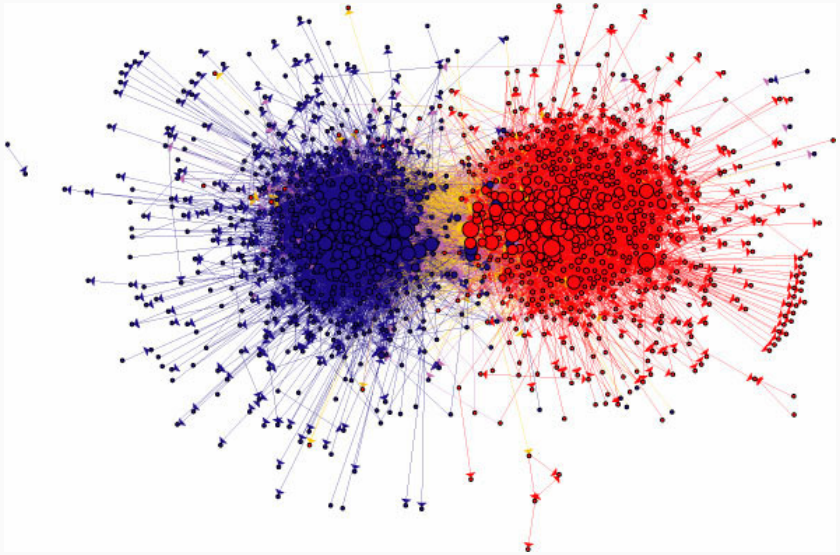
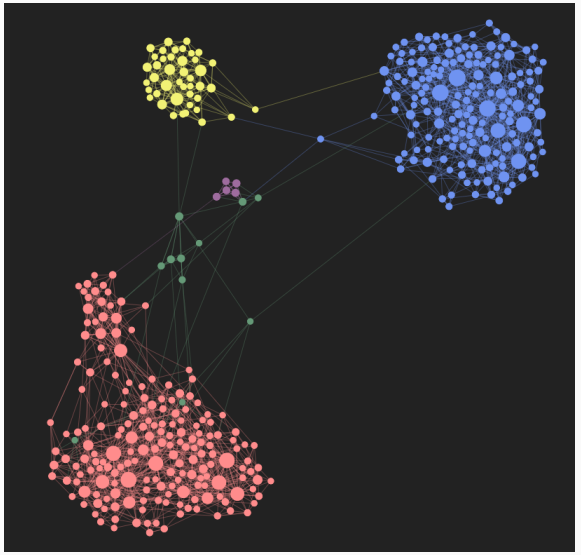




Modeling node popularity in networks with community structure

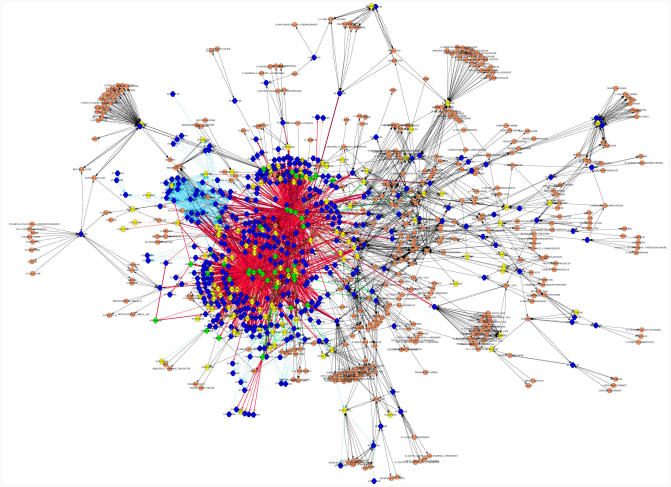
Srijan Sengupta
Assistant Professor of Statistics
Virginia Tech

Why study networks

- (1990-): Barabási, Newman, Watts...
- 100s of nodes $\rightarrow$ millions, sparse, power law
- Computer science, biology, internet, social media

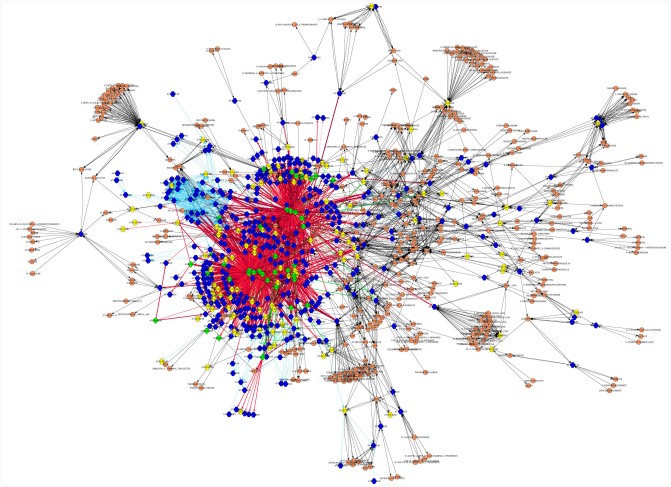
Why study networks

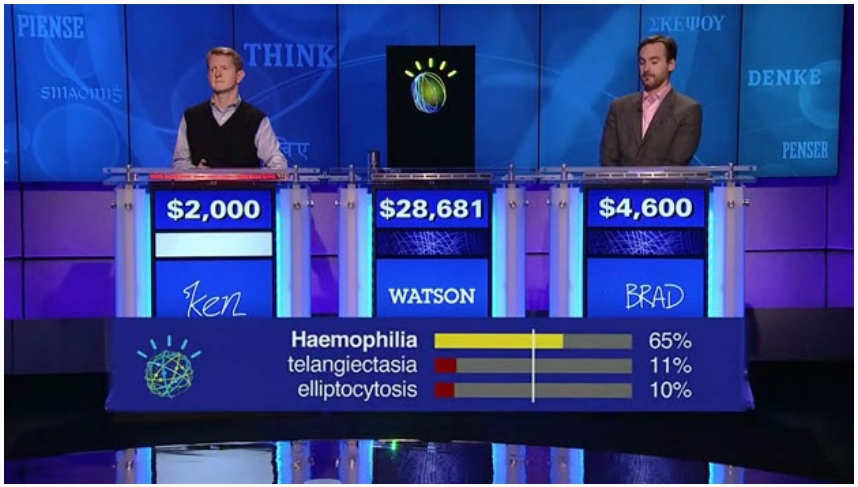
- (1990-): Barabási, Newman, Watts...
- 100s of nodes $\rightarrow$ millions, sparse, power law
- Computer science, biology, internet, social media



Why study networks

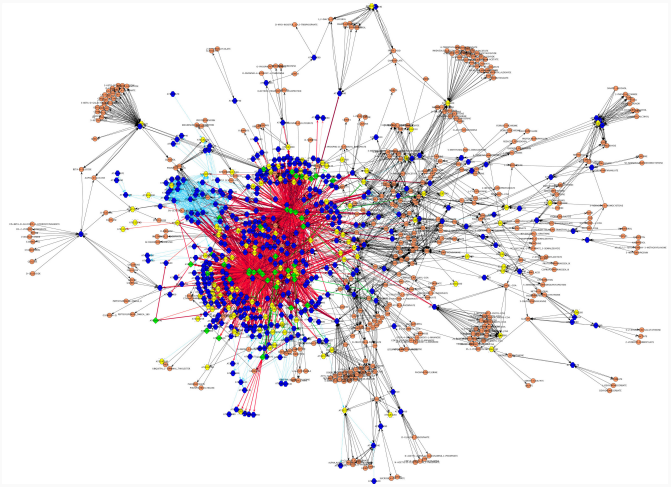
- (1990-): Barabási, Newman, Watts...
- 100s of nodes $\rightarrow$ millions, sparse, power law
- Computer science, biology, internet, social media

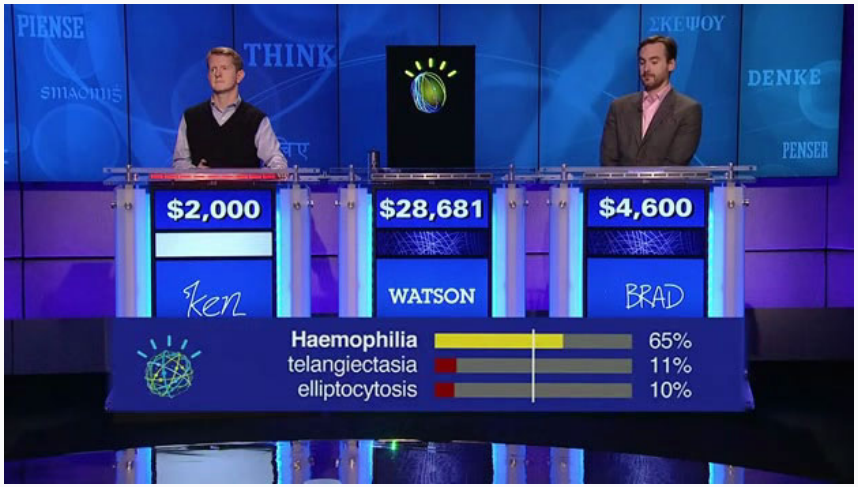


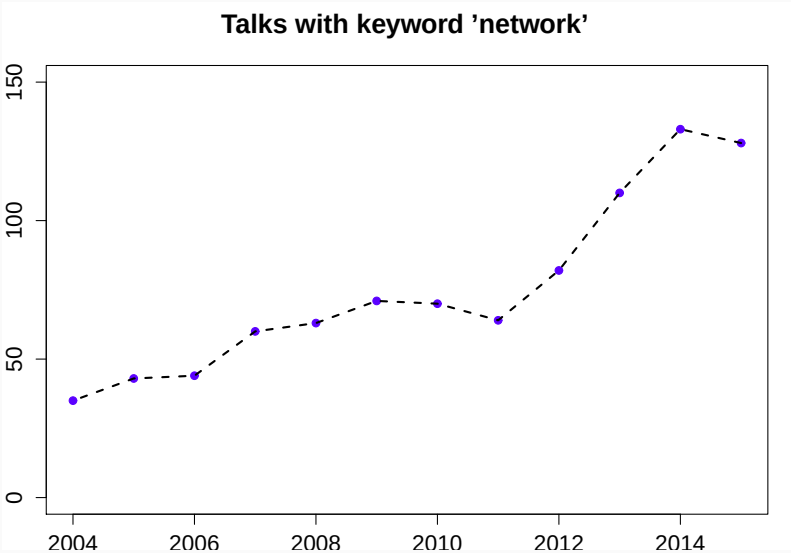


Why study networks

- (1990-): Barabási, Newman, Watts...
- 100s of nodes $\rightarrow$ millions, sparse, power law
- Computer science, biology, internet, social media

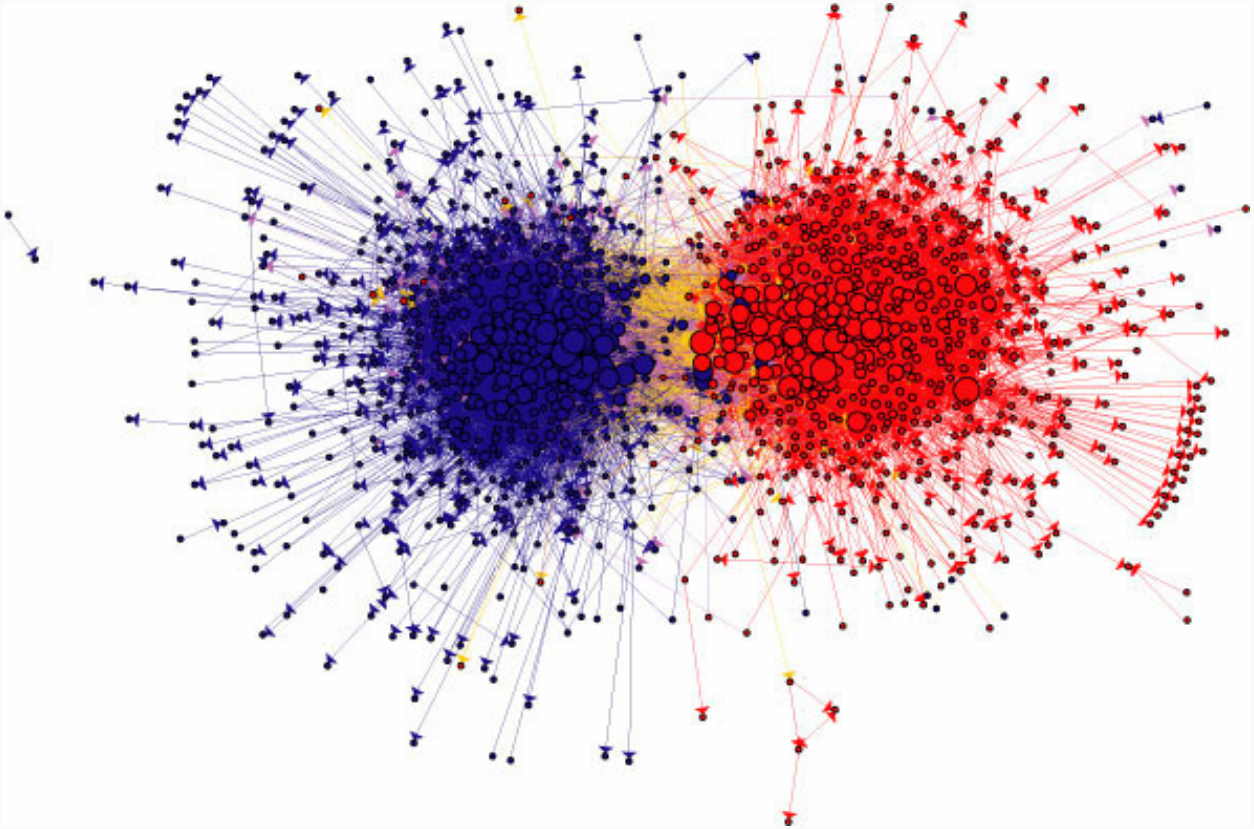






Community structure in networks

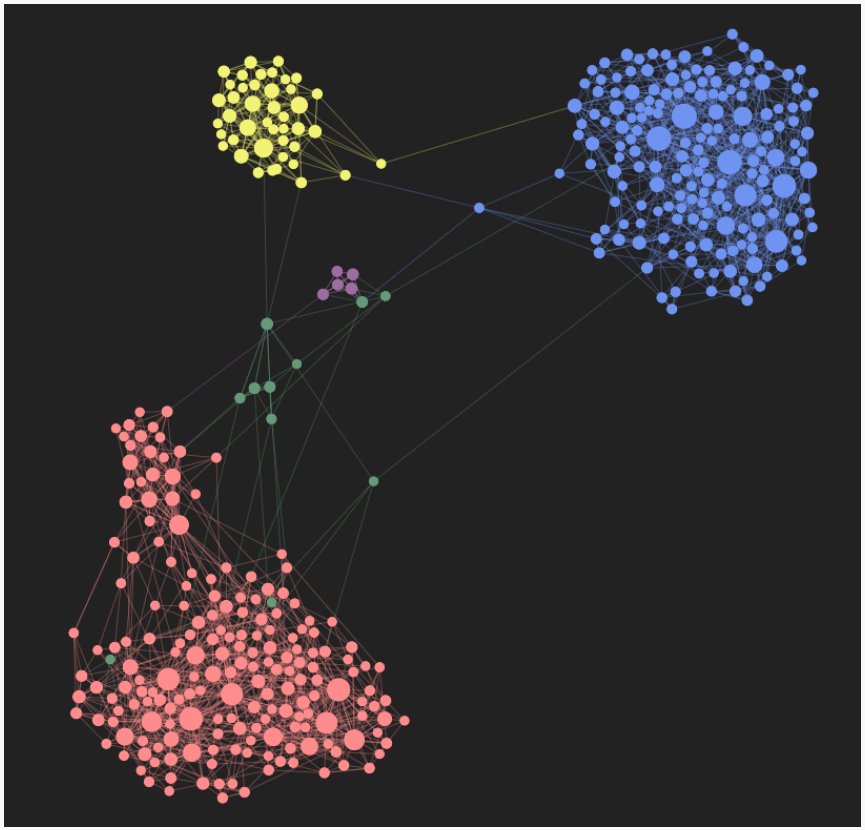
- Communities often correspond to important node features
- Community detection: discover communities from networks



Political blogs network (Adamic and Glance, 2005)

Community structure in networks

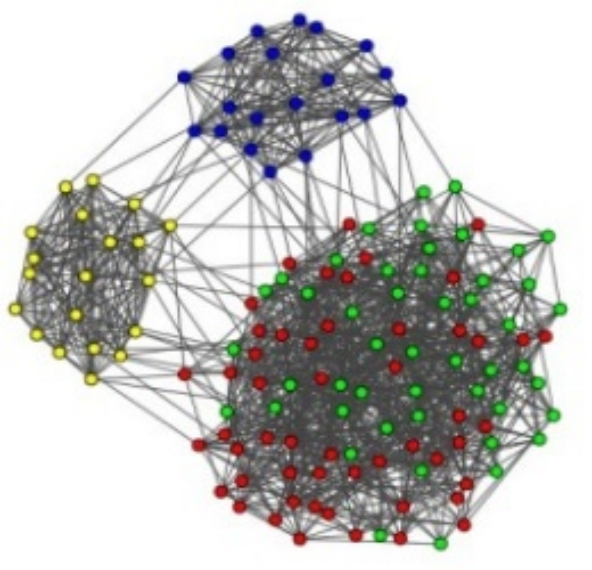
- Communities often correspond to important node features
- Community detection: discover communities from networks



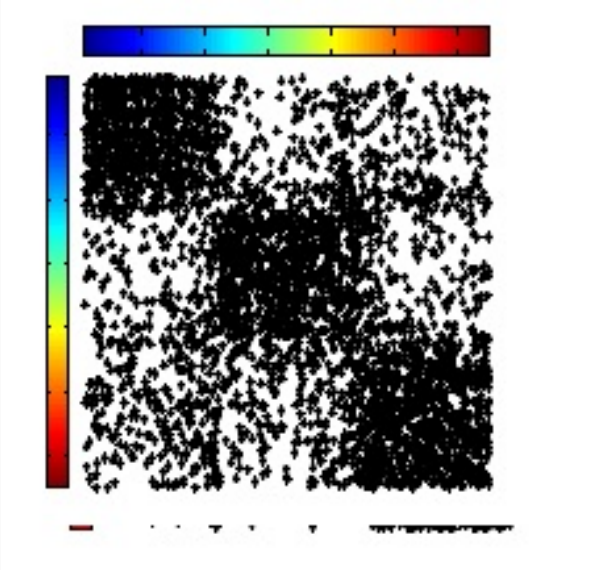
British MP Twitter network (Greene and Cunningham, 2013)

Community structure: statistical models

Network communities



Blockmodel approach



- Let $A_{n \times n}$ be the (binary, symmetric) adjacency matrix.
- $P = \mathbb{E}[A]$ has a **block matrix structure**.
- A two-step model-fitting process:

1. Community Detection (which vertex goes to which block)
2. Estimating other parameters of the statistical model

Stochastic Blockmodel (SBM)

For a K-block network, let $\boldsymbol{\omega}_{K\times K}$ be the (symmetric) matrix of block-block probabilities.

Let $\mathbf{c}$ denote the membership vector with $\mathbf{c}_i = r$ if the i^{th} node belongs to the r^{th} community.

Then for $i < j$

$$P_{ij} = \boldsymbol{\omega}_{\mathbf{c}_i\mathbf{c}_j}.$$

- Community structure: All nodes belonging to a community are *stochastically equivalent*.
- Expected degree is **identical** for all nodes in a community.

Stochastic Blockmodel (SBM)

For a K-block network, let $\boldsymbol{\omega}_{K\times K}$ be the (symmetric) matrix of block-block probabilities.

Let $\mathbf{c}$ denote the membership vector with $\mathbf{c}_i = r$ if the i^{th} node belongs to the r^{th} community.

Then for $i < j$

$$P_{ij} = \boldsymbol{\omega}_{\mathbf{c}_i \mathbf{c}_j}.$$

- Community structure: All nodes belonging to a community are *stochastically equivalent*.
- Expected degree is **identical** for all nodes in a community.

Degree Corrected Blockmodel (DCBM)

- Adds degree scaling parameters θ_i for each node to allow for a broad degree distribution.

$$P_{ij} = \boldsymbol{\theta_i} \omega_{\mathbf{c}_i \mathbf{c}_j} \boldsymbol{\theta_j}$$

  where $\sum_{i \in \mathcal{N}_r} \theta_i = 1 \; \forall r = 1, \ldots, K$.

- Allows a broad degree distribution, and nodes can have **different** expected degree.
- Popular nodes have higher value of θ.

Degree Corrected Blockmodel (DCBM)

- Adds degree scaling parameters θ_i for each node to allow for a broad degree distribution.

$$P_{ij} = \boldsymbol{\theta_i} \omega_{\mathbf{c}_i \mathbf{c}_j} \boldsymbol{\theta_j}$$

  where $\sum_{i \in \mathcal{N}_r} \theta_i = 1 \; \forall r = 1, \ldots, K$.
- Allows a broad degree distribution, and nodes can have **different** expected degree.
- Popular nodes have higher value of θ.

Popularity of nodes

- Network feature closely associated with community structure.
- Popularity of the i^{th} node in the r^{th} community

$$M_{ir} = \sum_{j \in \mathcal{N}_r} A_{ij}.$$

- Model version

$$\mu_{ir} = \mathbb{E}[M_{ir}]$$

- Under the DCBM, degree parameter θ_i inflates or deflates node popularity **uniformly** across all communities.
- For i, j in same community,

$$\frac{\mu_{ir}}{\theta_i} = \frac{\mu_{jr}}{\theta_j}$$

$\Rightarrow$ **popularity** $\propto$ **degree**, which is unrealistic.

Popularity of nodes

- Network feature closely associated with community structure.
- Popularity of the i^{th} node in the r^{th} community

$$M_{ir} = \sum_{j \in \mathcal{N}_r} A_{ij}.$$

- Model version

$$\mu_{ir} = \mathbb{E}[M_{ir}]$$

- Under the DCBM, degree parameter θ_i inflates or deflates node popularity **uniformly** across all communities.
- For i, j in same community,

$$\frac{\mu_{ir}}{\theta_i} = \frac{\mu_{jr}}{\theta_j}$$

$\Rightarrow$ **popularity** $\propto$ **degree**, which is unrealistic.

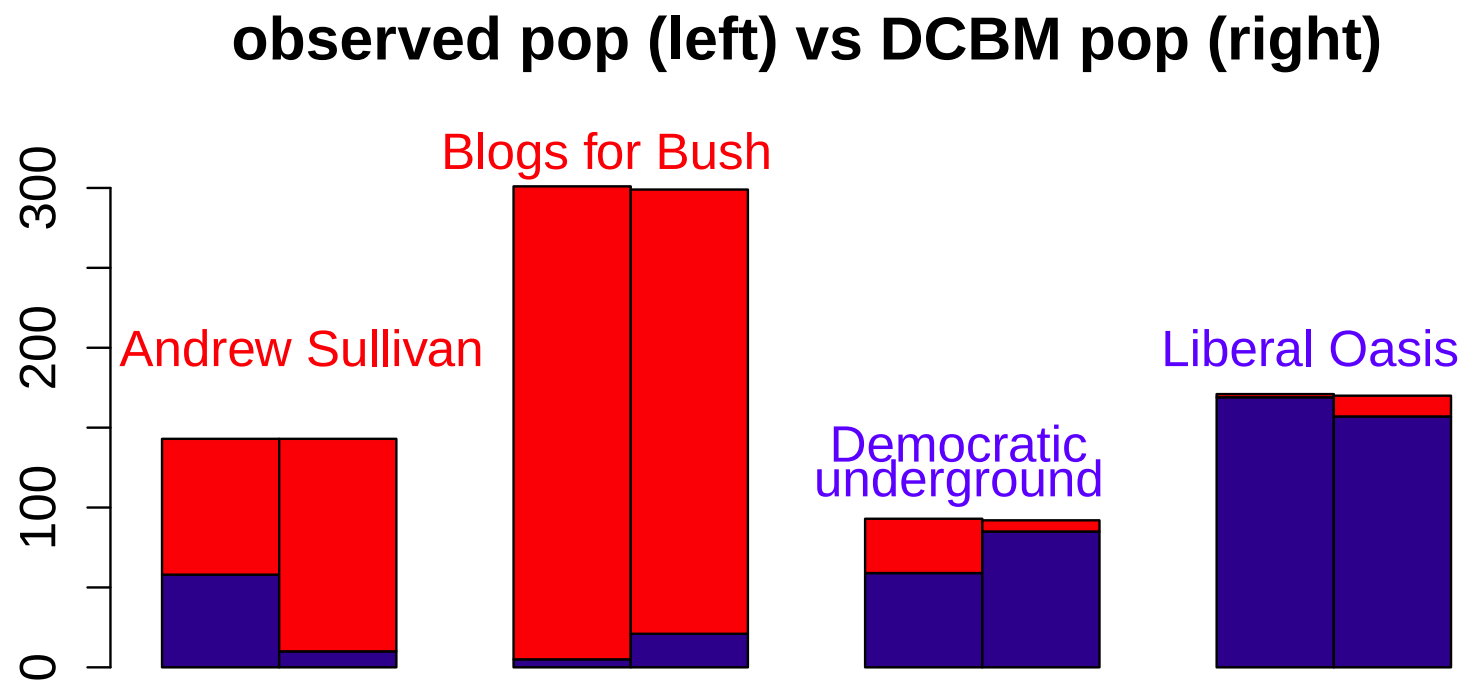
Popularity of nodes (pol blogs)

- Andrew Sullivan
- Conservative blogger
- Degree = 143
  Liberal = 58 (41%)
  Conservative = 85 (59%)

- Blogs for Bush
- Conservative blogger
- Degree = 301
  Liberal = 5 (2%)
  Conservative = 296 (98%)

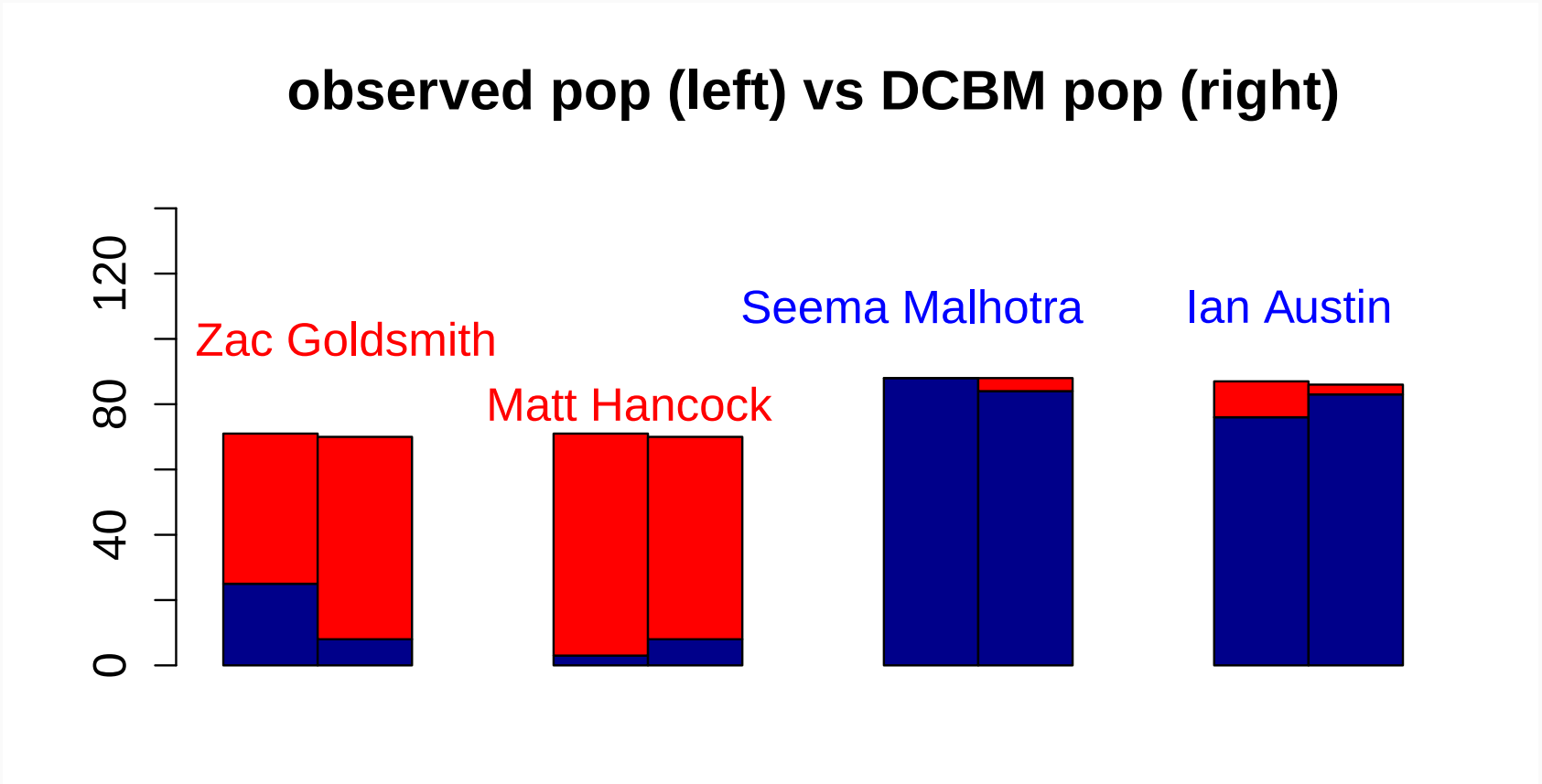
Popularity of nodes

- Under the DCBM, for two nodes in the same community, **popularity** $\propto$ **degree**
- DCBM fits node degrees and detects communities well, but inadequate for fitting node popularity.



Popularity of nodes

- Under the DCBM, for two nodes in the same community, **popularity** $\propto$ **degree**
- DCBM fits node degrees and detects communities well, but inadequate for fitting node popularity.



Proposed model

- **PABM** (popularity-adjusted blockmodel):
  For a K-block network, let $\boldsymbol{\lambda}_{n\times K}$ be popularity parameters.
  Let $\mathbf{c}$ denote the membership vector, then for $i < j$

$$P_{ij} = \boldsymbol{\lambda}_{\boldsymbol{ic_j}}\boldsymbol{\lambda}_{\boldsymbol{jc_i}}.$$

- DCBM is a special case of PABM: set $\lambda_{ir} = \theta_i\sqrt{\omega_{c_i r}}$, then

$$P_{ij} = \lambda_{ic_j}\lambda_{jc_i} = (\theta_i\sqrt{\omega_{c_i c_j}})(\theta_j\sqrt{\omega_{c_i c_j}}) = \theta_i\omega_{c_i c_j}\theta_j.$$

- Has the structural flexibility to model node popularity in a realistic way.

Proposed model

- **PABM** (popularity-adjusted blockmodel):
  For a K-block network, let $\boldsymbol{\lambda}_{n\times K}$ be popularity parameters.
  Let $\mathbf{c}$ denote the membership vector, then for $i < j$

$$P_{ij} = \boldsymbol{\lambda}_{\boldsymbol{ic_j}}\boldsymbol{\lambda}_{\boldsymbol{jc_i}}.$$

- DCBM is a special case of PABM: set $\lambda_{ir} = \theta_i\sqrt{\omega_{c_i r}}$, then

$$P_{ij} = \lambda_{ic_j}\lambda_{jc_i} = (\theta_i\sqrt{\omega_{c_i c_j}})(\theta_j\sqrt{\omega_{c_i c_j}}) = \theta_i\omega_{c_i c_j}\theta_j.$$

- Has the structural flexibility to model node popularity in a realistic way.

Community detection and model fitting

- **Likelihood modularity** (profile likelihood):

$$Q_{PABM}(b) = 2\sum_i \sum_r M_{ir} \log(M_{ir}) - \sum_{rs} O_{rs} \log(O_{rs}).$$

- **Community detection**: Let c be the (unknown) true community assignment and $\mathcal{B}$ be the set of all possible community assignments.

$$\hat{c} = \arg\max_{b \in \mathcal{B}} Q_{PABM}(A, b)$$

- **MLE of parameters:**

$$\hat{\lambda}_{ir} = \frac{M_{ir}(\hat{c})}{\sqrt{O_{\hat{c}_i r}(\hat{c})}}.$$

where $O_{sr} = \sum_{i \in \mathcal{N}_s} M_{ir}$ and $\mathcal{N}_s = \{i \leq n : \hat{c}_i = s\}$.

Extreme points (Le et al, AoS 2016)

$$\hat{c} = \arg\max_{b \in \mathcal{B}} Q_{PABM}(A, b), \qquad \hat{\lambda}_{ir} = \frac{M_{ir}(\hat{c})}{\sqrt{O_{\hat{c}_i r}(\hat{c})}}$$

- $|\mathcal{B}| = O(K^n)$, exhaustive search for maxima infeasible.
- Popular alternatives include greedy algorithms based Kernighan-Lin algorithm.
- EP algorithm: exhaustive search over the $O(n^{K-1})$ **extreme** points of $\mathcal{B}$.
- Extreme points of the projection of $[1, 2, ..., K]^n$ onto the space spanned by top K eigenvectors of A.

Extreme points (Le et al, AoS 2016)

$$\hat{c} = \arg\max_{b \in \mathcal{B}} Q_{PABM}(A, b), \qquad \hat{\lambda}_{ir} = \frac{M_{ir}(\hat{c})}{\sqrt{O_{\hat{c}_i r}(\hat{c})}}$$

- $|\mathcal{B}| = O(K^n)$, exhaustive search for maxima infeasible.
- Popular alternatives include greedy algorithms based Kernighan-Lin algorithm.
- EP algorithm: exhaustive search over the $O(n^{K-1})$ **extreme** points of $\mathcal{B}$.
- Extreme points of the projection of $[1, 2, ..., K]^n$ onto the space spanned by top K eigenvectors of A.

Model-fitting algorithm

Input: A (adjacency matrix), K (number of communities) [1]

1. Find $\mathcal{B}_{EP}$[2], the set of extreme points.
2. For each $b \in \mathcal{B}_{EP}$, compute $Q_{PABM}(A, b)$.
3. $$\hat{c} = \arg\max_{b \in \mathcal{B}_{EP}} Q_{PABM}(A, b), \qquad \hat{\lambda}_{ir} = \frac{M_{ir}(\hat{c})}{\sqrt{O_{\hat{c}_i r}(\hat{c})}}$$

[1] There are methods to estimate K that can be run prior to this.

[2] There should be $O(n^{K-1})$ extreme points.

Consistency of community detection: heuristics

Goal: $\hat{c} = \arg\max_e Q(e) \to c$ (true assignment).

$\tilde{Q}(e) =$ population version of $Q(e)$.

1. Show that $Q(e)$ is **uniformly** (w.r.t. e) close to $\tilde{Q}(e)$.
2. For $\tilde{Q}(e)$, show that
   - $c = \arg\max_e \tilde{Q}(e)$ **uniquely**
   - $\tilde{Q}(e)$ is **uniformly** continuous (w.r.t. e)
3. By 1, $Q(\hat{c})$ is close to $\tilde{Q}(\hat{c})$ and $Q(c)$ is close to $\tilde{Q}(c)$.
   But $Q(\hat{c}) > Q(c)$ and $\tilde{Q}(c) > \tilde{Q}(\hat{c})$ 2(i).
   Hence by 2(ii), $\hat{c}$ must be close to c.

Consistency of community detection: heuristics

Goal: $\hat{c} = \arg\max_e Q(e) \to c$ (true assignment).

$\tilde{Q}(e) =$ population version of $Q(e)$.

1. Show that $Q(e)$ is **uniformly** (w.r.t. e) close to $\tilde{Q}(e)$.
2. For $\tilde{Q}(e)$, show that
   - $c = \arg\max_e \tilde{Q}(e)$ **uniquely**
   - $\tilde{Q}(e)$ is **uniformly** continuous (w.r.t. e)
3. By 1, $Q(\hat{c})$ is close to $\tilde{Q}(\hat{c})$ and $Q(c)$ is close to $\tilde{Q}(c)$.
   But $Q(\hat{c}) > Q(c)$ and $\tilde{Q}(c) > \tilde{Q}(\hat{c})$ 2(i).
   Hence by 2(ii), $\hat{c}$ must be close to c.

Consistency of community detection: heuristics

Goal: $\hat{c} = \arg\max_e Q(e) \to c$ (true assignment).

$\tilde{Q}(e) =$ population version of $Q(e)$.

1. Show that $Q(e)$ is **uniformly** (w.r.t. e) close to $\tilde{Q}(e)$.
2. For $\tilde{Q}(e)$, show that
   - $c = \arg\max_e \tilde{Q}(e)$ **uniquely**
   - $\tilde{Q}(e)$ is **uniformly** continuous (w.r.t. e)
3. By 1, $Q(\hat{c})$ is close to $\tilde{Q}(\hat{c})$ and $Q(c)$ is close to $\tilde{Q}(c)$.
   But $Q(\hat{c}) > Q(c)$ and $\tilde{Q}(c) > \tilde{Q}(\hat{c})$ 2(i).
   Hence by 2(ii), $\hat{c}$ must be close to c.

Consistency of community detection

1. The number of communities K is fixed and known.
2. **Sparsity:** $\rho_n = \omega(\frac{\log(n)}{\sqrt{n}})$ [implies $\frac{n\rho_n^2}{\log^2(n)} \to \infty$ as $n \to \infty$.]

$$Q(e) = \frac{2}{n^2\rho_n} \sum_i \sum_r M_{ir} \log\left(\frac{M_{ir}}{\sqrt{O_{e_i r}}}\right)$$

$$\tilde{Q}(e) = \frac{2}{n^2\rho_n} \sum_i \sum_r \mu_{ir}(e) \log\left(\frac{\mu_{ir}(e)}{\sqrt{o_{re_i}(e)}}\right)$$

Lemma 1

Under Assumptions 1 and 2,

$$\max_e |Q(e) - \tilde{Q}(e)| \xrightarrow{\mathbb{P}} 0.$$

This establishes an **uniform** concentration bound for the modularity function and its population version.

Consistency of community detection

3 **Identifiability:** $\Lambda_{ab} = \Lambda_{ba}$ for any communities a, b, where

$$\Lambda_{ab} := \sum_{j \in \mathcal{N}_a} \lambda_{jb}.$$

.

4 **Detectability:** for any two **distinct** communities a, b, and any two nodes $j_1 \in \mathcal{N}_a$, $j_2 \in \mathcal{N}_b$, the set $\left\{ \frac{p_{ij_1}}{p_{ij_2}} \right\}_{i=1}^{n}$ takes at least $K + 1$ distinct values.

Lemma 2

Under Assumptions 3 and 4, $\tilde{Q}(e)$ is *'uniquely'* maximized at the **correct** assignment $\mathbf{c}$, i.e., for any candidate assignment e,

$$\tilde{Q}(e) \leq \tilde{Q}(c)$$

where equality holds if and only if $e \in \Pi(c)$, and Π is the symmetric group of all permutations of $\{1, \ldots, K\}$.

Consistency of community detection

We define error as

$$\xi_n(e) = \min_{e' \in \Pi(e)} \frac{1}{n} \sum_{i=1}^{n} I[e'_i \neq c_i],$$

where disagreement is minimized over all label permutations of the candidate assignment.

Theorem

Under Assumptions 1 - 4,

$$\xi_n(\hat{c}) \xrightarrow{\mathbb{P}} 0$$

where $\hat{c} = \arg\max_{b \in \mathcal{B}} Q_{PABM}(A, b)$.

Sengupta and Chen (2015)

Consistency of parameter estimation

5 Community size: all communities as per true assignment, c, are of size $\Theta(n)$, i.e., the same order as n.

6 Signal strength: If the average interaction level between any two large, distinct subsets of the vertex set is non-zero, it must be at least of the order of $\frac{\rho_n}{\log n}$. Formally, let $\Gamma_1, \Gamma_2 \subset \{1, \ldots, n\}$, $\Gamma_1 \cap \Gamma_2 = \phi$, $|\Gamma_1| = \Theta(n)$, $|\Gamma_2| = \Theta(n)$, then

$$\frac{1}{|\Gamma_1||\Gamma_2|} \sum_{(i,j) \in \Gamma_1 \times \Gamma_2} p_{ij} = \begin{cases} 0 & p_{ij} = 0 \; \forall \; (i,j) \in \Gamma_1 \times \Gamma_2 \\ \Omega\left(\frac{\rho_n}{\log n}\right) & \text{otherwise} \end{cases}$$

where $|S|$ is the cardinality of set S.

Consistency of parameter estimation

We define the parameter estimation error as

$$\Delta_n(\hat{c}) = \frac{1}{\sqrt{n}} ||\hat{\lambda}_{n\times K} - \lambda_{n\times K}||_F.$$

Theorem:

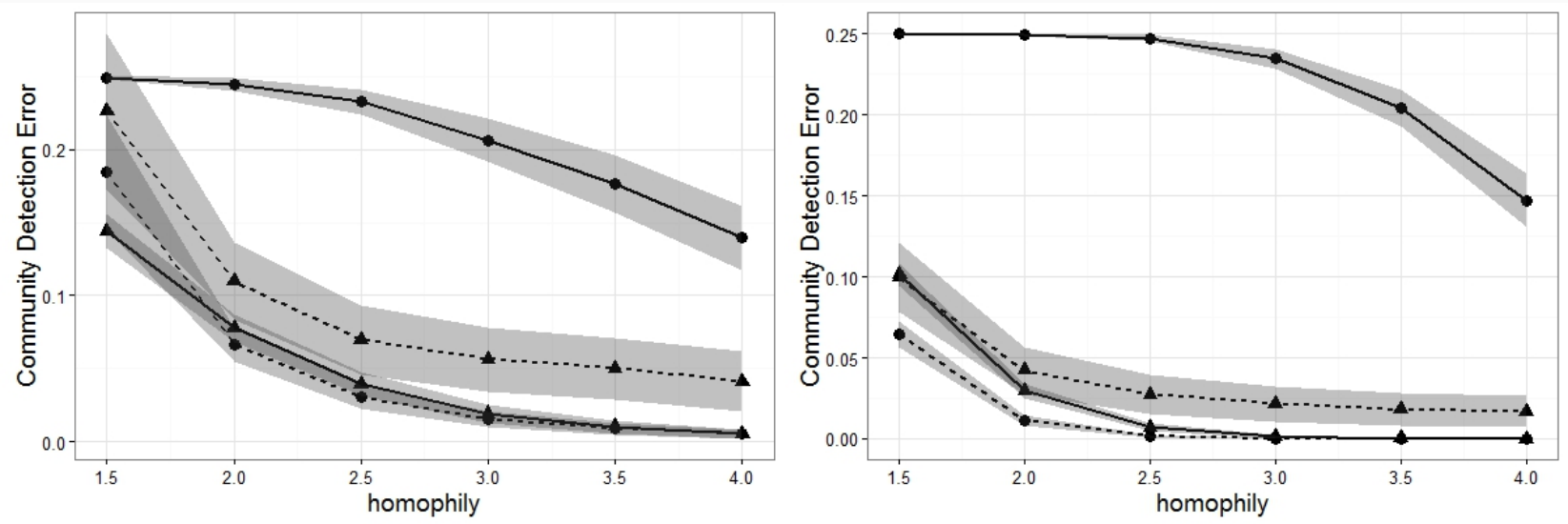
Under Assumptions 1 - 6,

$$\Delta_n(\hat{c}) \xrightarrow{P} 0.$$

Simulation Study

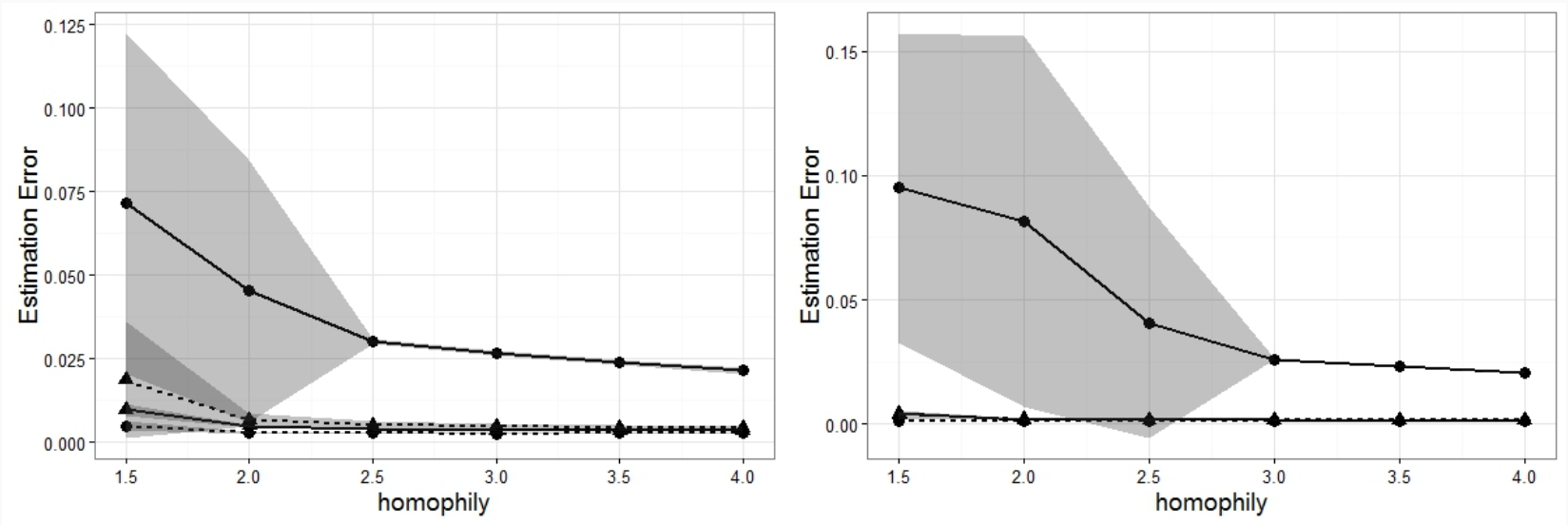
- $K = 2$, $n = 400, n_1 = n_2 = 200$ and $n = 1000, n_1 = n_2 = 500$.
- Node popularity varies:
  category 1 nodes have most (80% or more) of their neighbors in own community, category 2 nodes have neighbors uniformly in two communities.
- Homophily parameter h:
  expected number of intra-community edges is h times the expected number of inter-community edges (community detection easier with higher h).

Simulation Study: community detection



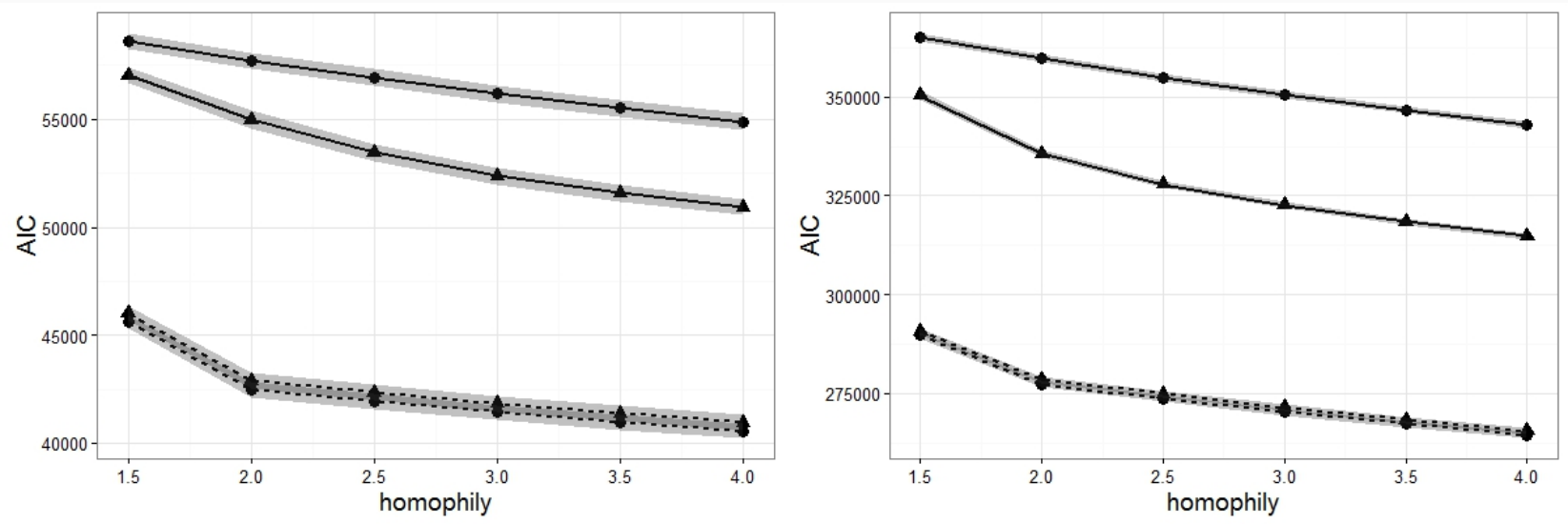
$n = 400$ (left) and $n = 1000$ (right). Solid lines represent PABM networks, dashed lines represent DCBM networks. Triangles represent results from PABM modularity and dots represent results from DCBM modularity. Shaded area represents standard deviation. For networks from PABM (solid lines), the PABM modularity (triangles) outperforms the DCBM modularity (dots) substantially. For networks from DCBM (dashed lines), the PABM modularity (triangles) performs only slightly worse than the DCBM modularity (dots).

Simulation Study: parameter estimation



$n = 400$ (left) and $n = 1000$ (right). Solid lines represent PABM networks, dashed lines represent DCBM networks. Triangles represent results from PABM modularity and dots represent results from DCBM modularity. Shaded area represents standard deviation. For networks from PABM (solid lines), the PABM modularity (triangles) outperforms the DCBM modularity (dots) substantially. For networks from DCBM (dashed lines), the PABM modularity (triangles) performs only slightly worse than the DCBM modularity (dots).

Simulation Study: Akaike Information Criterion (AIC)



$n = 400$ (left) and $n = 1000$ (right). Solid lines represent PABM networks, dashed lines represent DCBM networks. Triangles represent results from PABM modularity and dots represent results from DCBM modularity. Shaded area represents standard deviation. For networks from PABM (solid lines), the PABM modularity (triangles) outperforms the DCBM modularity (dots) substantially. For networks from DCBM (dashed lines), the PABM modularity (triangles) performs only slightly worse than the DCBM modularity (dots).

Real networks: community detection

Network	Nodes	PABM	DCBM
Political Blogs	1222	4.99% (61)	5.40% (66)
British MP	329	0.00% (0)	0.61% (2)
DBLP	2203	2.81% (62)	5.17% (114)

Community detection error rates (number of misclustered nodes in brackets)

Real networks: goodness of fit

$$F_1 = \frac{1}{2E} \sum_{i=1}^{n} \sum_{r=1}^{K} \left(\hat{\mu}_{ir}(\hat{c}) - M_{ir}(c) \right)^2 , \tag{1}$$

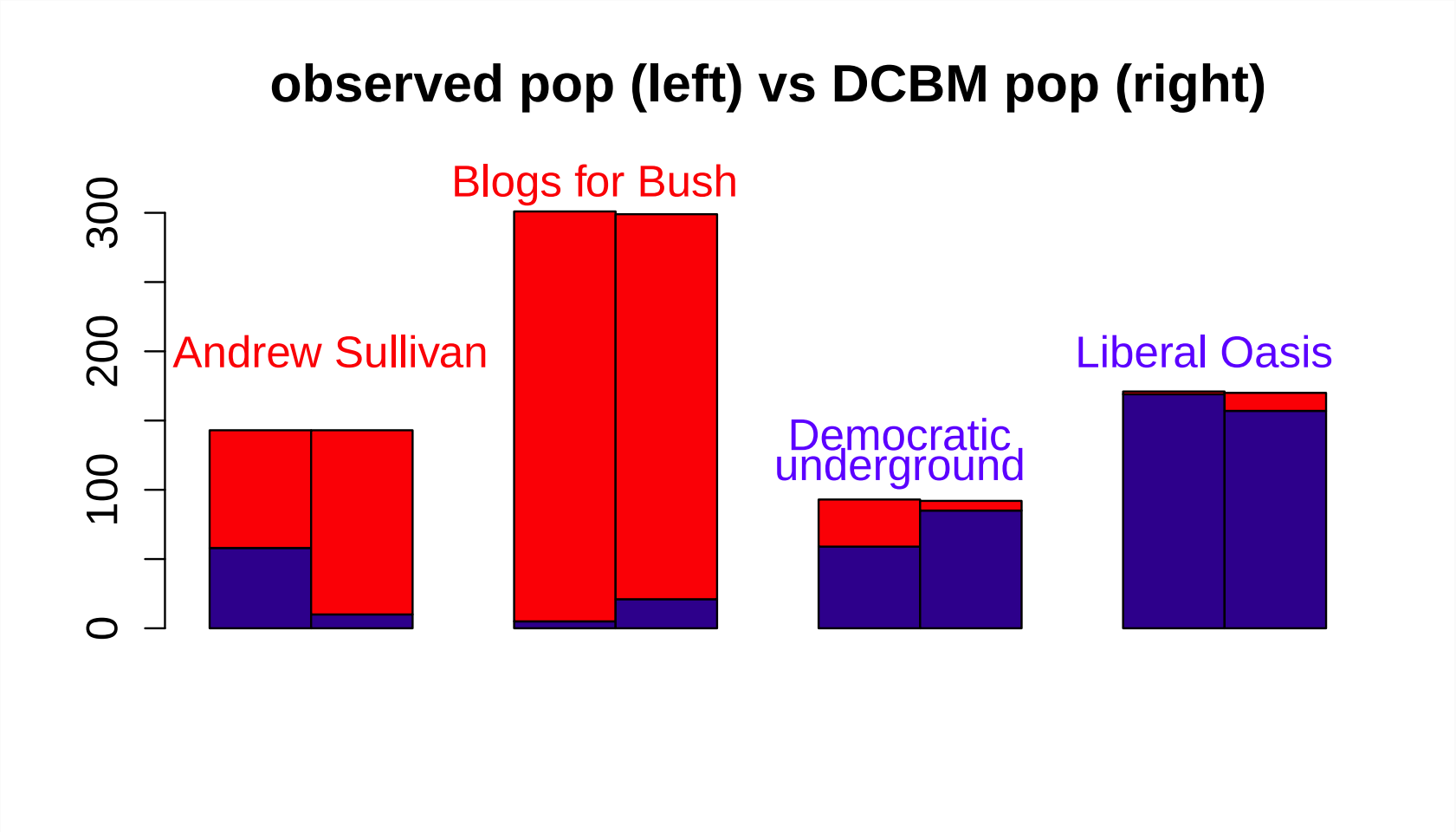
$$F_2 = \frac{1}{2E} \sum_{i=1}^{n} \sum_{r=1}^{K} \left(\hat{\mu}_{ir}(c) - M_{ir}(c) \right)^2 , \tag{2}$$

where E is the observed number of network edges.

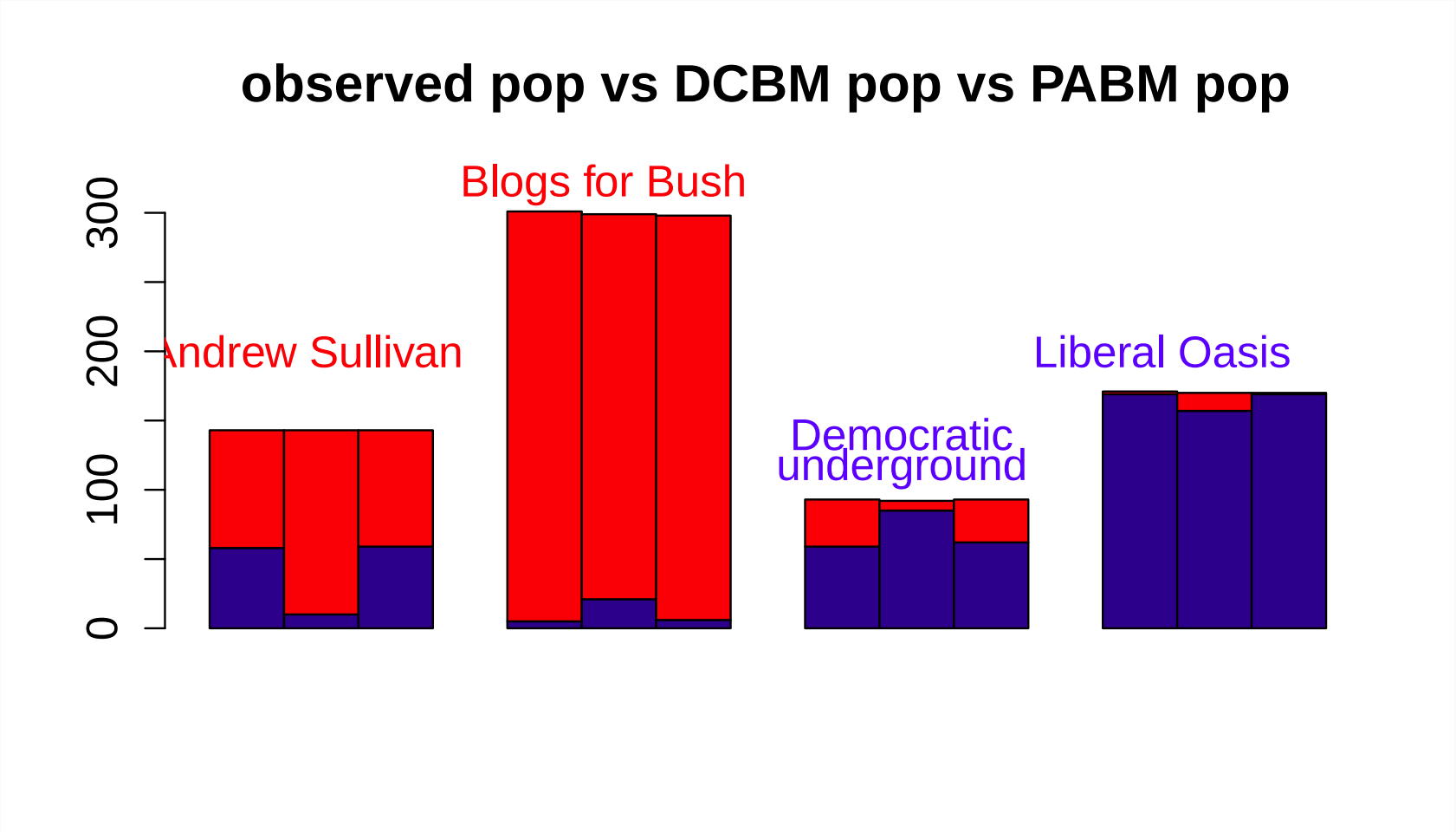
	F_1		F_2	
Network	PABM	DCBM	PABM	DCBM
Political Blogs	0.057	1.155	0.002	1.883
British MP	0.002	0.282	0.002	0.284
DBLP	2.255	52.430	0.000	61.425

Goodness of fit measures for node popularity

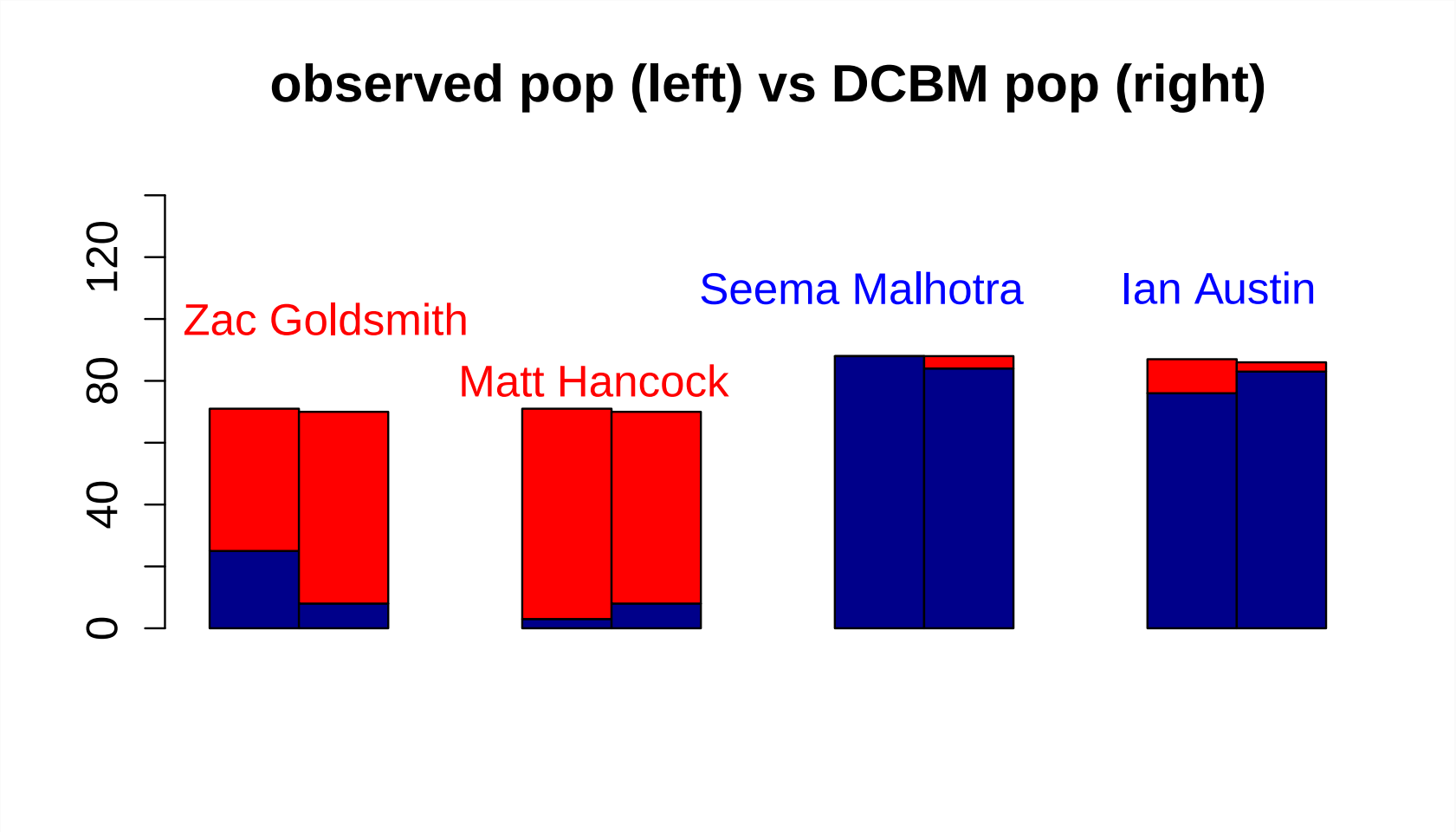
Popularity of nodes (pol blogs)



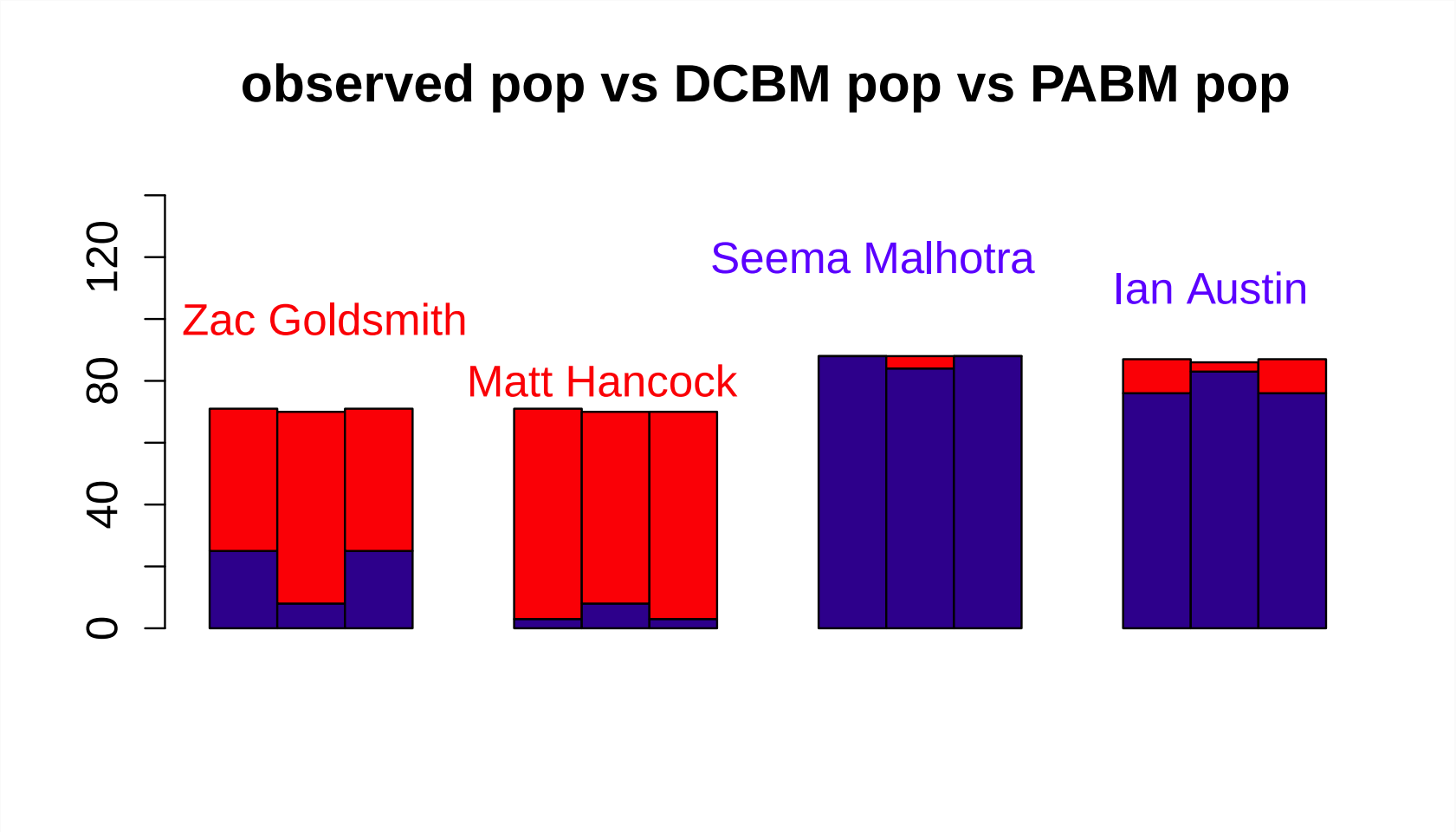
Popularity of nodes (pol blogs)



Popularity of nodes (British MP)



Popularity of nodes (British MP)



Summary

- PABM vastly improves modeling of node popularity in networks.
- Likelihood modularity is consistent and provides superior insights on well-studied real networks.
- The extreme points approach provides a computationally feasible method for model fitting and community detection.

Sengupta, S. and Chen, Y. (2017). A blockmodel for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Meethodology)*. Invited for minor revision.

Next steps: statistical questions

- Too many parameters: active and inactive node popularity for sparse networks
- Penalized estimation
- Thresholding
- Spectral clustering with node popularity
- Network monitoring using node popularities

Next steps: computational/algorithmic questions

- How to find extreme points for $K > 2$?
- How to scale up to large networks (millions of nodes)?
- How to make this efficient?

THANKS!