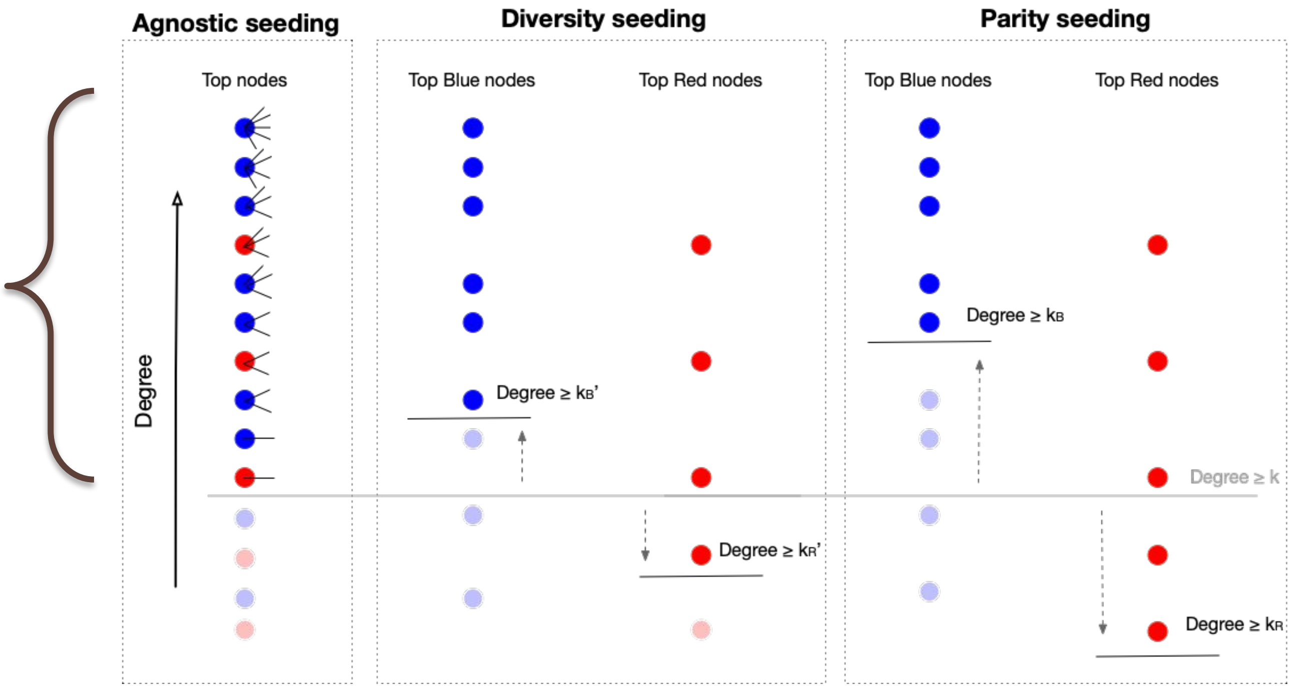
Color-agnostic v. Diversity Seeding



Keeping the same budget!

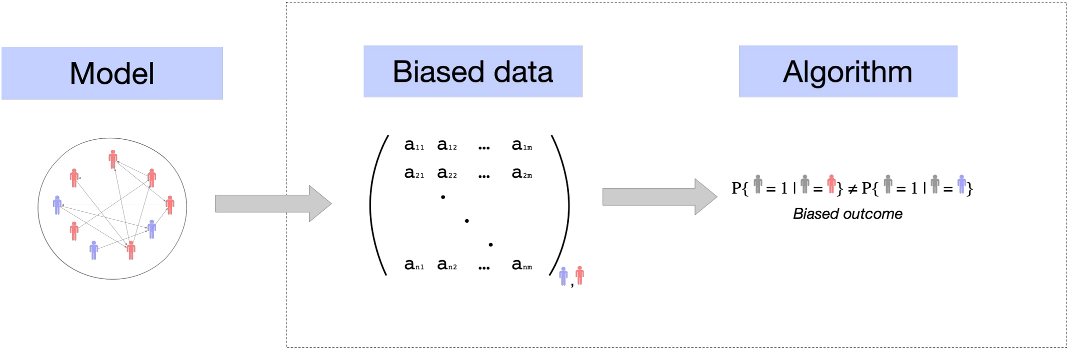
Color-agnostic v. Diversity Seeding

Early adopters (seedset)



Keeping the same budget!

Networks modeling for building more diverse and efficient heuristics



Models of network evolution:

- Explain where inequality or bias originates and how it propagates in an algorithm
- Useful to prove guarantees about interventions to mitigate bias

Where and **how** do we intervene to improve the gain of a minority group?

Biased preferential attachment model (BPAM)

Minority-majority: blue label and red label

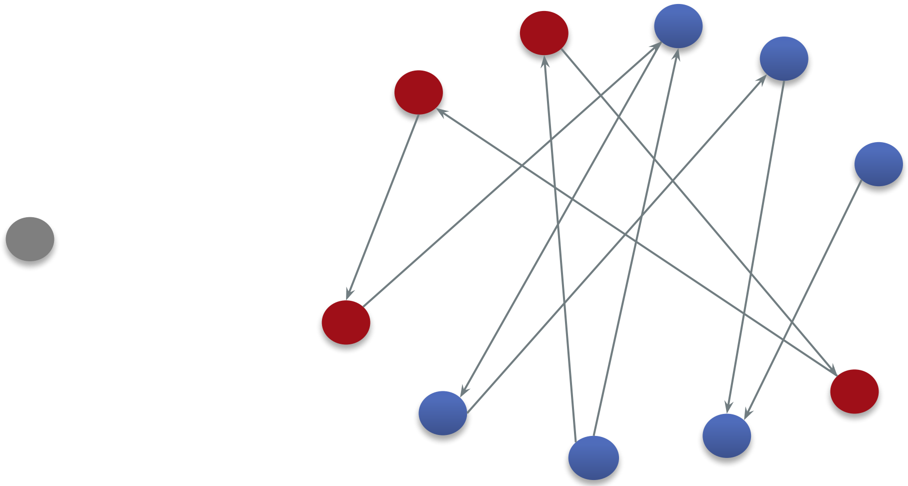
- Fraction of red nodes = r < ½

Preferential attachment (rich-get-richer): nodes connect w.p. proportional to degree

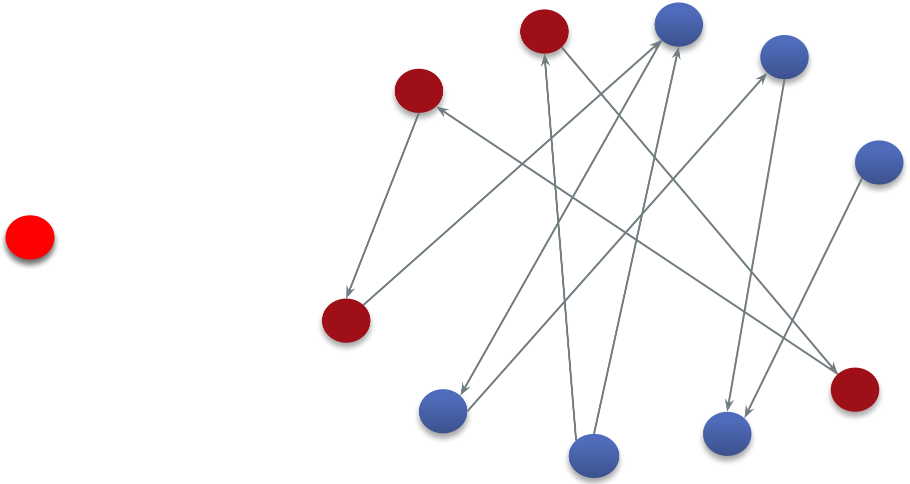
Homophily: if different labels, connection is accepted w.p. ρ

[4] Avin, Chen et al. "Homophily and the glass ceiling effect in social networks." ITCS. 2015

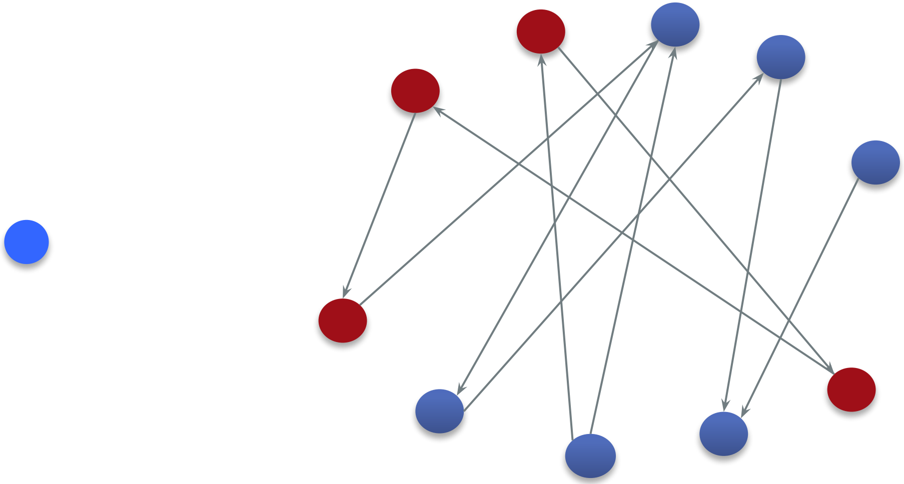
Preferential attachment with homophily



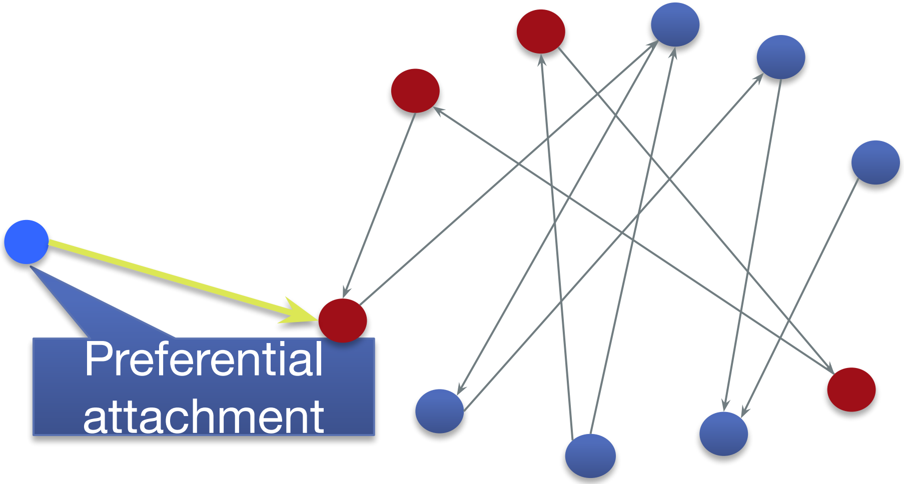
Preferential attachment with homophily



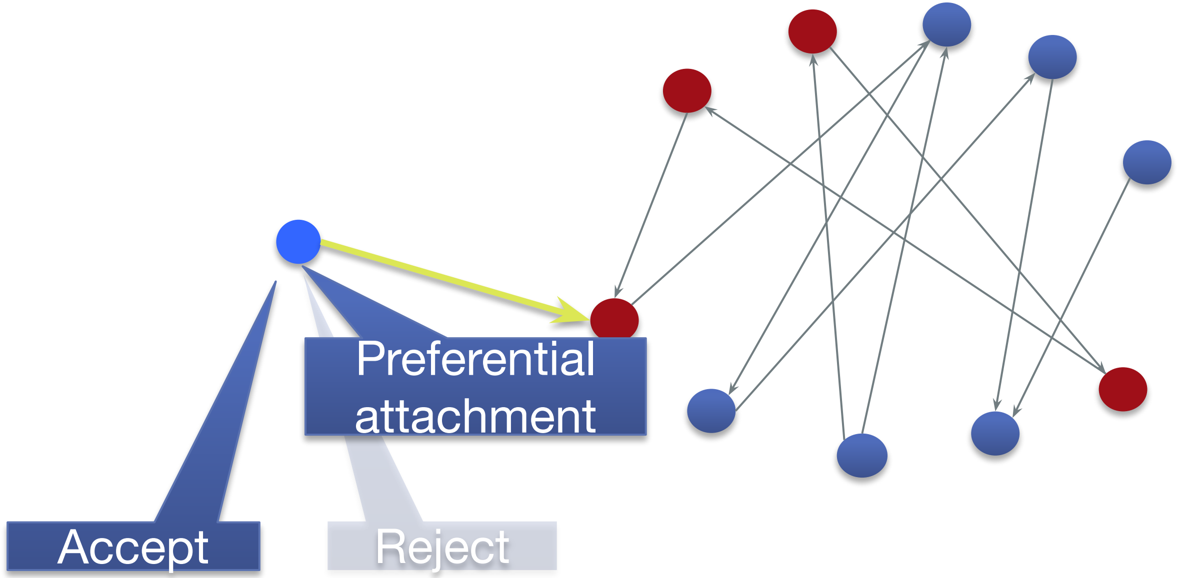
Preferential attachment with homophily



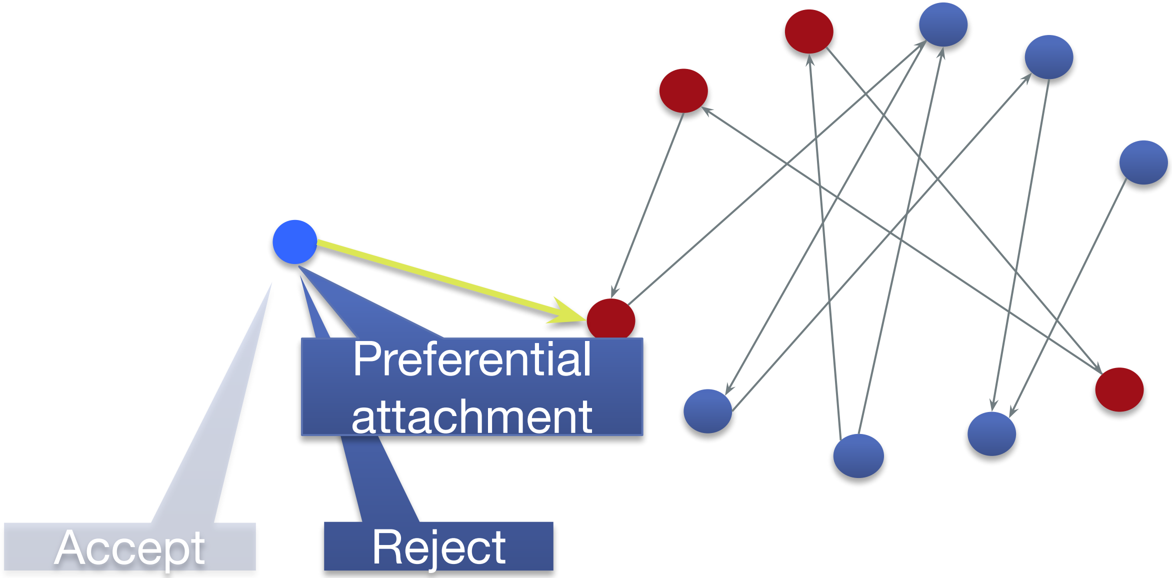
Preferential attachment with homophily



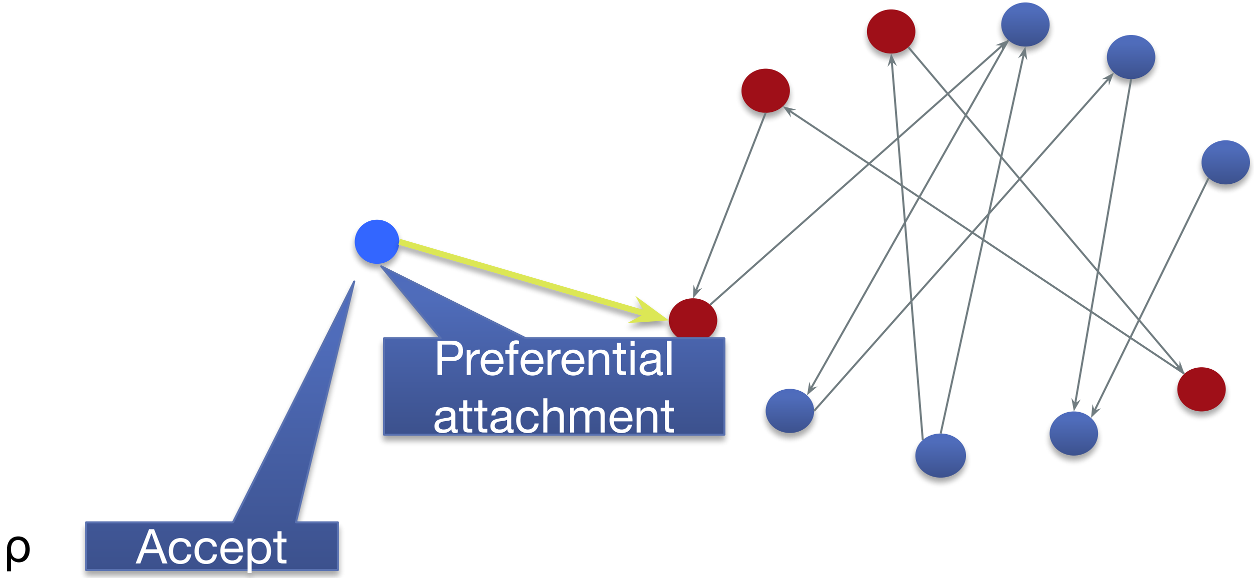
Preferential attachment with homophily



Preferential attachment with homophily



Preferential attachment with homophily



Biased preferential attachment model (BPAM)

Minority-majority: blue label and red label

- Fraction of red nodes = r < ½

Preferential attachment (rich-get-richer): nodes connect w.p. proportional to degree

Homophily: if different labels, connection is accepted w.p. ρ

⇒ known to exhibit inequality in the degree distribution of the two communities[4]

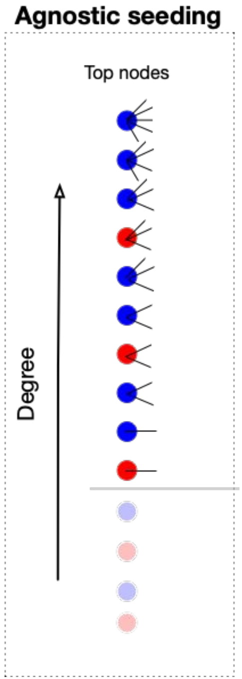
$$top_k(R) \sim k^{-\beta(R)}$$
$$top_k(B) \sim k^{-\beta(B)}$$

Thm [Avin et al]: $\beta(R) > 3 > \beta(B)$

[4]Avin, Chen et al. "Homophily and the glass ceiling effect in social networks." ITCS. 2015

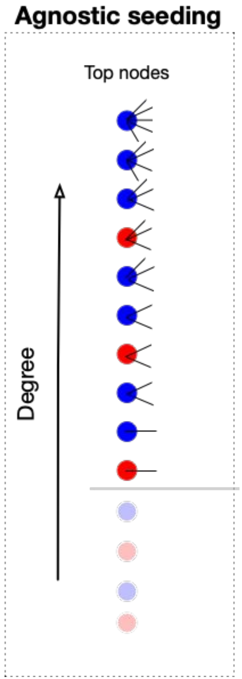
Color-agnostic v. Diversity Seeding

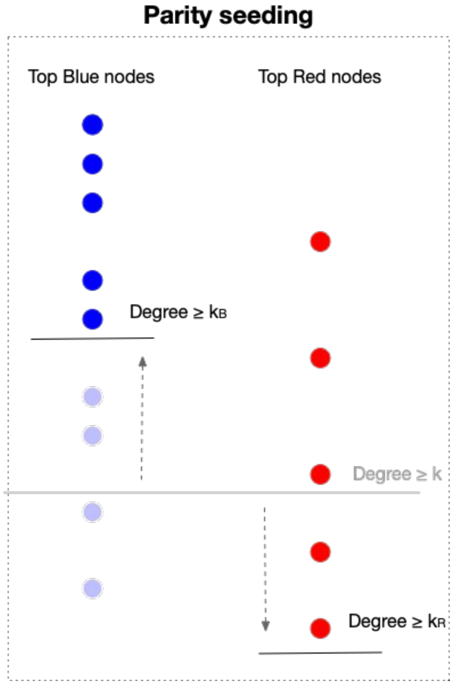
Early adopters (seedset)



Color-agnostic v. Diversity Seeding

Early adopters (seedset)

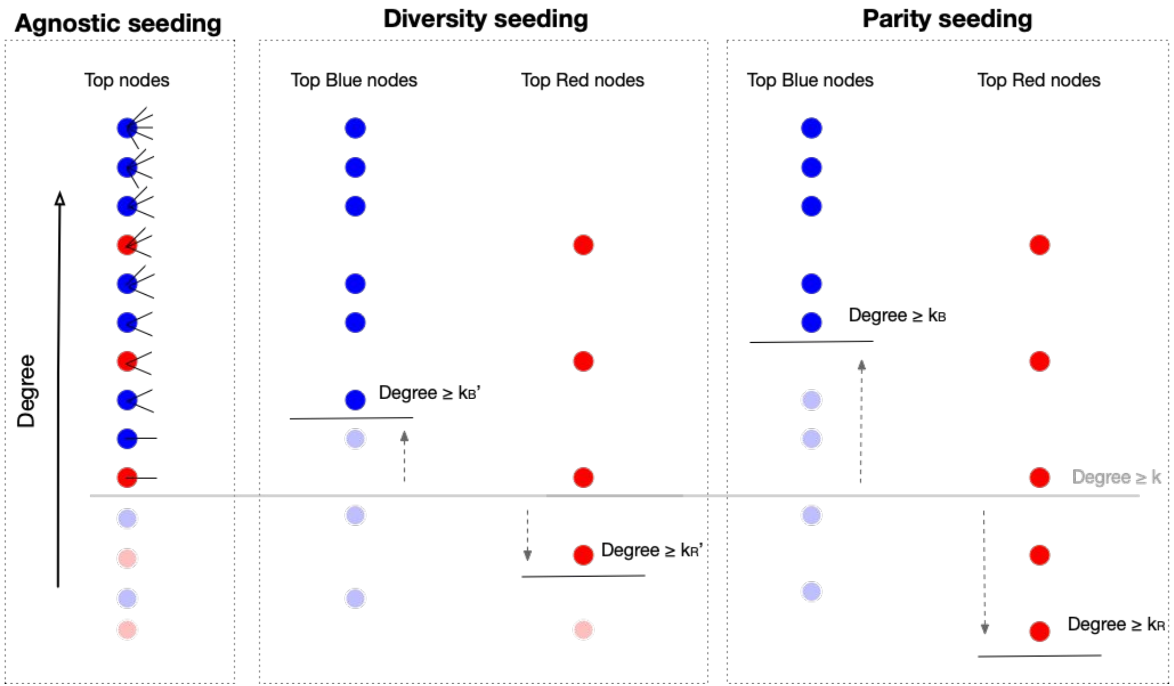




Keeping the same budget!

Color-agnostic v. Diversity Seeding

Early adopters (seedset)

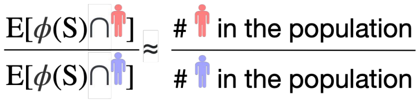


Keeping the same budget!

Theoretical analysis of diversity interventions

***Theorem**: for the graph sequences G(n) generated from the BPAM:*

1. *Diversity seeding and parity seeding leads to fairer outreach for the same budget*



2. *∃ k* (closed form) such that when k > k* diversity seeding and parity seeding can outperform agnostic seeding in outreach*

$$\mathbb{E}(\phi(S_{\text{diversity}})) > \mathbb{E}(\phi(S_{\text{parity}})) > \mathbb{E}(\phi(S_{\text{agnostic}})),$$
$$\text{given } |S_{\text{diversity}}| = |S_{\text{diversity}}| = |S_{\text{diversity}}|$$

Proof sketch

Our goal is to find two thresholds $k^R(n)$and $k^B(n)$that give in expectation the same amount of seeds as a general ("agnostic") threshold k(n) but better influence:

$$\mathbb{E}(\phi(S_{k(n)})) < \mathbb{E}(\phi(S_{k^R(n)} \cup S_{k^B(n)})),$$
$$\text{s.t. } \mathbb{E}(|S_{k(n)}|) = (|S_{k^R(n)} \cup S_{k^B(n)}|)$$

First step: estimate first-step influence size of $S_{k(n)} = \{v \in V | deg(v) \geq k(n)\}$

Second step: extend to an estimation of $\mathbb{E}(\phi(S_{k(n)}))$

Proof sketch

Our goal is to find two thresholds $k^R(n)$and $k^B(n)$that give in expectation the same amount of seeds as a general ("agnostic") threshold k(n) but better influence:

$$\mathbb{E}(\phi(S_{k(n)})) < \mathbb{E}(\phi(S_{k^R(n)} \cup S_{k^B(n)})),$$
$$\text{s.t. } \mathbb{E}(|S_{k(n)}|) = (|S_{k^R(n)} \cup S_{k^B(n)}|)$$

First step: estimate first-step influence size of $S_{k(n)} = \{v \in V | deg(v) \geq k(n)\}$

Second step: extend to an estimation of $\mathbb{E}(\phi(S_{k(n)}))$

Proof sketch

Our goal is to find two thresholds $k^R(n)$and $k^B(n)$that give in expectation the same amount of seeds as a general ("agnostic") threshold k(n) but better influence:

$$\mathbb{E}(\phi(S_{k(n)})) < \mathbb{E}(\phi(S_{k^R(n)} \cup S_{k^B(n)})),$$
$$\text{s.t. } \mathbb{E}(|S_{k(n)}|) = (|S_{k^R(n)} \cup S_{k^B(n)}|)$$

First step: estimate first-step influence size of $S_{k(n)} = \{v \in V | deg(v) \geq k(n)\}$

- We know $|S_{k(n)}|$ because the degree distribution follows a power law with coefficients $\beta(R), \beta(B)$
- Can compute first order influence for any threshold by computing $\mathbb{P}(v \text{ influenced by one edge} | v \in B)$ and $\mathbb{P}(v \text{ influenced by one edge} | v \in R)$

Proof sketch

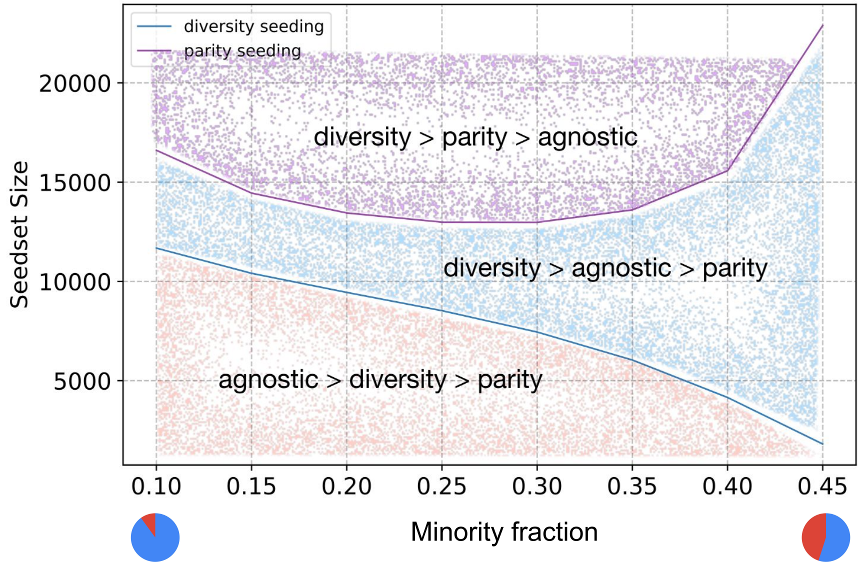
Our goal is to find two thresholds $k^R(n)$ and $k^B(n)$ that give in expectation the same amount of seeds as a general ("agnostic") threshold k(n) but better influence:

$$\mathbb{E}(\phi(S_{k(n)})) < \mathbb{E}(\phi(S_{k^R(n)} \cup S_{k^B(n)})),$$
$$\text{s.t. } \mathbb{E}(|S_{k(n)}|) = (|S_{k^R(n)} \cup S_{k^B(n)}|)$$

Set $k^B(n) = k(n) \cdot x$, compute $k^R(n)$ based on the budget constraint, and solve

$$F(x) = \mathbb{E}(\phi(S_{k^B(n)} \cup S_{k^R(n)})) - \mathbb{E}(\phi(S_{k(n)}))$$

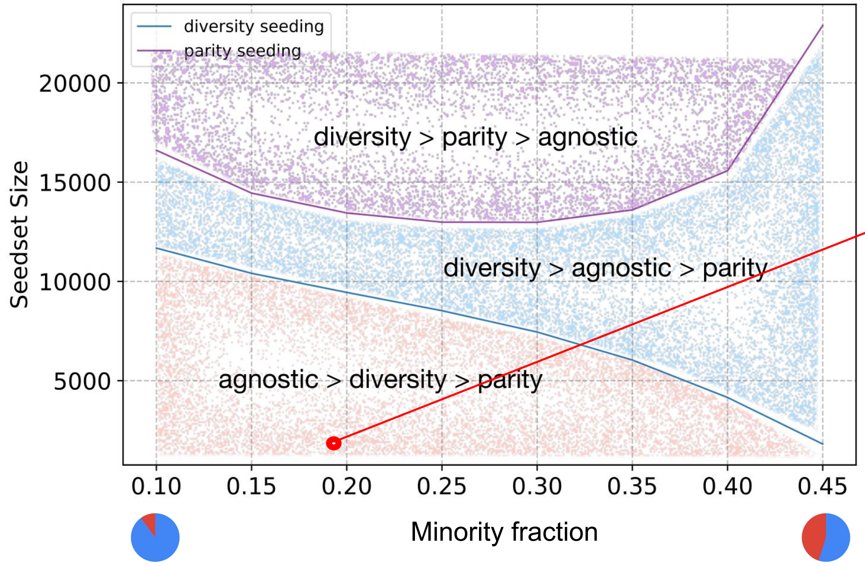
Theoretical analysis of diversity interventions



Network of ~53,000 nodes, 2 communities, homophily $\rho = 0.135$

- Compute regions where each heuristic performs better than the agnostic one
- As communities become more equal, need fewer seeds for diversity heuristic to be more efficient
- Not the same thing happens with the parity heuristic!

Theoretical analysis of diversity interventions

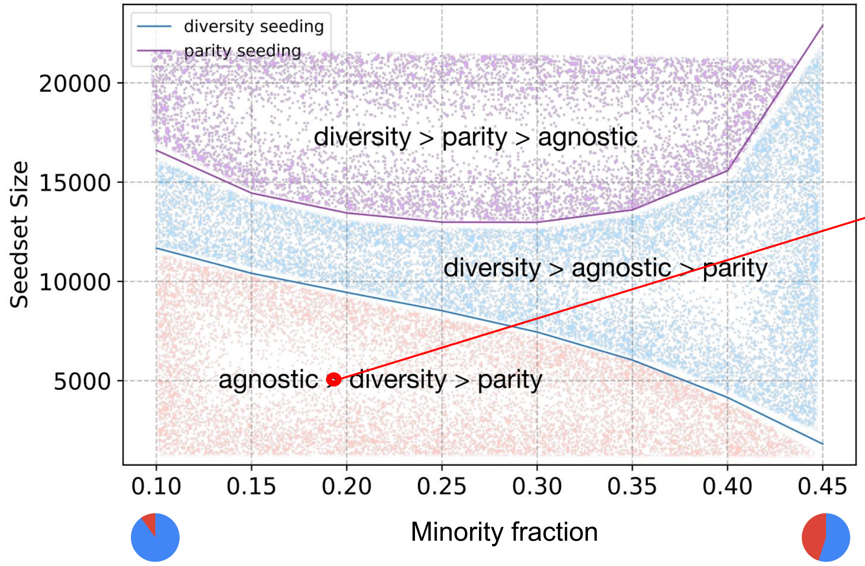


Network of ~53,000 nodes, 2 communities, homophily ρ = 0.135

DBLP citation dataset: men and women

	1,000 seeds		
$p = 0.01$	Agnostic seeding	Parity seeding	Diversity seeding
Total outreach	**1,149.15**	↓1,147.874	↓1,149.1
F outreach	191.95	↑**210.456**	↑196.6
M outreach	**957.2**	↓937.418	↓952.5
F % in outreach	0.167	↑**0.183**	↑0.171

Theoretical analysis of diversity interventions

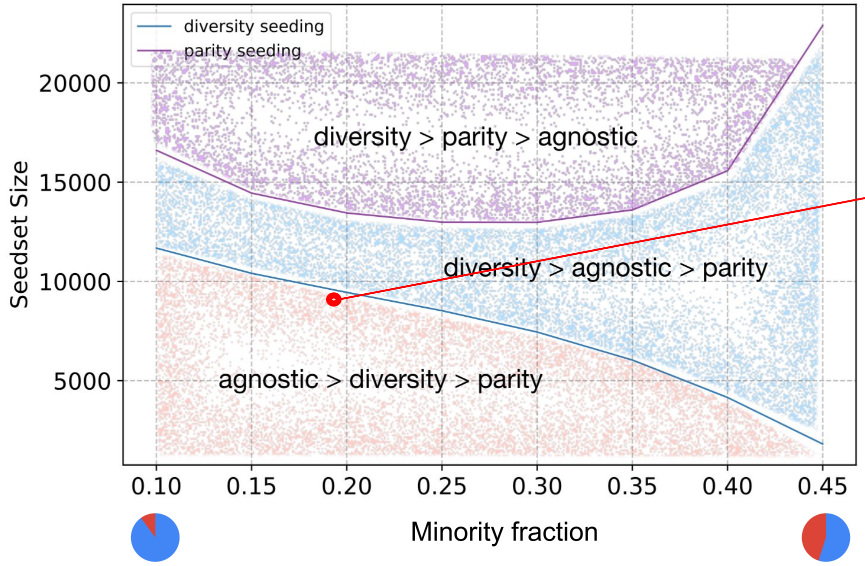


Network of ~53,000 nodes, 2 communities, homophily ρ = 0.135

DBLP citation dataset: men and women

$p = 0.01$	5,000 seeds		
	Agnostic seeding	Parity seeding	Diversity seeding
Total outreach	5,410.748	↓5,408.762	↑**5411.191**
F outreach	862.191	↑**1,004.232**	↑892.11
M outreach	**4,548.557**	↓4,404.53	↓4,519.081
F % in outreach	0.15934	↑**0.18567**	↑0.165

Theoretical analysis of diversity interventions



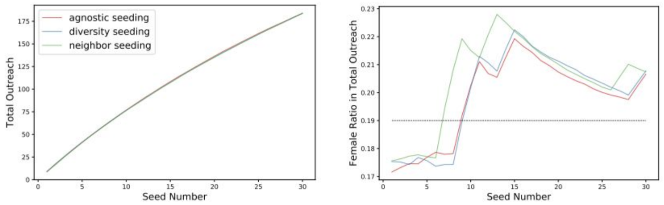
Network of ~53,000 nodes, 2 communities, homophily ρ = 0.135

DBLP citation dataset: men and women

	9,100 seeds		
p = 0.01	Agnostic seeding	Parity seeding	Diversity seeding
Total outreach	9,554.934	↑9,555.559	↑**9,556.349**
F outreach	1,581.842	↑**1,776.037**	↑1,679.423
M outreach	**7,973.092**	↓7,779.522	↓7,876.926
F % in outreach	0.16555	↑**0.186**	↑0.176

What about other seeding heuristics?

- Extend diversity seeding to neighbor seeding (NS): the neighbor set of the seeds has statistical parity



(a) (b)

Figure 5: Outreach (a) and female ratio in outreach (b) for the degree heuristics in DBLP, $p = 0.01$.

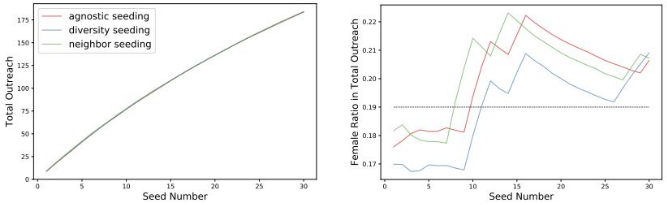
DBLP citation dataset: men and women

1000 seeds:

	Degree		
$p = 0.01$	AS	DS	NS
F outreach	37.984	↑**38.174**	↑38.151
M outreach	145.746	↓145.71	↑**145.786**

What about other seeding heuristics?

- Extend diversity seeding to neighbor seeding (NS): the neighbor set of the seeds has statistical parity



(a) (b)

Figure 6: Outreach (a) and female ratio in outreach (b) for the degree discount heuristics in DBLP, $p = 0.01$.

DBLP citation dataset: men and women

1000 seeds:

	Degree discount		
$p = 0.01$	AS	DS	NS
F outreach	37.901	↑**38.484**	↑38.135
M outreach	145.728	↓145.523	↑**145.901**

What about other seeding heuristics?

- Extend diversity seeding to neighbor seeding (NS): the neighbor set of the seeds has statistical parity

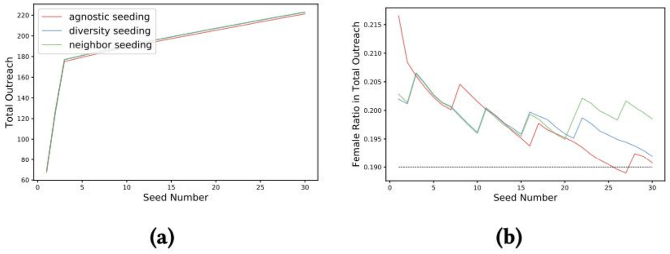


Figure 7: Outreach (a) and female ratio in outreach (b) for the greedy heuristics in DBLP, $p = 0.01$.

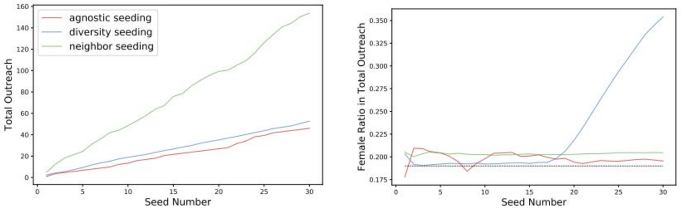
DBLP citation dataset: men and women

1000 seeds:

$p = 0.01$	Greedy AS	Greedy DS	Greedy NS
F outreach	42.215	↑42.793	↑**44.251**
M outreach	179.093	↑**180.128**	↓178.67

What about other seeding heuristics?

- Extend diversity seeding to neighbor seeding (NS): the neighbor set of the seeds has statistical parity



(a) (b)

Figure 9: Outreach (a) and female ratio in outreach (b) for the random heuristics in DBLP, $p = 0.01$.

DBLP citation dataset: men and women

1000 seeds:

$p = 0.01$	Random AS	Random DS	Random NS
F outreach	8.99	↑18.646	↑**31.433**
M outreach	36.988	↓34.021	↑**122.246**

Future directions

- Other models beyond independent cascade?
  - Linear threshold model[5]
- Theoretical analysis for different centrality metrics?
- Diversify modeling choices?
- Causality questions
  - Am I friends with people because we influenced each other or the other way around?[6]

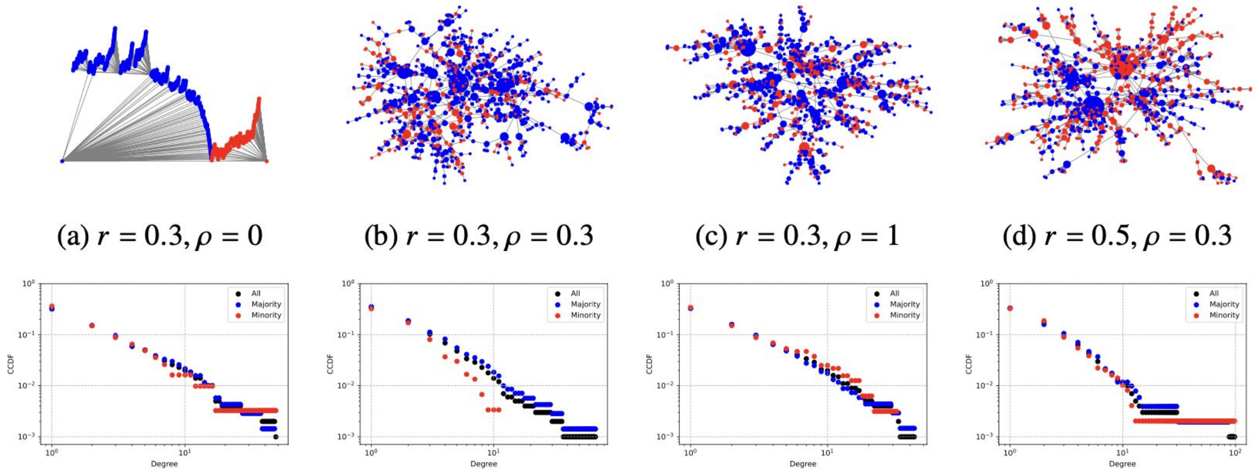
[5] Ali, J., Babaei, M., Chakraborty, A., Mirzasoleiman, B., Gummadi, K.P. and Singla, A., 2022, May. On the fairness of time-critical influence maximization in social networks. *ICDE*. 2022.
[6] Cristali I, Veitch V. Using Embeddings for Causal Estimation of Peer Influence in Social Networks. arXiv preprint arXiv:2205.08033. 2022.

Thank you!

Questions?

Additional slides

Preferential attachment with homophily

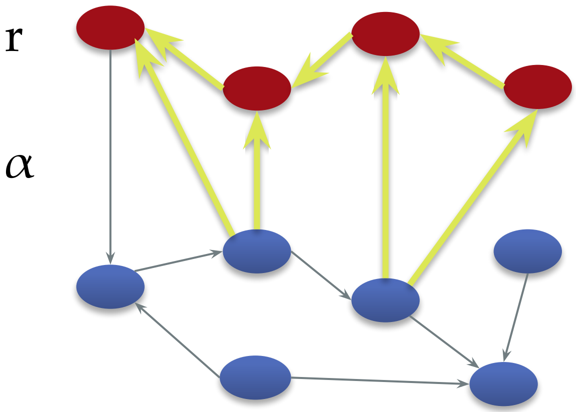


(a) $r = 0.3, \rho = 0$ (b) $r = 0.3, \rho = 0.3$ (c) $r = 0.3, \rho = 1$ (d) $r = 0.5, \rho = 0.3$

(e) $r = 0.3, \rho = 0$ (f) $r = 0.3, \rho = 0.3$ (g) $r = 0.3, \rho = 1$ (h) $r = 0.5, \rho = 0.3$

Figure 3.4: Networks generated from the Biased Preferential Attachment model (top row) and their respective cumulative complementary distribution functions, by community (bottom row), for different parameters.

Proof sketch



‘Wealth’ of red nodes:

- Fraction of edges towards R

$$\alpha_t = \sum_{v \in R} in\deg(v) / t$$

Define a function F as the rate of growth of α_t

- F has a fixed point α $\Rightarrow \alpha_t \to \alpha < r$

Proof sketch

Evolution equation:

- When does a node of degree k get a new link

Randomly

Preferential attachment

$T_t^R =$ rate at which R nodes receive edges through **randomness**

$k \cdot C_t^R =$ rate at which R nodes receives edges through **preferential attachment**

$$top_k(R) \sim k^{-\beta(R)} \qquad \beta(R) = 1 + \frac{1}{C^R}$$

$$top_k(B) \sim k^{-\beta(B)} \qquad \beta(B) = 1 + \frac{1}{C^B}$$

$$C_B = \frac{r\rho}{2\alpha + 2(1-\alpha)\rho} + \frac{(1-r)}{2\alpha\rho + 2(1-\alpha)}$$

$$C_R = \frac{(1-r)\rho}{2(\alpha\rho + 1 - \alpha)} + \frac{r}{2(\alpha + (1-\alpha)\rho))}.$$