A Random Walk Around The Block

Johan Ugander
Stanford University

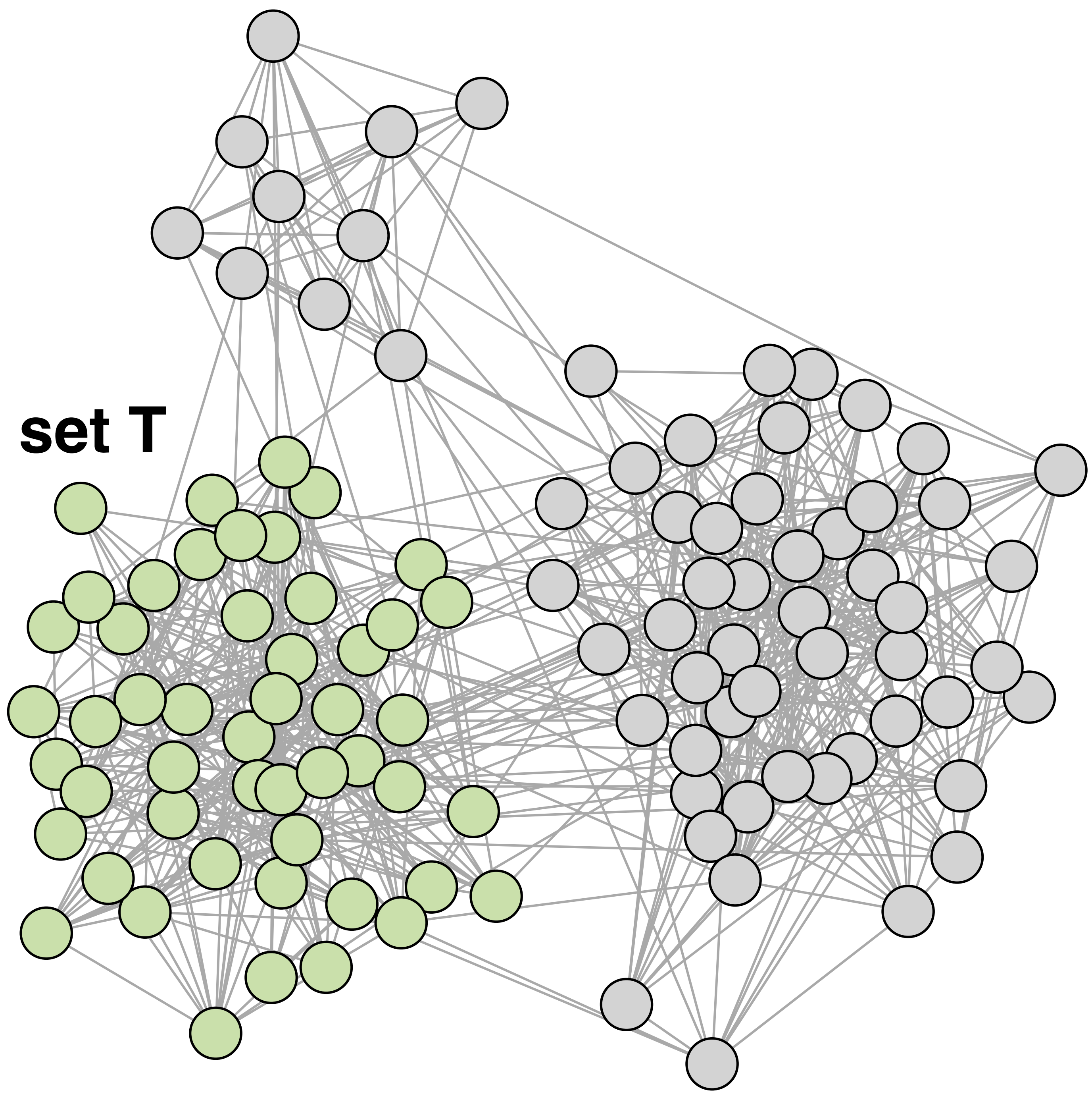
Joint work with:
Isabel Kloumann (Facebook)
& Jon Kleinberg (Cornell)

Google Mountain View
August 17, 2016

Seed set expansion

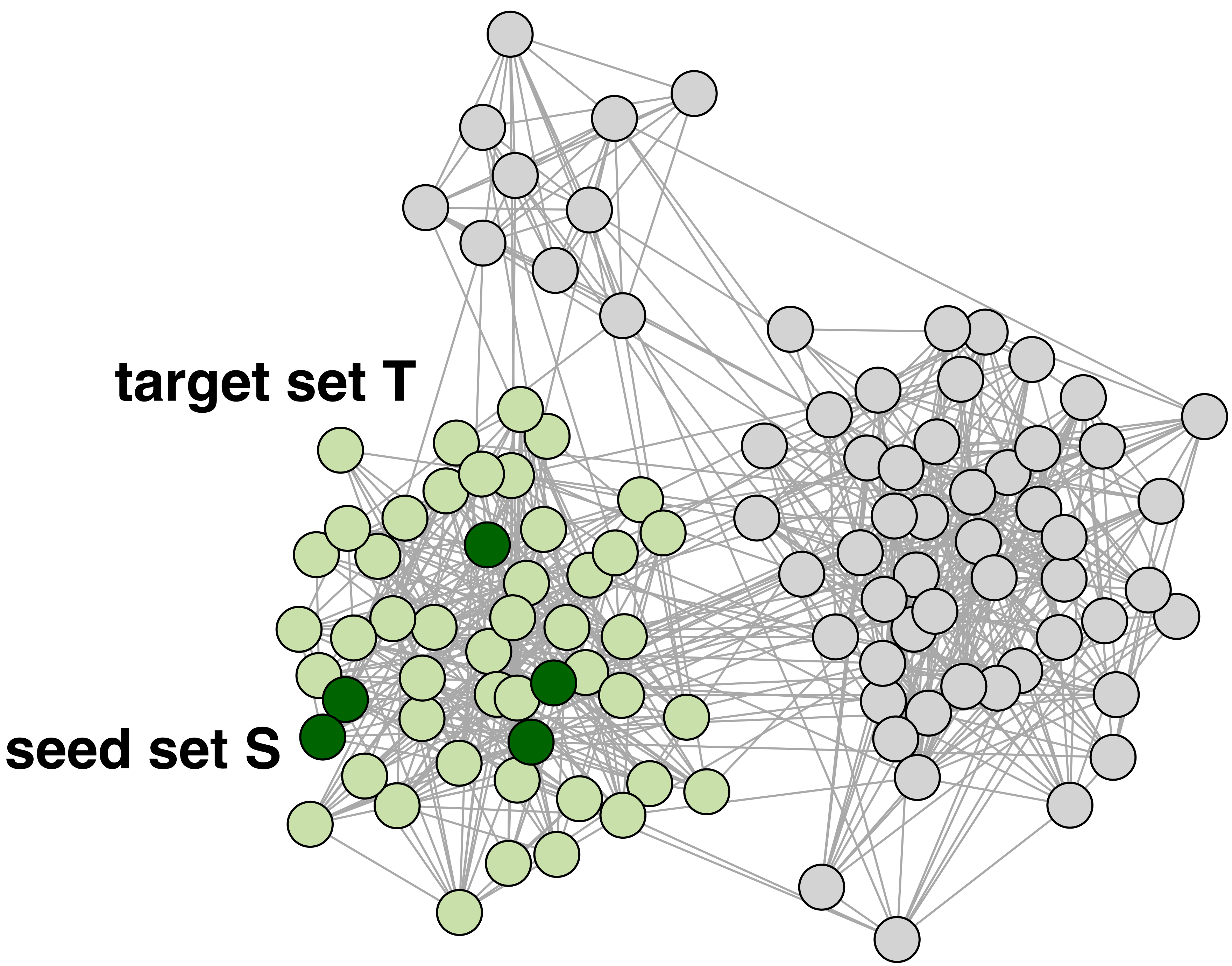
- Given a graph G=(V, E), goal is to accurately identify a **target set T** ⊂ **V** from a smaller **seed set S** ⊂ **T**.

target set T



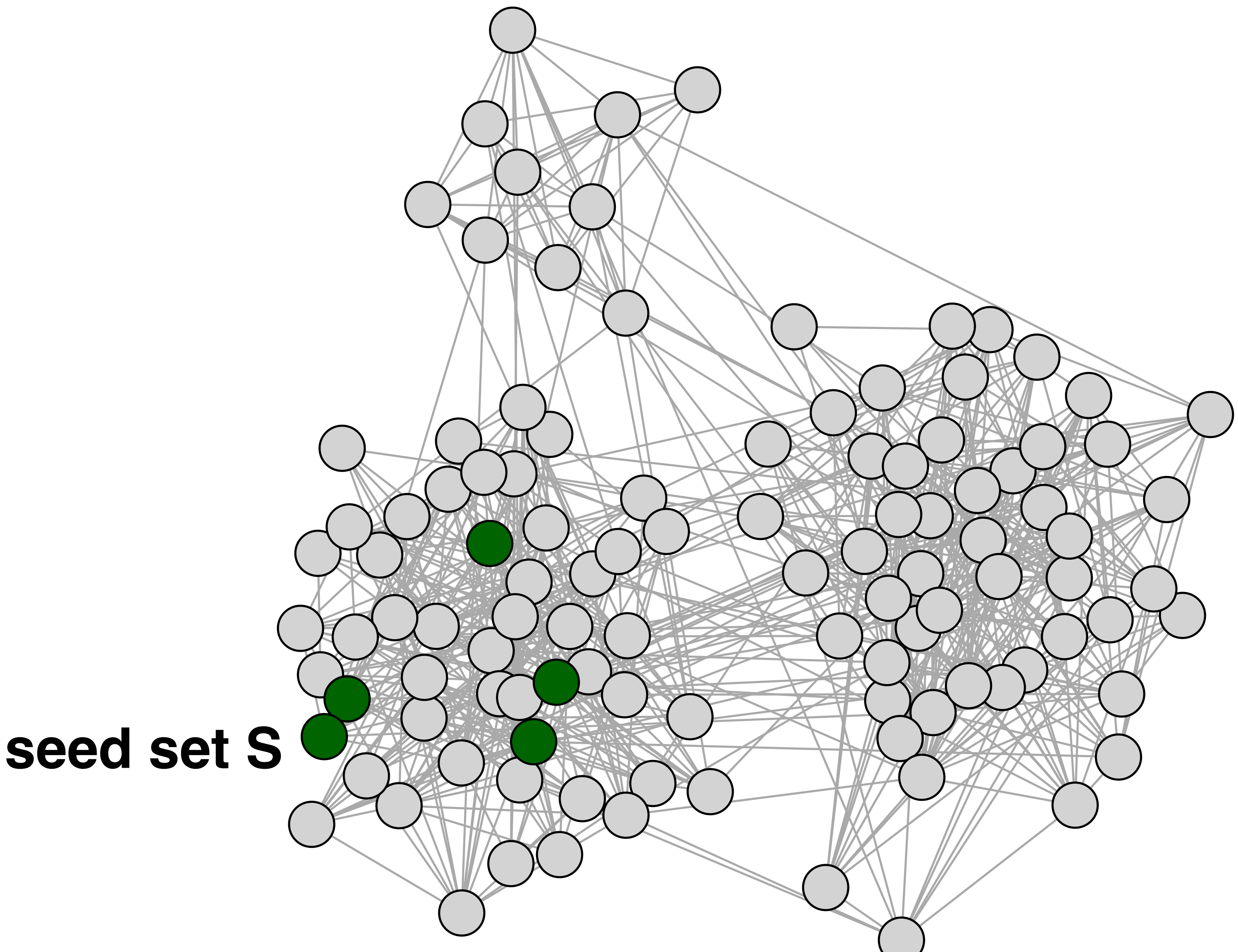
Seed set expansion

- Given a graph G=(V, E), goal is to accurately identify a **target set T** ⊂ **V** from a smaller **seed set S** ⊂ **T**.



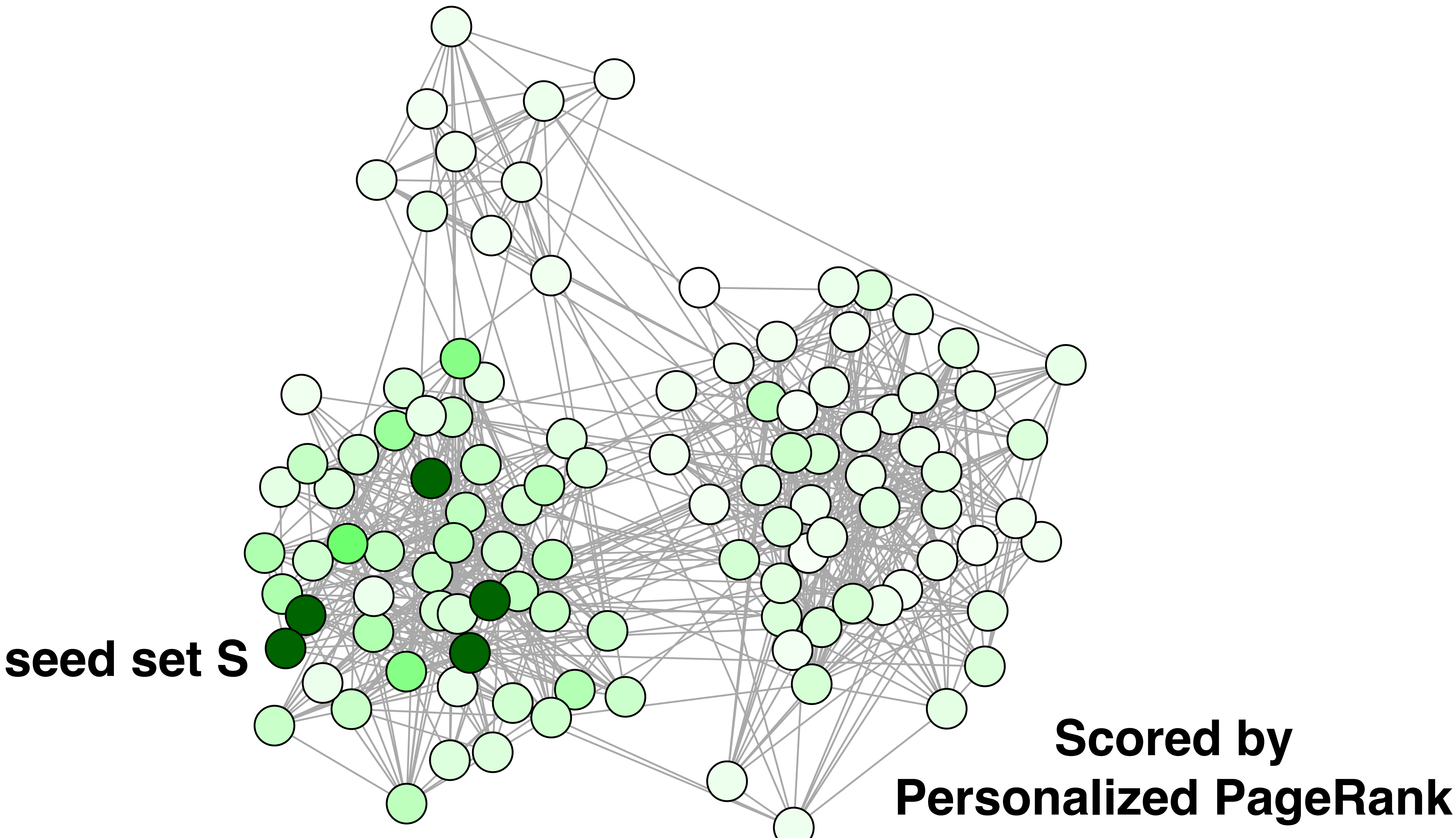
Seed set expansion

- Given a graph G=(V, E), goal is to accurately identify a **target set T** ⊂ **V** from a smaller **seed set S** ⊂ **T**.



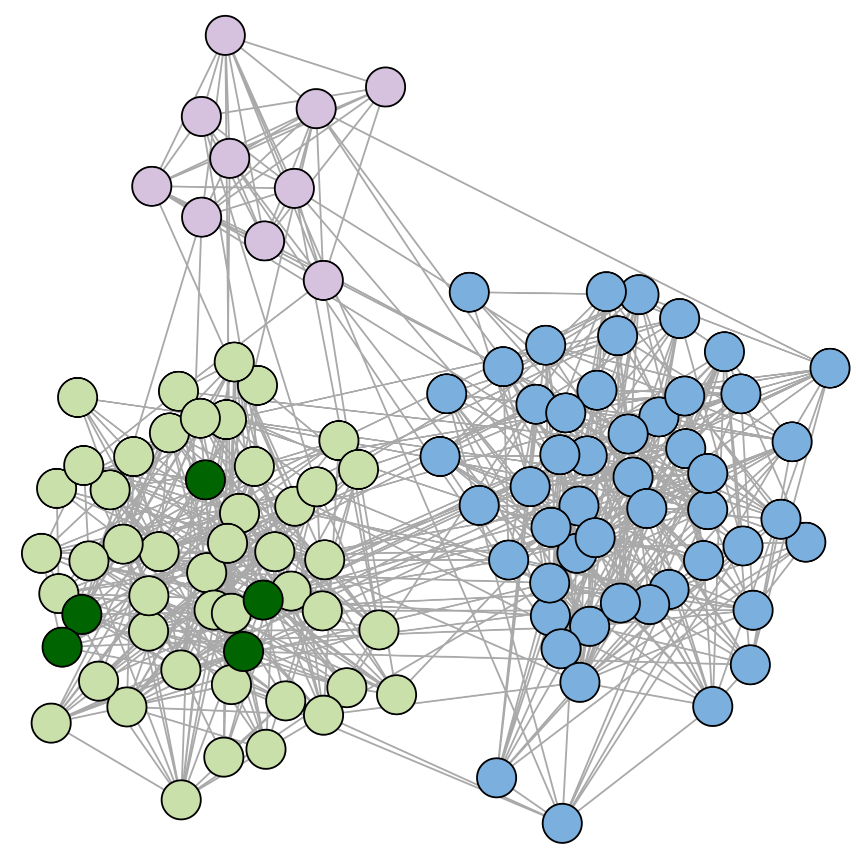
Seed set expansion

- Given a graph G=(V, E), goal is to accurately identify a **target set T** ⊂ **V** from a smaller **seed set S** ⊂ **T**.



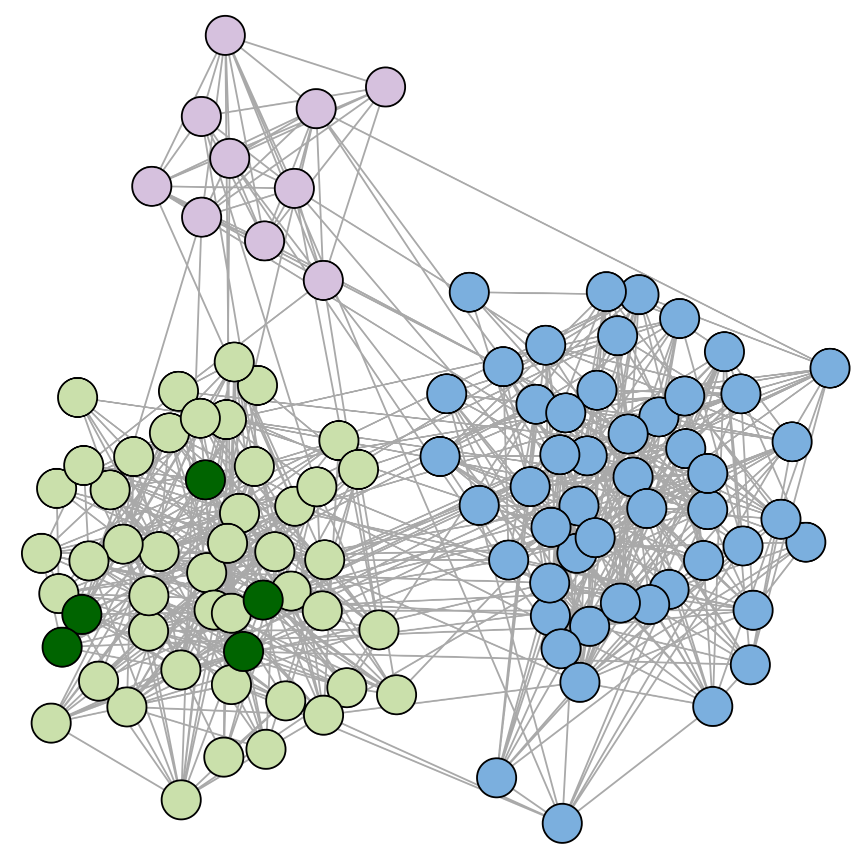
Seed set expansion

- Given a graph G=(V, E), goal is to accurately identify a **target set T ⊂ V** from a smaller **seed set S ⊂ T**.
- Applications:
  - **Broadly:** ranking on graphs, recommendation systems
  - Spam filtering (Wu & Chellapilla ’07)
  - Community detection (Weber et al. ’13)
  - Missing data inference (Mislove et al. ’14)
- Common methods:
  - Semi-supervised learning (Zhu et al. ’03)
  - Diffusion-based classification (Jeh & Widom ’03, Kloster & Gleich ’14)
  - Outwardness, modularity and more (Bagrow ’08, Kloumann & Kleinberg ’14)

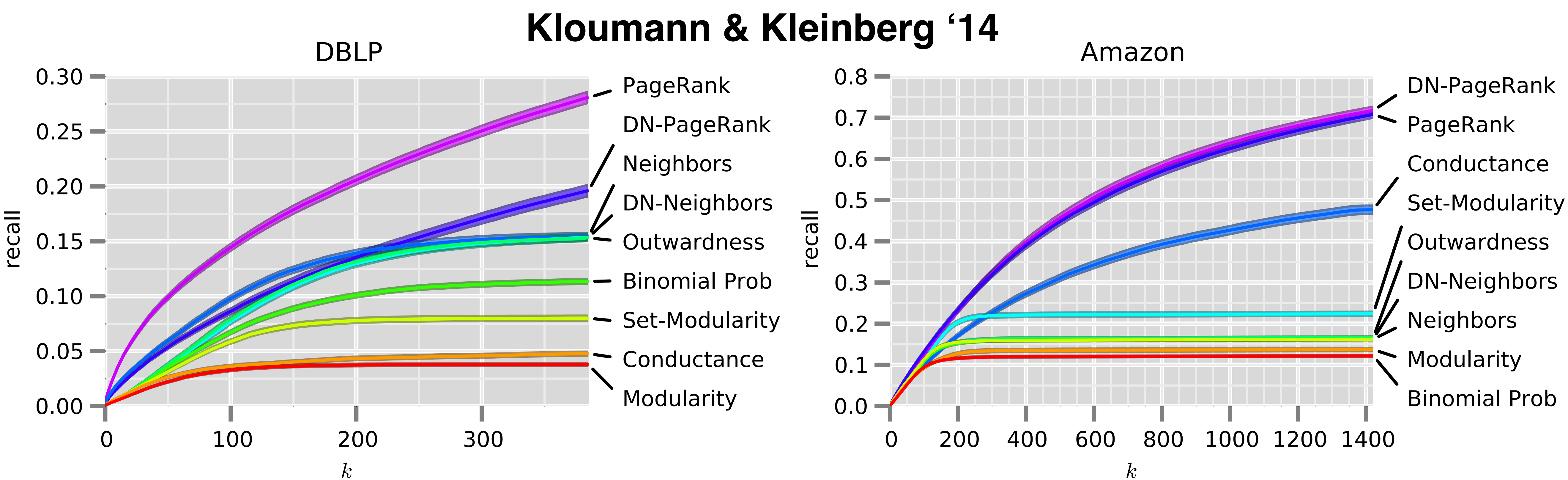


Seed set expansion

- Given a graph G=(V, E), goal is to accurately identify a **target set T** ⊂ **V** from a smaller **seed set S** ⊂ **T**.
- Applications:
  - **Broadly:** ranking on graphs, recommendation systems
  - Spam filtering (Wu & Chellapilla ’07)
  - Community detection (Weber et al. ’13)
  - Missing data inference (Mislove et al. ’14)
- Common methods:
  - Semi-supervised learning (Zhu et al. ’03)
  - **Diffusion-based classification** (Jeh & Widom ’03, Kloster & Gleich ’14)
  - Outwardness, modularity and more (Bagrow ’08, **Kloumann & Kleinberg ’14**)

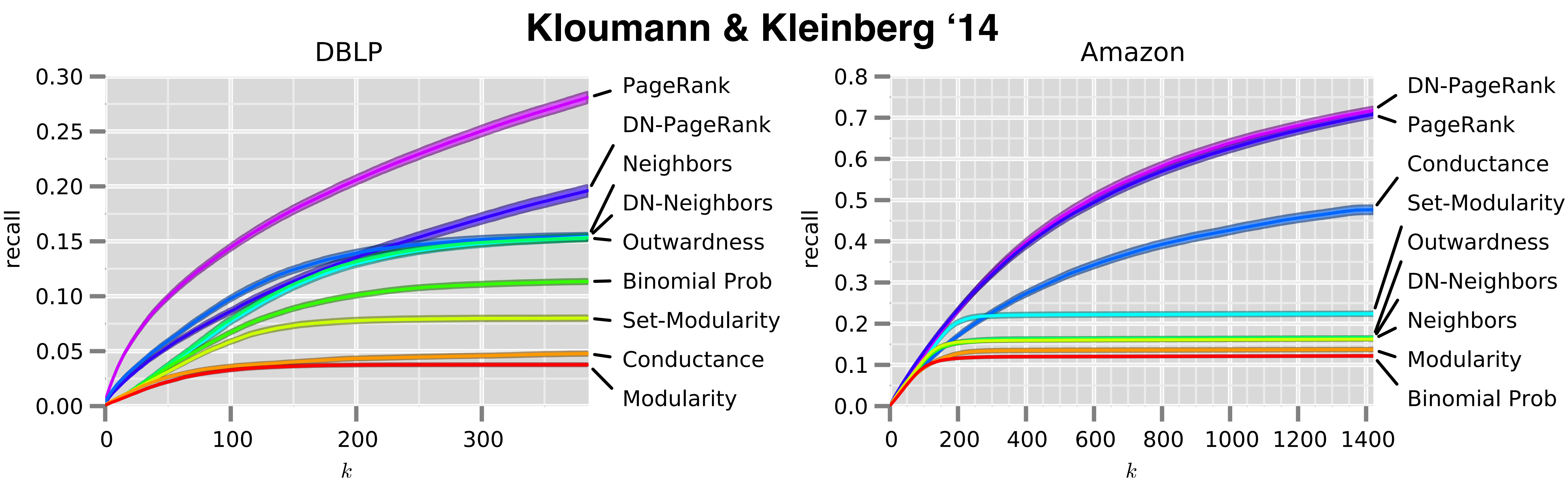


Recall curves for seed set expansion



- **Recall curve:** true positive rate, as a function of the number of items returned based on small uniformly random seed set.
- Kloumann & Kleinberg '14 tested many different methods on data, broadly found **Personalized PageRank** to be best.

Recall curves for seed set expansion



- **Recall curve:** true positive rate, as a function of the number of items returned based on small uniformly random seed set.
- Kloumann & Kleinberg '14 tested many different methods on data, broadly found **Personalized PageRank** to be best.
- **Truncated PPR** (first K steps) comparable to PPR from K=4.
- **Heat Kernel** later found comparable to PPR.

Diffusion-based node classification

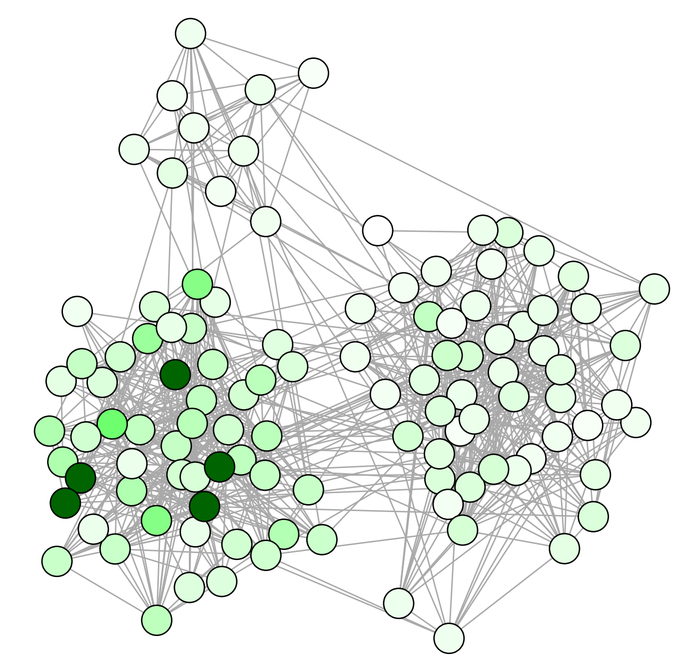
- Classification based on **random walk landing probabilities**
- r_k^v, probability that a random walk starting in **S** is at **v** after **k** steps.
- $(r_1^v, r_2^v, ..., r_K^v)$, truncated vector of landing probabilities.

- **Personalized PageRank** and **Heat Kernel** ranking:

$$\mathrm{PPR}(v) \propto \sum_{k=1}^{\infty} (\alpha^k) r_k^v \qquad \mathrm{HK}(v) \propto \sum_{k=1}^{\infty} \left(\frac{t^k}{k!}\right) r_k^v$$

- **General diffusion score function:**

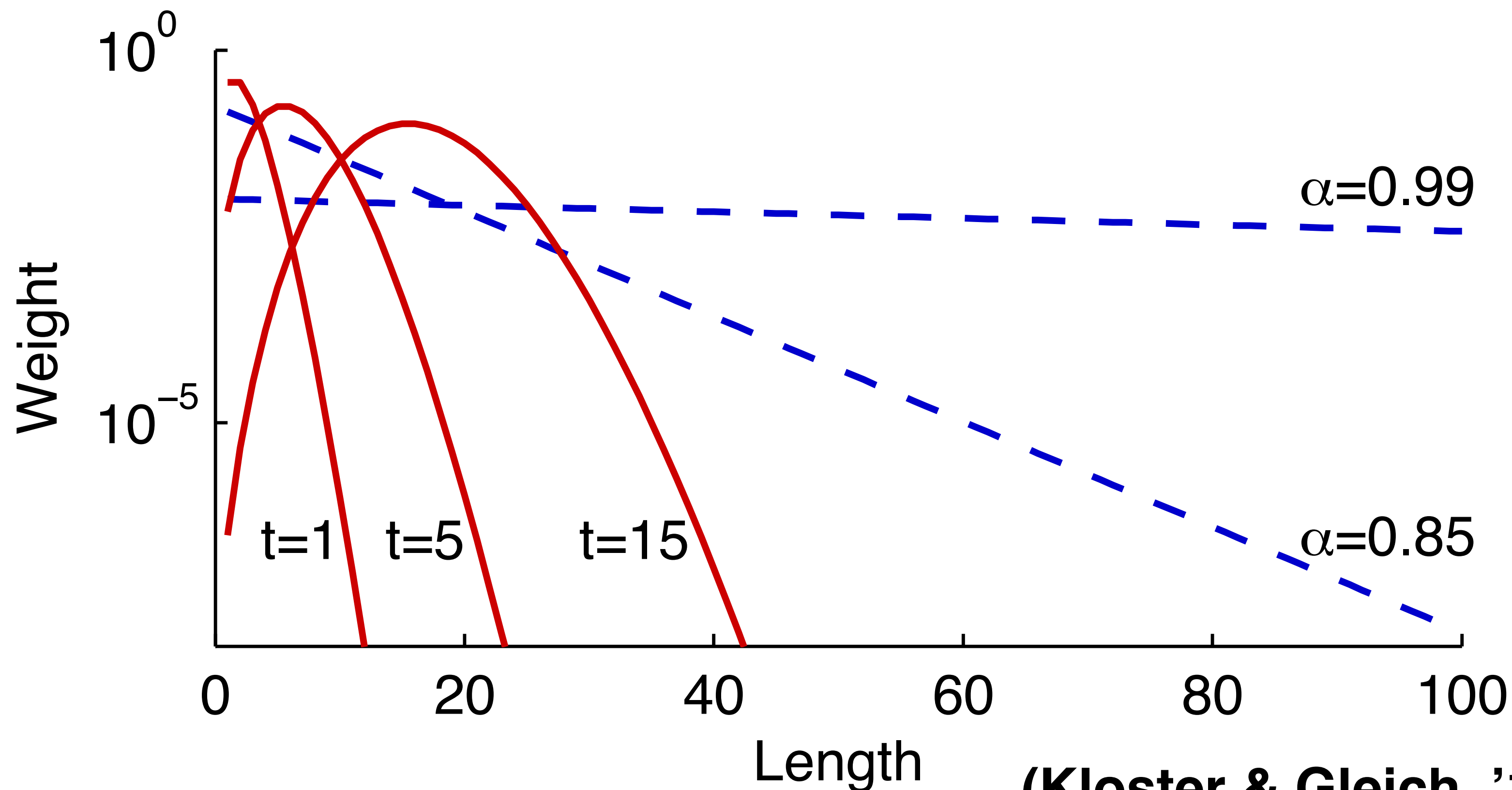
$$\mathrm{score}(v) = \sum_{k=1}^{\infty} w_k r_k^v$$



Diffusion-based node classification

- **Personalized PageRank** and **Heat Kernel** = two parametric families of linear weights

$$\text{score}(v) = \sum_{k=1}^{K} w_k r_k^v$$



(Kloster & Gleich, '14)

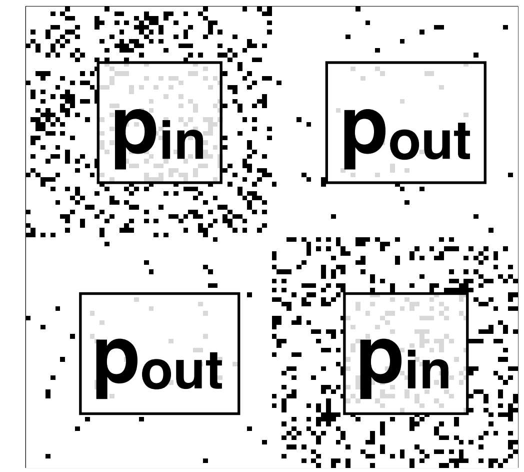
PPR $w_k = \alpha^k$

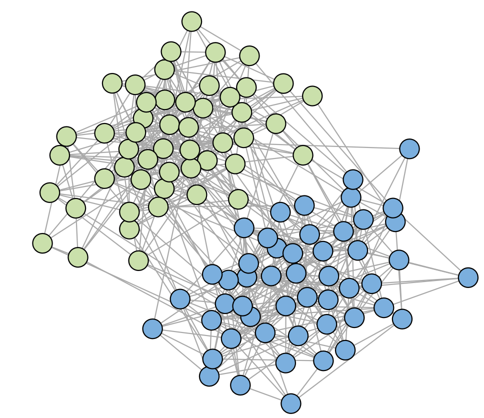
HK $w_k = t^k / k!$

- **Question in this work:**
  What weights are "optimal" for diffusion-based classification?

The stochastic block model

- **C** blocks
  - **Focus on C=2 blocks: 1="Target", 2="Other"**
- $\mathbf{n_1}$, $\mathbf{n_2}$ nodes in blocks
- Independent edge probabilities:
  - Edge probability within a block = $\mathbf{p_{in}}$
  - Edge probability across blocks = $\mathbf{p_{out}}$
- (Results for C>2 as well, see paper)

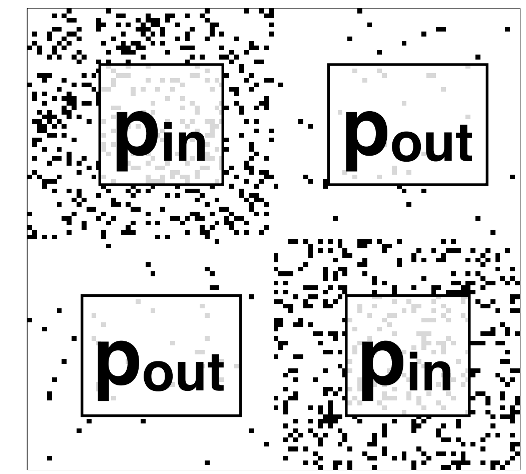


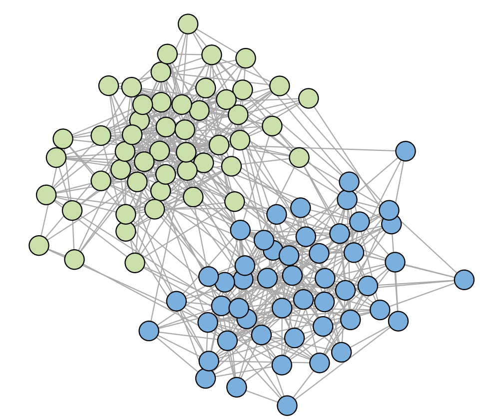


- Model with many names:
  - Stochastic Block Model (Holland et al. '83)
  - Affiliation Model (Frank-Harary '82)
  - Planted Partition Model (Dyer-Frieze '89)

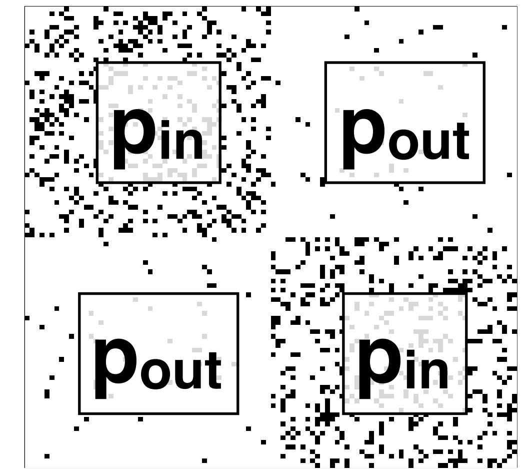
The SBM resolution limit

- **Find true partition in poly(n) time w.h.p. as n→∞ :**
  - Dyer-Frieze '89: If $p_{in} - p_{out} = O(1)$
  - Condon-Karp '01: If $p_{in} - p_{out} \geq \Omega(n^{-1/2})$
  - McSherry '01: If $p_{in} - p_{out} \geq \Omega((p_{out}(\log n)/n)^{-1/2})$

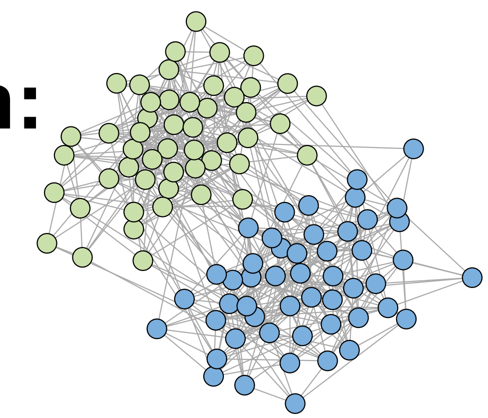




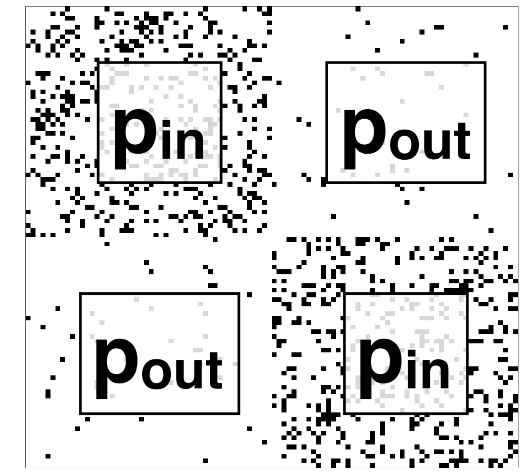
The SBM resolution limit



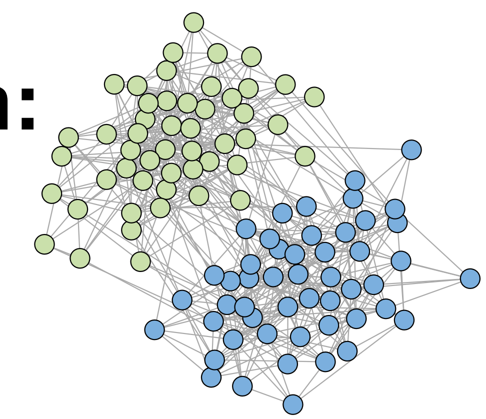
- **Find true partition in poly(n) time w.h.p. as $n \to \infty$:**
  - Dyer-Frieze '89: If $p_{in} - p_{out} = O(1)$
  - Condon-Karp '01: If $p_{in} - p_{out} \geq \Omega(n^{-1/2})$
  - McSherry '01: If $p_{in} - p_{out} \geq \Omega((p_{out}(\log n)/n)^{-1/2})$
- **Find partition positively correlated with true partition:**
  - Coja-Oghlan '06: If $p_{in} - p_{out} \geq \Omega((p_{out}/n)^{-1/2})$,



The SBM resolution limit



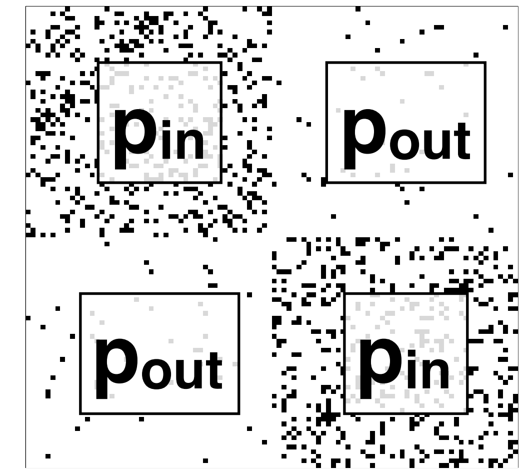
- **Find true partition in poly(n) time w.h.p. as $n \rightarrow \infty$:**
  - Dyer-Frieze '89: If $p_{in} - p_{out} = O(1)$
  - Condon-Karp '01: If $p_{in} - p_{out} \geq \Omega(n^{-1/2})$
  - McSherry '01: If $p_{in} - p_{out} \geq \Omega((p_{out}(\log n)/n)^{-1/2})$



- **Find partition positively correlated with true partition:**
  - Coja-Oghlan '06: If $p_{in} - p_{out} \geq \Omega((p_{out}/n)^{-1/2})$,
  - **If and only if** $(a-b)^2 > 2(a+b)$ ($p_{in} = a/n$, $p_{out} = b/n$):
    - Decelle et al '11: Conjecture and belief propagation numerics
    - Mossel et al '12,'13, Massoulié '13, Abbe et al. '14: Proven

- **Recent extensions:**
  - More than two blocks (e.g. Neeman-Netrapalli '14)
  - Unequal block sizes (e.g. Zhang et al. '16)

The SBM resolution limit



- **Is block recovery/classification over? No!**
  - Unsupervised vs. semi-supervised
  - Empirical graphs != SBMs
  - Optimal algorithms not practical
  - Beyond asymptotic limits, what are decay rates?

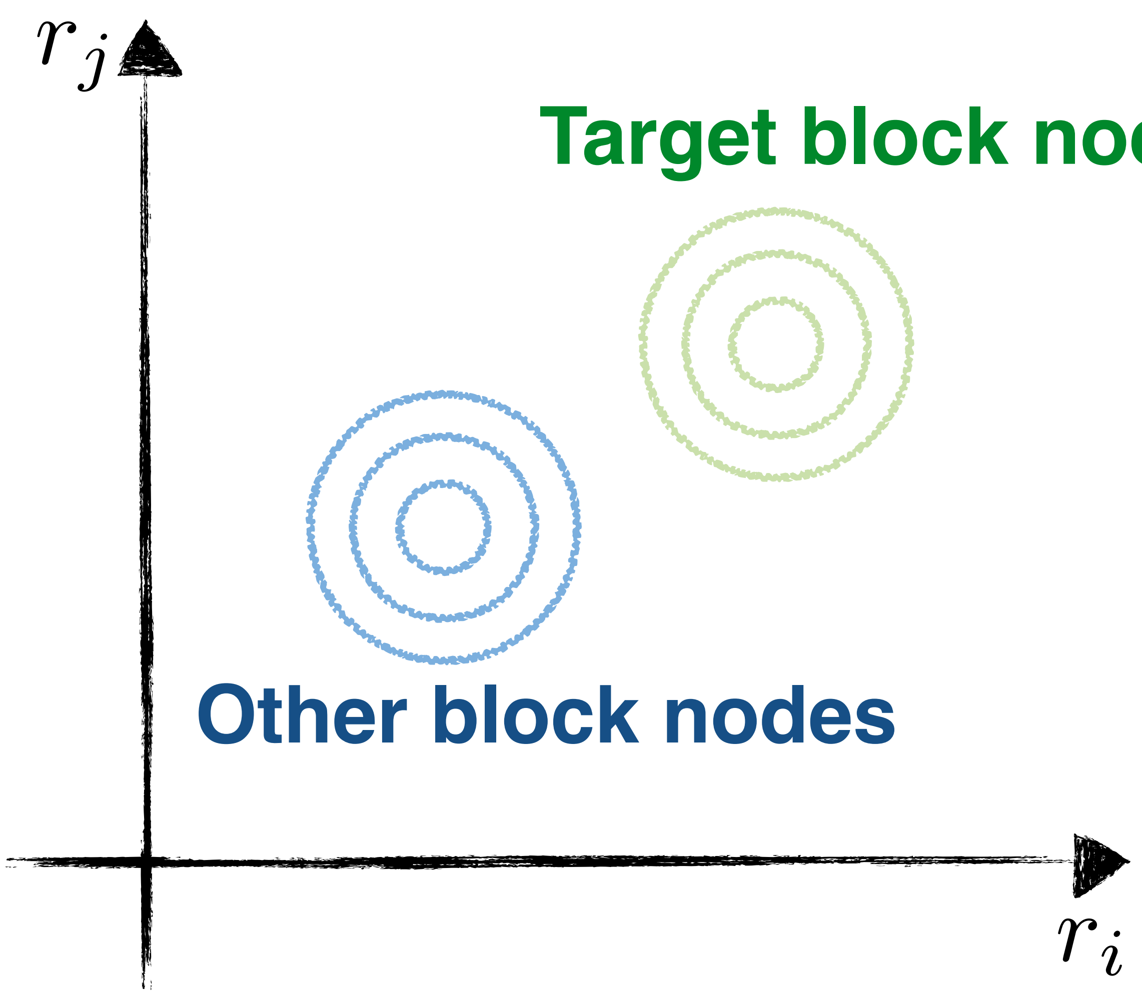
- Rather than being "**problem down**" (SBM classification), this talk will be "**method up**": how to tune diffusion weights to find seed sets?

$$\mathrm{score}(v) = \sum_{k=1}^{K} w_k r_k^v$$

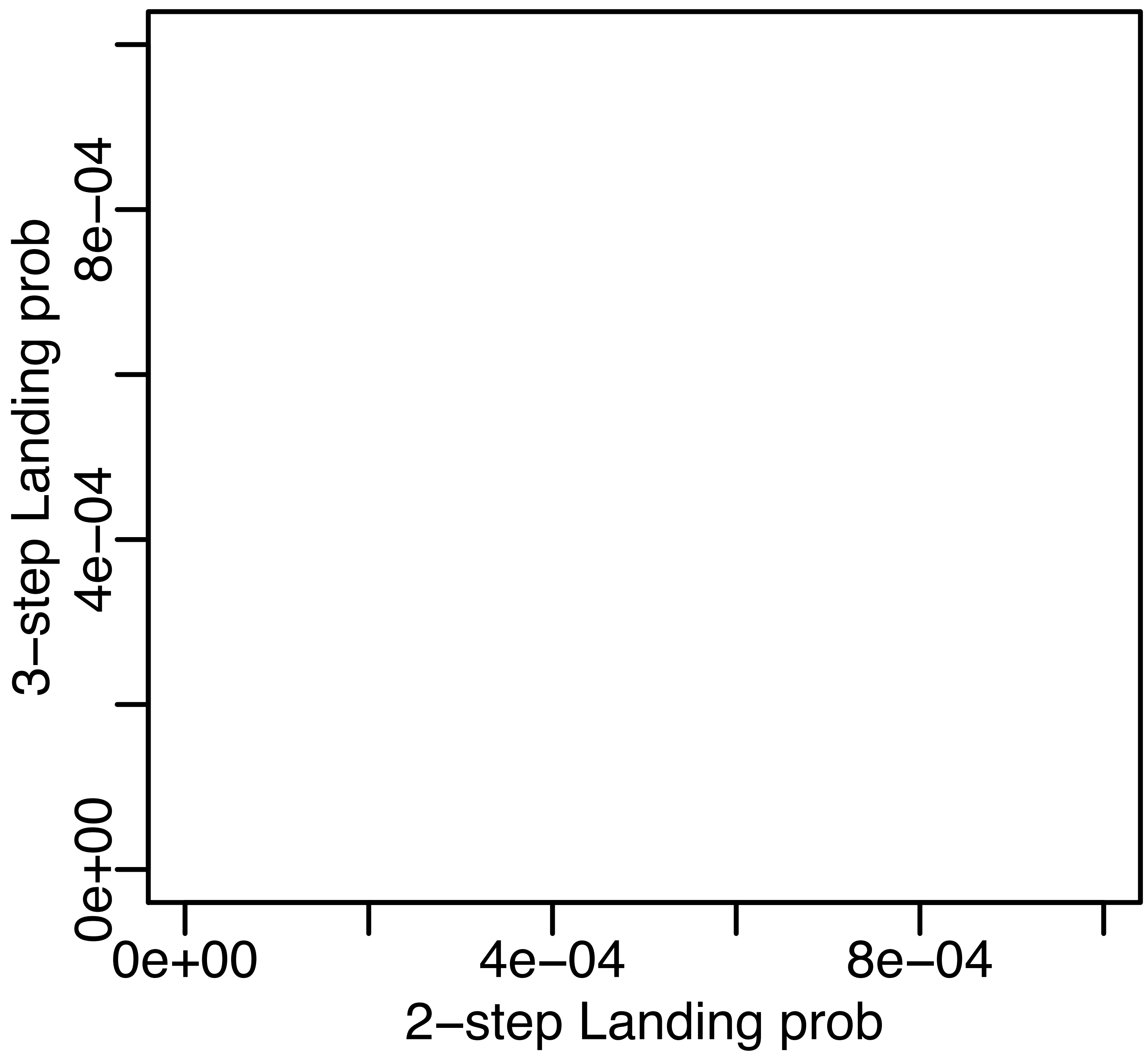
- Possible variations: Diffusion weights for seed set expansion in core-periphery models? Latent space models (Hoff et al. 2002)? Etc.

Diffusion-based classification in SBMs

- SBMs present a natural binary classification problem.
- Recall notation:
  - r_k^v , probability that a random walk starting in **S** is at **v** after **k** steps.
  - $(r_1^v, r_2^v, ..., r_K^v)$, truncated vector of landing probabilities.
- Choices of $(w_1, ..., w_K)$ define sweep directions through space.
- **Optimistically:**

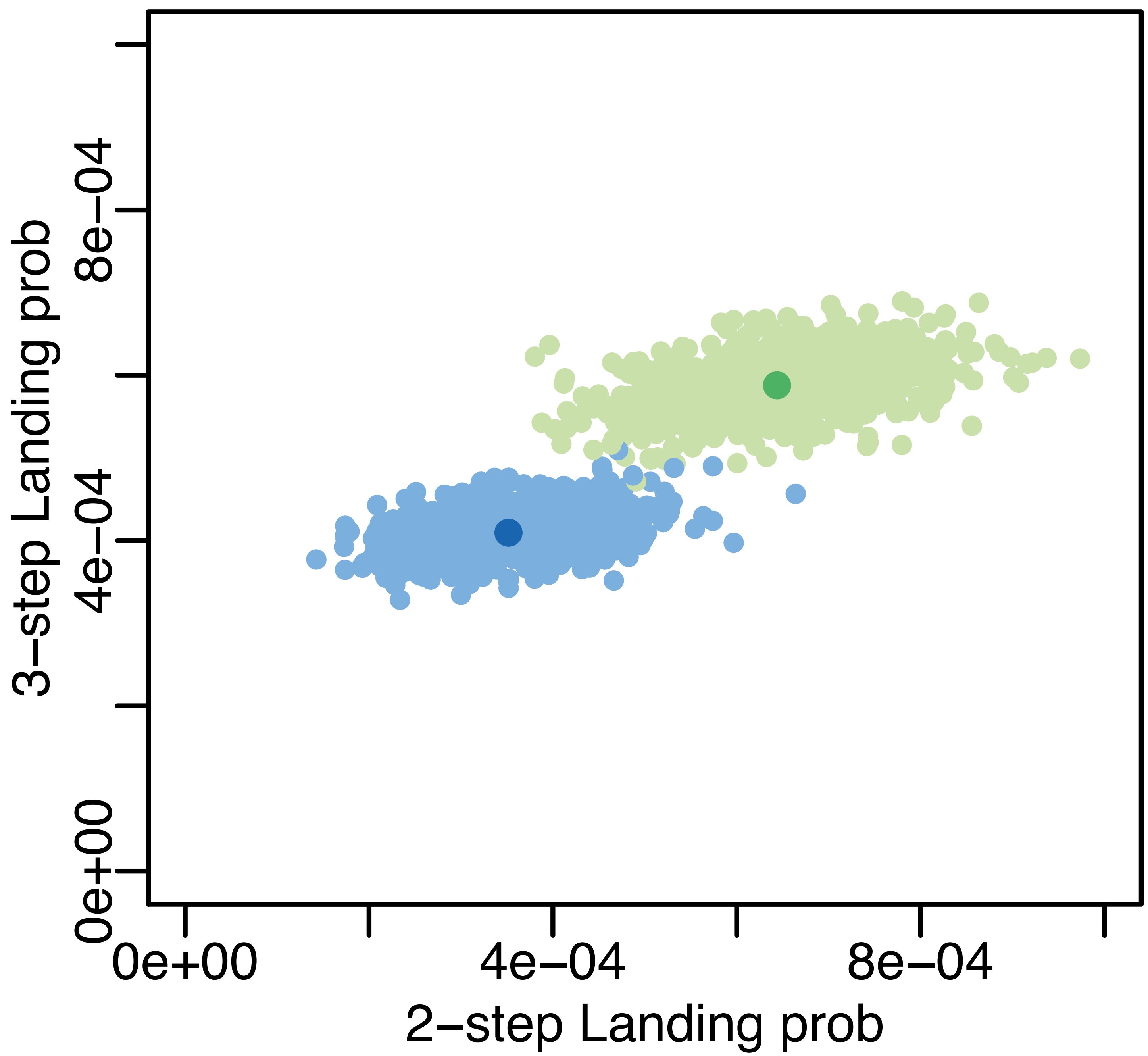


The space of landing probabilities



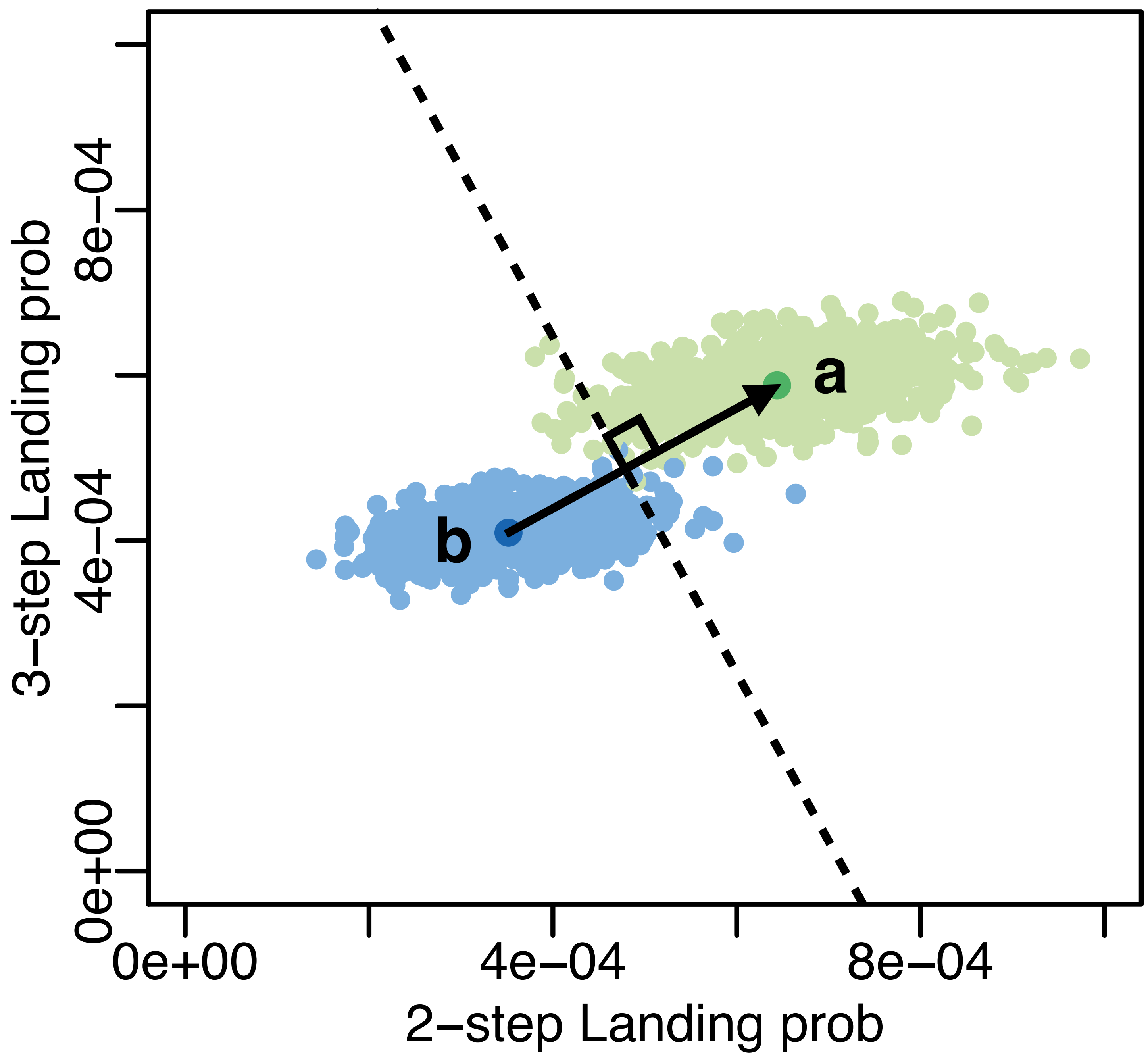
- SBM: 2000 nodes, **Target** & **Other** blocks, $p_{in} = 0.2$, $p_{out} = 0.05$
- One seed node (uniformly at random from Target set)

The space of landing probabilities



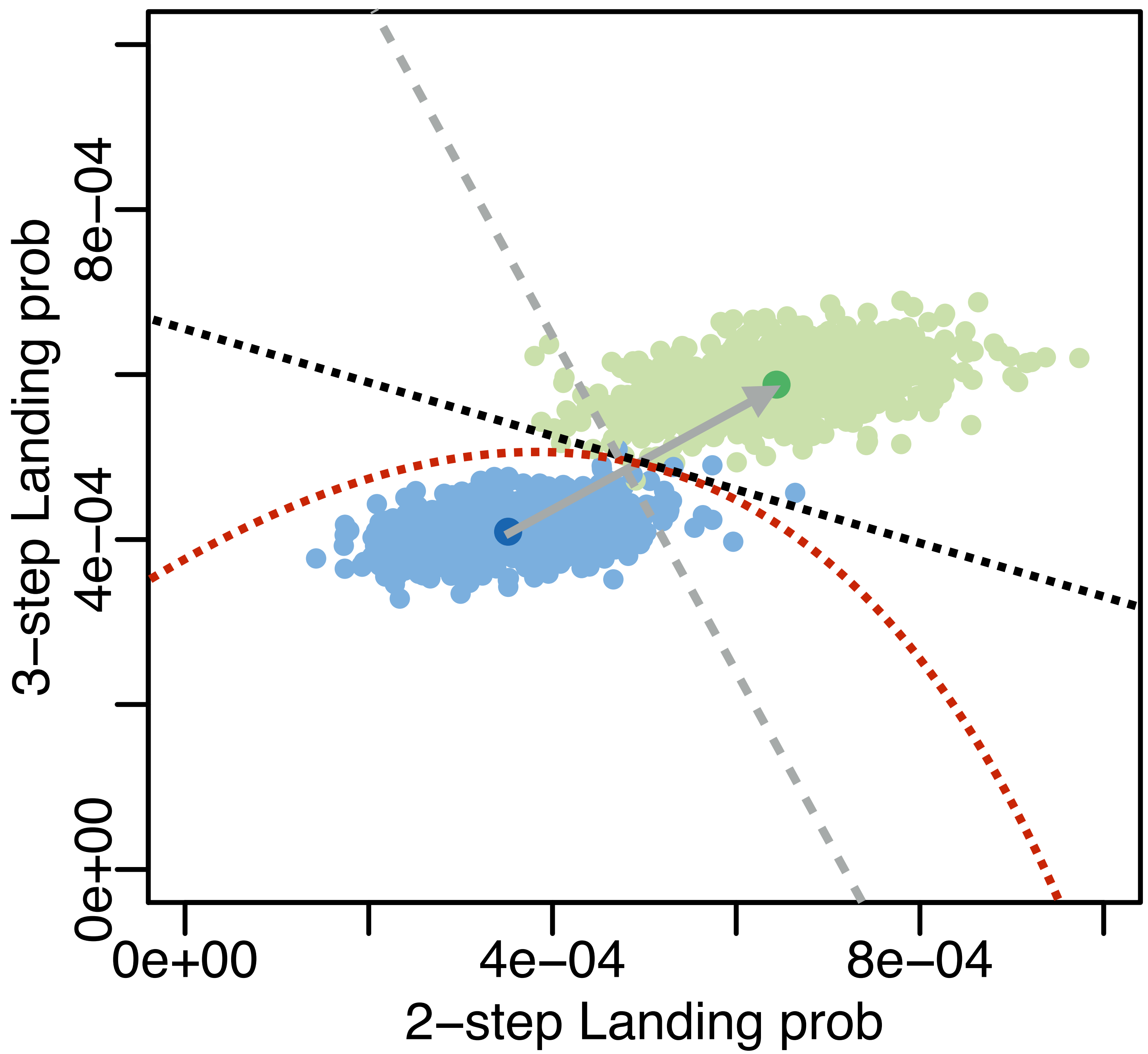
- SBM: 2000 nodes, **Target** & **Other** blocks, $p_{in} = 0.2$, $p_{out} = 0.05$
- One seed node (uniformly at random from Target set)

The space of landing probabilities



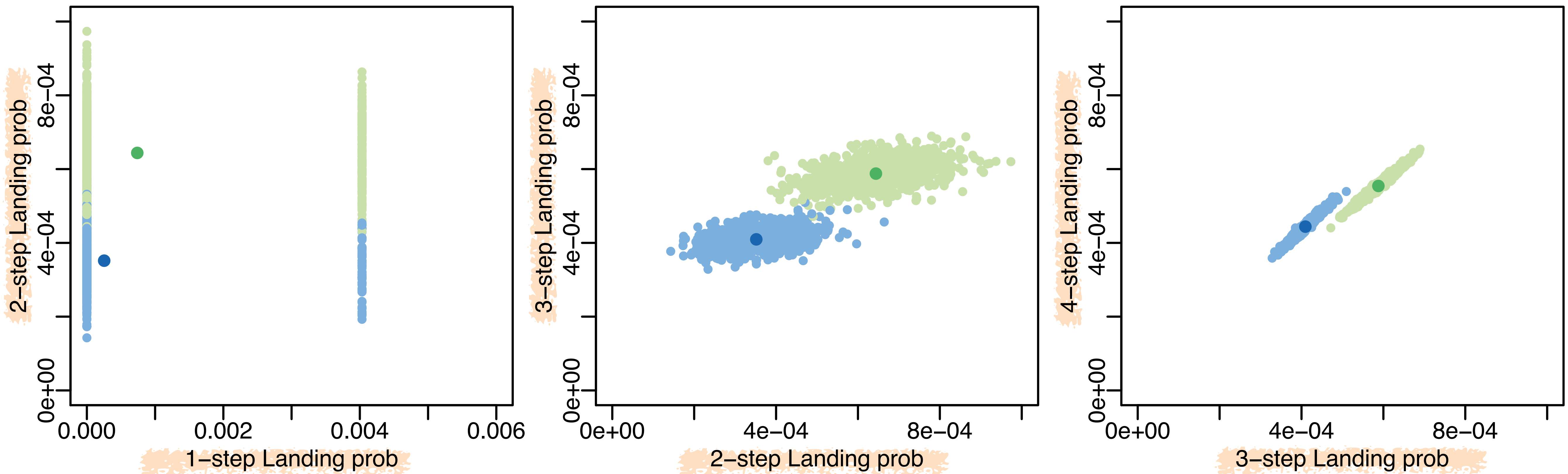
- **Geometric discriminant function:** sweeps through the space of landing probabilities following vector from **b** to **a**.

The space of landing probabilities



- **Fisher discriminant functions:** Clearly exist better **linear** and **quadratic** functions. Forward pointer, will return.

The space of landing probabilities

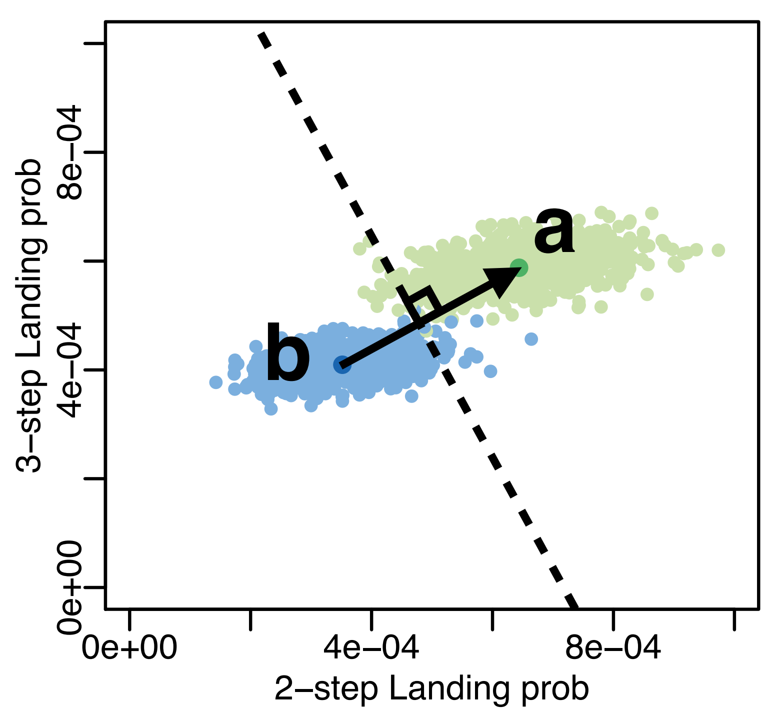


- Focus on deriving optimal **Geometric discriminant function** first.

Geometric discriminant functions

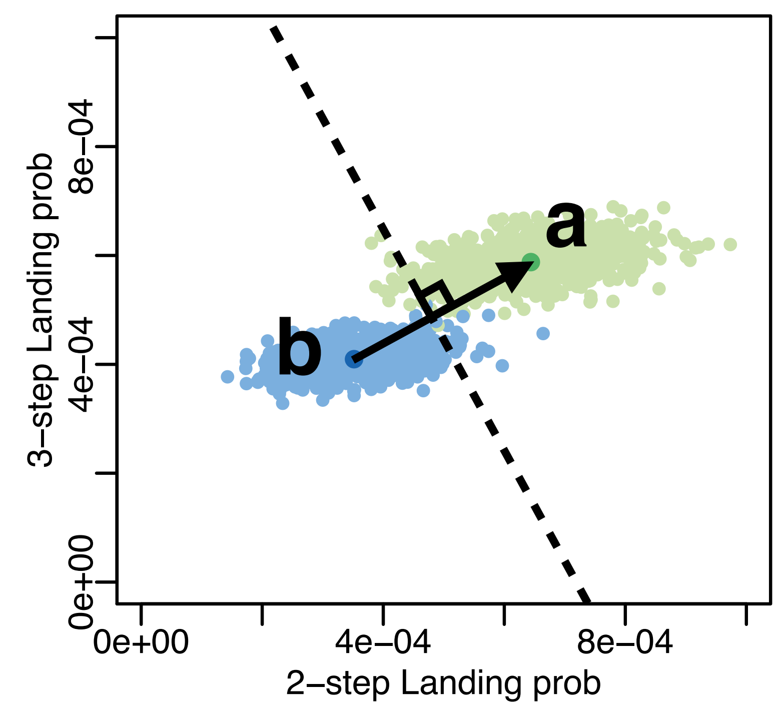
- Let $\mathbf{r} = (r_1, \ldots, r_K)$ be the landing probabilities of a node
- Let $\mathbf{a} = (a_1, \ldots, a_K)$ be the **Target** class centroid
- Let $\mathbf{b} = (b_1, \ldots, b_K)$ be the **Other** class centroid
- Then $f(\mathbf{r}) = (\mathbf{a} - \mathbf{b})^T \mathbf{r}$ is the geometric discriminant function.

- Notice: $f(\mathbf{r})$ increases when **r** moves in direction of **a - b**.
- Can classify nodes based on thresholds of $f(\mathbf{r})$.



Personalized PageRank is “optimal”

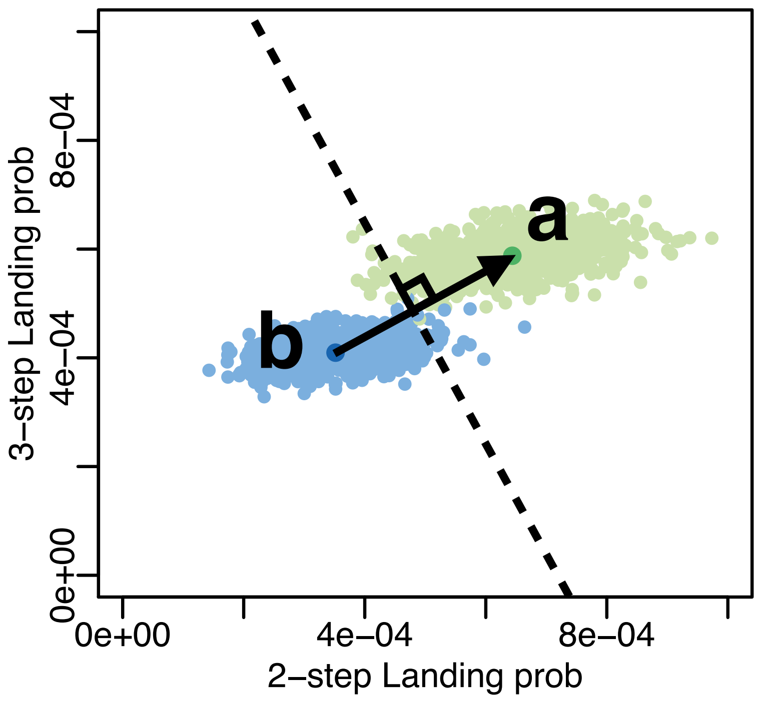
- **Main Theorem (informal version).**
  For 2-block SBM with equal sized blocks and edge densities p_{in}, p_{out}:
  $$a_k - b_k = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right)^k,$$
  and the optimal geometric classifier is therefore: $\sum_{k=1}^{K} (\alpha_*)^k r_k$.
  which is PPR(!) with $\alpha_* = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right)$.



Personalized PageRank is "optimal"

- **Main Theorem (informal version).**
For 2-block SBM with equal sized blocks and edge densities p_{in}, p_{out}:
$$a_k - b_k = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right)^k,$$
and the optimal geometric classifier is therefore: $\sum_{k=1}^{K} (\alpha_*)^k r_k.$
which is PPR(!) with $\alpha_* = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right).$

- **Two main parts:**
  1. Centroids **a, b** concentrate on quantities determined by the solution to a linear recurrence relation.
  2. That linear recurrence relation can be solved and yields PPR.



PPR is “optimal”: Proof idea

- **Part 1: Concentration of landing probabilities**

Lemma 1. For any $\epsilon, \delta > 0$, there is an n sufficiently large such that the random landing probabilities $(\hat{a}_1, ..., \hat{a}_K)$ and $(\hat{b}_1, ..., \hat{b}_K)$ for a uniform random walk on G_n starting in the seed block satisfy the following conditions with probability at least $1 - \delta$ for all $k > 0$:

$$N\hat{a}_k \in \left[(1-\epsilon)\frac{A_k}{A_k + B_k}, (1+\epsilon)\frac{A_k}{A_k + B_k}\right] \text{ and} \tag{1}$$

$$N\hat{b}_k \in \left[(1-\epsilon)\frac{B_k}{A_k + B_k}, (1+\epsilon)\frac{B_k}{A_k + B_k}\right], \tag{2}$$

where A_k, B_k are the solutions to the matrix recurrence relation

$$\begin{cases} A_k = N(p_{in}A_{k-1} + p_{out}B_{k-1}) \\ B_k = N(p_{out}A_{k-1} + p_{in}B_{k-1}), \end{cases}$$

with $A_0 = 1$, $B_0 = 0$.

PPR is “optimal”: Proof idea

- **Part 1: Concentration of landing probabilities**

Lemma 1. For any $\epsilon, \delta > 0$, there is an n sufficiently large such that the random landing probabilities $(\hat{a}_1, ..., \hat{a}_K)$ and $(\hat{b}_1, ..., \hat{b}_K)$ for a uniform random walk on G_n starting in the seed block satisfy the following conditions with probability at least $1 - \delta$ for all $k > 0$:

$$N\hat{a}_k \in \left[(1-\epsilon)\frac{A_k}{A_k + B_k}, (1+\epsilon)\frac{A_k}{A_k + B_k}\right] \text{ and} \tag{1}$$

$$N\hat{b}_k \in \left[(1-\epsilon)\frac{B_k}{A_k + B_k}, (1+\epsilon)\frac{B_k}{A_k + B_k}\right], \tag{2}$$

where A_k, B_k are the solutions to the matrix recurrence relation

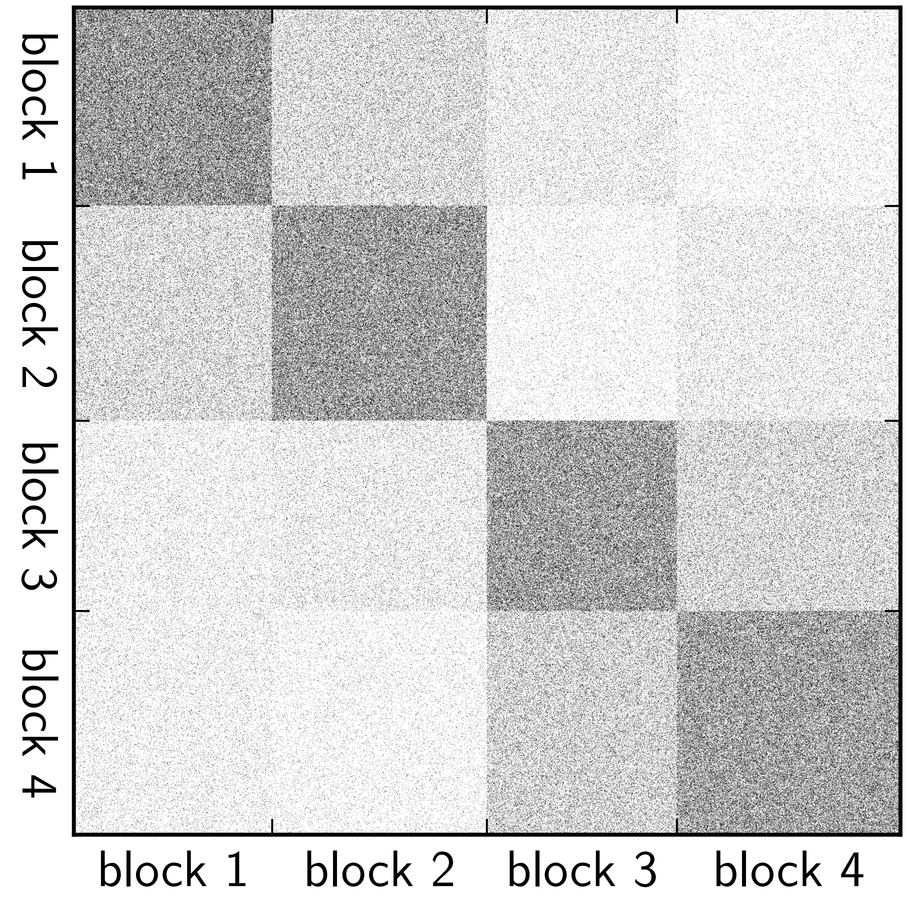
$$\boxed{\begin{cases} A_k = N(p_{in}A_{k-1} + p_{out}B_{k-1}) \\ B_k = N(p_{out}A_{k-1} + p_{in}B_{k-1}), \end{cases}}$$

with $\boxed{A_0 = 1,\ B_0 = 0.}$

- A_k, B_k interpretable as length-k walk count to nodes in block 1 vs. 2.
- For large **n**, block walk counts increase by factors of ~E[degree].

More general SBMs

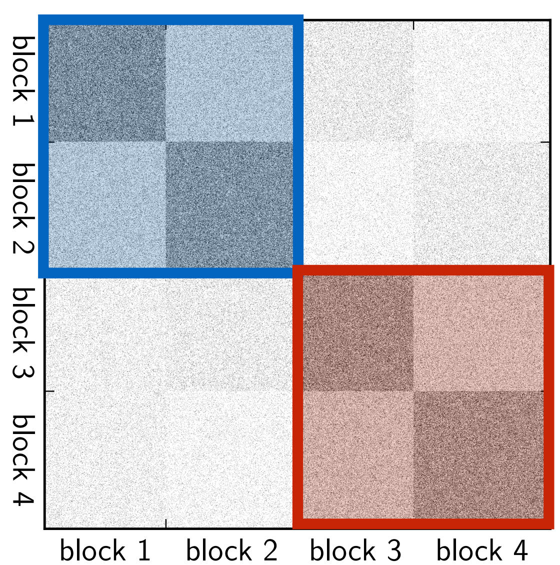
- For SBMs with **C>2** blocks and/or with arbitrary **P:**
  - Seed set expansion asks: identify nodes in a **target block set**.
  - With conditions on equal expected degrees, PPR(!).
  - Without conditions, still:
    - Asymptotically optimal weights for geometric classification still **obtainable from solutions to a matrix recurrence relation**.

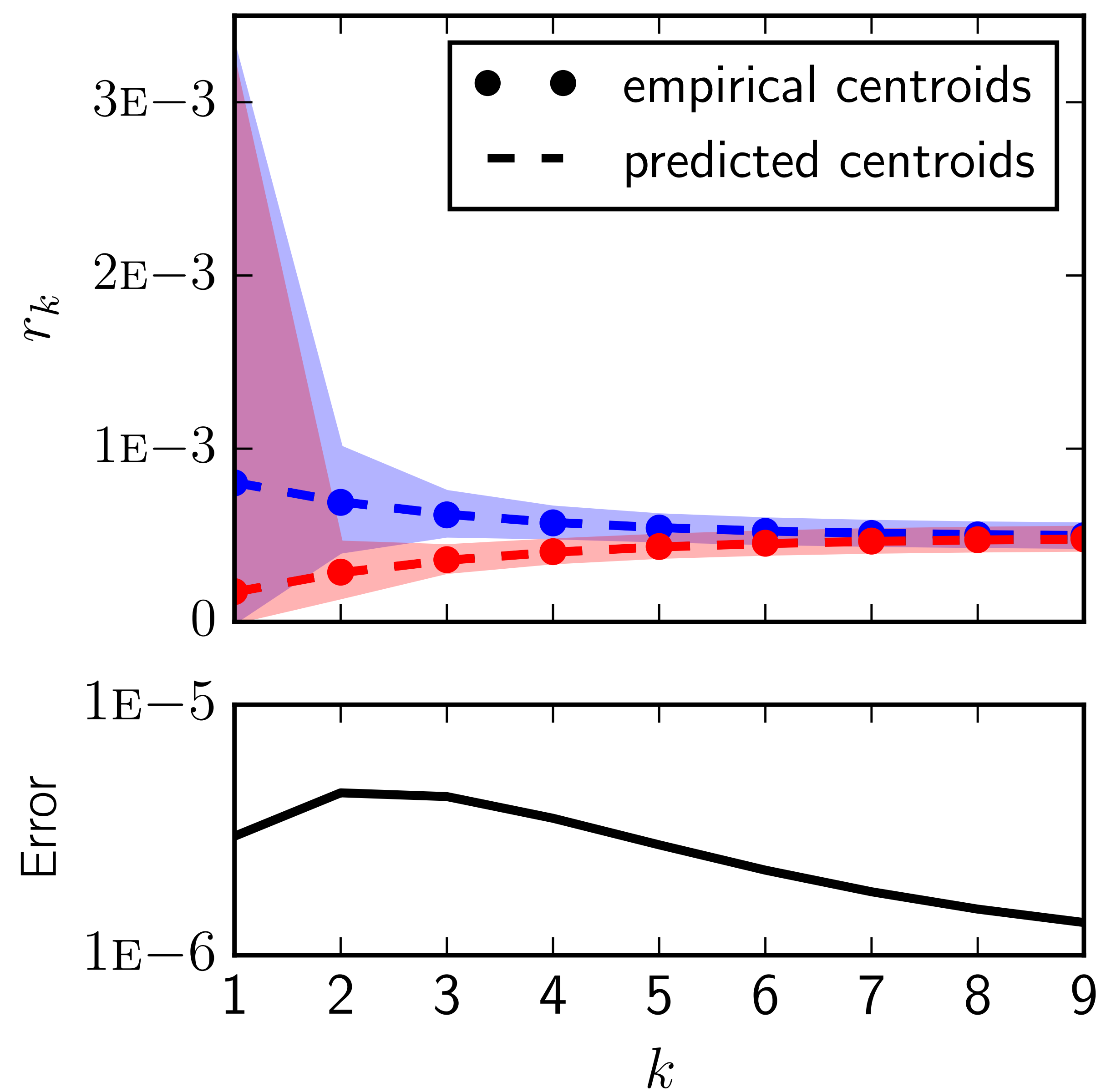


Empirical vs. theoretical centroids

- 2048-node, 4-block SBM, empirical class centroids vs. theory:

- **a, Target blocks**
- **b, Other blocks**

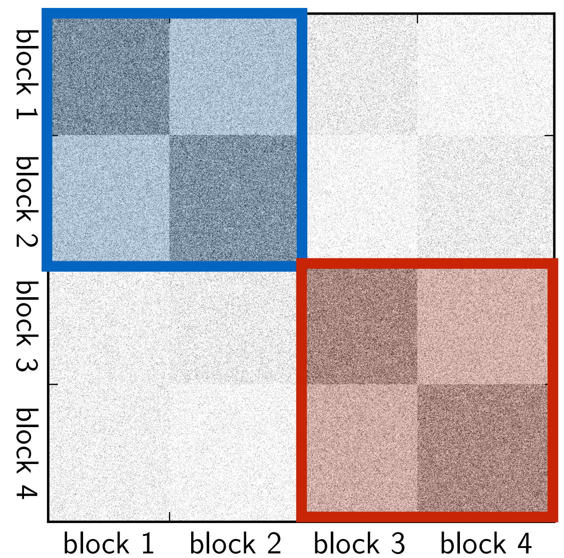


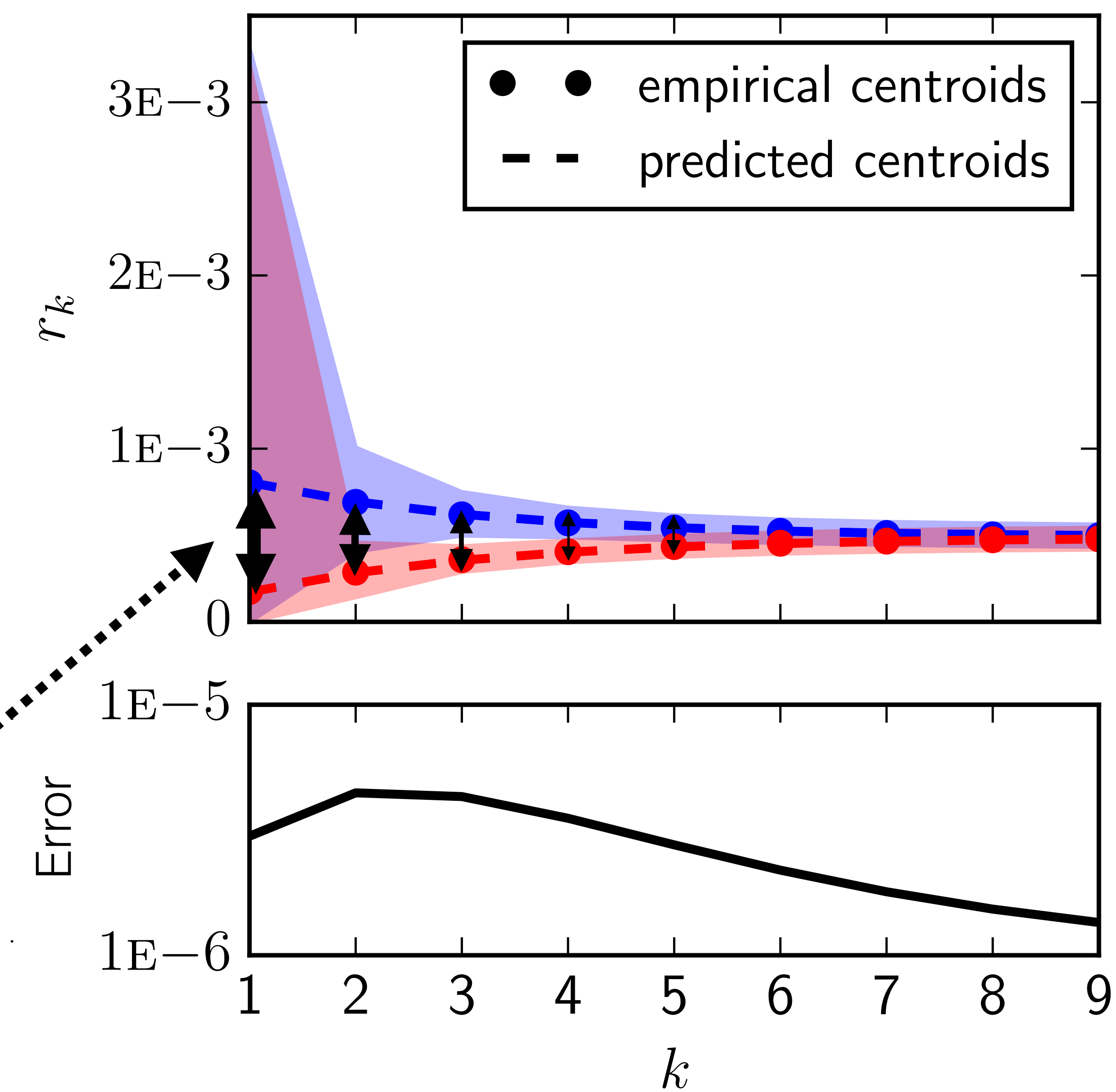


Empirical vs. theoretical centroids

- 2048-node, 4-block SBM, empirical class centroids vs. theory:

- **a, Target blocks**
- **b, Other blocks**

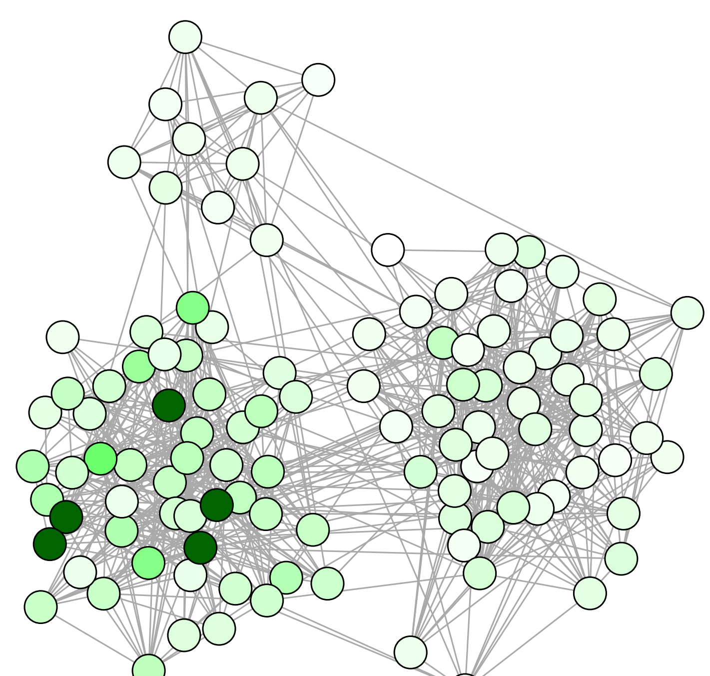




From matrix recurrence relation

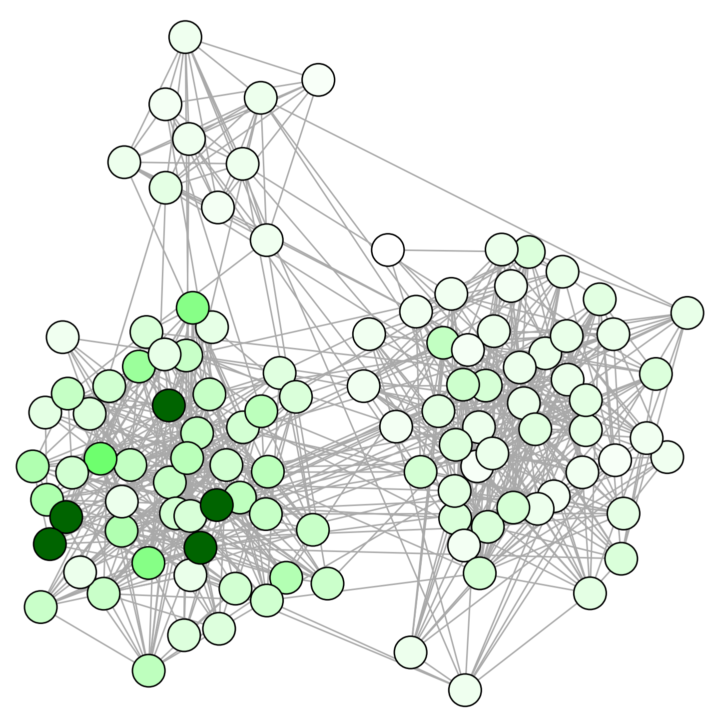
Theories of graph diffusion

- Other motivations for PPR:
  - Random Surfer Model (Brin-Page ’98)
  - Cheeger inequalities for PPR, HK (Andersen et al ’06, Chung ’09)
  - Local spectral algorithm with regularization (Mahoney et al. ’12)
- Our work shows PPR can be derived as “optimal” geometric classifier.
- Also motivates how to choose PPR α, as $\alpha = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right)$.

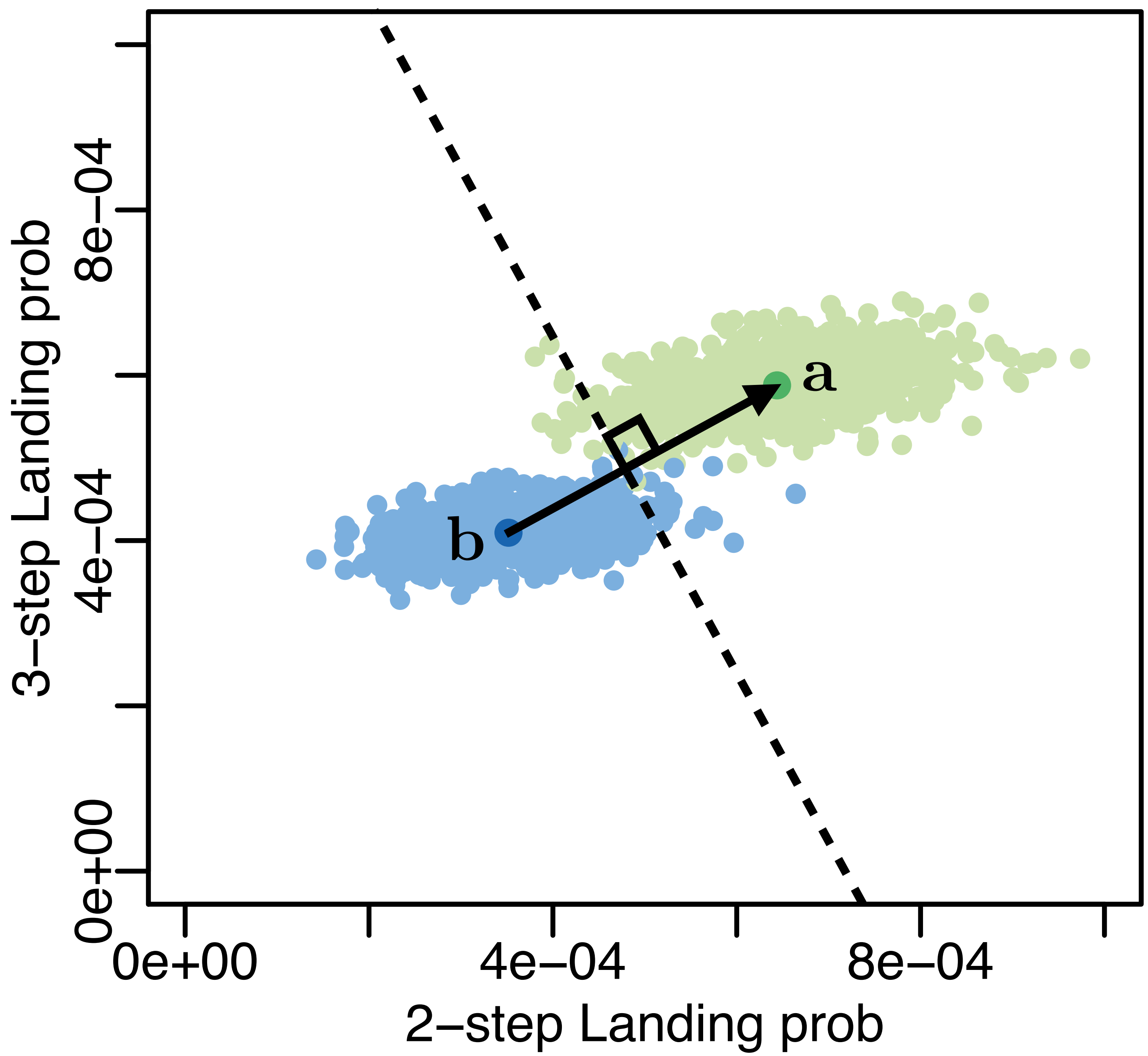


Theories of graph diffusion

- Other motivations for PPR:
  - Random Surfer Model (Brin-Page ’98)
  - Cheeger inequalities for PPR, HK (Andersen et al ’06, Chung ’09)
  - Local spectral algorithm with regularization (Mahoney et al. ’12)
- Our work shows PPR can be derived as “optimal” geometric classifier.
- Also motivates how to choose PPR α, as $\alpha = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right)$.
- **Most importantly: also opens door to methods beyond PPR.**

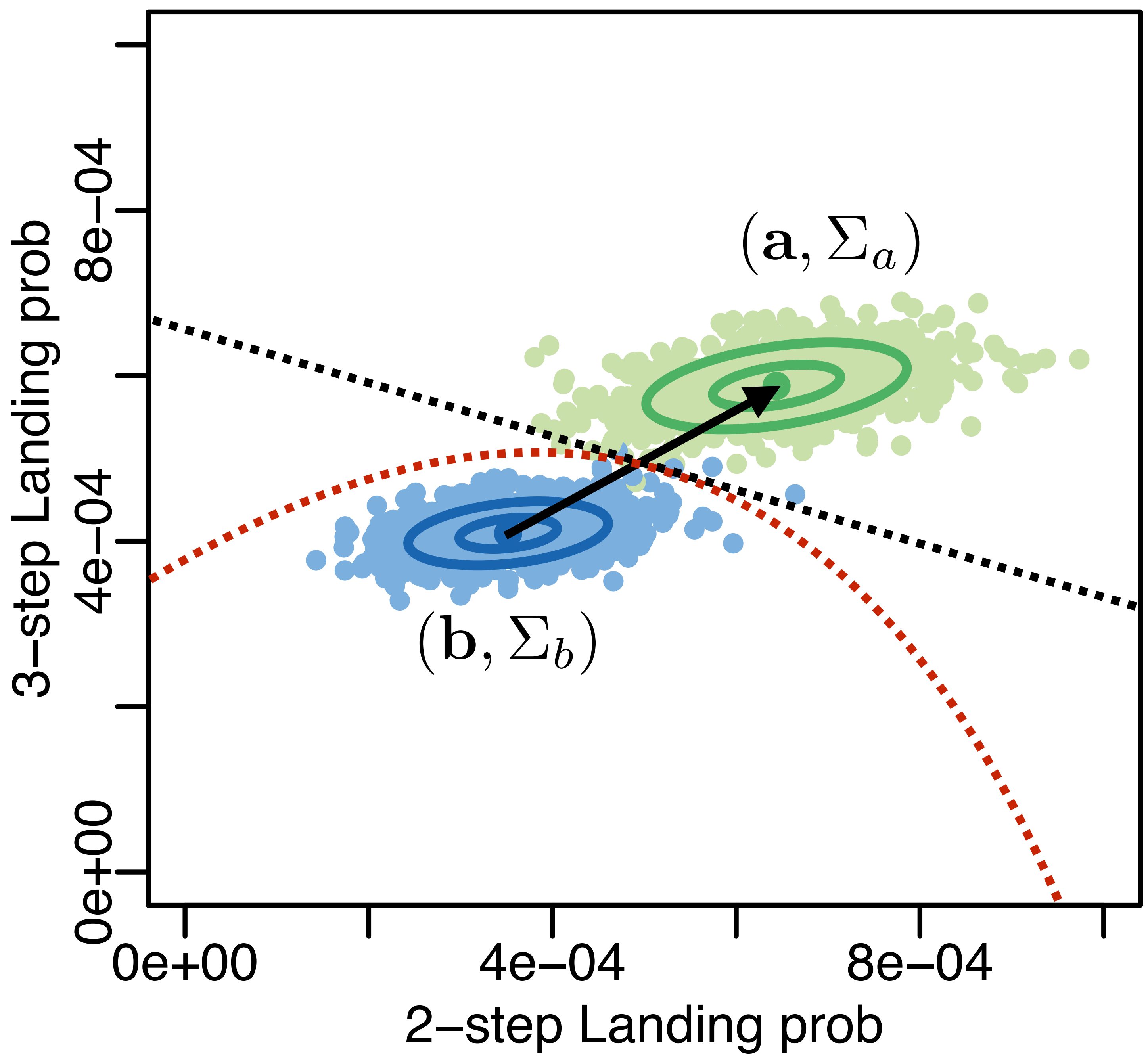


PPR is "optimal" in a narrow sense



- Discriminant functions that model higher moments of point clouds?

Fisher discriminant functions



- Discriminant functions that model higher moments of point clouds.

Fisher discriminant functions

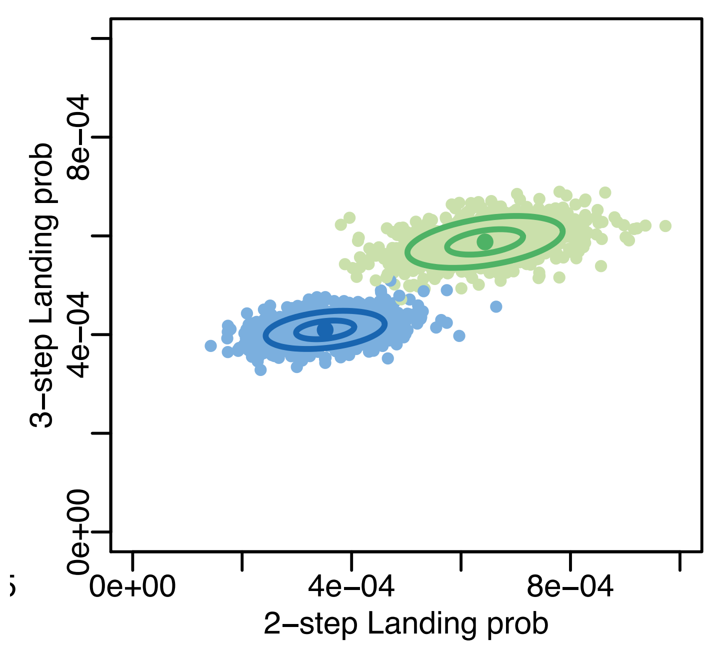
- Let **z** be the latent class of each node.
- Capture (mean, variance) of class point clouds:

$$\Pr(\mathbf{r}|z=1) \propto |\Sigma_a|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{r}-\mathbf{a})^T \Sigma_a^{-1} (\mathbf{r}-\mathbf{a})\right)$$

$$\Pr(\mathbf{r}|z=0) \propto |\Sigma_b|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{r}-\mathbf{b})^T \Sigma_b^{-1} (\mathbf{r}-\mathbf{b})\right)$$

- Log-likelihood ratio as discriminant function:

$$g(\mathbf{r}) = \log \frac{\Pr(\mathbf{r}|z=1)\Pr(z=1)}{\Pr(\mathbf{r}|z=0)\Pr(z=0)}$$



Fisher discriminant functions

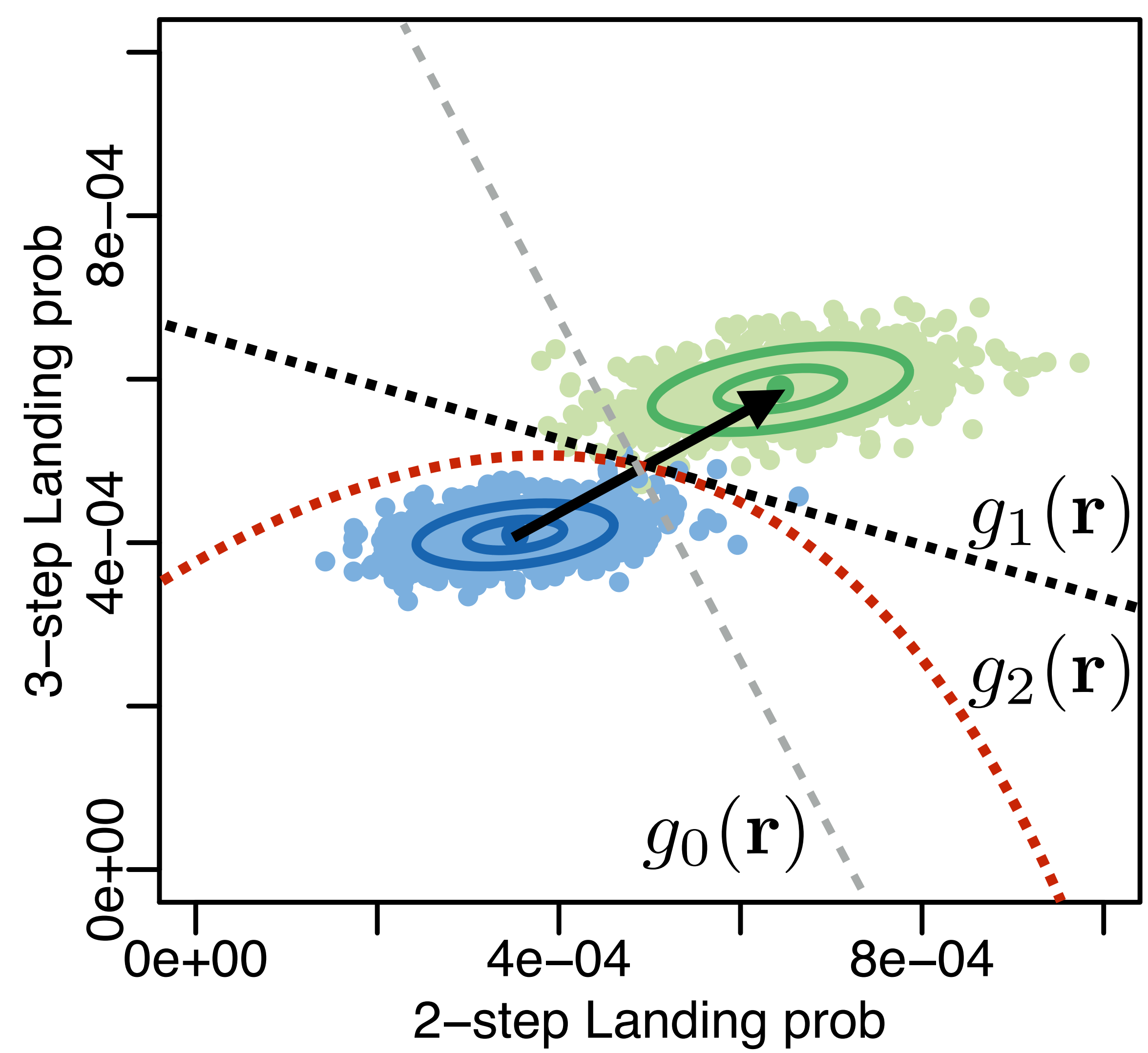
- **Three approaches:**

$$\text{General}: \quad g_2(\mathbf{r}) \propto \left(\Sigma_a^{-1}\mathbf{a} - \Sigma_b^{-1}\mathbf{b}\right)^T \mathbf{r} + \tfrac{1}{2}\mathbf{r}^T \left(\Sigma_b^{-1} - \Sigma_a^{-1}\right)\mathbf{r}$$

$$\text{Assume } \Sigma_a = \Sigma_b = \Sigma: \quad g_1(\mathbf{r}) \propto \Sigma^{-1}(\mathbf{a} - \mathbf{b})^T \mathbf{r}$$

$$\text{Assume } \Sigma_a = \Sigma_b = I: \quad g_0(\mathbf{r}) \propto (\mathbf{a} - \mathbf{b})^T \mathbf{r}$$

- We call the first two methods **QuadSBMRank**, **LinSBMRank**.
- Perhaps reasonable to assume equal covariances; effective.
- PPR follows from an assumption of uniform variance, no covariance.



Fisher discriminant functions

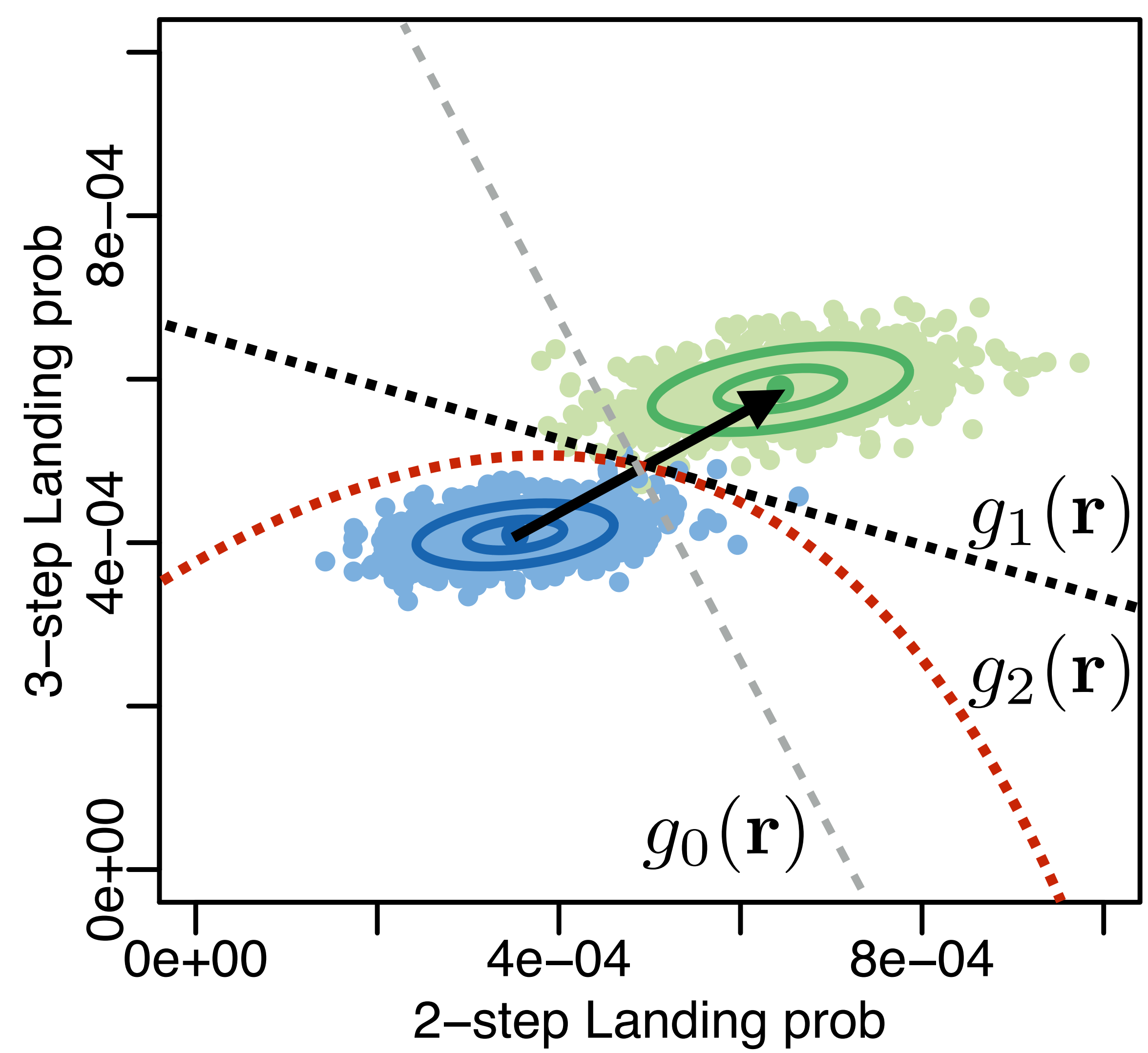
- **Three approaches:**

$$\text{General}: \quad g_2(\mathbf{r}) \propto \left(\Sigma_a^{-1}\mathbf{a} - \Sigma_b^{-1}\mathbf{b}\right)^T \mathbf{r} + \tfrac{1}{2}\mathbf{r}^T \left(\Sigma_b^{-1} - \Sigma_a^{-1}\right)\mathbf{r}$$

$$\text{Assume } \Sigma_a = \Sigma_b = \Sigma: \quad g_1(\mathbf{r}) \propto \Sigma^{-1}(\mathbf{a} - \mathbf{b})^T \mathbf{r}$$

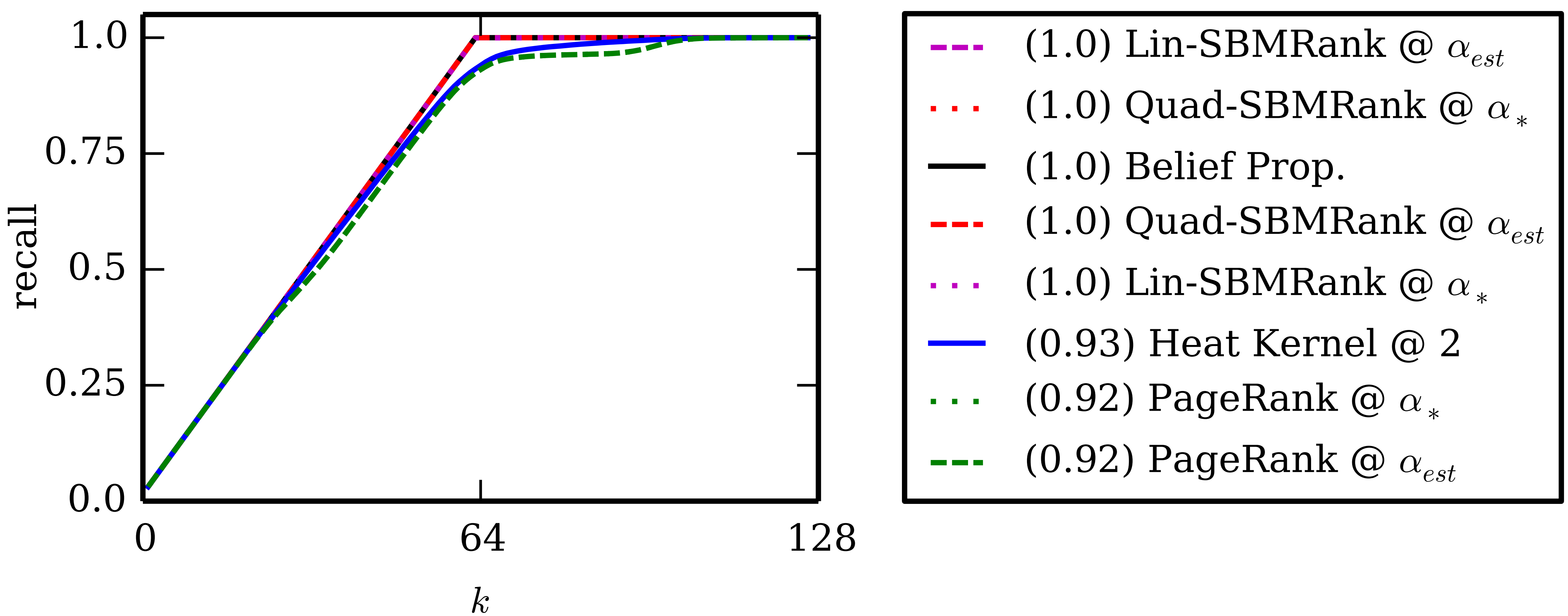
$$\text{Assume } \Sigma_a = \Sigma_b = I: \quad g_0(\mathbf{r}) \propto (\mathbf{a} - \mathbf{b})^T \mathbf{r}$$

- We call the first two methods **QuadSBMRank**, **LinSBMRank**.
- Perhaps reasonable to assume equal covariances; effective.
- PPR follows from an assumption of uniform variance, no covariance.
- **Open challenge:** Possible to show asymptotic normality and characterize covariance matrices?



Evaluation: recall curves

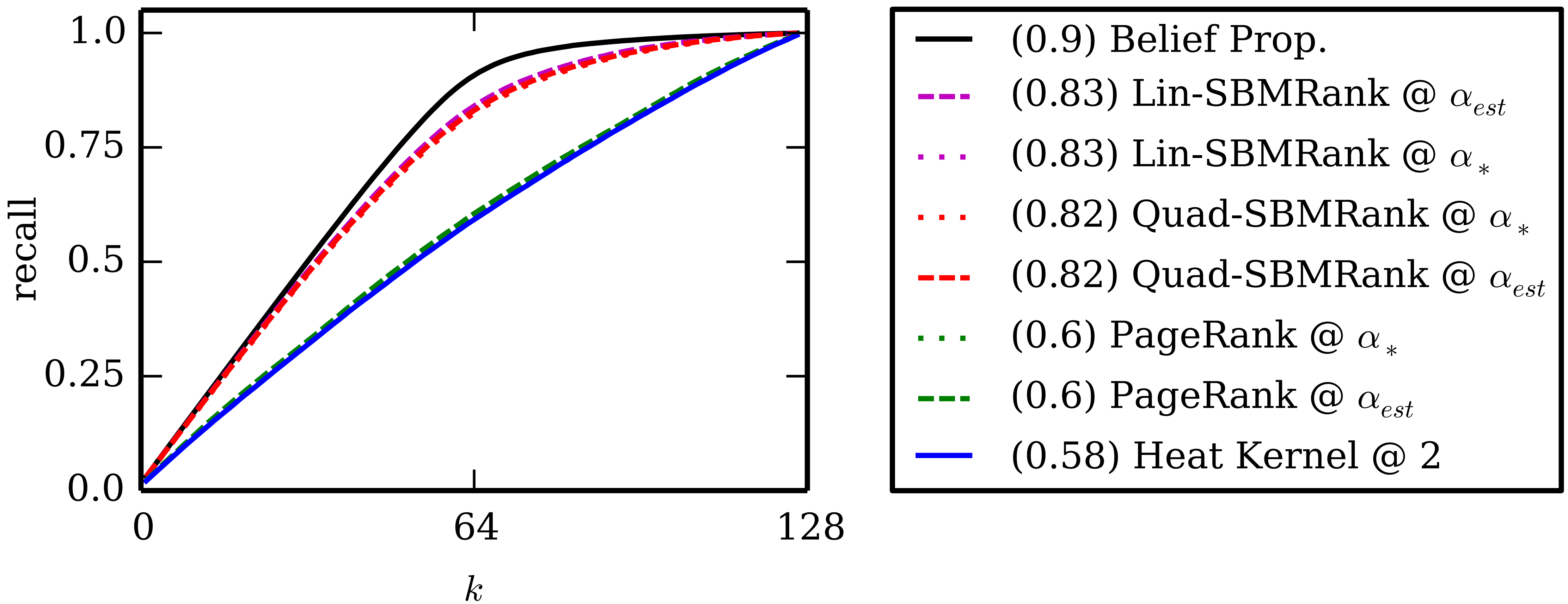
- SBM with 2 blocks, 64 nodes/block, 1 seed node.
- Recall that Belief Propagation reaches resolution limit.



- Easy instance ($p_{in} >> p_{out}$):
  - Everything does well.

Evaluation: recall curves

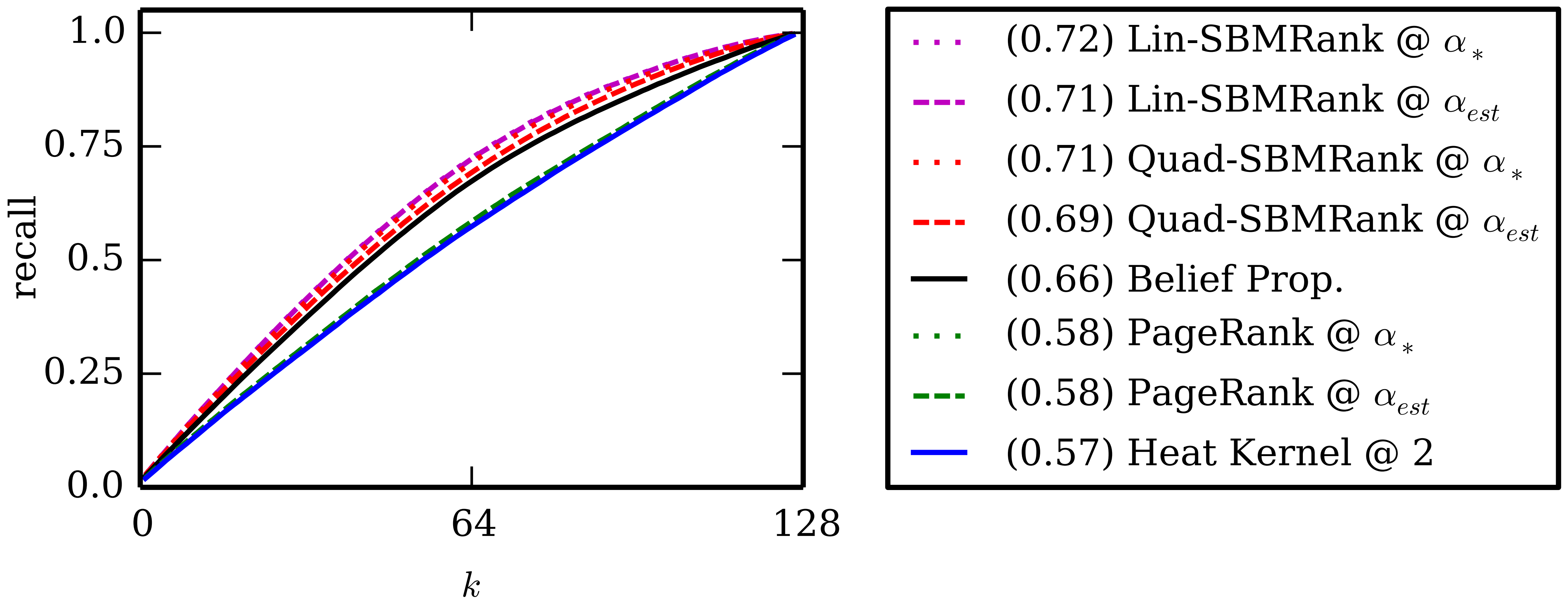
- SBM with 2 blocks, 64 nodes/block, 1 seed node.
- Recall that Belief Propagation reaches resolution limit.



- Hard instance…
  - PPR/HK lost all recall, LinSBMRank and QuadSBMRank near BP.

Evaluation: recall curves

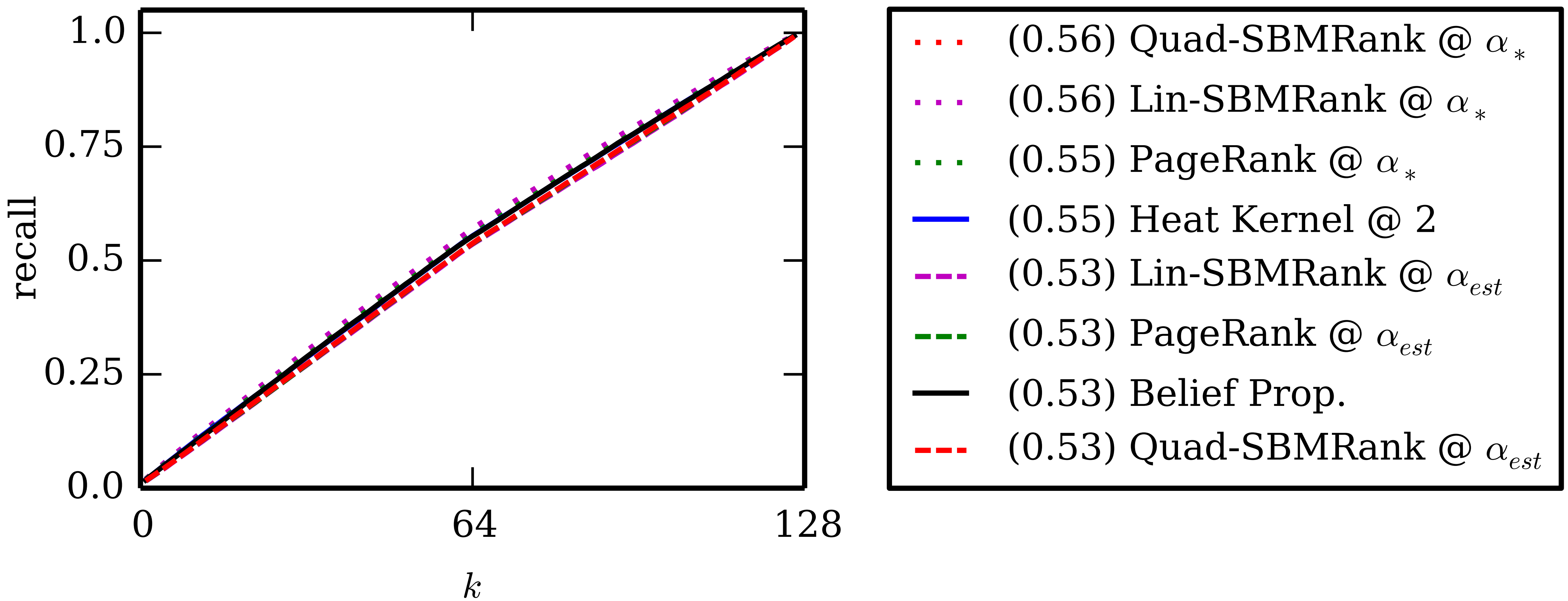
- SBM with 2 blocks, 64 nodes/block, 1 seed node.
- Recall that Belief Propagation reaches resolution limit.



- **Even** harder instance…
  - LinSBMRank and QuadSBMRank outperforming BP by a hair…?

Evaluation: recall curves

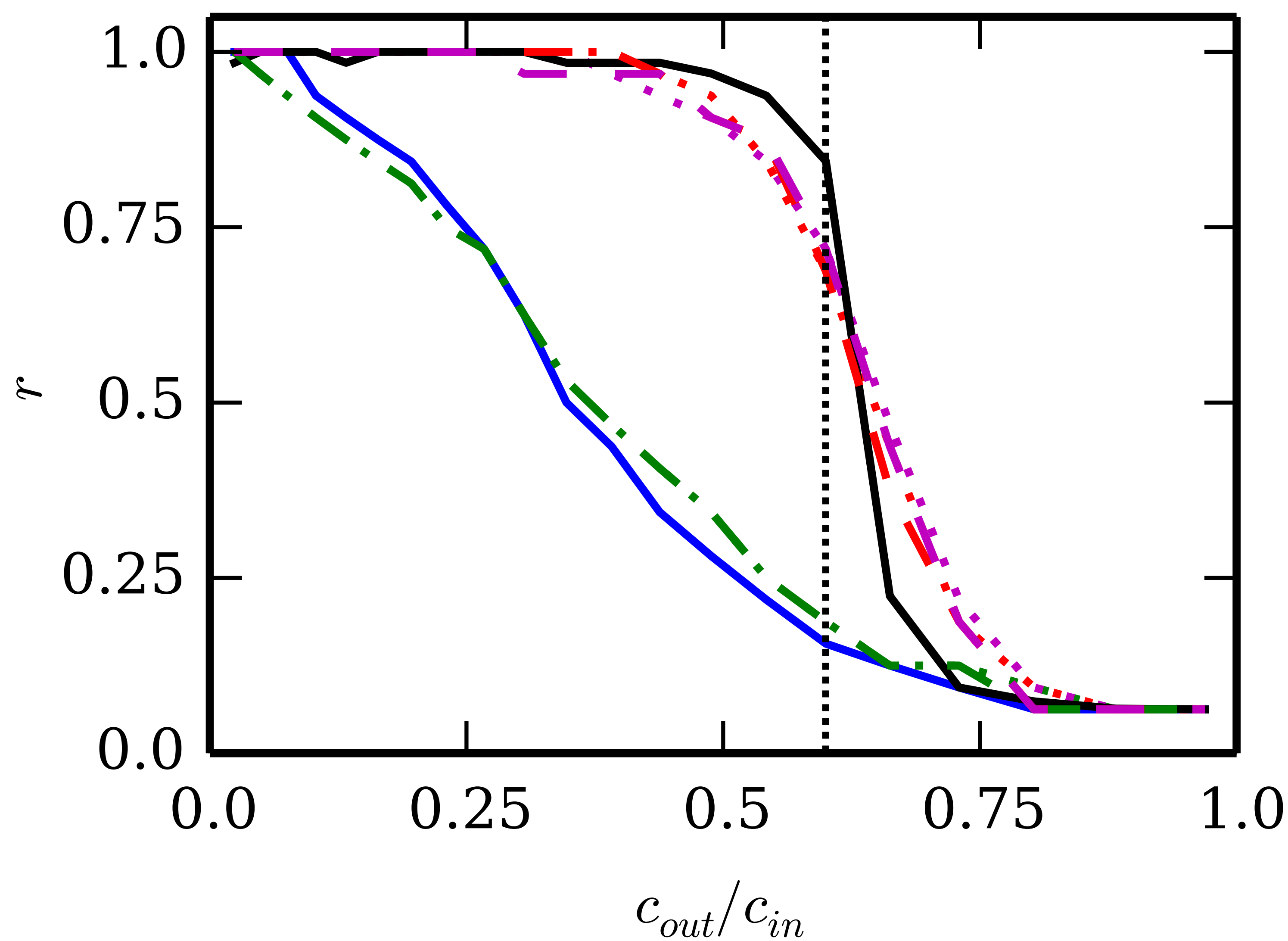
- SBM with 2 blocks, 64 nodes/block, 1 seed node.
- Recall that Belief Propagation reaches resolution limit.



- Impossible ($p_{in} = p_{out}$):
  - Nothing works.

Evaluation: resolution limit

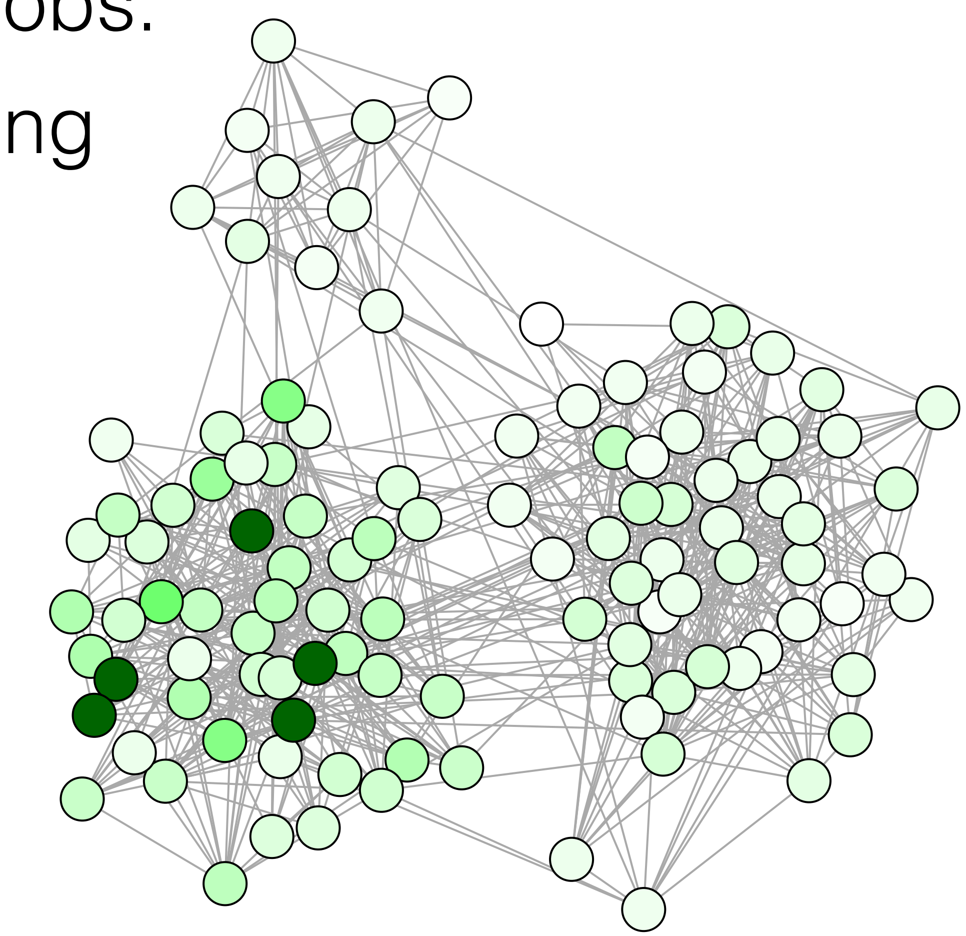
- Pearson correlation **r** between true partition and inferred partition.
- Empirically, we see LinSBMRank and QuadSBMRank get very close to resolution limit (dotted line), with slower decay rate.



PPR, **HK**, **LinSBMRank**, **QuadSBMRank**, **BP**

Conclusions

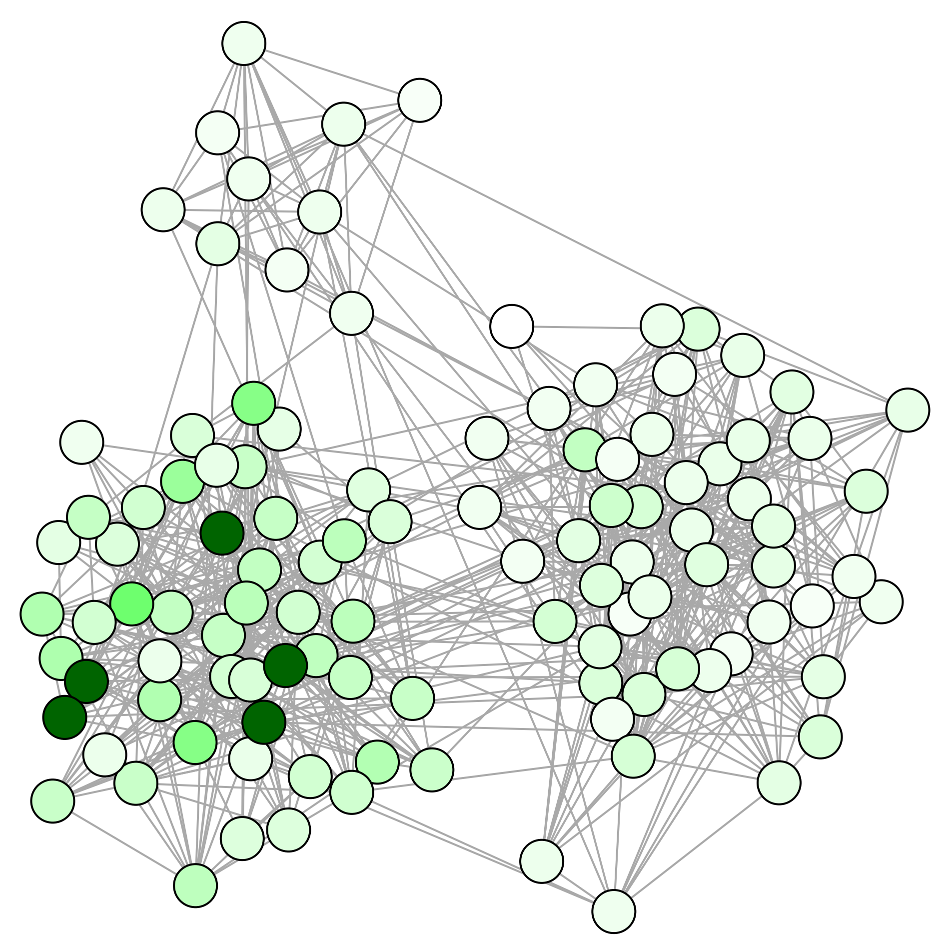
- Personalized PageRank with $\alpha = \left(\frac{p_{in} - p_{out}}{p_{in} + p_{out}}\right)$ is optimal geometric discriminant function for balanced 2-block SBM.
- Geometric discriminant functions for more general block models follow from recurrence relation.
- Landing probabilities are correlated; correcting for higher moments in the space of landing probabilities greatly improves classification.
- In practice: fit GMMs in space of landing probs.
- A new perspective on diffusion-based ranking that can hopefully open new doors.



- Pre-print:
  Isabel Kloumann, Johan Ugander, Jon Kleinberg
  "Block Models and Personalized PageRank"
  arXiv:1607.03483

Open directions

- Model covariance of landing probabilities?
- Currently requires at least ~logarithmic degrees (we think); possible to derive weights for bounded degree SBMs?
- Better classifiers in the space of landing probabilities for other random walks? (Non-backtracking, etc.)
- Not just SBM? Optimal weights for dcSBM, core-periphery, Hoff latent space model, etc, etc.
- Slow decay beyond resolution limit?



- Pre-print:
  Isabel Kloumann, Johan Ugander, Jon Kleinberg
  "Block Models and Personalized PageRank"
  arXiv:1607.03483