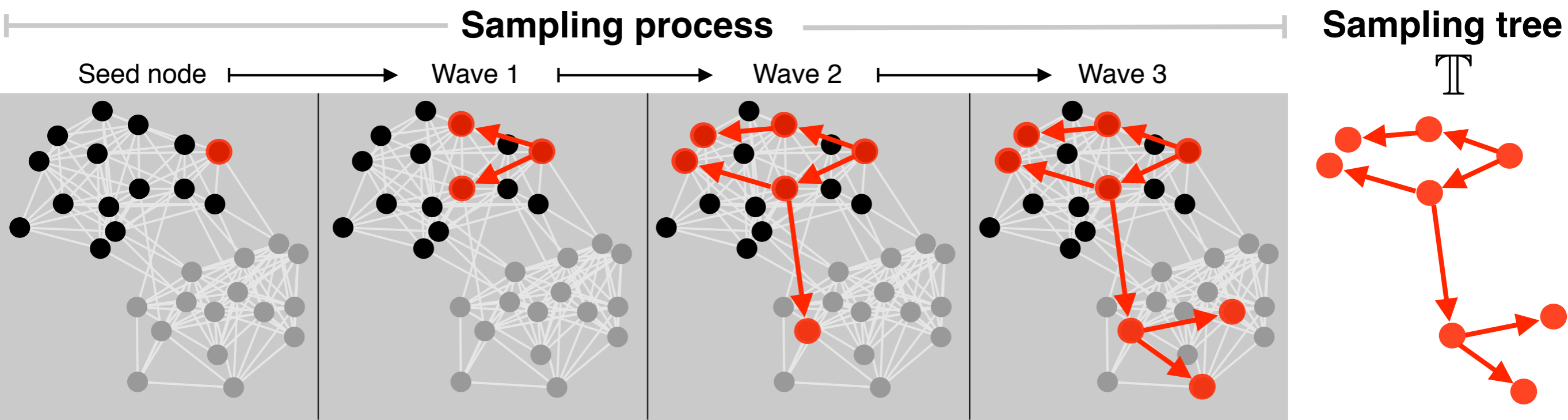


social driven
transition Colorado design Markov
bottleneck threshold Volz-Heckathorn
correlation random Galton-Watson
Variance **network**
graph political ^{sex} blogs **Sampling**
spectral tree eigenvectors
Theorem respondent
snowball
friends

A critical threshold for network driven sampling

Karl Rohe; karlrohe@stat.wisc.edu
UW-Madison Department of Statistics



$$\text{Var}_{SRS}(\hat{\mu}) = \frac{\sigma^2}{n}$$

$$\text{Var}_{RDS}(\hat{\mu}) = \sum_{\ell=2}^N \langle y, f_{\ell} \rangle_{\pi}^2 \mathbb{G}(\lambda_{\ell})$$

Network driven sampling does not require a sampling frame.

- Standard sampling techniques like random digit dialing require a sampling frame
“simple random sample”
- No sampling frame for:
 - homeless, twitter discussions, refugees, sex workers, jazz musicians.
- Even with a sampling frame, low response rates!

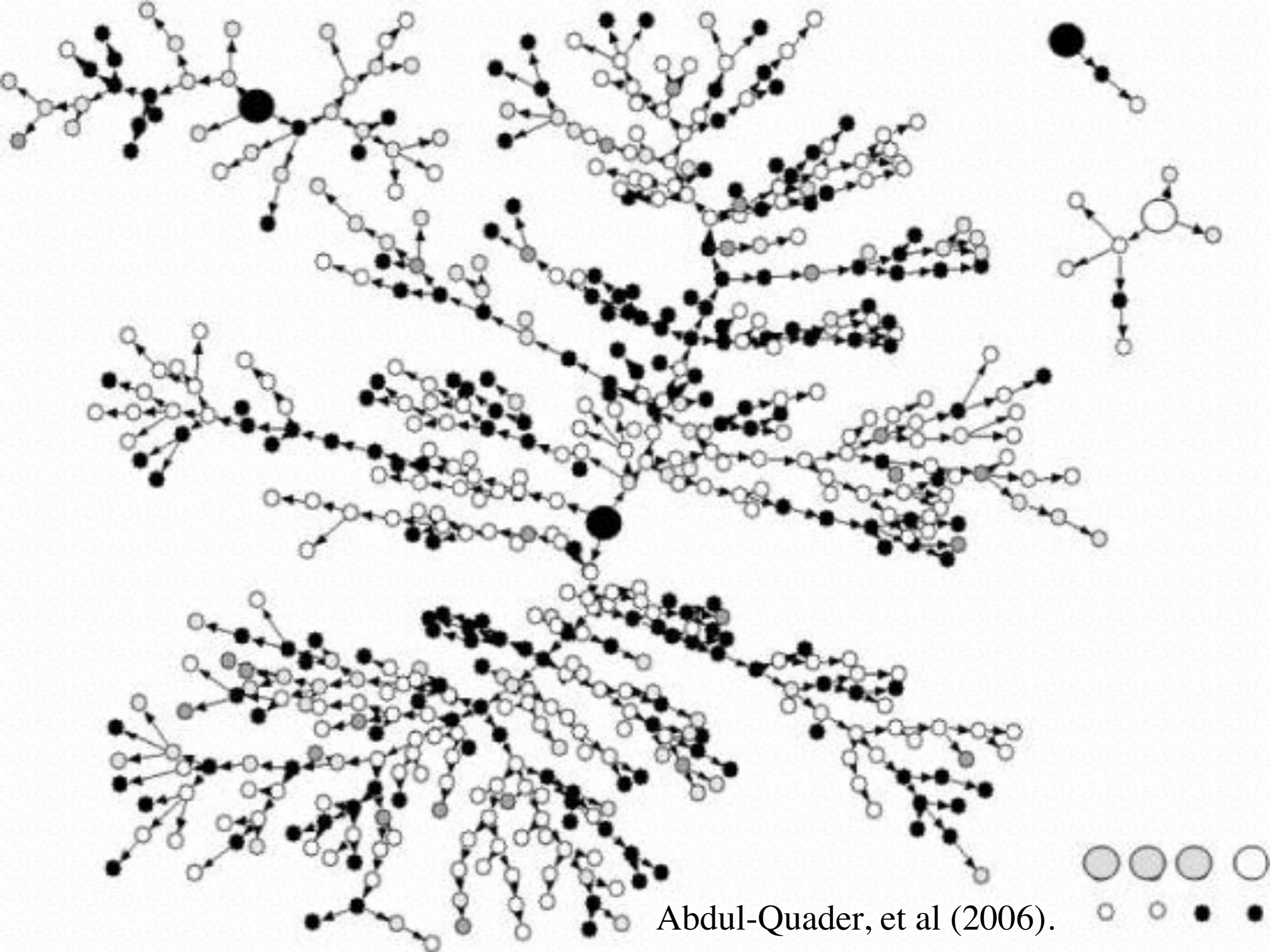
Respondent driven sampling (RDS) and snowball sampling are often used in social science research.

- To study the hard-to-reach and/or marginalized populations.
- Particularly prevalent in HIV research.
- Three risk populations:
Men who have sex with men (MSM),
people who inject drugs (PWID), and
Female sex workers (FSW)

RDS for HIV is the motivating example, but network driven sampling appears elsewhere.

RDS relies on friends passing coupons.

- Find seeds from a convenience sample.
- Give each seed three coupons to refer friends.
- The coupons have a dual incentive structure
 - Pay the person for making a referral
 - Pay the person being referred
- Iterate through referral tree



Abdul-Quader, et al (2006).

RDS is increasingly popular

- The CDC manages the National HIV Behavioral Surveillance System
 - RDS every few years in 20 major metropolitan areas.
- The WHO and UNAIDS have published an extensive manual for how to properly implement RDS

We wish to estimate the
proportion of the
population that is HIV+.

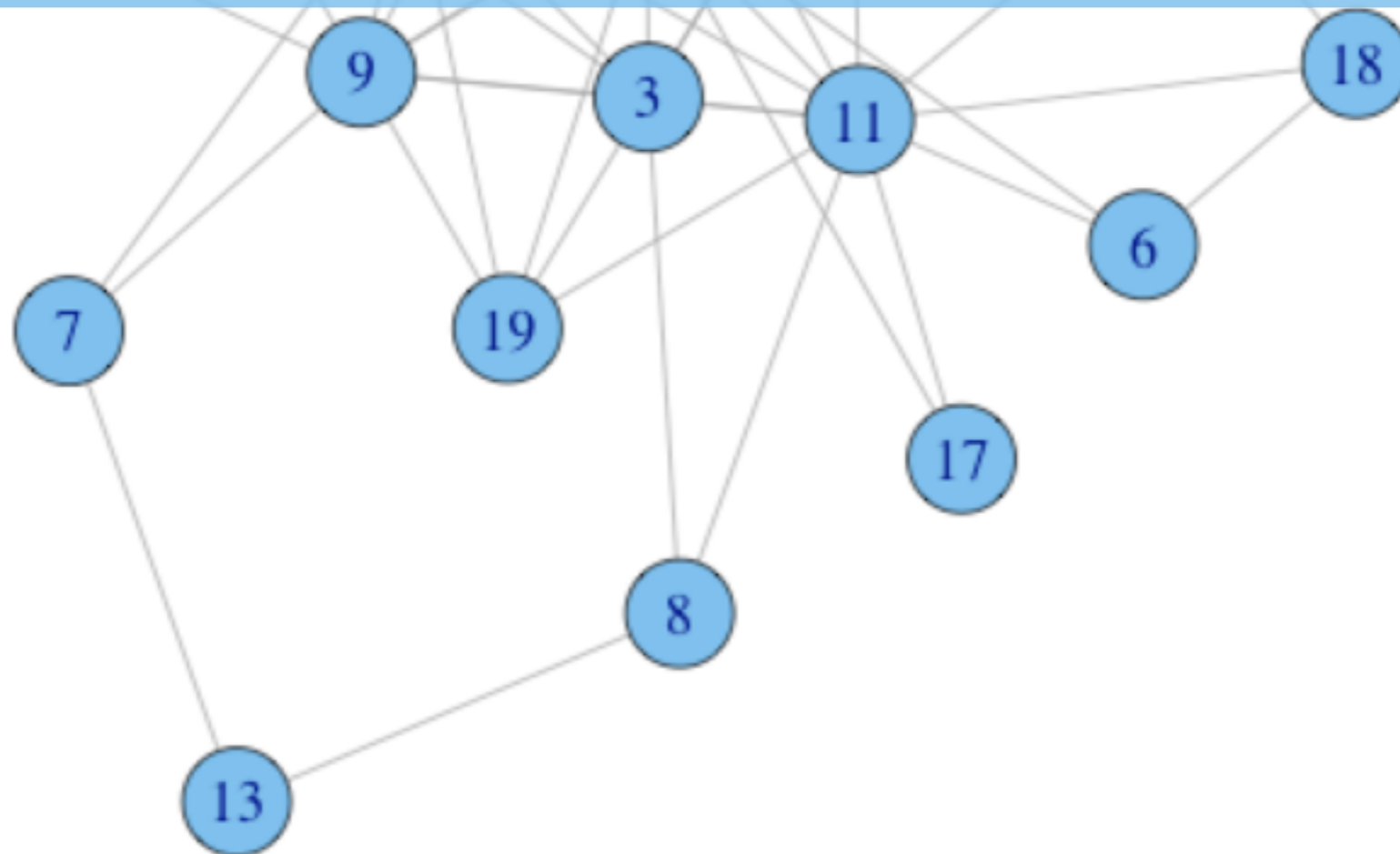
A Major Assumption

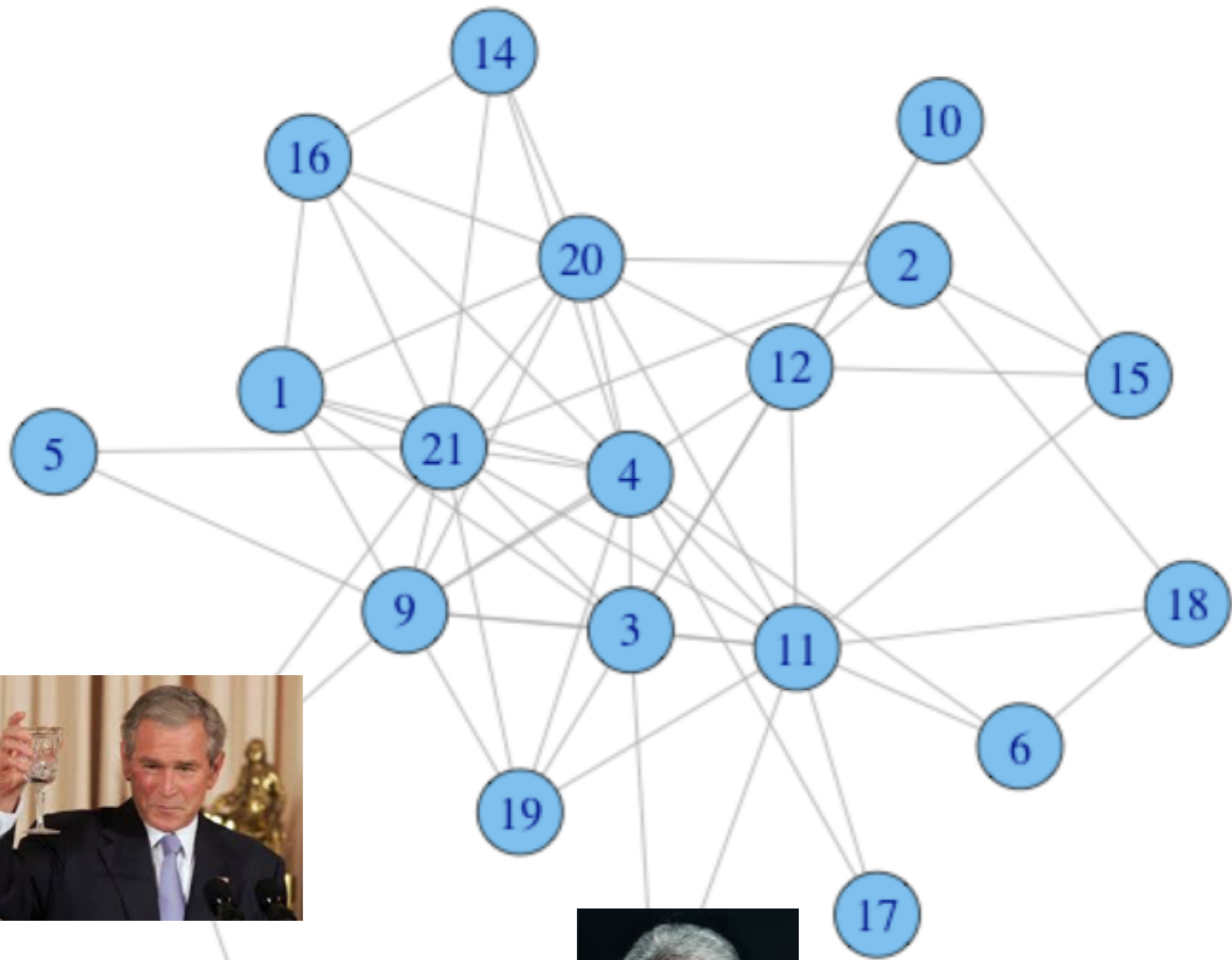
- If people make random referrals then we can make statistical inferences using probabilistic tools and techniques
 - confidence intervals and p-values.
- Suppose you have X friends in the target population. The friend you refer is chosen uniformly at random from these X people.

A Major Assumption

- Under this model, the passing of the coupon is a random walk on the social network.
- For simplicity, temporarily suppose that there is only one coupon.

**Friendships can be plotted as a graph.
The circles are “nodes” or people.
The lines are “edges” or friendships.**





Unbiased estimation requires the inclusion probabilities.

- People with large degrees are more likely to be sampled.

Using the random walk (and other) assumptions, we can approximate the inclusion probabilities.

The stationary distribution:

$$\pi_j = \lim_{t \rightarrow \infty} P(\text{referral } t \text{ is person } j | \text{seed is person } i)$$

Using the random walk (and other) assumptions, we can approximate the inclusion probabilities.

$$\pi_j \propto \text{number of friends of person } j$$

- Assumptions:
 - uniform selection of friends, with replacement.
 - reversible Markov chain (i.e. “symmetric”).
 - social network is connected and aperiodic
- Asymptotic

Volz-Heckathorn (2008) estimator uses the stationary distribution to construct a Horvitz-Thompson estimator.

$$y_t = \begin{cases} 1 & \text{if person } t \text{ is HIV positive} \\ 0 & \text{if person } t \text{ is HIV negative} \end{cases}$$

$deg(t)$ = # of friends of person t

$$\hat{\mu}_{VH} = \sum_t^n w_t y_t \quad w_t = \frac{1/deg(t)}{\sum_j^n 1/deg(j)}$$

Asymptotically unbiased!

What is the standard error of
this estimator?

Computing the standard error is an essential step for creating confidence intervals and testing hypotheses.

What is the standard error of this estimator?

Classical standard error for simple random sample.

$$\frac{\sigma}{\sqrt{n}}$$

Usually, sigma is unknown.

What is the standard error of this estimator?

Classical standard error for simple random sample.

$$\frac{\sigma}{\sqrt{n}}$$

You can estimate it!

$$\frac{\hat{\sigma}}{\sqrt{n}}$$

In this talk, we will find the analogue to the first formula for RDS.

Key difficulty with RDS data: samples are dependent.

- Friends are similar in many ways, including HIV status.
- If Fred refers Bill, they are likely to have similar HIV status.

Need to decide what to model as “random”

- Salganik and Heckathorn (and co-authors) typically model *referrals* as random.
- Giles and Handcock and others make an additional assumption that the network is random.
- There are currently three bootstrap techniques that build off of these assumptions.

Previous standard error estimators have modeled traits y as a first order Markov Chain

- Salganik (2006). “Variance estimation, design effects, and sample size calculations for respondent-driven sampling.” *Journal of Urban Health*.
- This is an additional modeling assumption beyond what is needed for the unbiased-ness of VH.
- Underestimates variance. Neely [2009], Verdery et al. [2013]
- In Q/A, I can address when the “first order assumption” holds and when it does not.

Recap

- RDS is a network link tracing technique
- Three bits: social network, referral tree, HIV status
- Assumption of random referrals allow for statistical inference
- VH estimator is asymptotically unbiased
- We need to understand the variance.

Outline

I. Model and notation.

Markov transitions, sampling tree,
node features.

II. Key mathematical pieces.

eigenvectors of P

The G function

III. The true sampling variance

$$se(\hat{\mu}_{SRS}) = \frac{\sigma}{\sqrt{n}}$$

A. A scary story

IV. Designed RDS

The Markov transition matrix describes how coupons are passed along the social network.

- Markov transition matrix P is $N \times N$.
 N = population size. (n will be sample size)
- $P(\text{person } i \text{ refers person } j | \text{person } i \text{ has 1 coupon}) = P_{ij}$
- e.g. VH assumption: i chooses friend uniformly at random, so that

$$P_{ij} = \frac{\mathbf{1}(i, j \text{ friends})}{\text{deg}(i)}$$

Regularity conditions on P

- Assume P is reversible wrt the stationary distribution.

$$\pi_i P_{ij} = \pi_j P_{ji}$$

- For random walk, equivalent to assuming an undirected network.

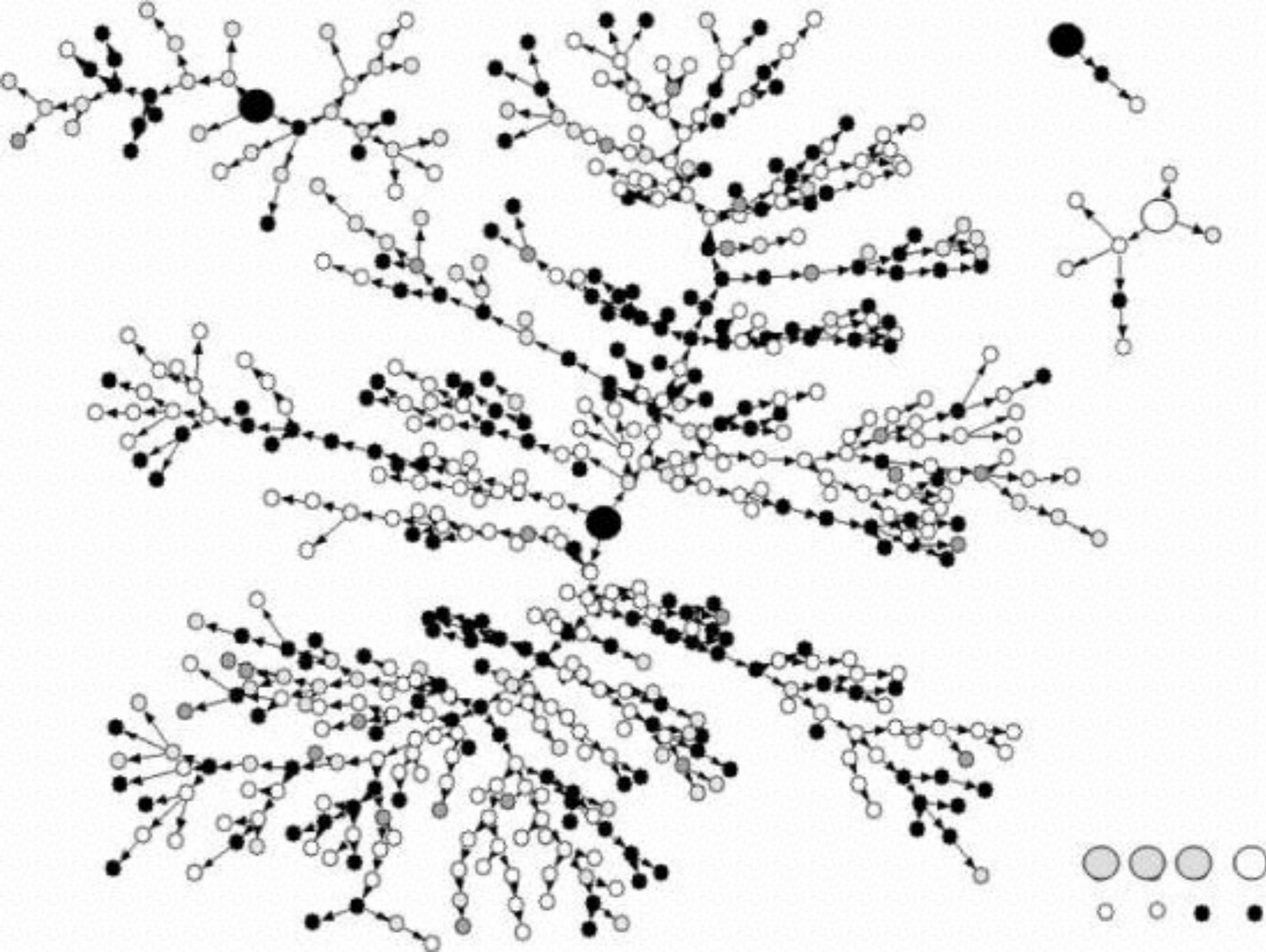
- Assume $|\lambda_2(P)| < 1$

(akin to connected and aperiodic)

$$\pi_j = \lim_{t \rightarrow \infty} P(\text{referral } t \text{ is person } j | \text{seed is person } i)$$

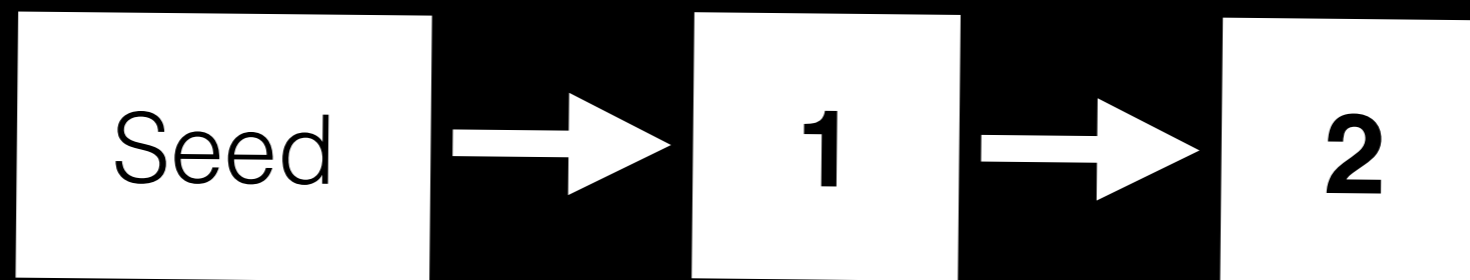
Each person can refer up to three (sometimes five) future participants.

- To represent the referral process, we need a “tree”
- Call this object T . It is a graph, that contains elements $1, \dots, n$ corresponding to the n samples.
- If i refers j into the study, then $i \rightarrow j$.



Standard Markov chain

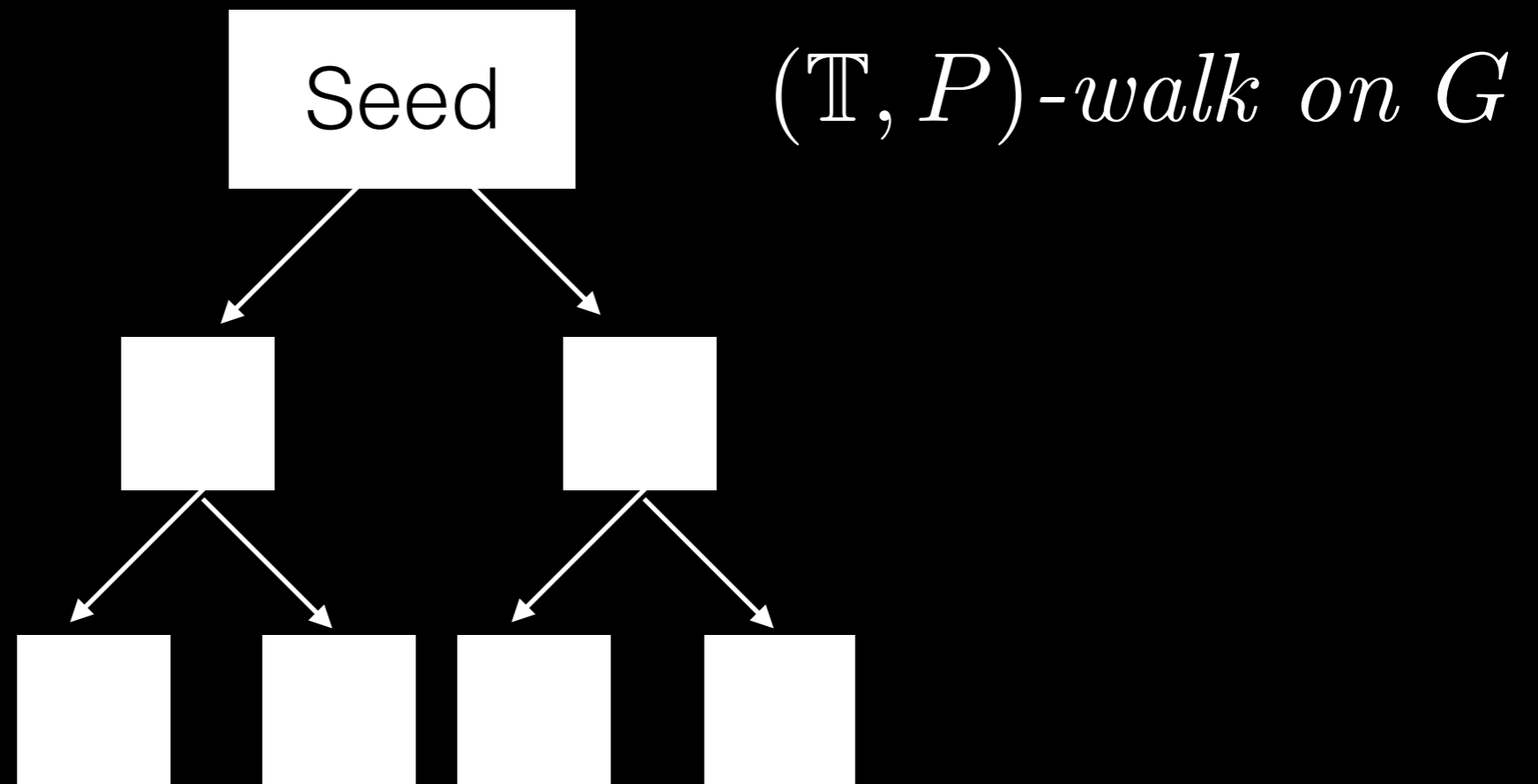
$$\{X(i) \in \text{people} : i = 1, \dots, n\}$$

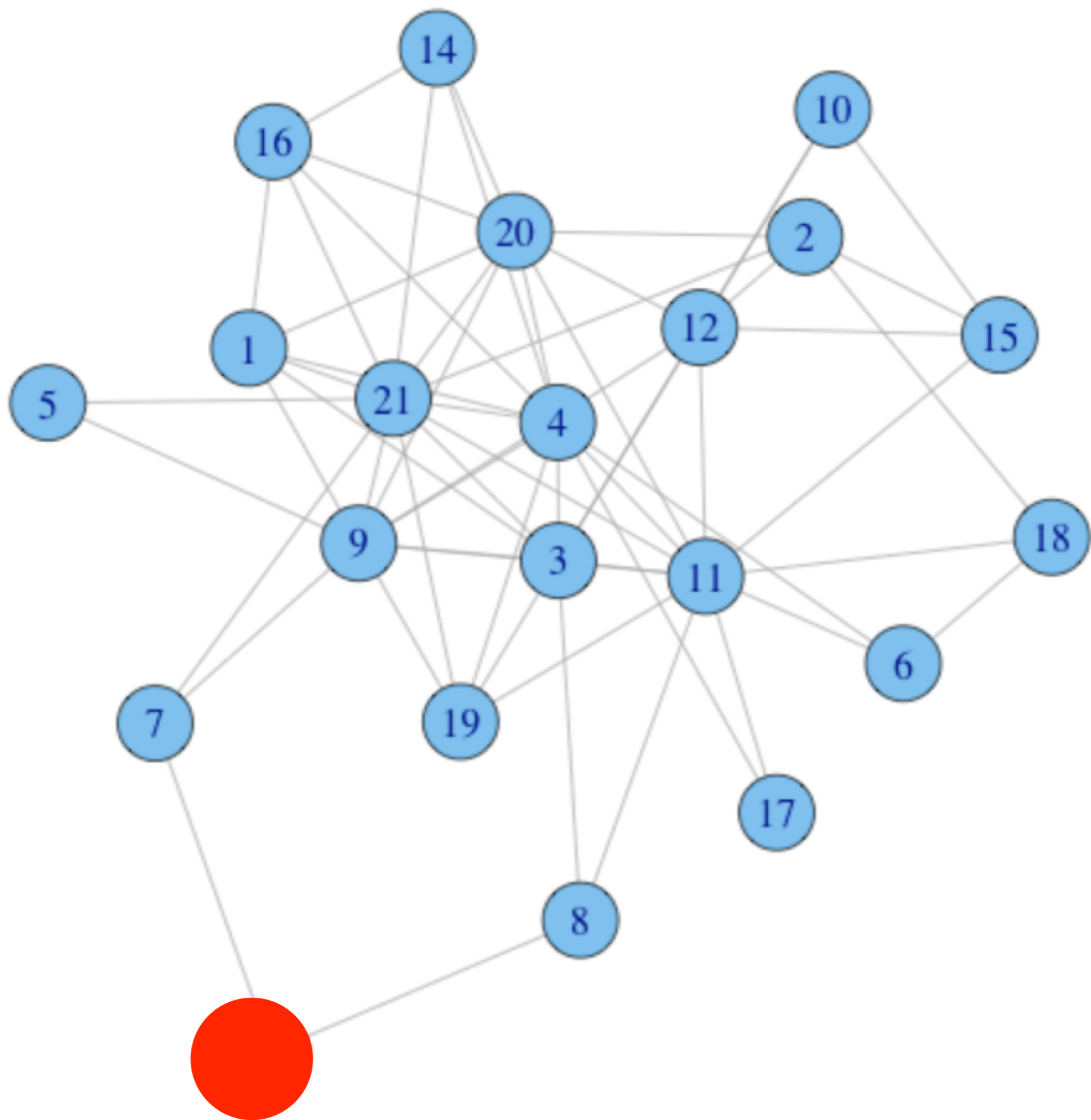


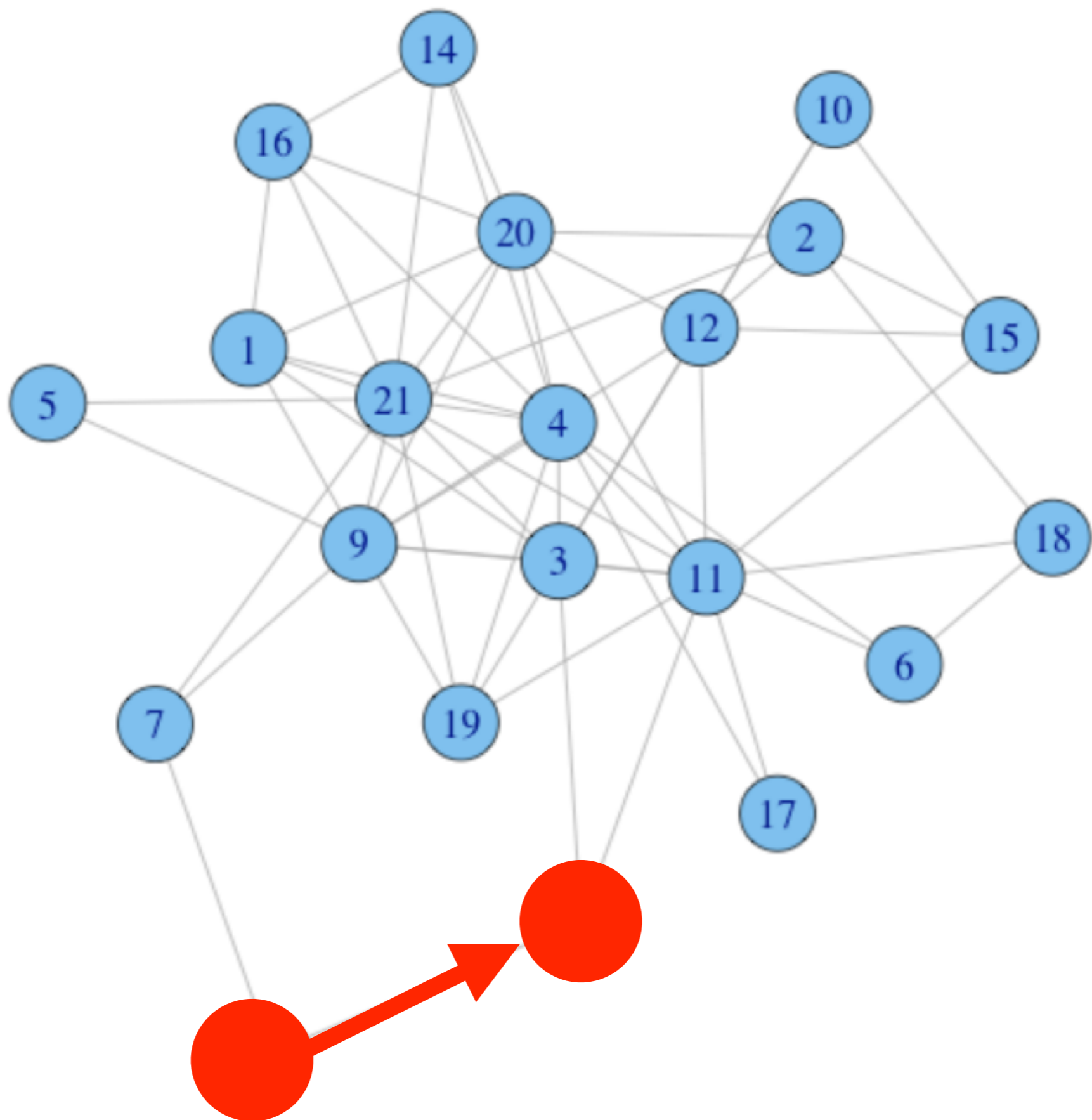
Markov transitions on the referral tree.

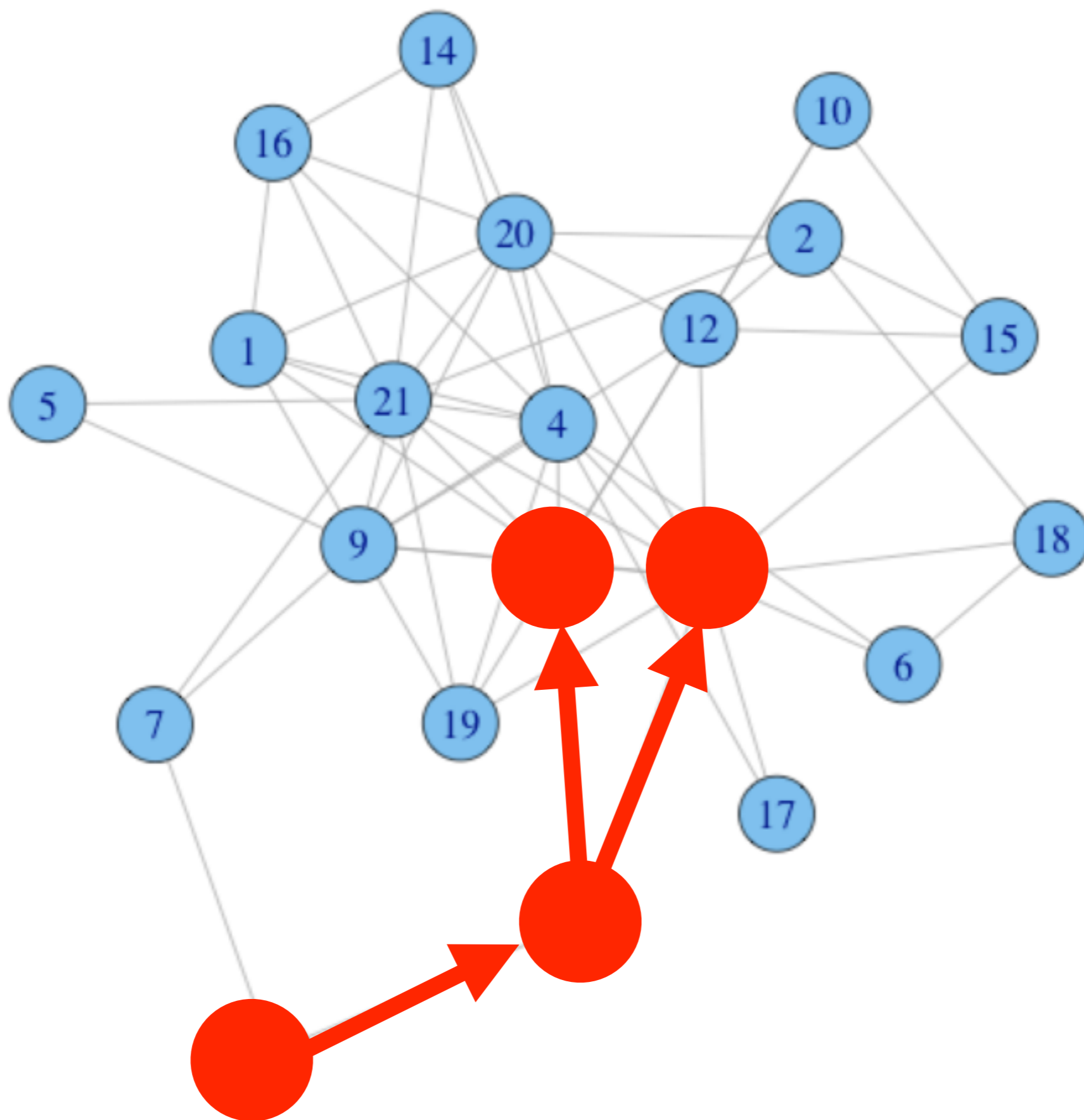
$$\{X(i) \in \text{people} : i = 1, \dots, n\}$$

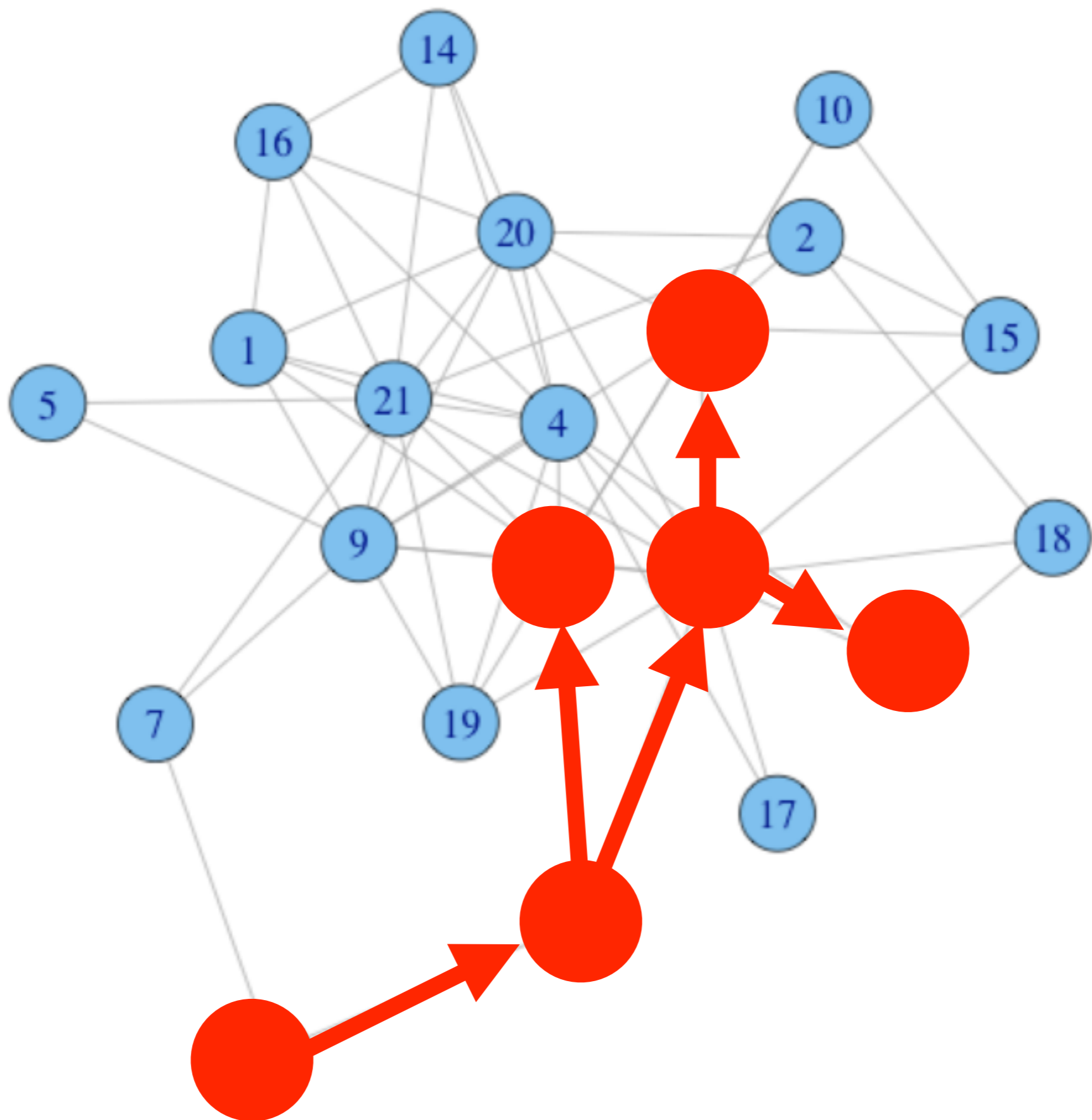
$$\{X_\tau \in \text{people} : \tau \in \mathbb{T}\}$$

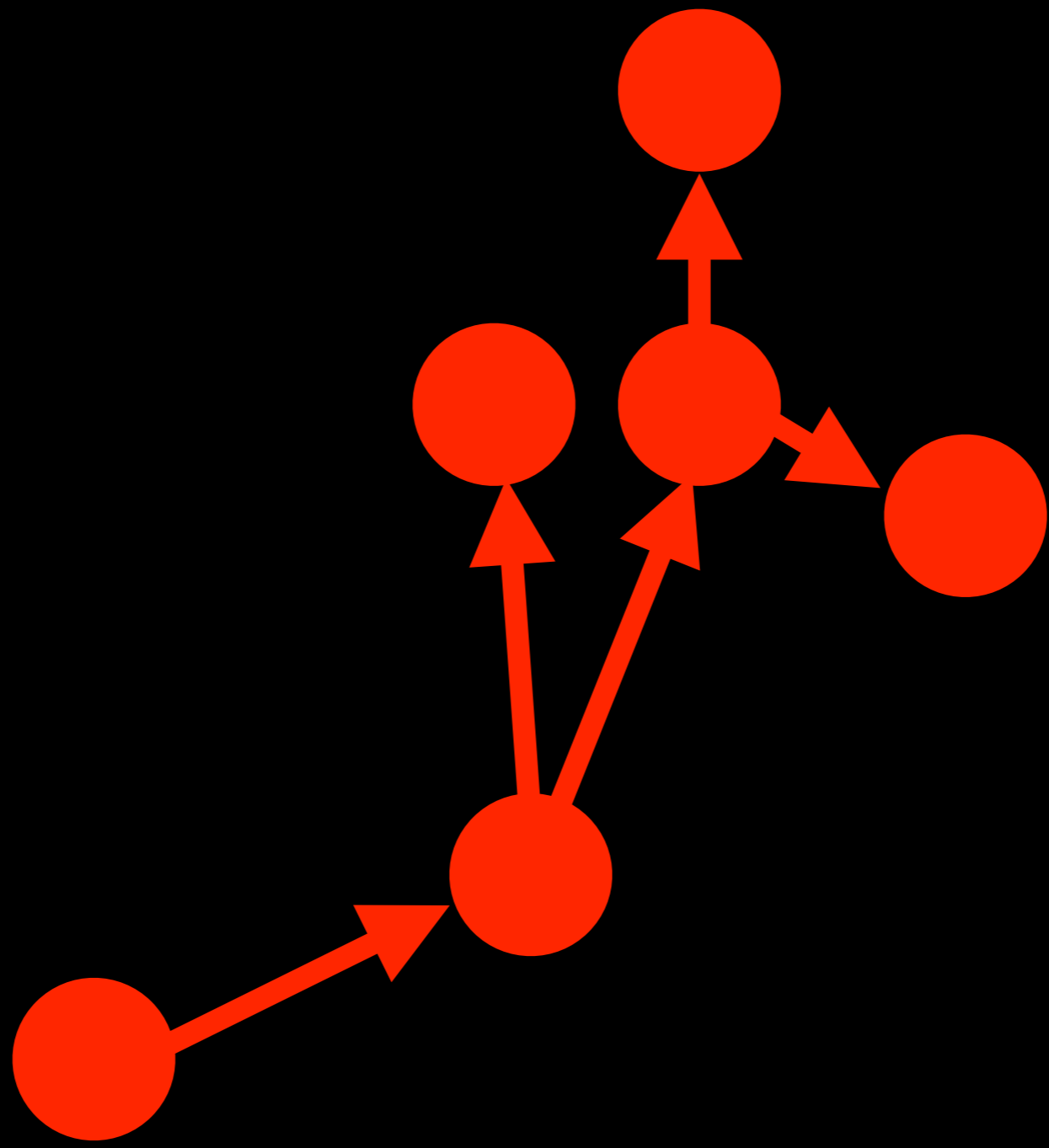












Each node in the graph is either infected or not.

- For person $i = 1, \dots, N$

$$y_i = \begin{cases} 1 & i \text{ sick} \\ 0 & i \text{ not sick} \end{cases}$$

We want to estimate y averaged across all nodes $1, \dots, N$ in the social network.

Everything holds if y is continuous.

These terms constitute the core of the model and the notation.

I. Markov transition matrix P
(underneath P , there is a social network)

II. Referral tree T , $\{1, \dots, n\} \subset T$

III. Health status y_j , $j = 1, \dots, N$

Outline

I. Model and notation.

Network, Markov transitions, sampling tree,
node features.

II. Key mathematical pieces.

eigenvectors of P

The G function

$$\text{Var}(\hat{\mu}) = \sum_{\ell=2}^N \langle y, f_{\ell} \rangle_{\pi}^2 \mathbb{G}(\lambda_{\ell})$$

III. The true sampling variance

A. A scary story

IV. Designed RDS

“Bottlenecks” can prevent representative samples.

- Suppose a town with two communities: EAST and WEST.
- All seeds belong to EAST.
- Few friendships cross the town.
- Will you collect enough data from WEST?
- What if EAST and WEST have same incidence of HIV?
- We need a mathematical way to express these “bottlenecks” ...

Lemma 1.1. *Let P be reversible with respect to the stationary distribution π . The eigenvectors of P , denoted as f_1, \dots, f_N , are real valued functions of the nodes $i \in V$ and orthonormal with respect to the inner product*

$$\langle f_a, f_b \rangle_\pi = \sum_{i \in V} f_a(i) f_b(i) \pi_i. \quad (2)$$

When $|\lambda_2| < 1$, the leading eigenvector is constant vector of ones, $f_1 = 1$. Moreover, the probability of a transition from $i \in V$ to $j \in V$ in t steps can be written as

$$\mathbb{P}(X(t) = j | X(0) = i) = P_{ij}^t = \pi_j + \pi_j \sum_{\ell=2}^N \lambda_\ell^t f_\ell(i) f_\ell(j).$$

Eigenvectors are a mathematical refinement for the concept of “bottlenecks”

Eigenvectors indicate the bottlenecks


$$f_\ell, \lambda_\ell \quad \ell = 1, \dots, N$$



eigenvalues indicate the “strength”
of the bottleneck.

These are well studied objects in linear algebra
and spectral graph theory.

Eigenvectors are a mathematical refinement for the concept of “bottlenecks”

$f_\ell(i)$

An eigenvector assigns a value to each node.

IF: $sign(f_\ell(i)) = sign(f_\ell(j))$

THEN: i and j are on the same side of this bottleneck.

For example, the EAST and WEST bottleneck would be represented by one eigenvector.

- The two communities are EAST and WEST.
- If this is the biggest bottleneck in the network, then the second eigenvector will have opposite signs on nodes from EAST and WEST.

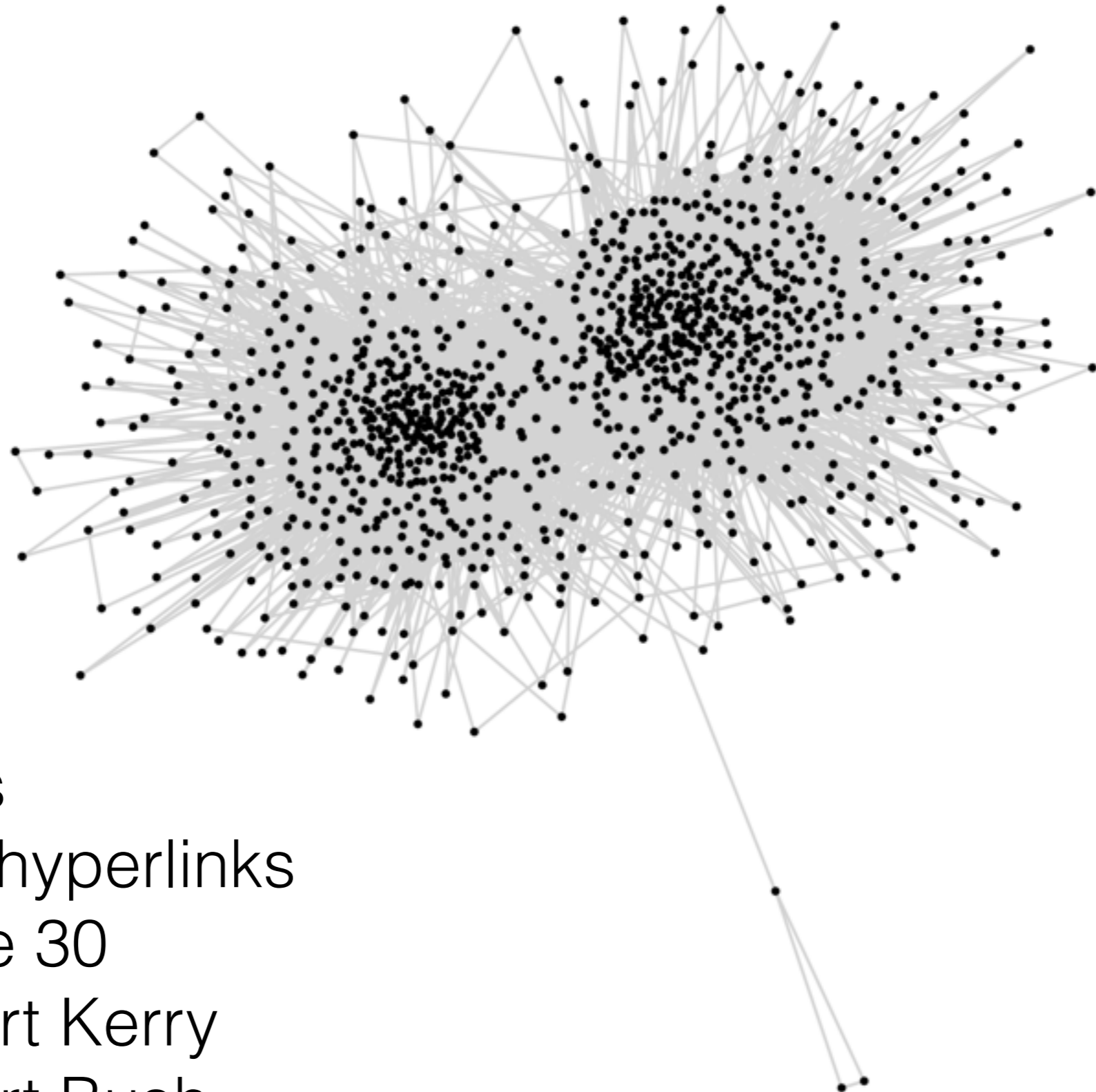
Lemma 1.1. *Let P be reversible with respect to the stationary distribution π . The eigenvectors of P , denoted as f_1, \dots, f_N , are real valued functions of the nodes $i \in V$ and orthonormal with respect to the inner product*

$$\langle f_a, f_b \rangle_\pi = \sum_{i \in V} f_a(i) f_b(i) \pi_i. \quad (2)$$

When $|\lambda_2| < 1$, the leading eigenvector is constant vector of ones, $f_1 = 1$. Moreover, the probability of a transition from $i \in V$ to $j \in V$ in t steps can be written as

$$\mathbb{P}(X(t) = j | X(0) = i) = P_{ij}^t = \pi_j + \pi_j \sum_{\ell=2}^N \lambda_\ell^t f_\ell(i) f_\ell(j).$$

Political Blogs



1084 blogs

edges are hyperlinks

avg degree 30

Half support Kerry

Half support Bush

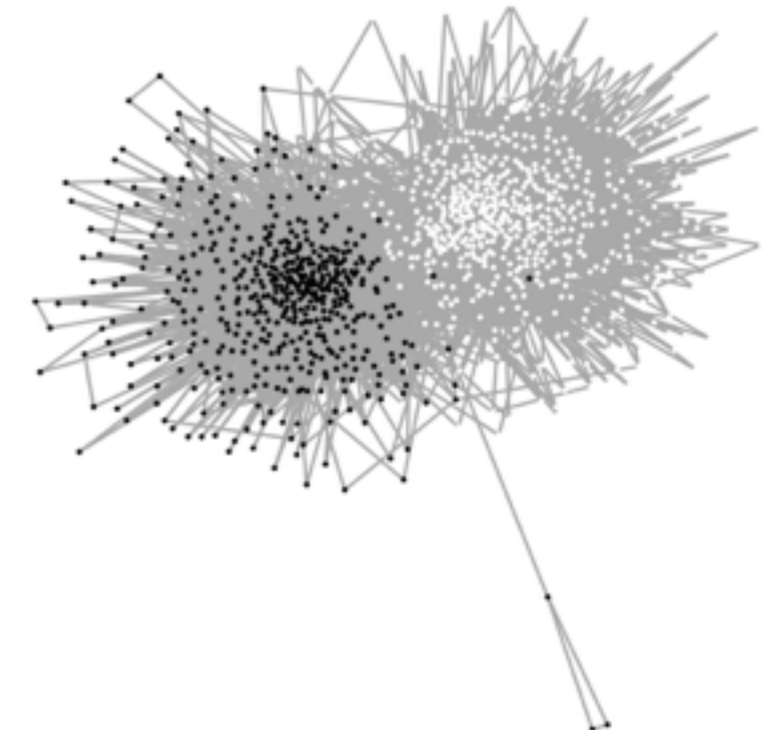
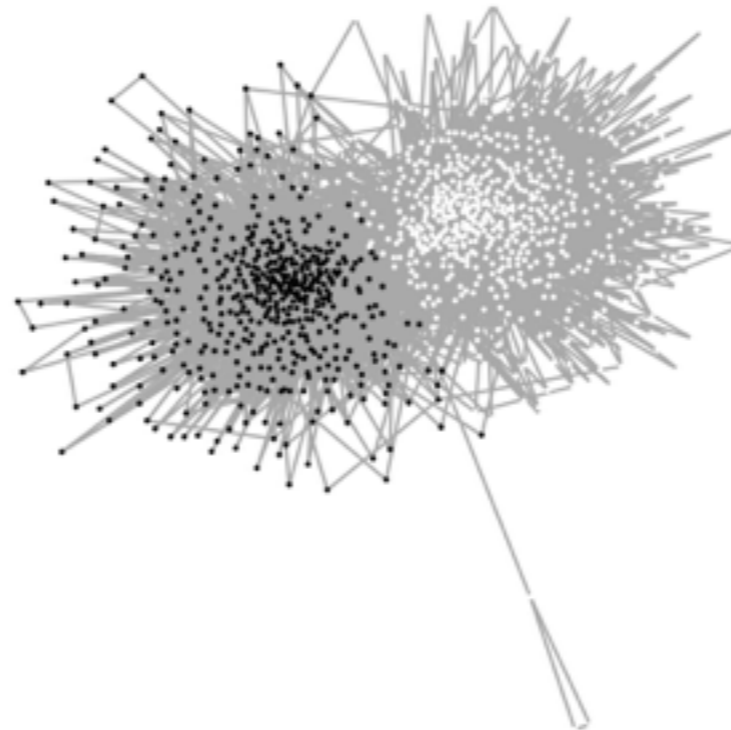
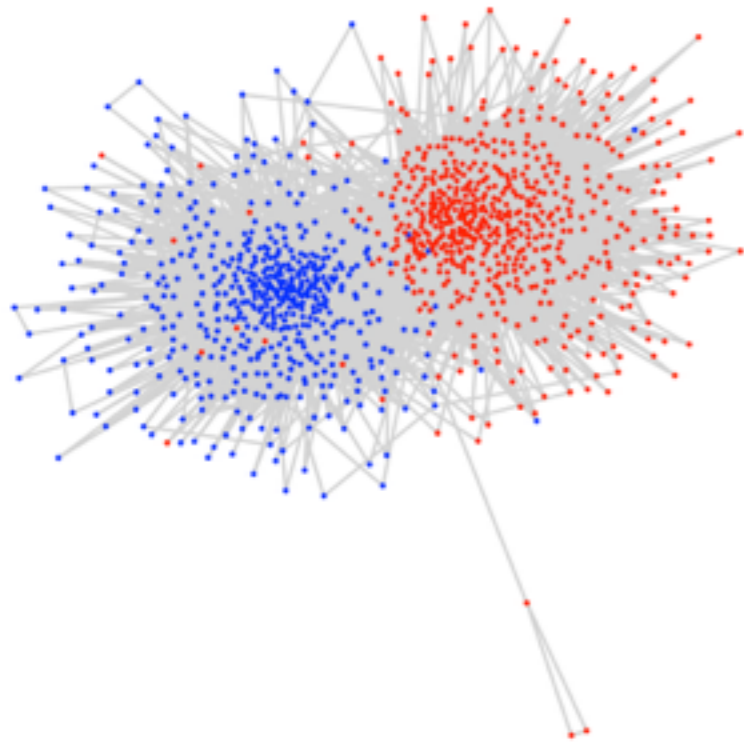
Eigenvectors correspond to bottlenecks.
Each one has a value on each node.
Bottlenecks are problematic when they correlate with y .

$$\langle y, f_\ell \rangle_\pi^2$$

Political Blogs

$\lambda_2 = 0.89$

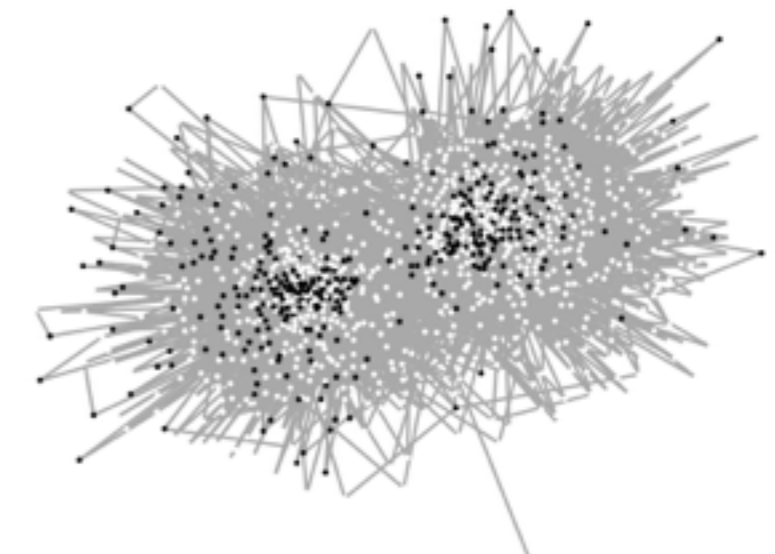
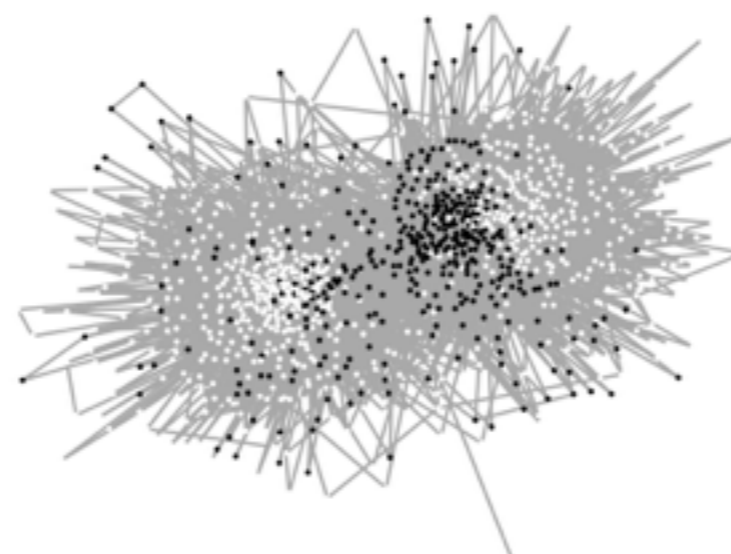
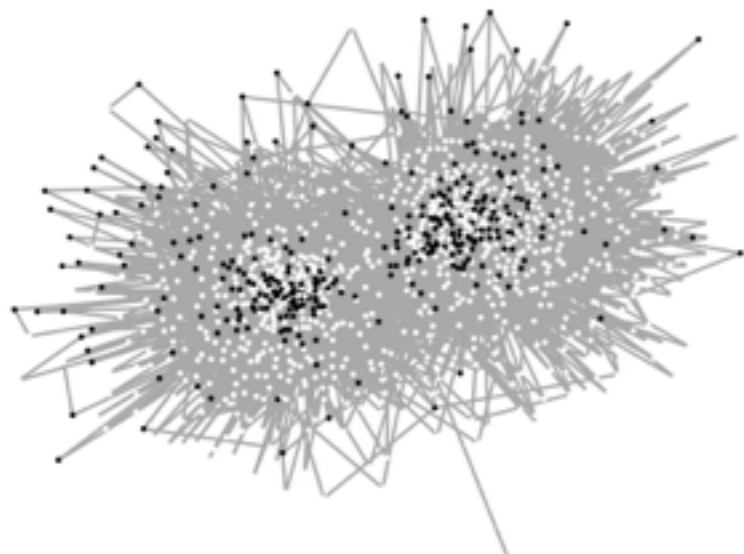
$\lambda_3 = 0.89$



$\lambda_4 = 0.55$

$\lambda_5 = 0.53$

$\lambda_6 = 0.53$



Eigenvalues indicate the strength of the bottleneck.

For all ℓ , $-1 \leq \lambda_\ell \leq 1$

$\lambda_1 = 1$ Don't worry about the first eigenvalue.

If $\lambda_2 = 1$, then the graph is not connected!

If λ_2 is close to one, then the graph has a strong bottleneck.

$$\mathit{Var}(\hat{\mu}) = \sum_{\ell=2}^N \langle y, f_{\ell} \rangle_{\pi}^2 \mathbb{G}(\lambda_{\ell})$$

The G function measures the “stringiness” vs the “bushy-ness” of the tree T.

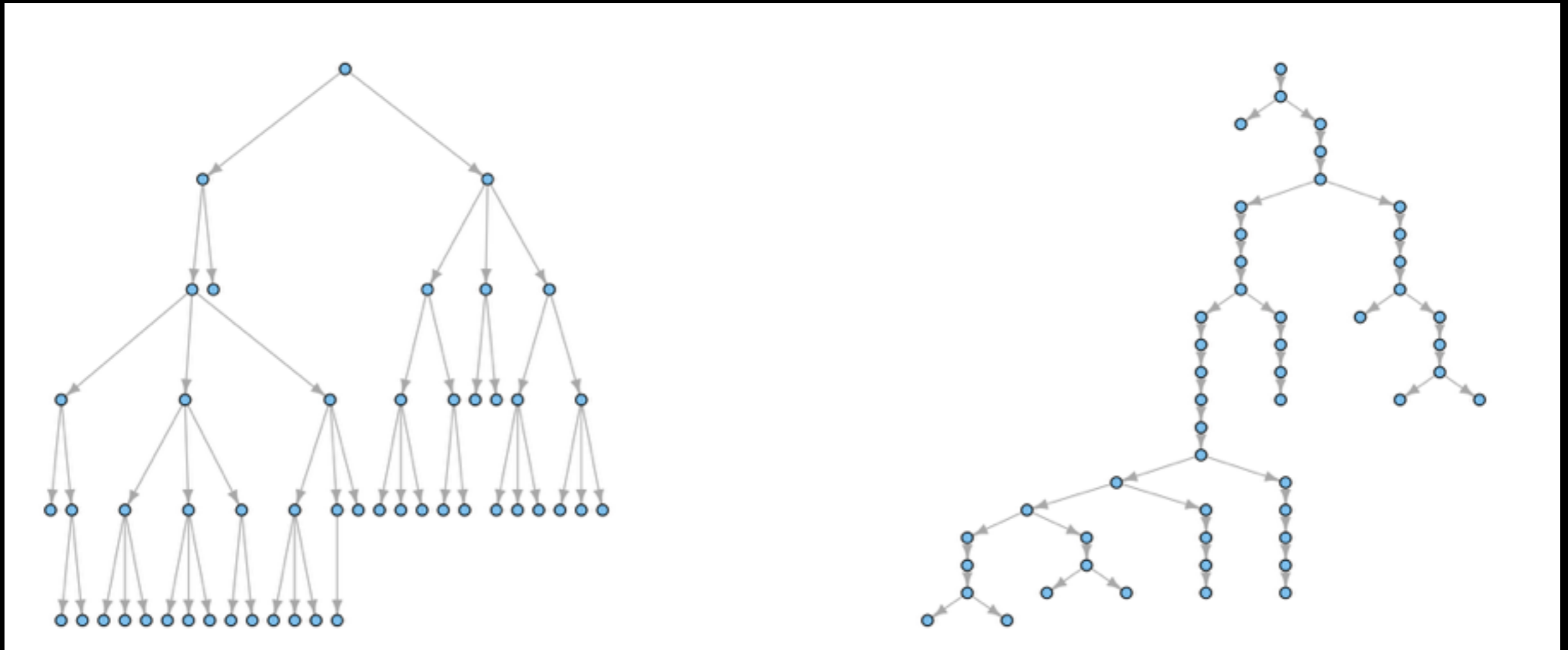
- Draw two observations I and J uniformly at random from T.
- $D = d(I, J)$, graph distance in T.
- Then,

$$G(z) = E(z^D) \text{ for } |z| < 1$$

Bushy trees have larger referral rates.

“Bushy”

“Stringy”



$m \sim 2$

$m \sim 1$

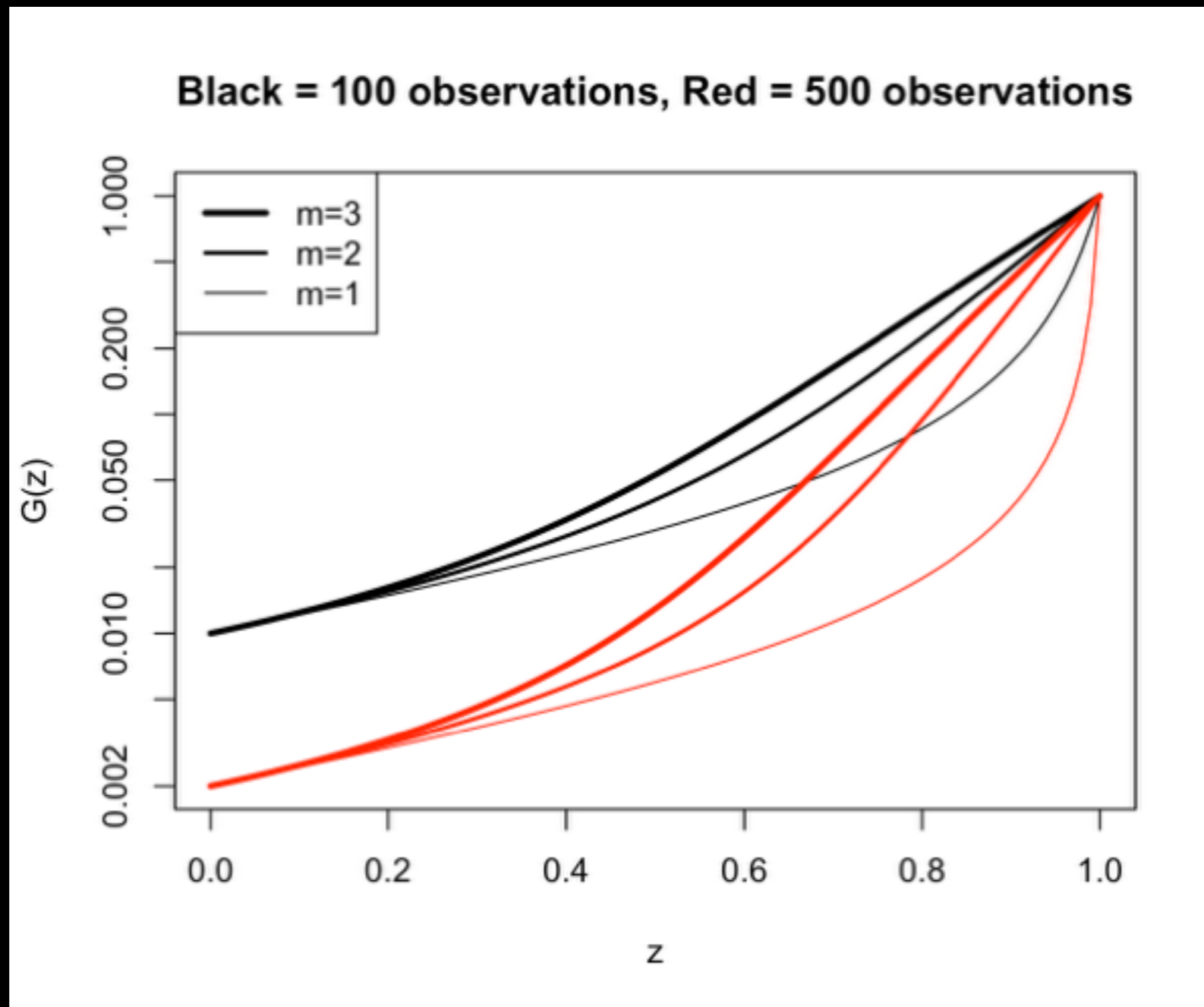
average number of referrals = m

$$G(z) = E(z^D) \text{ for } |z| < 1$$

G is an increasing function.

G decreases when n increases.

G is larger for bushy trees.



Outline

I. Model and notation.

Network, Markov transitions, sampling tree,
node features.

II. Key mathematical pieces.

eigenvectors of P

The G function

III. The true sampling variance

A. A scary story

IV. Designed RDS

Sampling variance theorem

- Suppose that (1) the Markov chain satisfies regularity conditions and that (2) the seed node is sampled from the stationary distribution.

$$\hat{\mu} = \frac{1}{n} \sum_{i \in T} y_i$$

Sampling variance theorem

- Suppose that (1) the Markov chain satisfies regularity conditions and that (2) the seed node is sampled from the stationary distribution.

$$\hat{\mu} = \frac{1}{n} \sum_{i \in T} y_i$$

$$\text{Var}(\hat{\mu}) = \sum_{\ell=2}^N \langle y, f_{\ell} \rangle_{\pi}^2 \mathbb{G}(\lambda_{\ell})$$

Sampling variance theorem

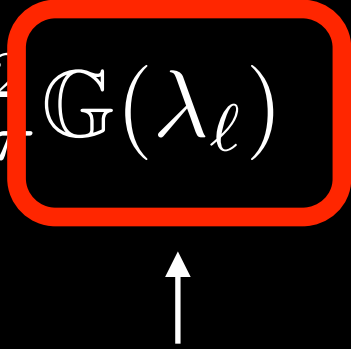
$$\hat{\mu} = \frac{1}{n} \sum_{i \in T} y_i$$

$$\text{Var}(\hat{\mu}) = \sum_{\ell=2}^N \langle y, f_{\ell} \rangle_{\pi}^2 \mathbb{G}(\lambda_{\ell})$$

Squared correlation between y and bottleneck ℓ

Sampling variance theorem

$$\hat{\mu} = \frac{1}{n} \sum_{i \in T} y_i$$

$$\text{Var}(\hat{\mu}) = \sum_{\ell=2}^N \langle y, f_{\ell} \rangle^2 \mathbb{G}(\lambda_{\ell})$$


G function evaluated at each eigenvalue.
Recall: eigenvalue is large if bottleneck is strong.

Proof sketch

By reversibility:

For $\sigma, \tau \in \mathbb{T}$, $d(\sigma, \tau) = t \implies (X_\sigma, X_\tau) \stackrel{d}{=} (X(0), X(t))$

By spectral representation:

$$\text{Cov}_\pi(Y_\sigma, Y_\tau) = \sum_{\ell=2}^N \lambda_\ell^{d(\sigma, \tau)} \langle y, f_\ell \rangle_\pi^2$$

Summing over all (σ, τ) and exchanging summations yields the result.

The theorem is stated for sample average.
Variance of Horvitz-Thompson is a slight
adjustment.

$$y^\pi(i) = \frac{y(i)}{\pi_i N}$$

$$\hat{\mu}_{HT} = \frac{1}{n} \sum_{i \in T} y_i^\pi$$

$$\hat{\mu}_{VH} = \sum_t^n w_t y_t$$

$$w_t = \frac{1/\deg(t)}{\sum_j^n 1/\deg(j)}$$

The theorem is stated for sample average.
Variance of Horvitz-Thompson is a slight adjustment.

$$y^\pi(i) = \frac{y(i)}{\pi_i N} \quad \hat{\mu}_{HT} = \frac{1}{n} \sum_{i \in T} y_i^\pi$$

$$Var(\hat{\mu}_{HT}) = \sum_{\ell=2}^N \langle y, f_\ell \rangle^2 \mathbb{G}(\lambda_\ell)$$

$$\langle y, f_\ell \rangle^2 = \sum_{i=1}^N y(i) f_\ell(i) \frac{1}{N}$$

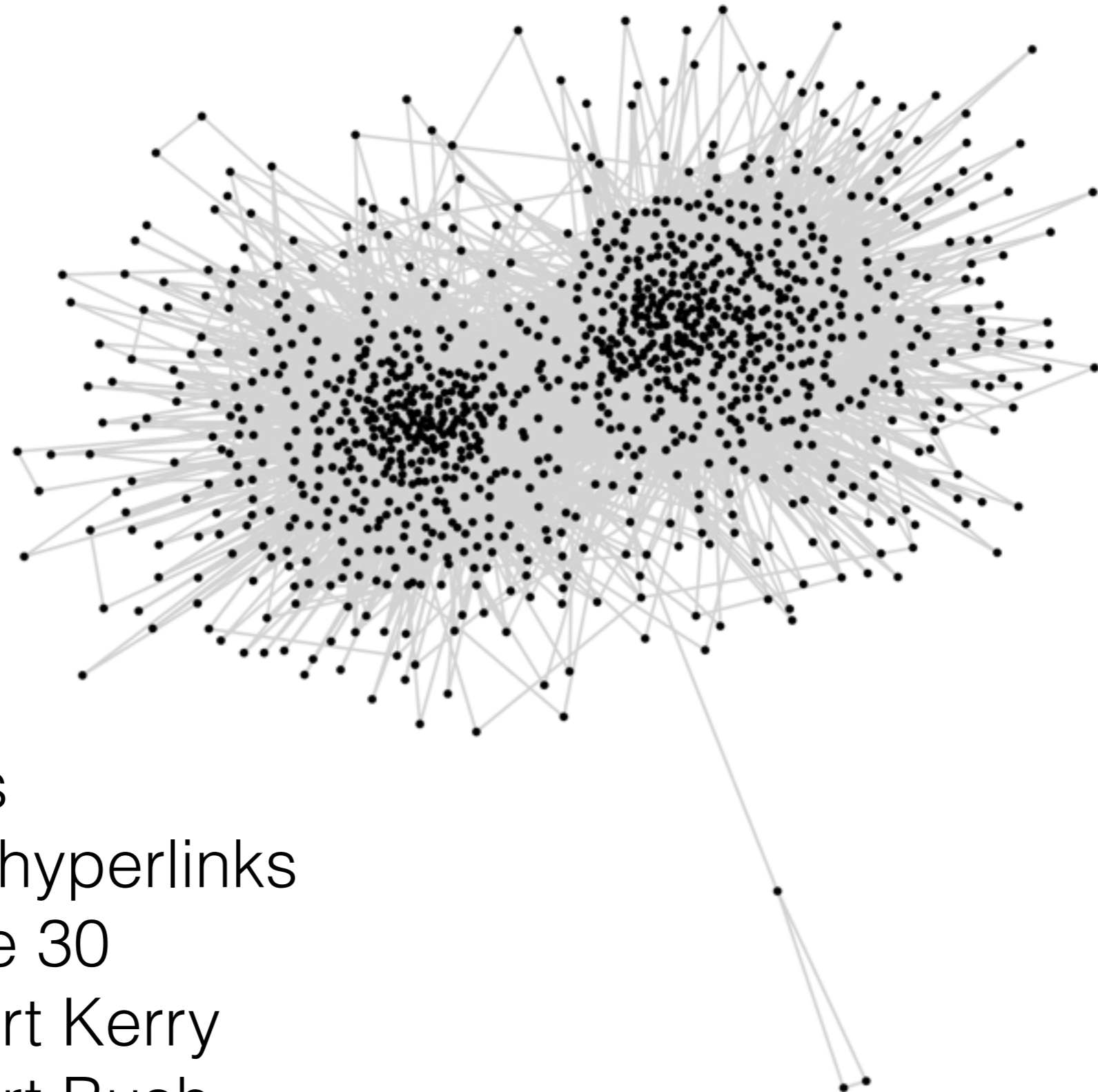


no pi!

Two example networks to study this formula.

- Political Blog data from 2004 US presidential election
- Colorado Project 90 data. Census of heterosexuals at risk for HIV, living around Colorado Springs in ~1990.
- Both study the 2-core of the largest connected component.

Political Blogs

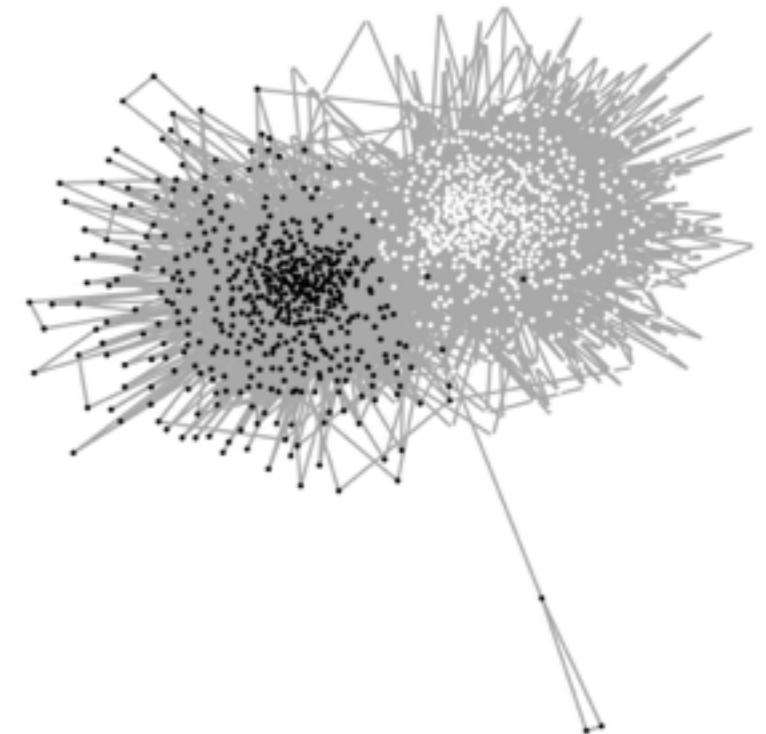
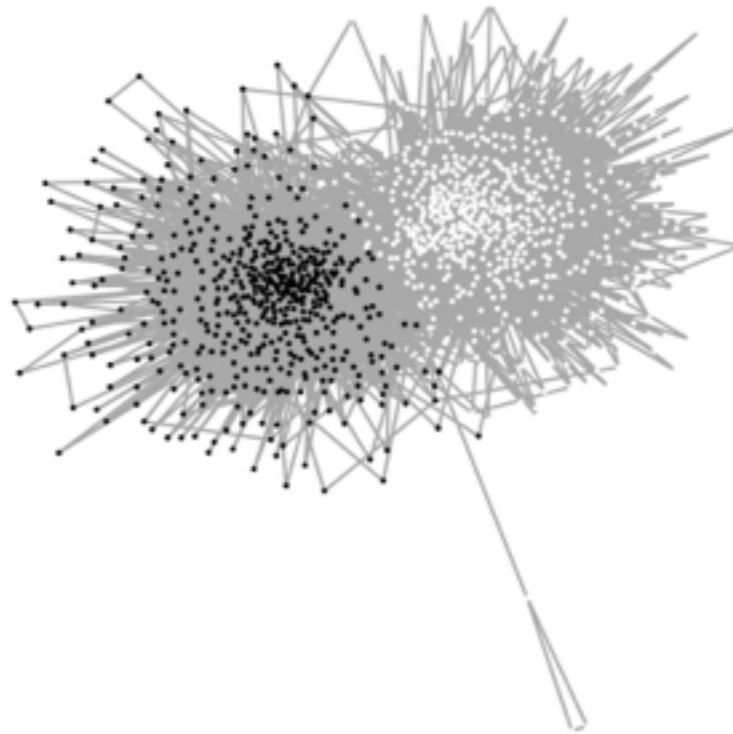
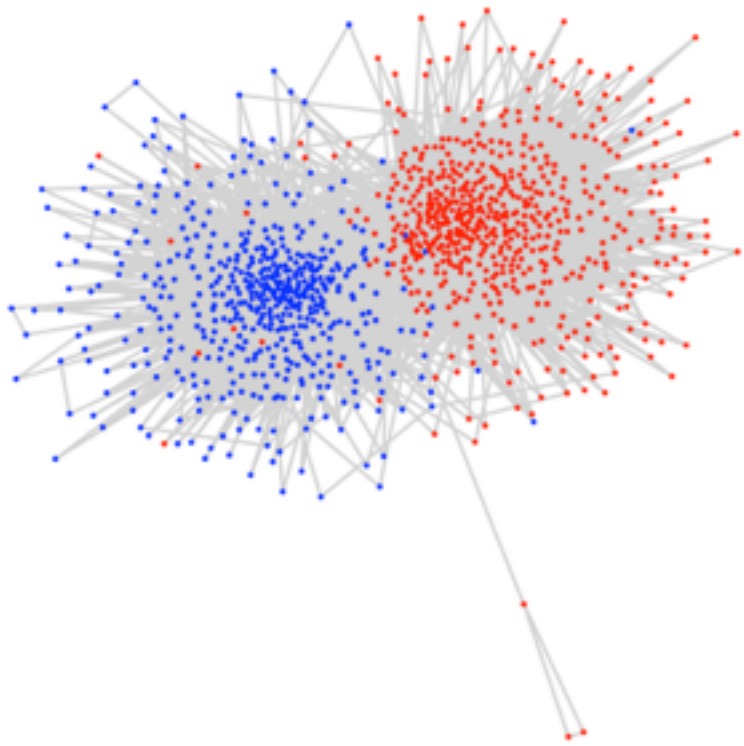


1084 blogs
edges are hyperlinks
avg degree 30
Half support Kerry
Half support Bush

Political Blogs

$\lambda_2 = 0.89$

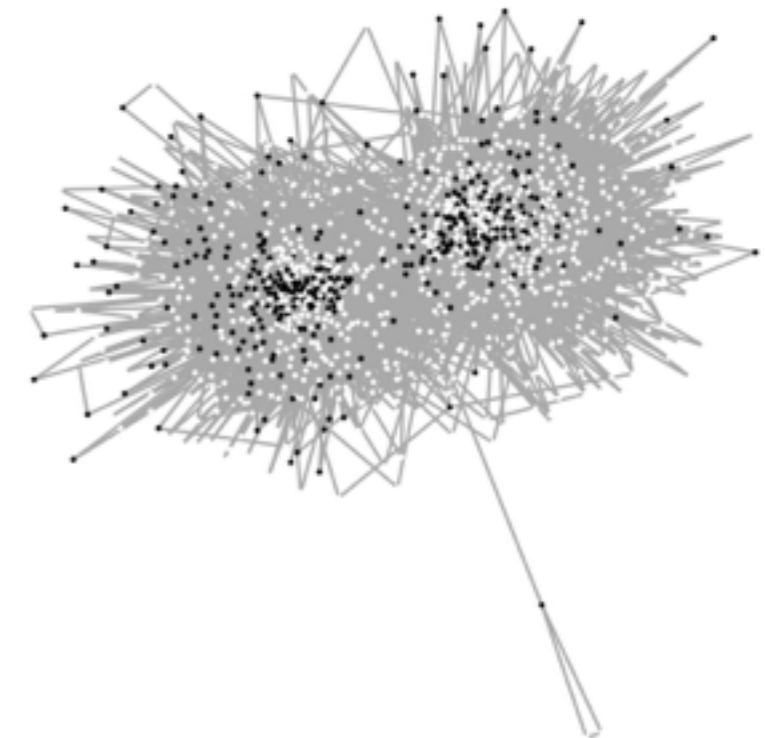
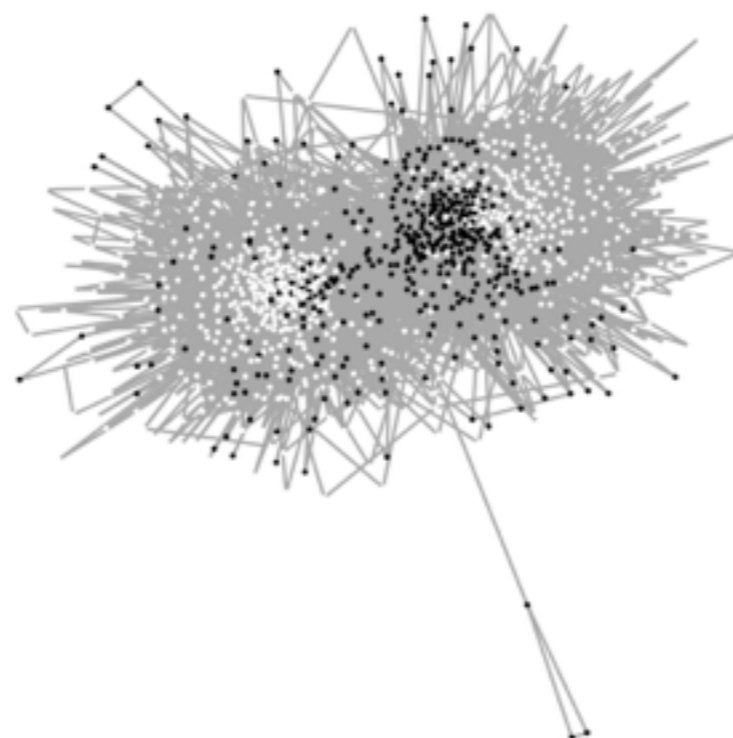
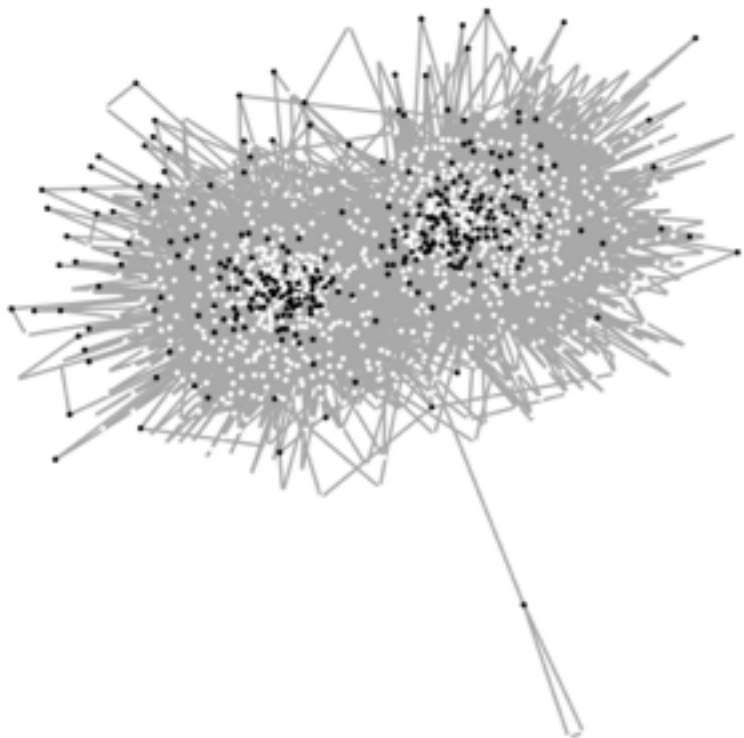
$\lambda_3 = 0.89$



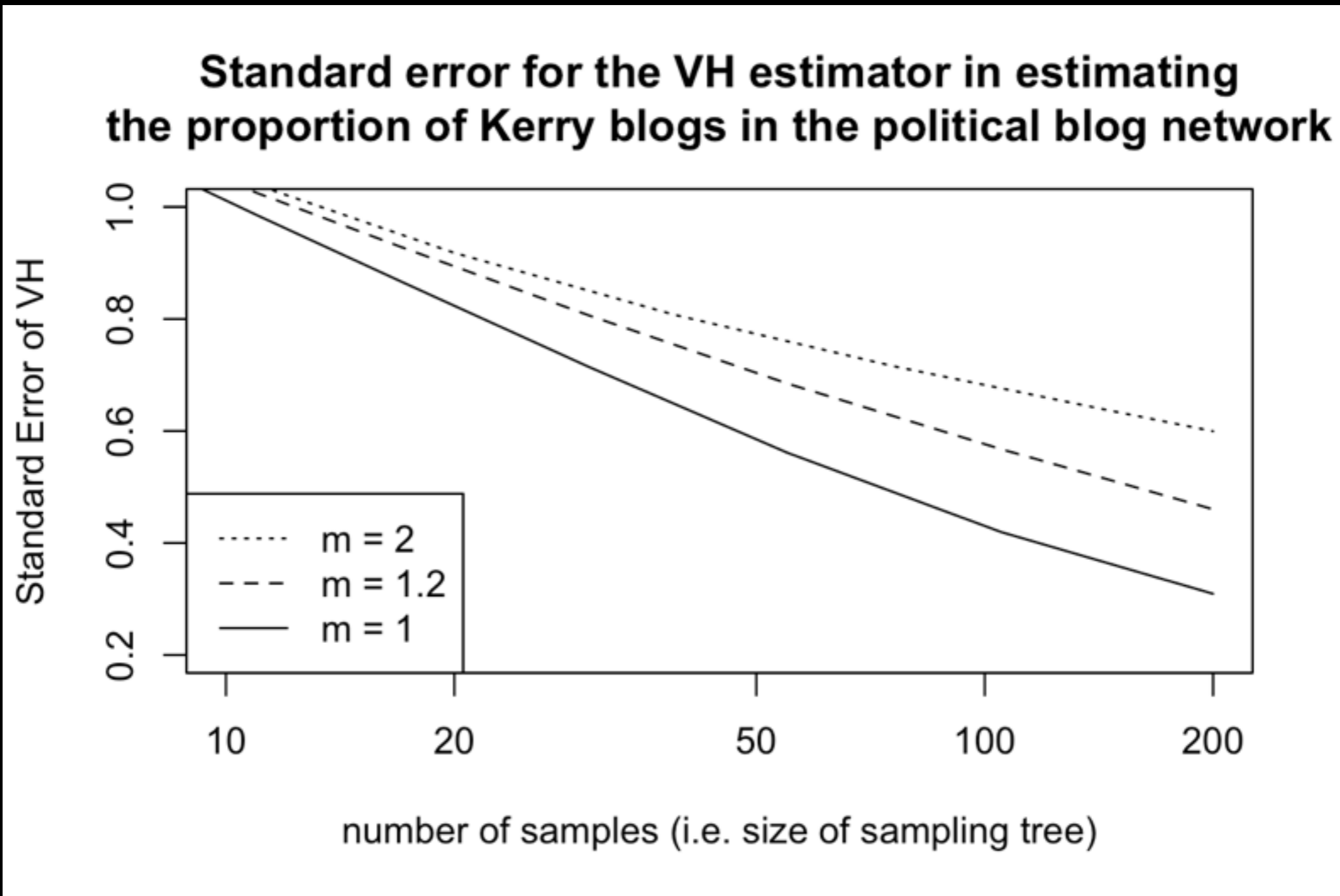
$\lambda_4 = 0.55$

$\lambda_5 = 0.53$

$\lambda_6 = 0.53$

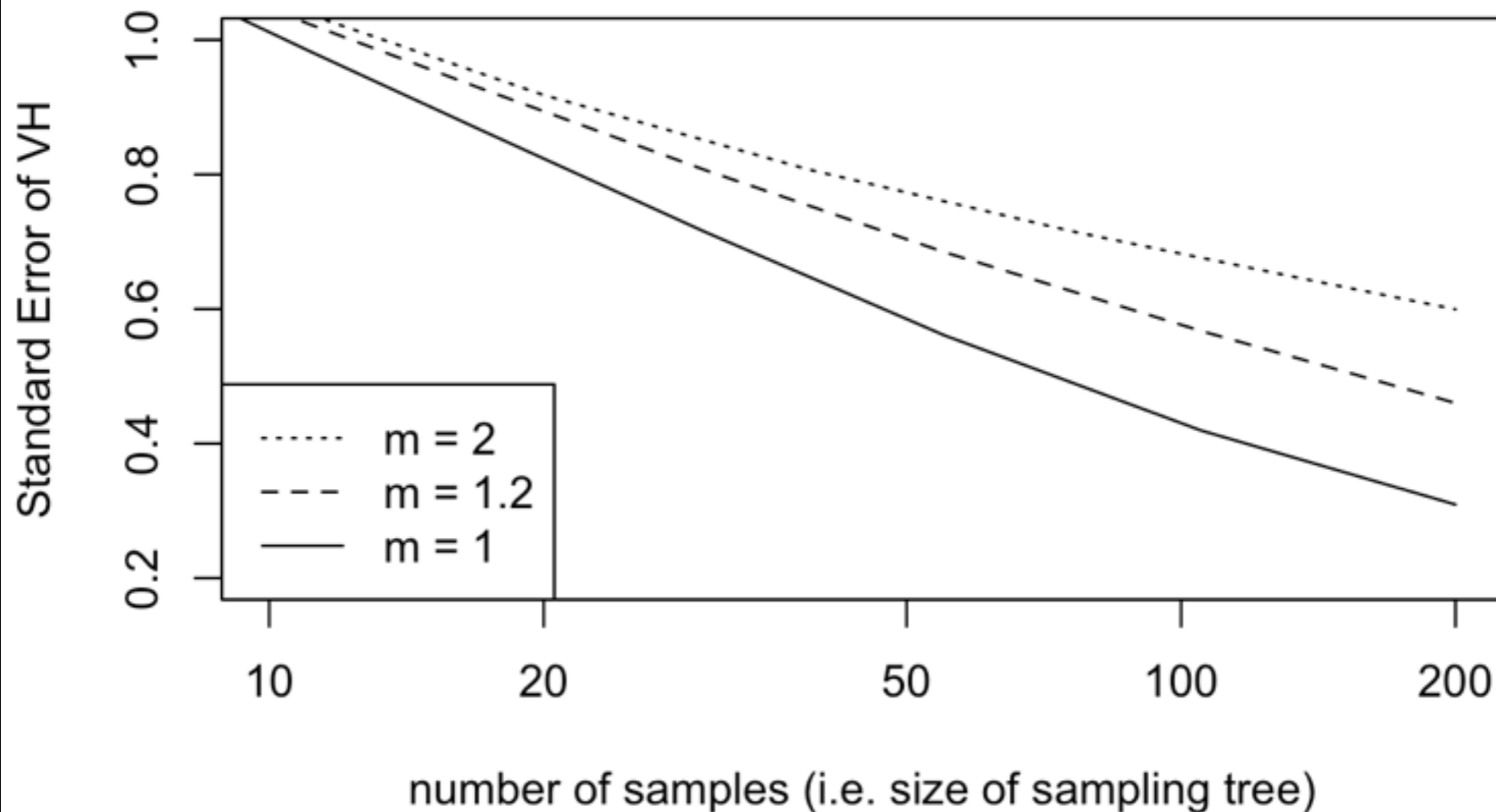


The standard error is large because the bottleneck aligns with the outcome of interest.



Bushy tree creates much larger standard errors.

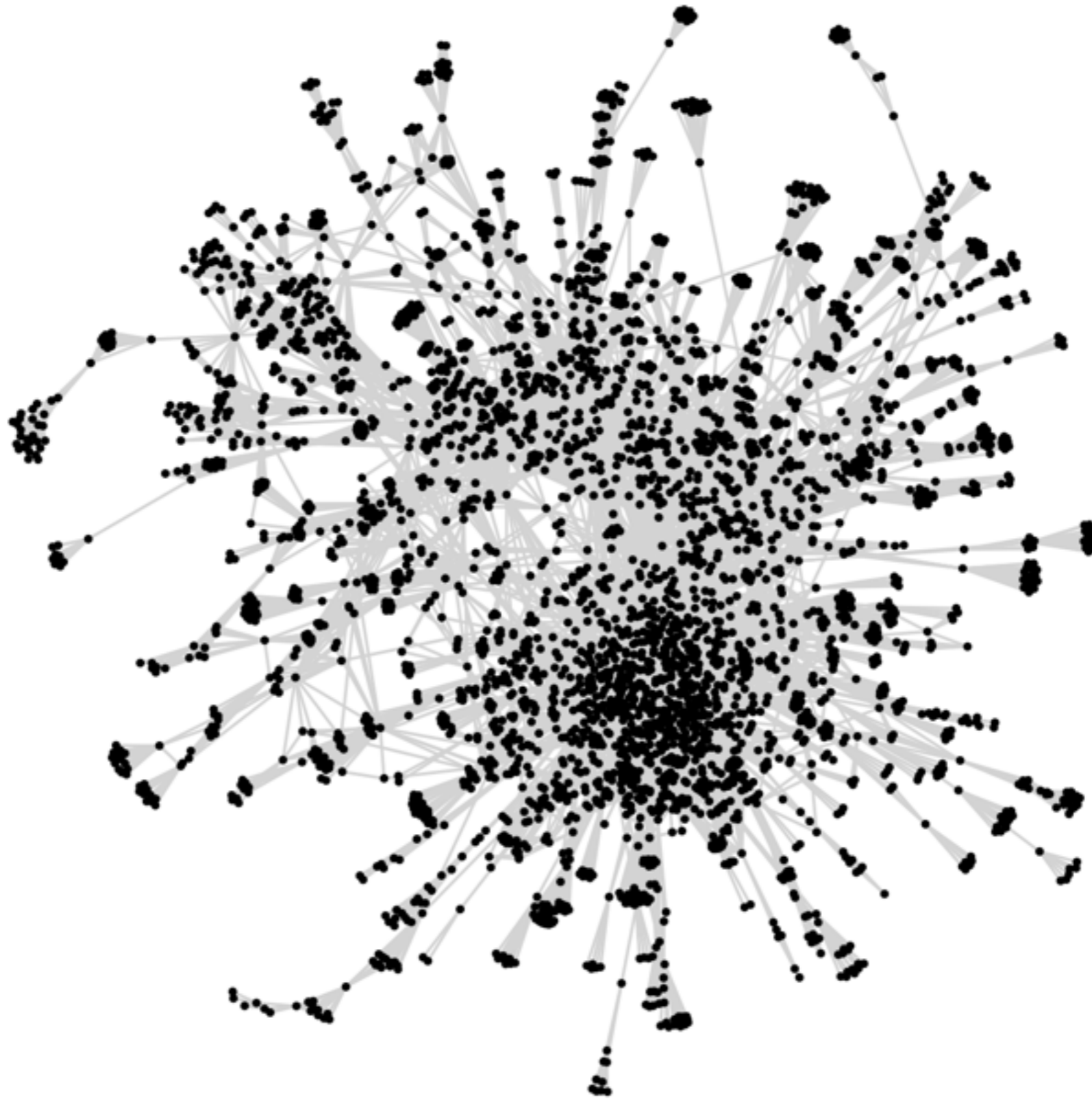
Standard error for the VH estimator in estimating the proportion of Kerry blogs in the political blog network



Project 90

- “Fully observed” network on at risk and marginalized population.
- CDC funded census of heterosexuals at risk for HIV transmission & living around Fort Collins, CO in ~1990.
- 2-core of largest connected component
 - $N = 3615$ people, mean degree = 9.7

**Project 90 collected the network of heterosexuals
at risk for HIV in Colorado Springs circa 1990**

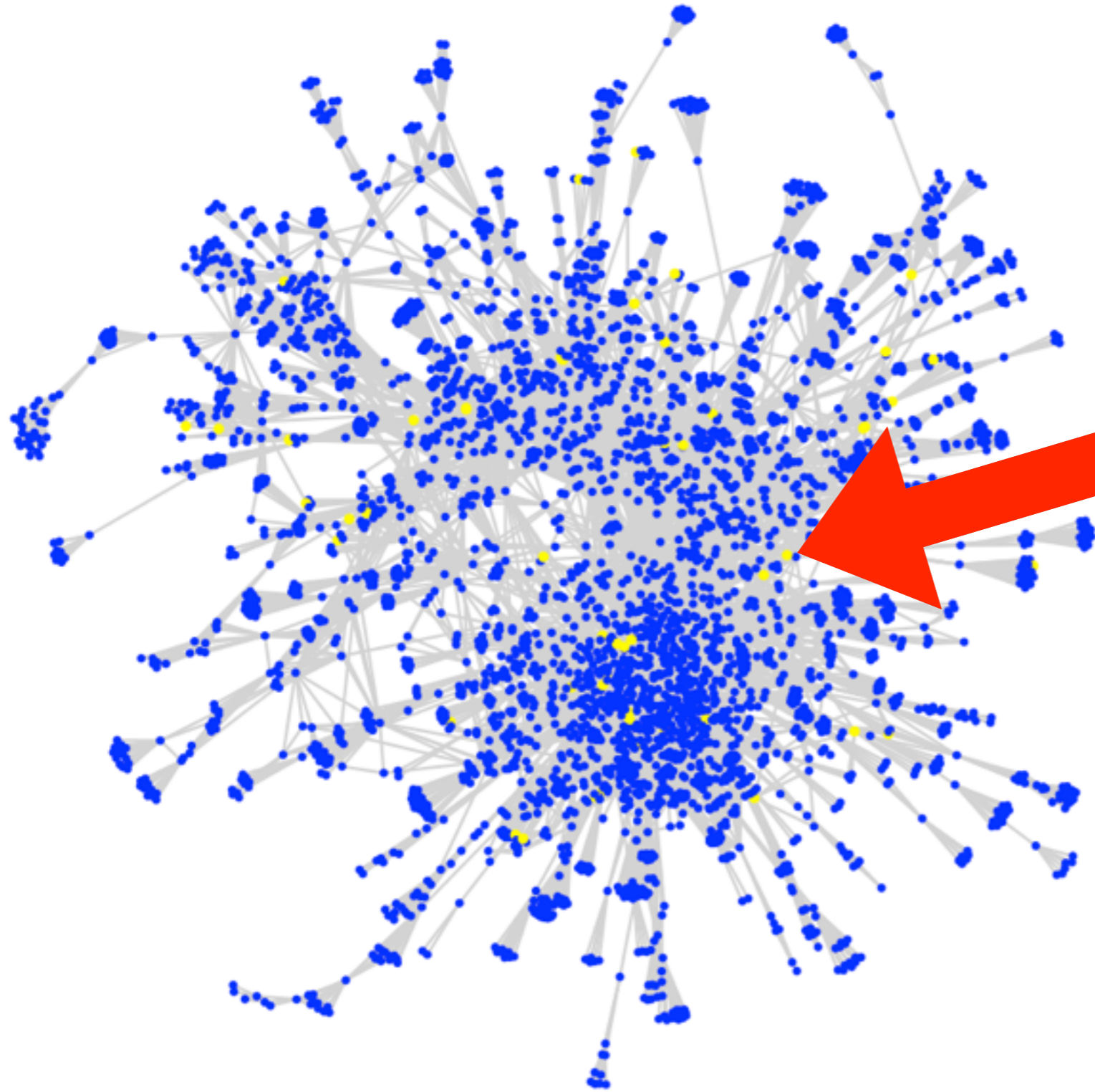


Data includes several
covariates on the nodes.

gender
sex.worker
pimp
sex.work.client
drug.dealer
drug.cook
thief
retired
housewife

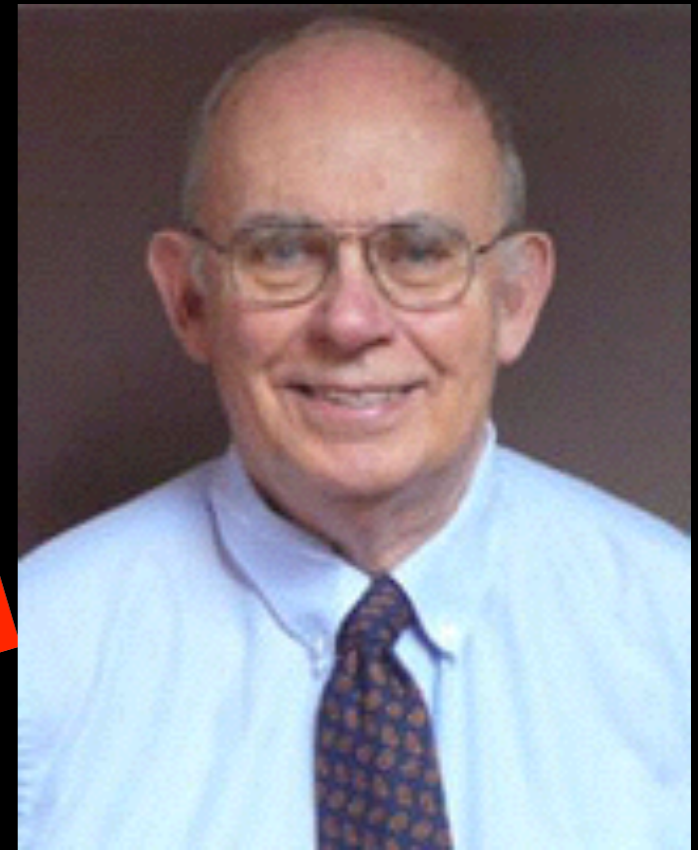
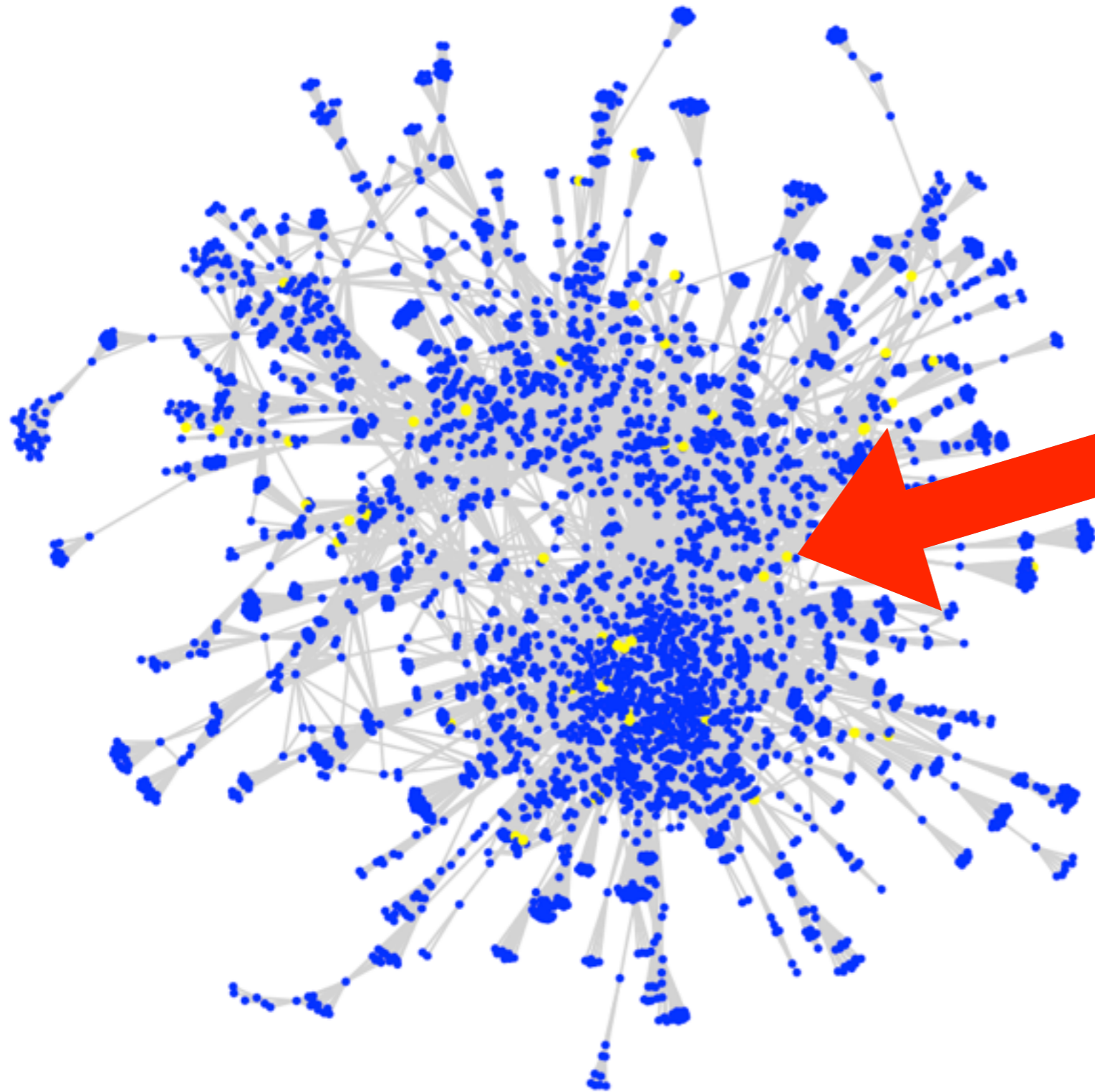
disabled
unemployed
homeless
Native American
Black
Asian/Pacific Islander
White
Other

homeless



*Breaches in data
privacy have
allowed the
identification of some
individuals!*

homeless

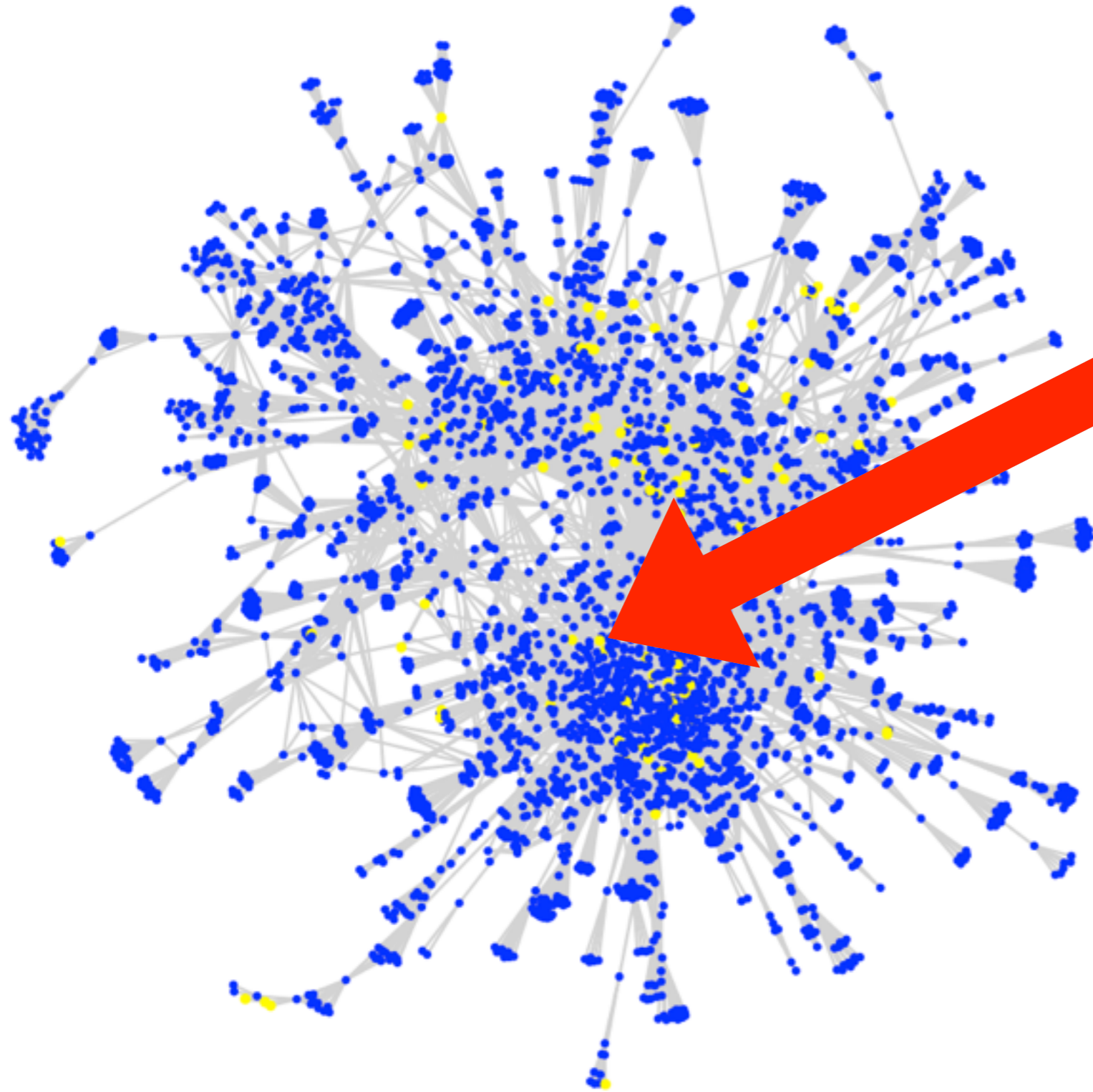


Jim?!

thief



thief

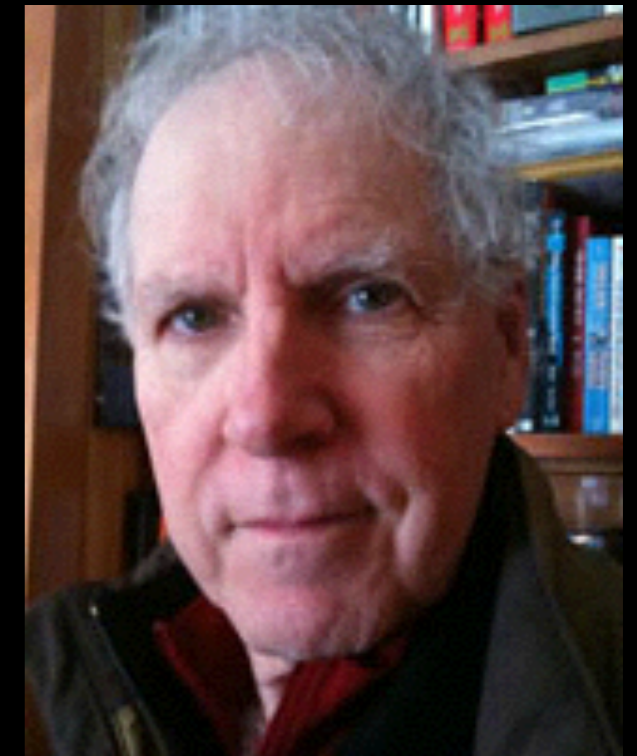
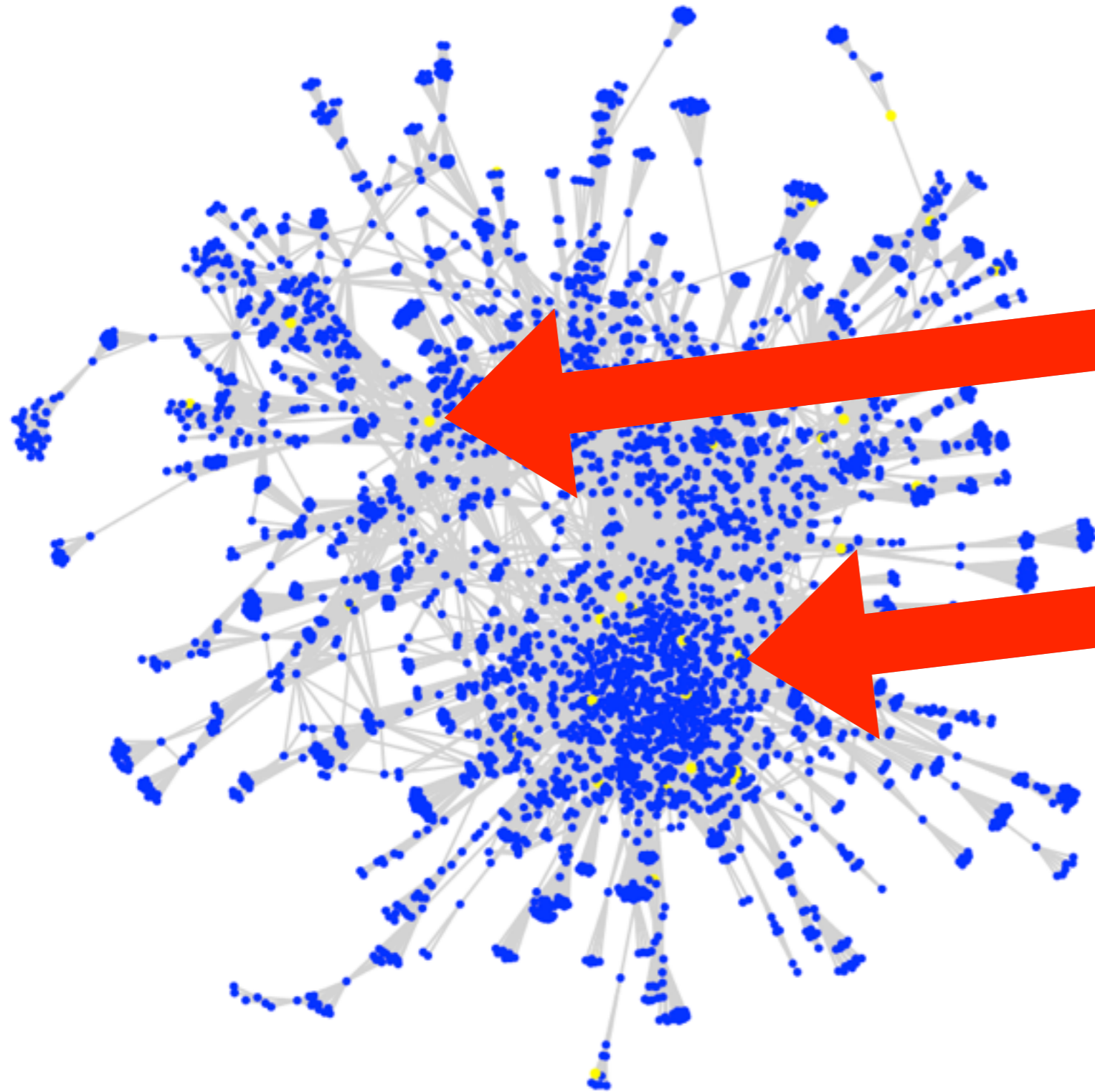


Gov Engler?!

drug.cook

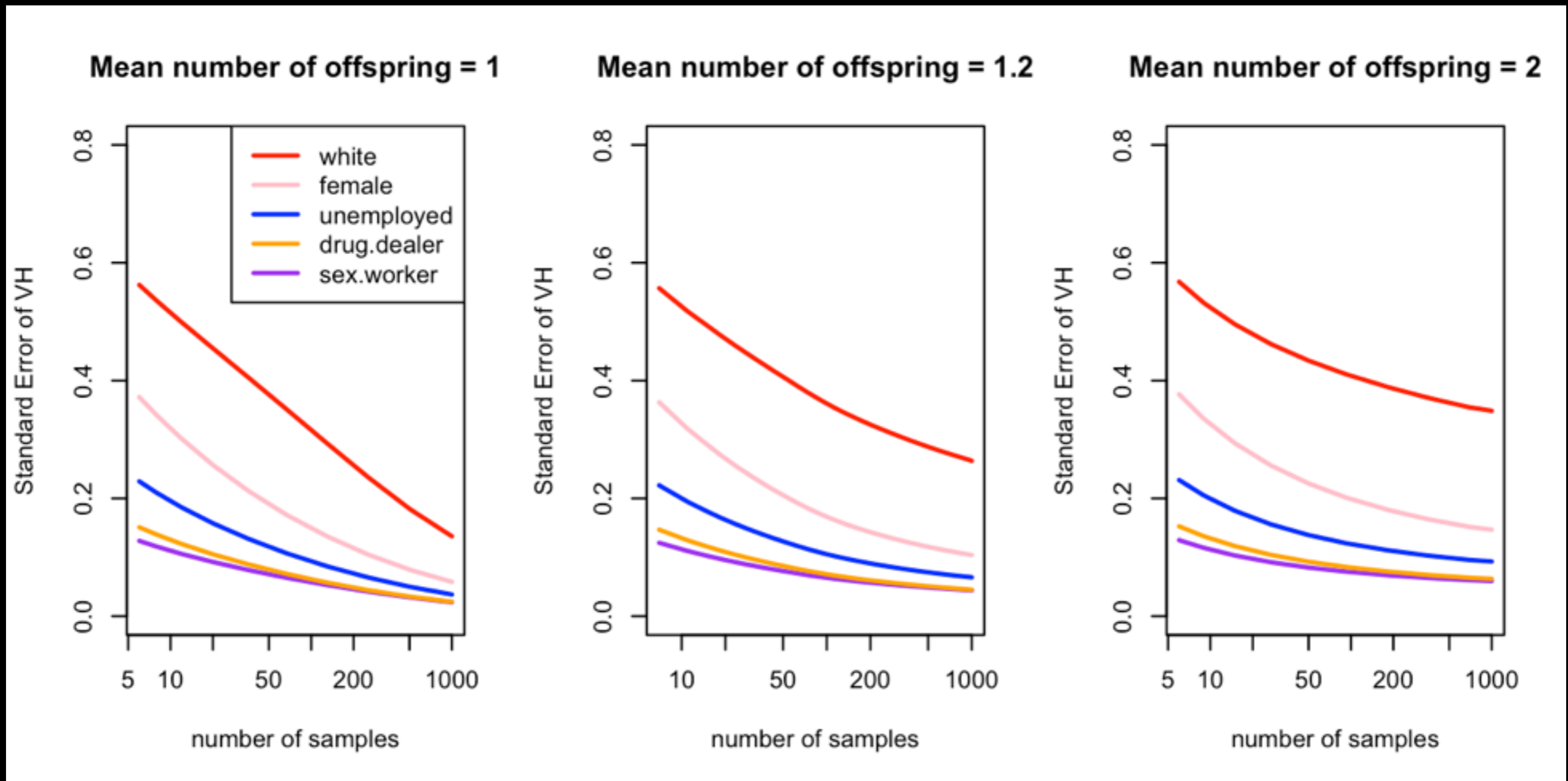


drug.cook

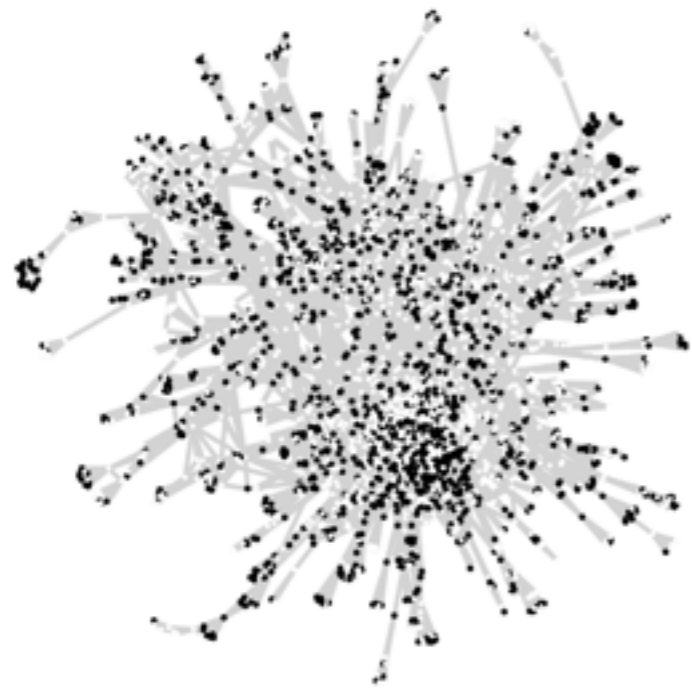


Hira and Raoul??

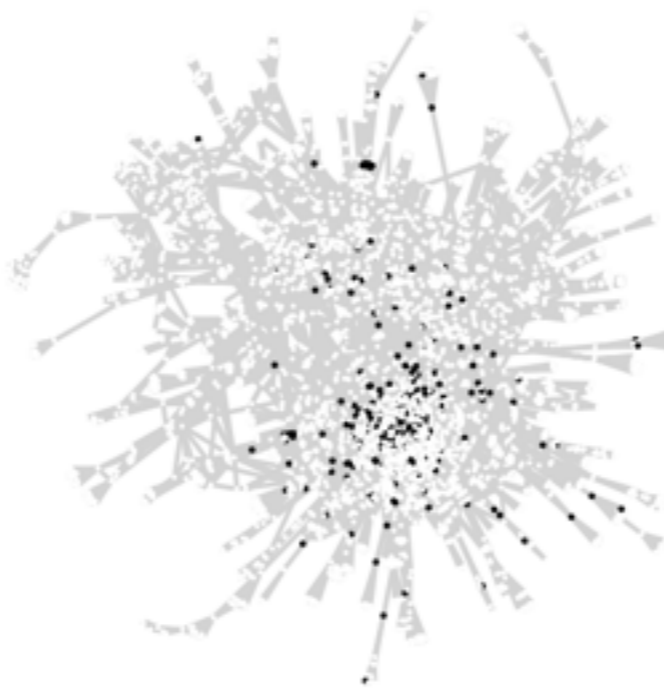
Different covariates can have drastically different standard errors.



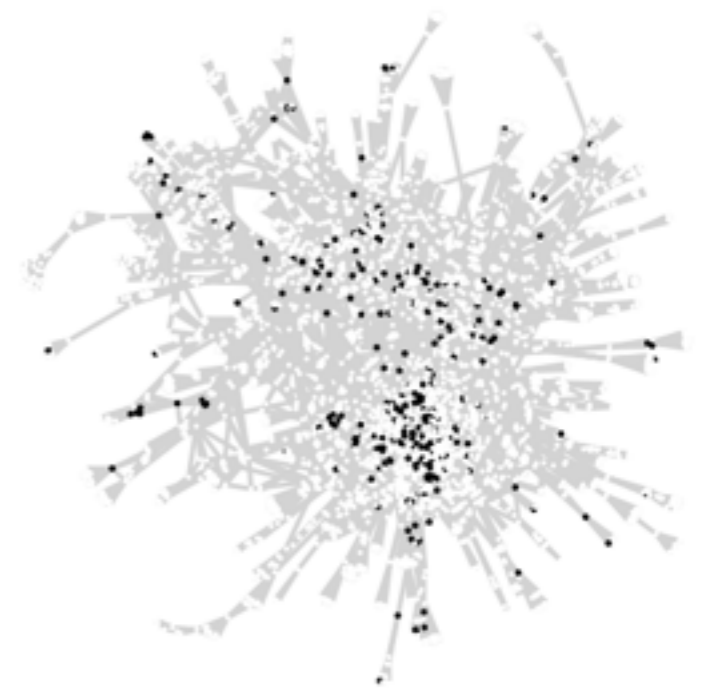
female



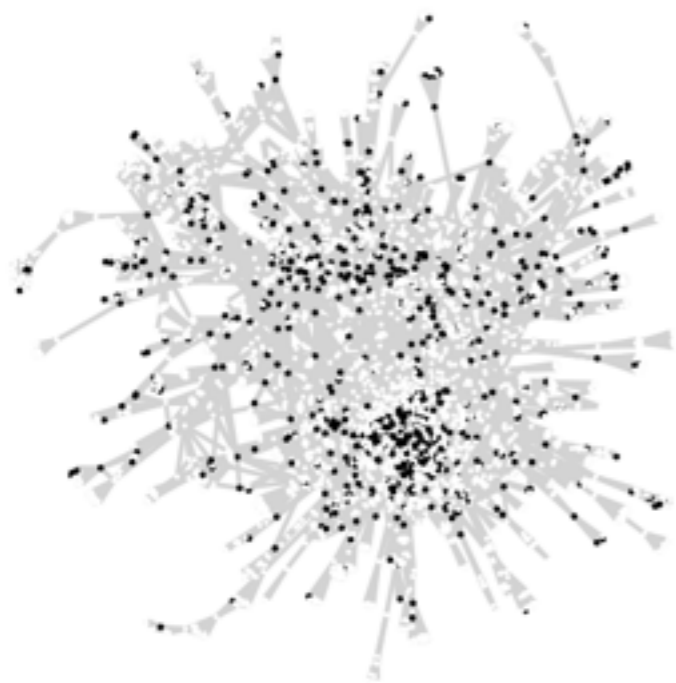
sex.worker



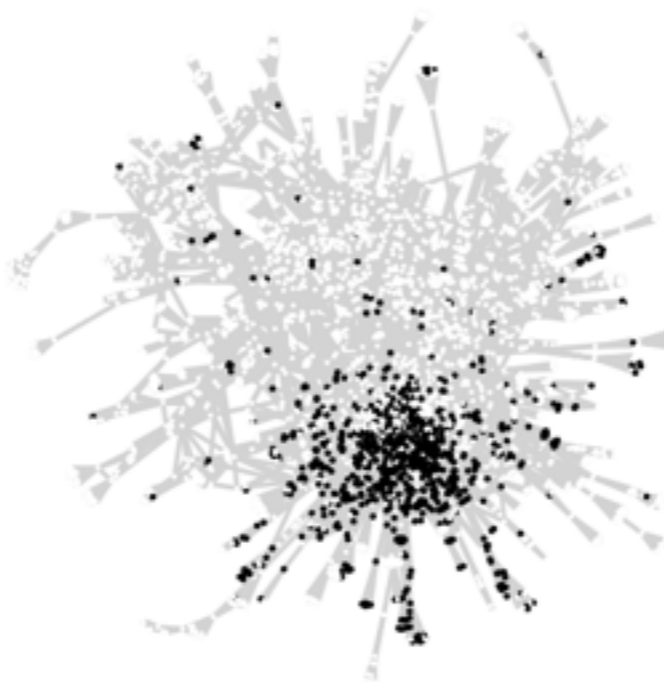
drug.dealer



unemployed



black



random



Outline

I. Model and notation.

Network, Markov transitions, sampling tree,
node features.

II. Key mathematical pieces.

eigenvectors of P

The G function

III. The true sampling variance

A. A scary story, the critical threshold

IV. Designed RDS

For standard rates, we need the tree to grow in a way that G is order $1/n$.

If y is correlated with f_2 , then under a certain technical condition,

$$c \mathbb{G}_n(\lambda_2) + o(\mathbb{G}_n(\lambda_2)) \leq \text{Var}_{RDS}(\hat{\mu}) \leq c \mathbb{G}_n(\lambda_2)$$

Design effect

How much larger is an RDS confidence interval compared to an SRS confidence interval using the same number of samples?

If $y = f_2$, then

$$DE = \frac{\text{Var}_{RDS}(\hat{\mu})}{\text{Var}_{SRS}(\hat{\mu})} = \frac{\mathbb{G}(\lambda_2)}{1/n} = n\mathbb{G}(\lambda_2)$$

Because G characterizes the rates of convergence, we would like upper and lower bounds on G that depend on the sample size n (i.e. the number of nodes in the referral tree)

Under an m -tree, there is an easy lower bound.

The height of the m -tree is bounded:

$$h(\mathbb{T}) \leq \log_m n$$

$$\mathbb{G}(\lambda_2) = \mathbb{E}\lambda_2^D \geq \lambda_2^{2h(\mathbb{T})} \geq \lambda_2^{2\log_m n} = n^{-\log_m 1/\lambda_2^2}$$

Under an m -tree
(i.e. everyone refers m people),
there is an easy lower bound.

The height of the m -tree is bounded:

$$h(\mathbb{T}) \leq \log_m n$$

$$\mathbb{G}(\lambda_2) = \mathbb{E}\lambda_2^D \geq \lambda_2^{2h(\mathbb{T})} \geq \lambda_2^{2 \log_m n} = n^{-\log_m 1/\lambda_2^2}$$

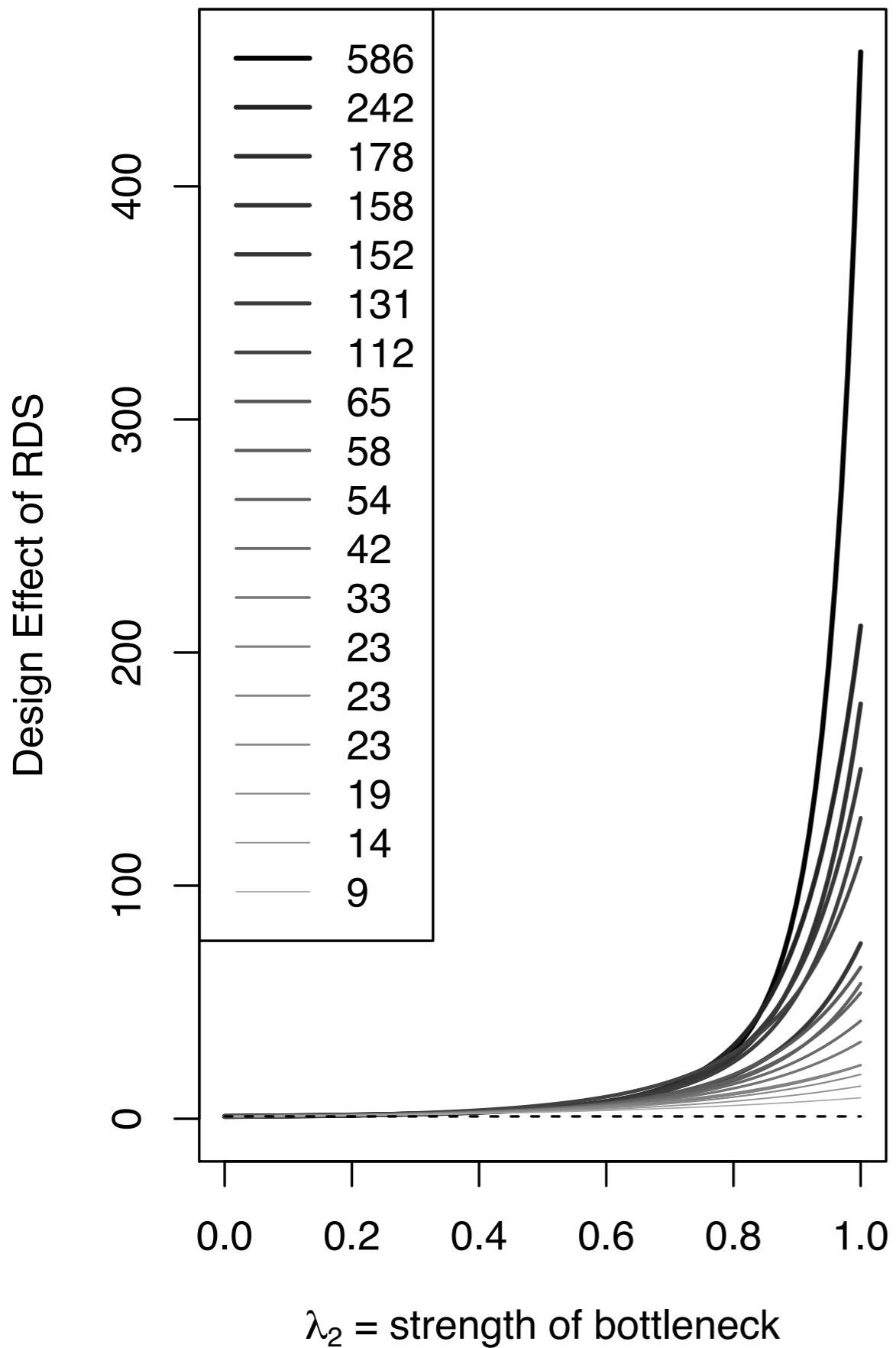
This fails to converge
at the desired rate if

$$m > 1/\lambda_2^2$$

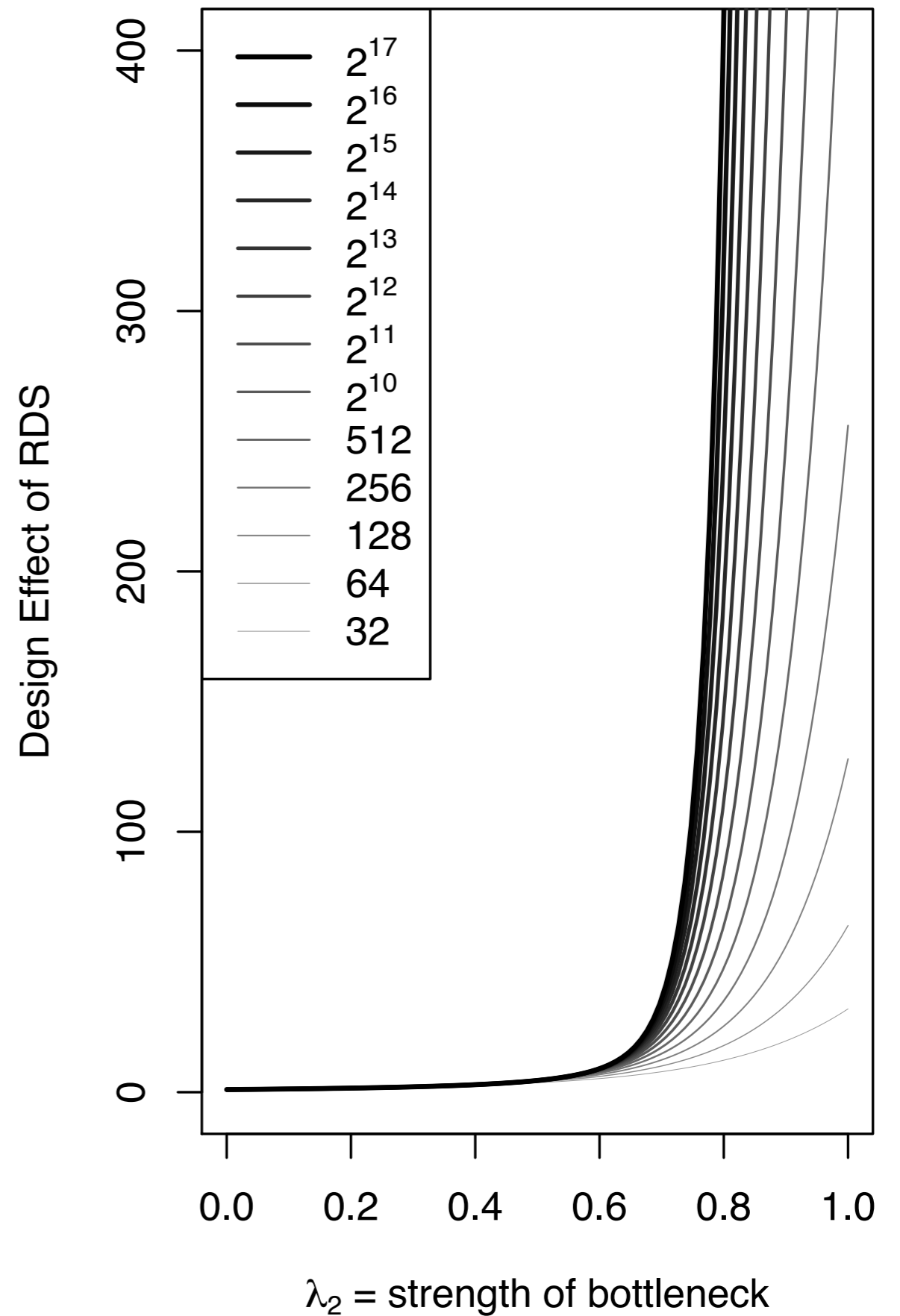
Threshold:	$m < 1/\lambda_2^2$	$m > 1/\lambda_2^2$
standard error:	$n^{-1/2}$	$n^{\log_m \lambda_2}$

- Lower bounds hold when tree is Galton Watson under the $N \log N$ assumption. $m =$ expected # referrals.
- Upper bounds require more work and an additional assumption.
 - gives a matching threshold. rate matches (up to log terms).
 - Galton Watson trees satisfy this additional assumption under a bounded fourth moment assumption.

Design effects for 18 different empirical referral trees



Design effects under the 2-tree



A recap of the scary story

$$\lambda_j: -1 < \lambda_j < 1$$

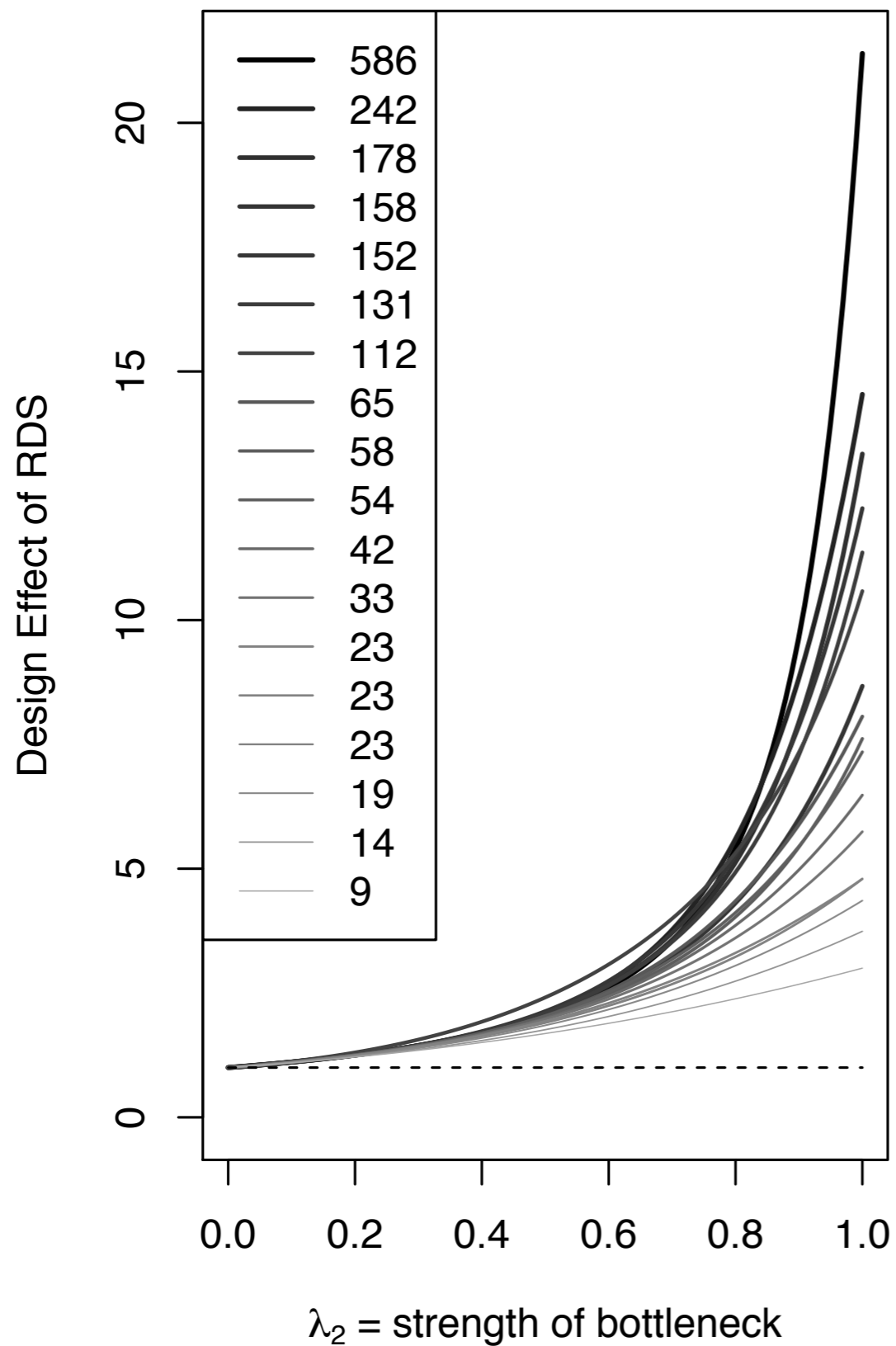
- Close to 1 when bottleneck is strong
- Values larger than 0.9 not uncommon

- Say $\lambda_j = 0.7 \Rightarrow \frac{1}{\lambda_j^2} \approx 2$

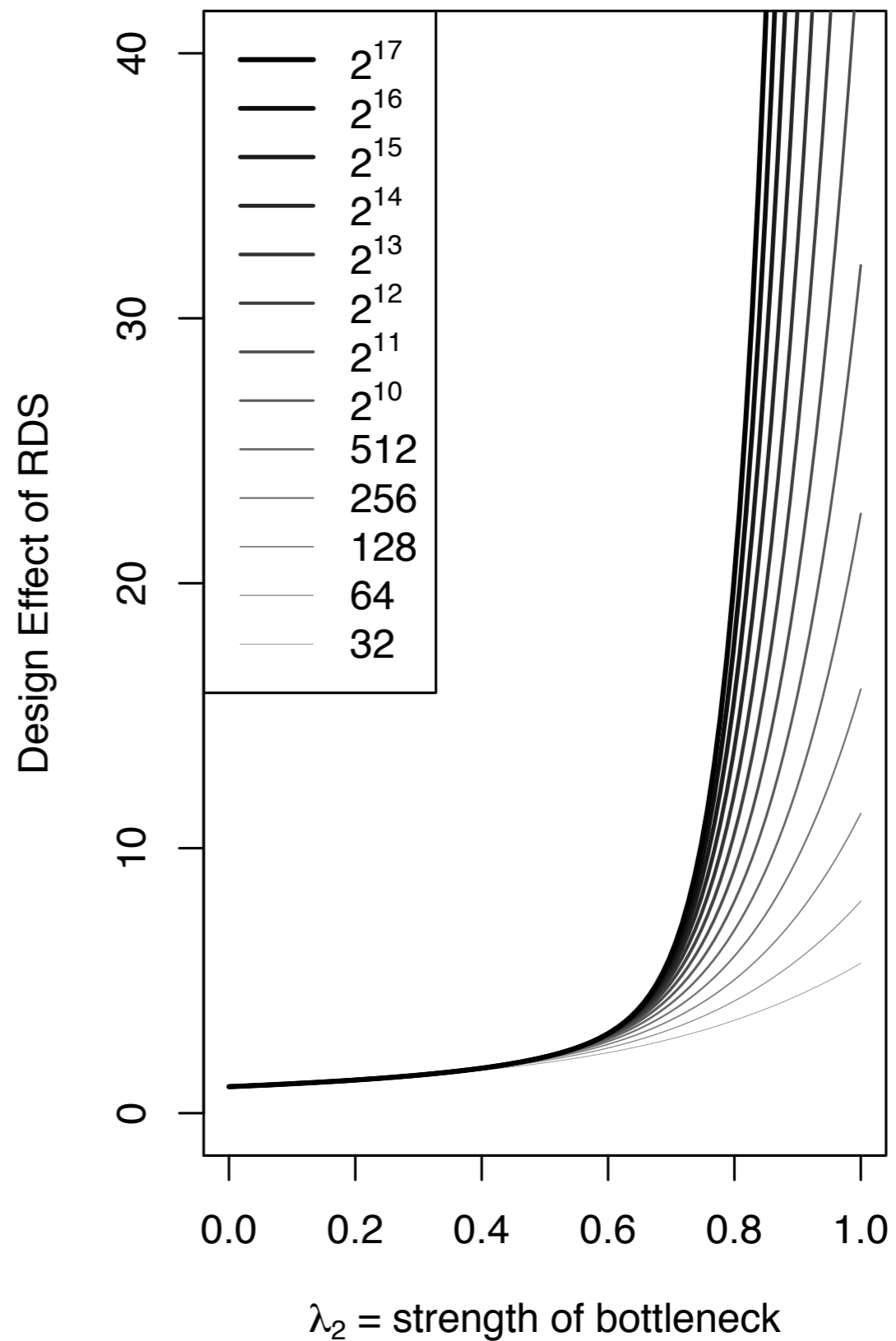
$$\lambda_j = 0.9 \Rightarrow \frac{1}{\lambda_j^2} \approx 1.23$$

- If the average person refers more people than this, then the DE grows with n .

Design effects for 18 different empirical referral trees



Design effects under the 2-tree



There is a fundamental conflict between obtaining “enough data” and an unbiased sample.

- Volz-Heckathorn is *asymptotically unbiased*.
- Want *longer* chain
- Chains often die.
- Need several referrals to prevent chain death.
- This gives you a bushy tree.
 - As you gain more samples, standard errors decrease, but design effect can grow.

Outline

I. Model and notation.

Network, Markov transitions, sampling tree,
node features.

II. Key mathematical pieces.

eigenvectors of P

The G function

III. The true sampling variance

A. A scary story

IV. Designed RDS

- We derived variance of VH estimator under the assumptions needed for unbiasedness & initialization from stationary distribution.
- The variance depends on
 - (1) the bottlenecks in the referral process,
 - (2) how y correlates with the bottlenecks,
 - (3) the tree structure.
- There is a fundamental conflict between obtaining “enough data” and an unbiased sample. If referral rate is too large, then design effect grows with n .

social driven
transition Colorado design Markov
bottleneck threshold Volz-Heckathorn
correlation random Galton-Watson
Variance **network**
graph political ^{sex} blogs **Sampling**
spectral tree eigenvectors
Theorem respondent
snowball
friends

Bottlenecks prevent accurate estimates

- If the bottleneck is too strong, you get a growing design effect!

Designed RDS

- Joint research with Mohammad Khabbazian, Zoe Russek, and Bret Hanlon.

Designed RDS

- Seeks to minimize bottlenecks
- Classically, a sample design is defined as a way of assigning a probability to everyone in the population
- Designed RDS seeks referrals that cross bottlenecks

How do you do this?

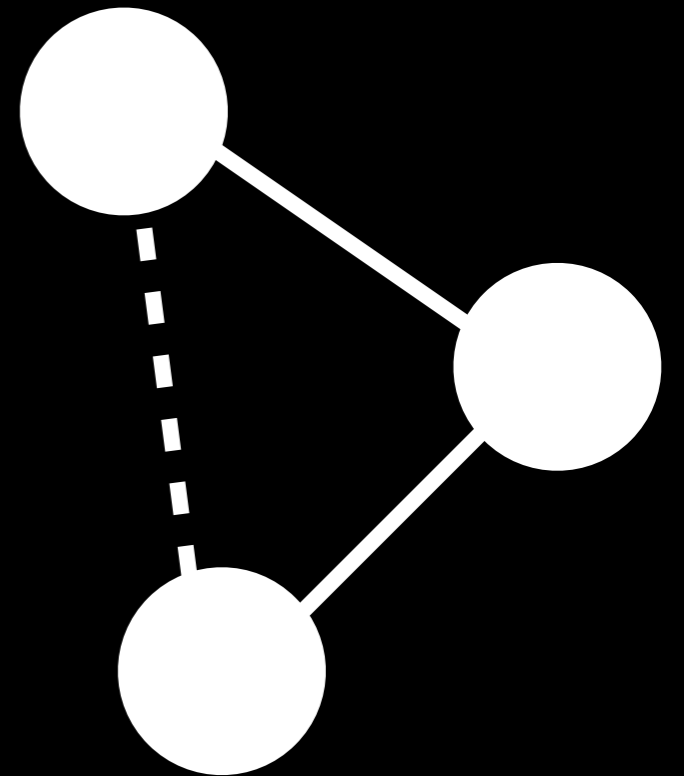
- You could ask for it,
 - “Please refer someone from neighborhood x”
- Problems with this:
 - Need to know which bottlenecks are important
 - Hard to model as random

Anti-cluster sampling

“Please refer two people who don’t know each other”

and/or

“Please refer someone who doesn’t know the person who referred you”



**Disclaimer:
This is work
in progress.**



Anti-cluster sampling

- This can be modeled as a new Markov transition matrix P for the same graph G
- Random walk is “choose a friend uniformly at random”
- Anti-cluster can be modeled as “from your set of friends, find all pairs that don’t know each other and select one pair uniformly at random.”
- These sampling probabilities can be computed with matrix multiplication.

Under a “balanced” Stochastic Blockmodel with “within-block” probabilities larger than “out-of-block” probabilities, AC-RDS has smaller bottlenecks.

Lemma 5 (Spectral gap of the population graph). Let $\mathcal{A} := E[A] = ZBZ^T$ under the stochastic block model with k blocks of equal sizes. Let $B_{ii} = p$ and $B_{ij} = q$ for $i \neq j$. If $0 < q < p < 1$, then

$$0 < \lambda_2(\mathcal{P}^{AC}) < \lambda_2(\mathcal{P}^{RW}) < 1.$$

Work in progress

- Our current theory presumes that the underlying network is a stochastic block model
- Does it work in practice?
 - Need experiments!

- We derived variance of VH estimator under the assumptions needed for unbiasedness & initialization from stationary distribution.
- The variance depends on (1) the tree structure and (2) how y correlates with the bottlenecks.
- There is a fundamental conflict between obtaining “enough data” and an unbiased sample. If referral rate is too large, then design effect grows with n .

social driven
transition Colorado design Markov
bottleneck threshold Volz-Heckathorn
correlation random Galton-Watson
Variance **network**
graph political ^{sex} blogs **Sampling**
spectral tree eigenvectors
Theorem respondent
snowball
friends