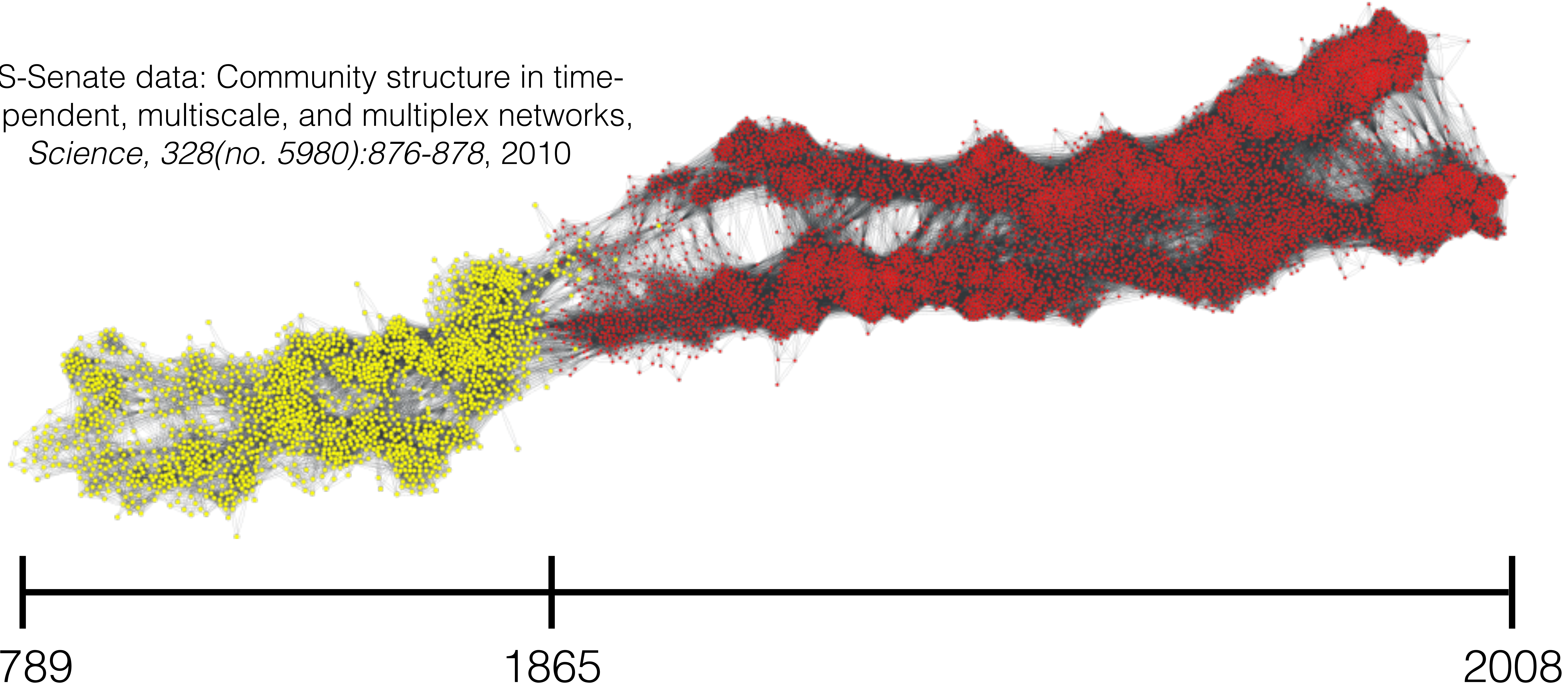
Variational perspective on local graph clustering

Kimon Fountoulakis

Postdoctoral Fellow at UC Berkeley, International Computer Science Institute
Assistant Professor at CS Waterloo

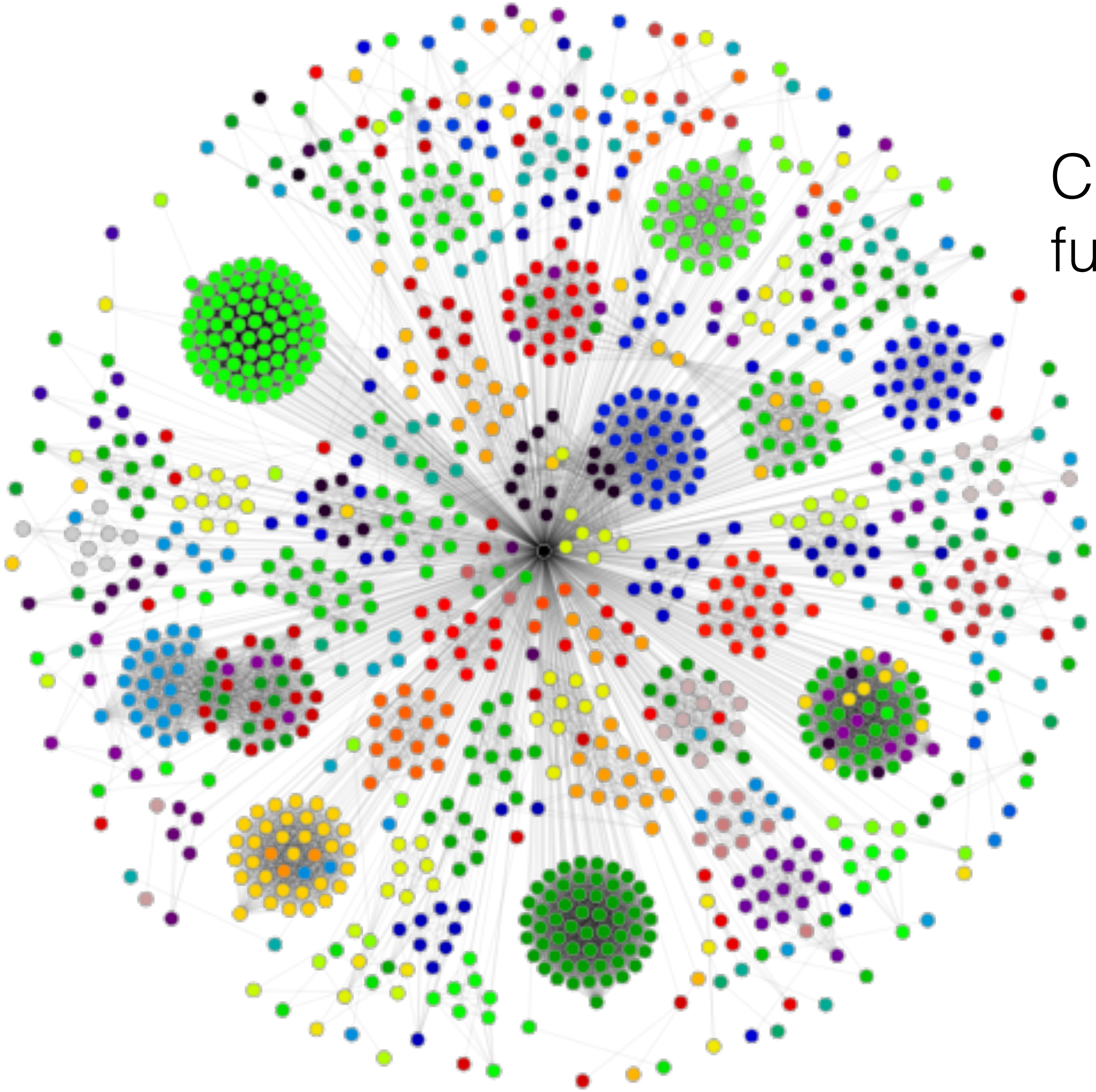
Past and present studies focus on **global** trends of the data

US-Senate data: Community structure in time-dependent, multiscale, and multiplex networks, *Science, 328(no. 5980):876-878*, 2010



The American civil war ended in 1865

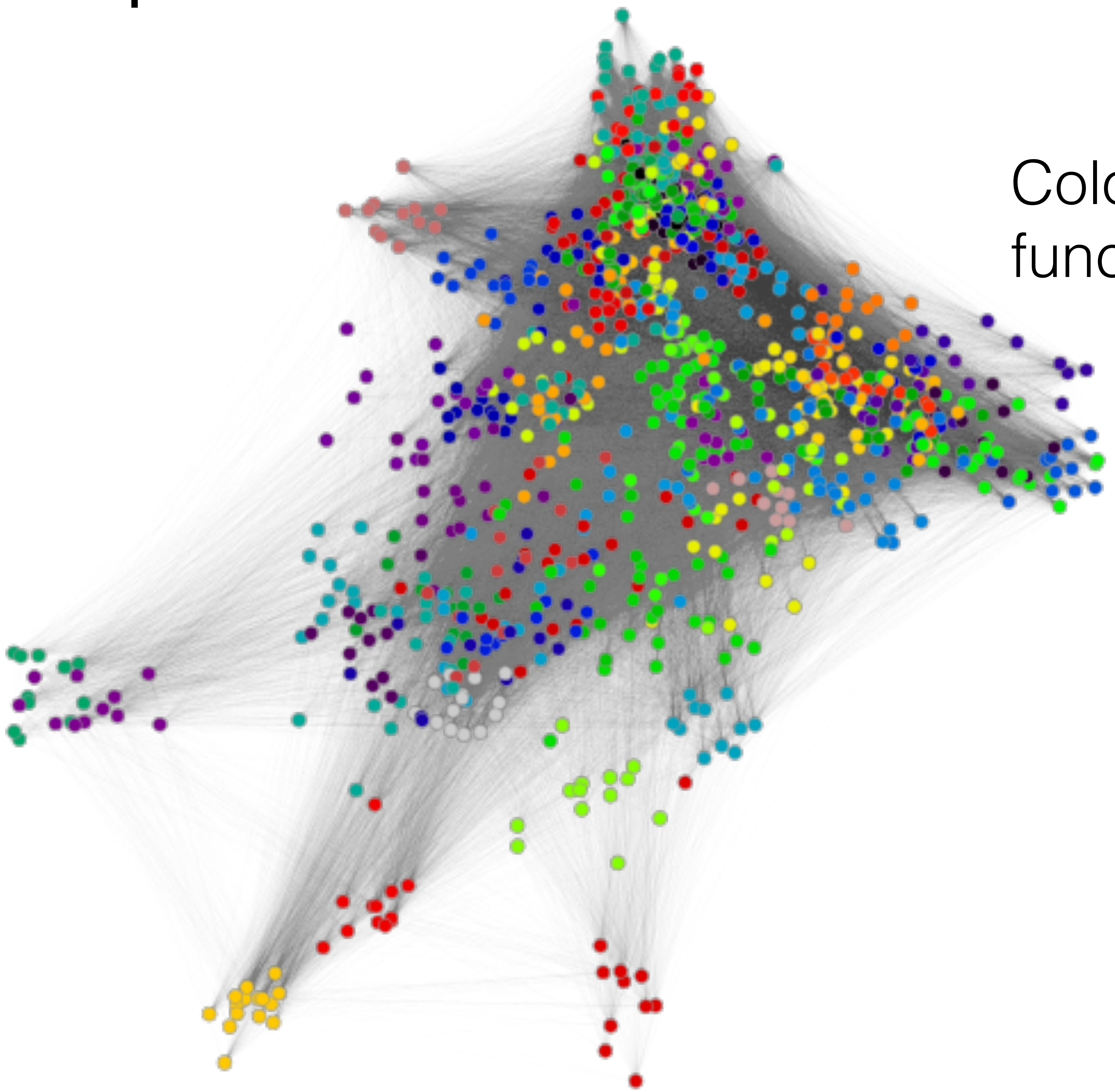
But, most real data have rich **local** structure



Color denotes similar function

Data: The MIPS mammalian protein-protein interaction database. *Bioinformatics, 21(6):832-834*, 2005

And can be very complex



Color denotes similar function

Data: The MIPS mammalian protein-protein interaction database. *Bioinformatics, 21(6):832-834*, 2005

Outline

1. Local graph clustering, definition, examples
2. Example of a state-of-the-art method
3. Variational model
4. Proximal gradient descent

What is local graph clustering and why is it useful?

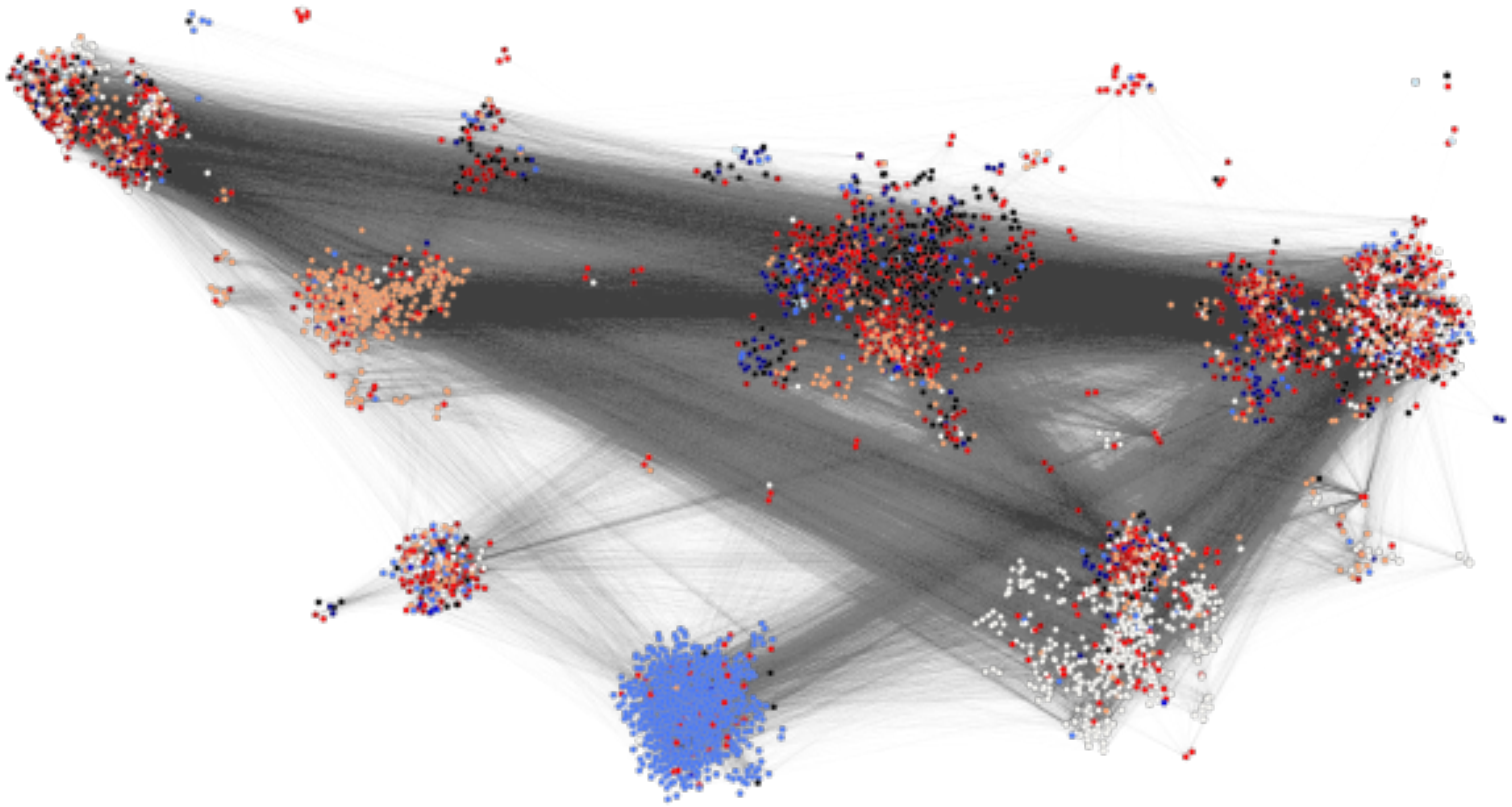
-Definition: find set of nodes A given a seed node in set B

- -Set A has good precision/recall w.r.t set B
- -The running time depends on A instead of the whole graph

-Scalable to graphs with billions of edges

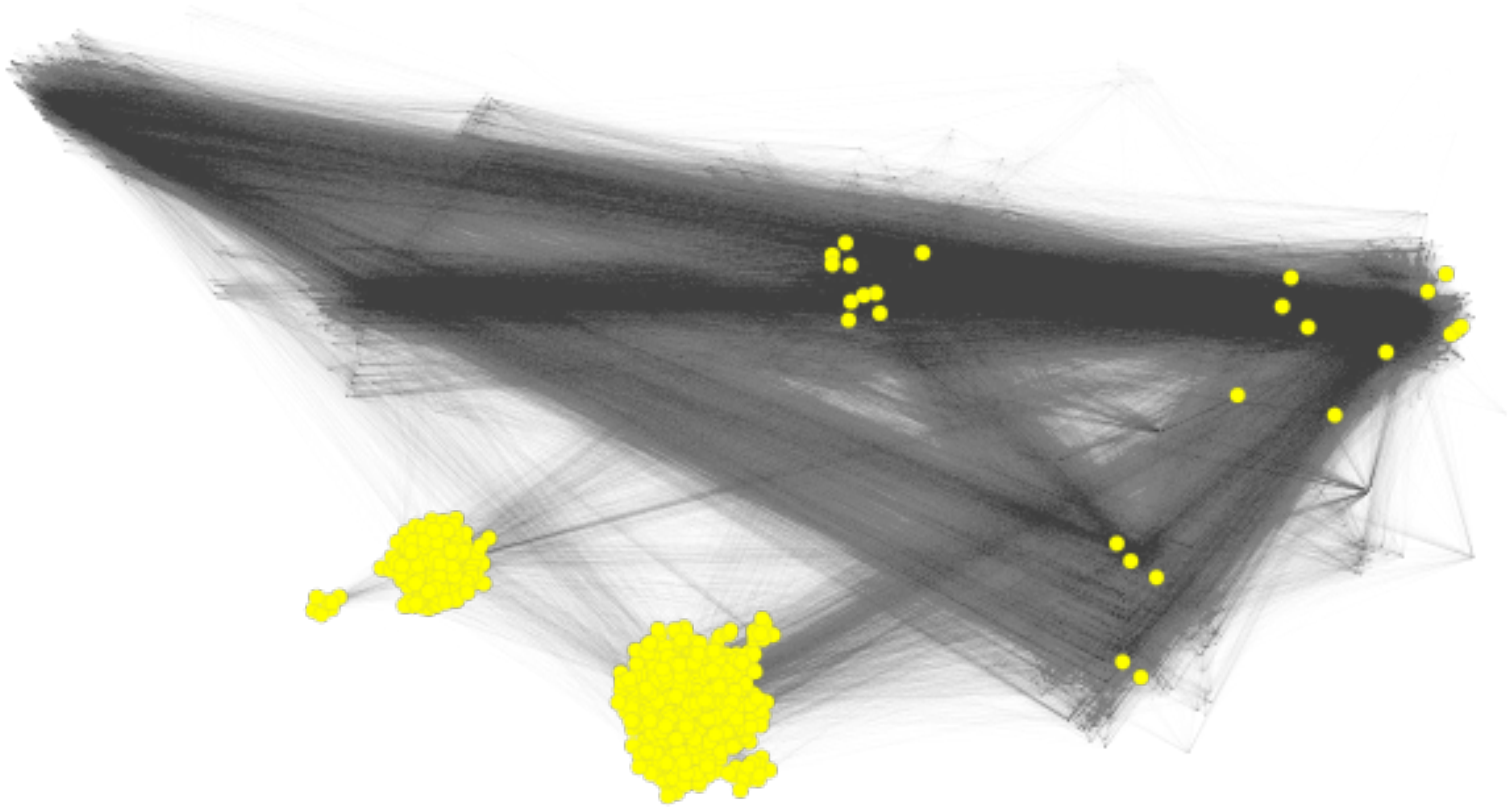
-Ideal for finding small clusters and small neighborhoods

Facebook social network: colour denotes class year



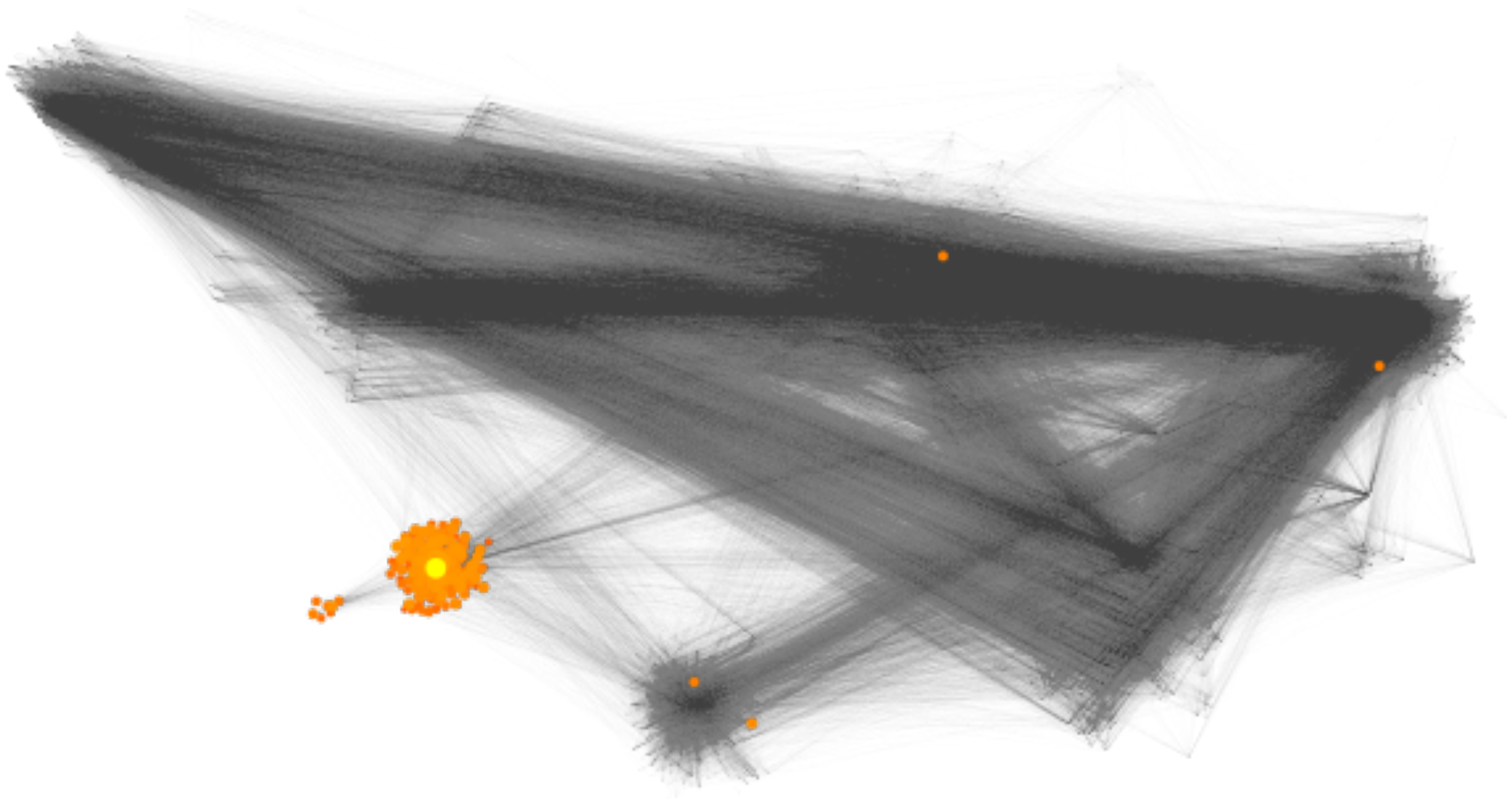
Data: Facebook John Hopkins, A. L. Traud, P. J. Mucha and M. A. Porter, Physica A, 391(16), 2012

Global spectral: finds 20% of the graph



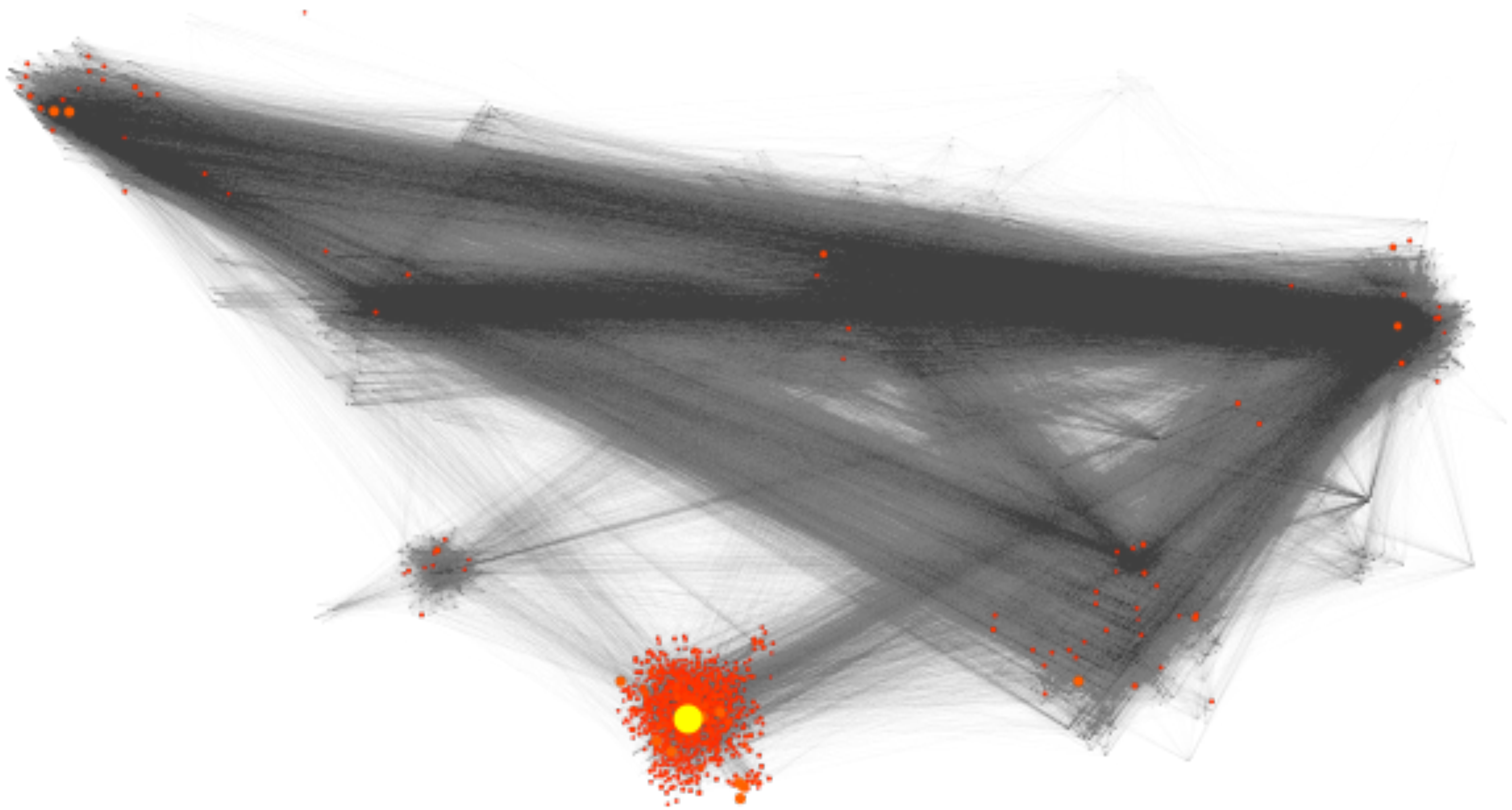
Data: Facebook John Hopkins, A. L. Traud, P. J. Mucha and M. A. Porter, Physica A, 391(16), 2012

Local graph clustering: finds 3% of the graph



Data: Facebook John Hopkins, A. L. Traud, P. J. Mucha and M. A. Porter, Physica A, 391(16), 2012

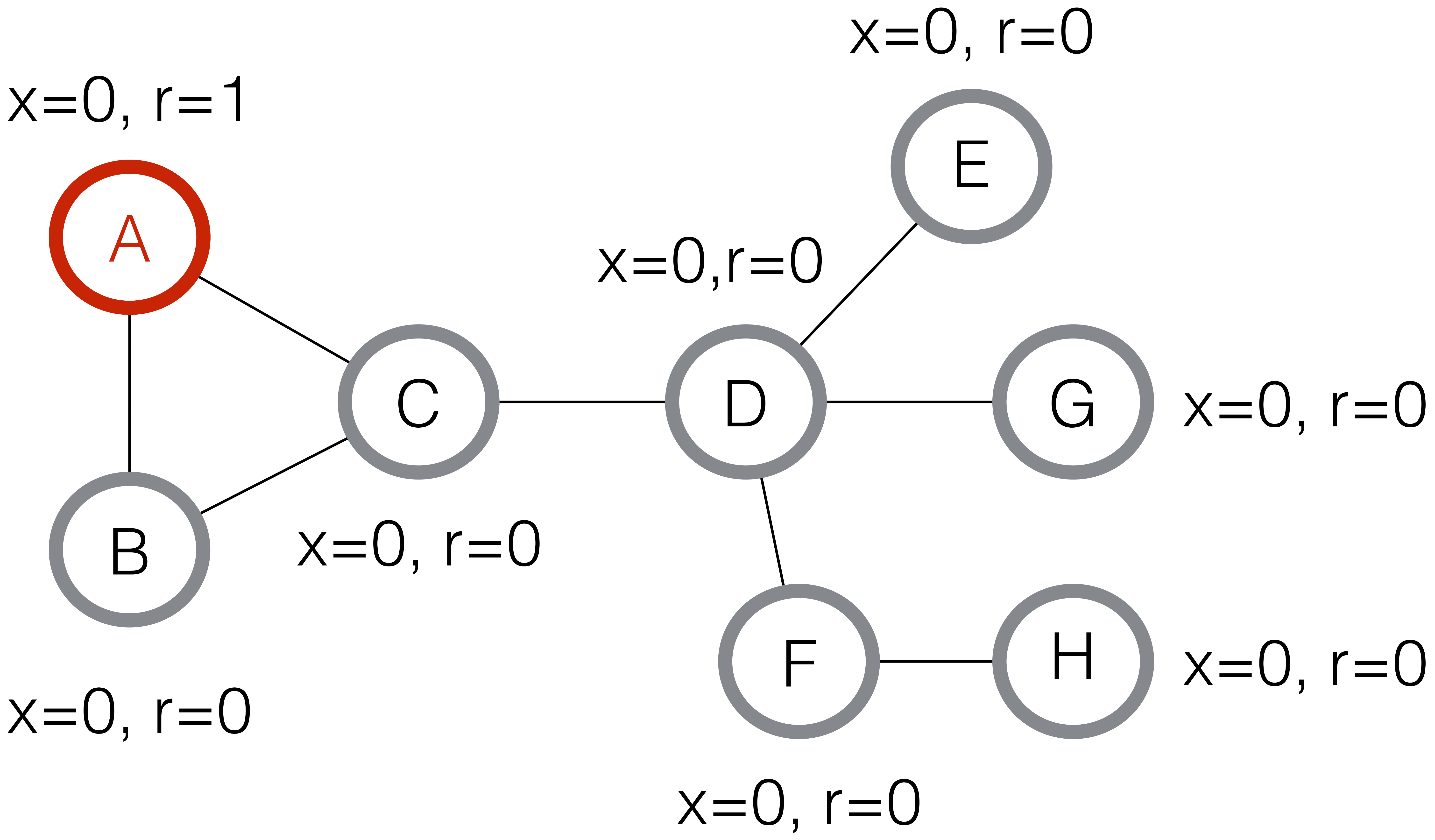
Local graph clustering: finds 17% of the graph



Data: Facebook John Hopkins, A. L. Traud, P. J. Mucha and M. A. Porter, Physica A, 391(16), 2012

Approximate Personalized PageRank

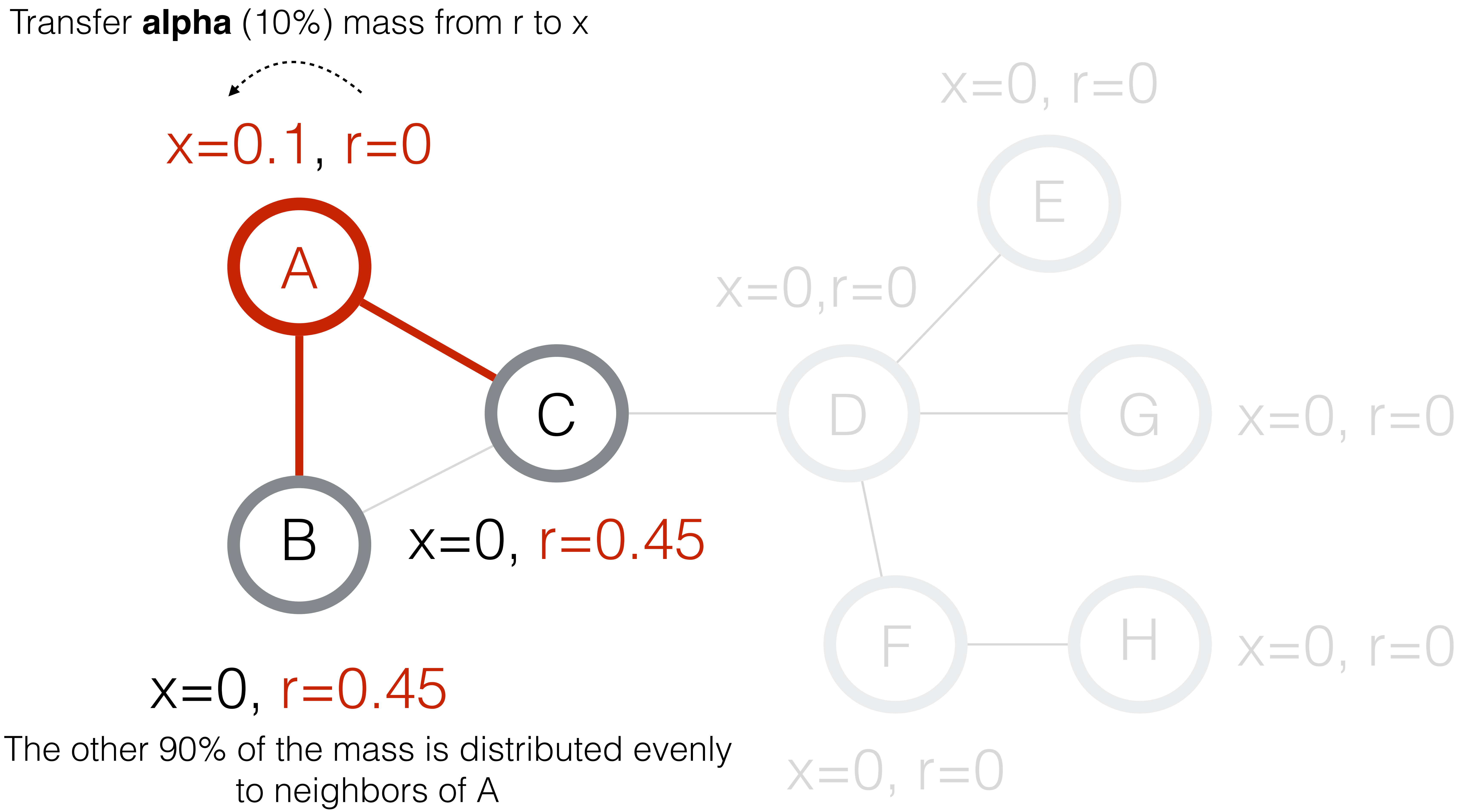
Algorithm idea: iteratively spread probability mass around the graph.



R. Andersen, F. Chung and K. Lang. Local graph partitioning using Page-Rank, FOCS, 2006

Approximate Personalized PageRank

Algorithm idea: iteratively spread probability mass around the graph.



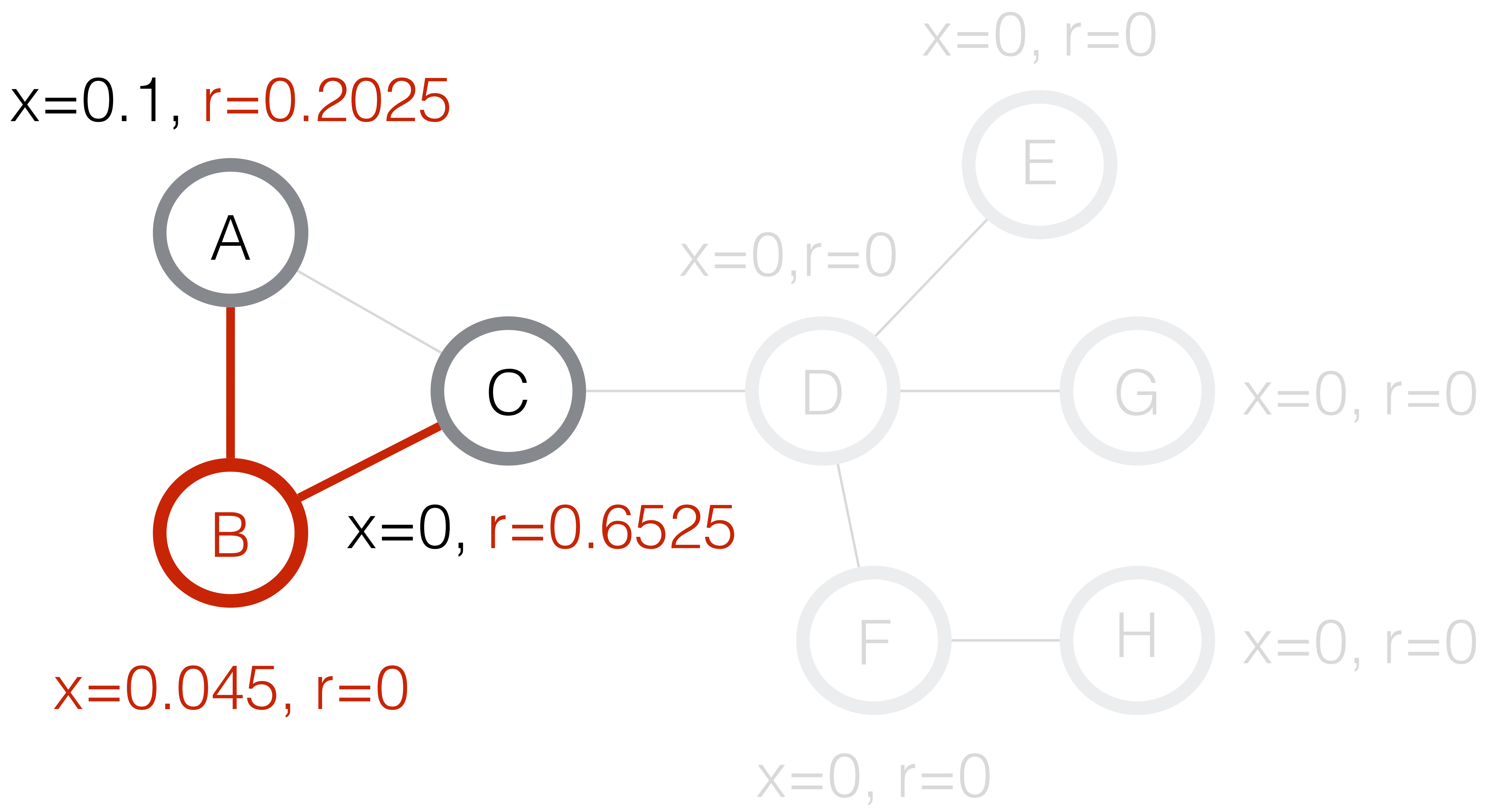
$$\alpha = 0.1$$

R. Andersen, F. Chung and K. Lang. Local graph partitioning using Page-Rank, FOCS, 2006

Approximate Personalized PageRank

Algorithm idea: iteratively spread probability mass around the graph.

$$\alpha = 0.1$$

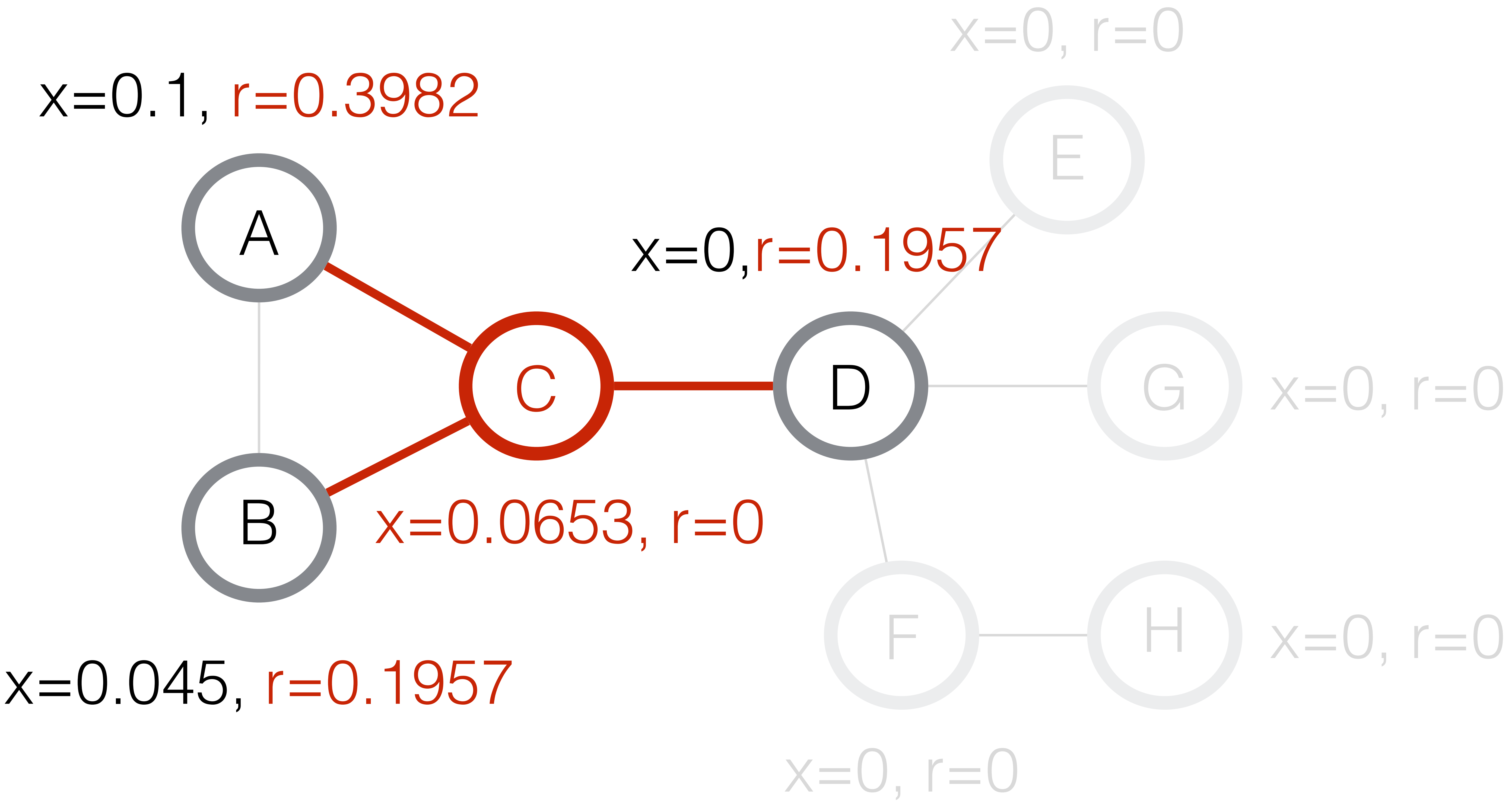


R. Andersen, F. Chung and K. Lang. Local graph partitioning using Page-Rank, FOCS, 2006

Approximate Personalized PageRank

Algorithm idea: iteratively spread probability mass around the graph.

$$\alpha = 0.1$$



R. Andersen, F. Chung and K. Lang. Local graph partitioning using Page-Rank, FOCS, 2006

Approximate Personalized PageRank

Algorithm idea: iteratively spread probability mass around the graph until

$$\max_i \frac{r_i}{d_i} \leq \rho\alpha$$

- ρ: termination parameter
- d_i: number of edges of node i

Variational model of APPR

$$\text{minimize } \frac{1-\alpha}{2}\|Bx\|_2^2 + \alpha\|H(\mathbf{1}-x)\|_2^2 + \alpha\|Zx\|_2^2 + \rho\alpha\|Dx\|_1$$

where

- B: is the incidence matrix
- D: Degree matrix
- H = diag(initial prob. dist. over nodes)
- Z = D - H

- α: teleportation parameter
- ρ: l1-reg. hyper-parameter

Observation: The optimality conditions of the l1-regularized version of the problem imply the early termination criterion of APPR.

Termination conditions vs optimality conditions

Termination criteria of Approximate Personalized PageRank

$$\max_i \frac{r_i}{d_i} \leq \rho\alpha$$

Optimality conditions of the variational model

$$\frac{r_i}{d_i} = \rho\alpha, \;\; x_i \neq 0$$

$$\frac{r_i}{d_i} \leq \rho\alpha, \;\; x_i = 0$$

Properties of the variational problem

- **Theorem:** The volume of the optimal solution is bounded by 1/ρ

$$\mathrm{vol}(S_*) = \sum_{i \in S_*} d_i =\leq \frac{1}{\rho}$$

- **Theorem:** Same combinatorial theoretical guarantees for local graph clustering

- **Crucial:** The model decouples the output from the algorithm.

Sketch of proof

- **Theorem:** The volume of the optimal solution is bounded by 1/ρ

$$\mathrm{vol}(S_*) = \sum_{i \in S_*} d_i = \leq \frac{1}{\rho}$$

Result 1: Negative partial derivatives are bounded from below

$$\rho \alpha d_i^{1/2} \leq -\nabla_i f(x_*) \quad \forall i \in S_* \Rightarrow \mathrm{vol}(S_*)\rho\alpha \leq \|D^{1/2}\nabla f(x_*)\|_1$$

Result 2: Gradients are bounded from above

$$\|D^{1/2}\nabla f(x_*)\|_1 \leq \alpha$$

Results 1 + 2 use proof-by-algorithm + strong convexity, Results 1 + 2 give the final result

Proximal gradient descent for local graph clustering

$$f(x) := \frac{1-\alpha}{2}\|Bx\|_2^2 + \alpha\|H(\mathbf{1}-x)\|_2^2 + \alpha\|Zx\|_2^2 \qquad g(x) := \rho\alpha\|Dx\|_1$$

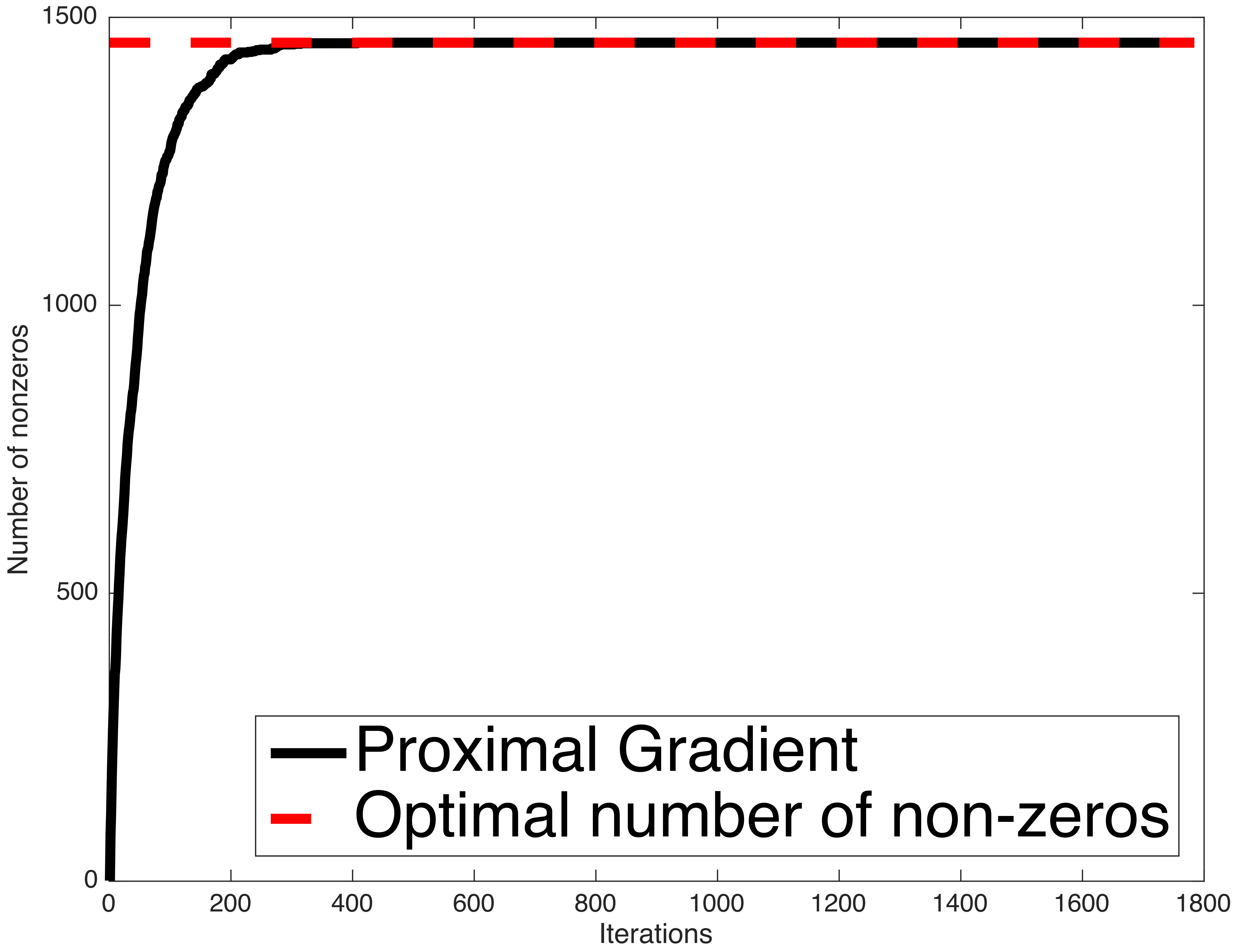
Proximal gradient descent

$$x_{k+1} := \text{argmin } g(x) + \underbrace{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle}_{\text{first-order Taylor approximation}} + \underbrace{\frac{1}{2}\|x - x_k\|_2^2}_{\substack{\text{upper bound on the} \\ \text{approximation error}}}$$

Requires careful implementation to avoid excessive running time

- Need to maintain a set of non-zero nodes
- Update x and gradient only for non-zero nodes and their neighbors at each iteration

Theorem: non-decreasing non-zero nodes



Sketch of proof

-**Theorem:** non-decreasing non-zero nodes

Result 1: Using induction we get that negative partial derivatives are bounded

$$-\nabla_i f(x_k) \geq \rho\alpha d_i^{1/2} \ \forall i \in S_k \ \text{ and } \ -\nabla_i f(x_k) < \rho\alpha d_i^{1/2} \ \forall i \in [n]\backslash S_k \quad \forall k$$

$\Big\downarrow$ Using the definition of a proximal step

Result 2: The mass of the variables is non-decreasing

$$x_k \leq x_{k+1} \quad \forall k$$

Results 1 + 2 give $S_k \subseteq S_{k+1}$

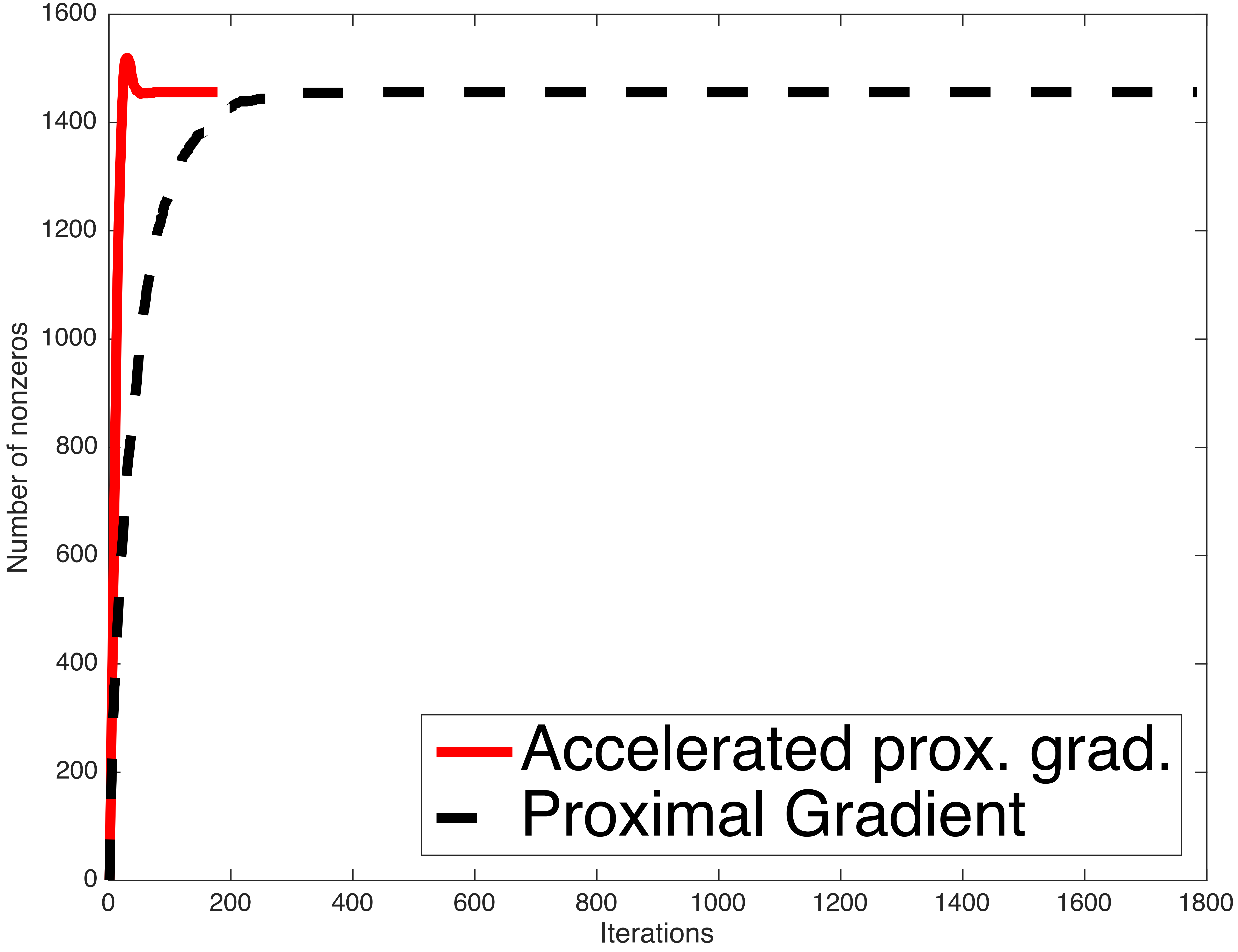
Worst-case running times

	Weighted graphs	Unweighted graphs
Prox. grad.	$\mathcal{O}\left(\frac{(\lvert\mathcal{S}_*\rvert + \widehat{\mathrm{vol}}(\mathcal{S}_*))}{\mu}\log\left(\frac{2}{\epsilon^2\rho^2\alpha^2\min_j d_j}\right)\right)$	$\mathcal{O}\left(\frac{2}{\rho\mu}\log\left(\frac{2}{\epsilon^2\rho^2\alpha^2\min_j d_j}\right)\right).$
APPR	✗	$\frac{1}{\alpha\rho}$

$$\mu := \alpha + \frac{1-\alpha}{4}\lambda_{min}(\mathcal{L}_{\mathcal{S}_*})$$

$\mathcal{L}_{\mathcal{S}_*}$: sub-matrix of normalized Laplacian

Open problem: is accelerated prox. grad. a local algorithm?



Gradient descent running time

$$\mathcal{O}\left(\frac{(|\mathcal{S}_*| + \widehat{\mathrm{vol}}(\mathcal{S}_*))}{\mu} \log\left(\frac{2}{\epsilon^2\rho^2\alpha^2 \min_j d_j}\right)\right)$$

Accel. gradient descent

$$\mathcal{O}\left(\frac{\mathrm{vol}(\mathcal{G})}{\sqrt{\mu}} \log\left(\frac{2}{\epsilon^2\rho^2\alpha^2 \min_j d_j}\right)\right)$$

$$\Downarrow$$

$$\mathcal{O}\left(\frac{|\mathcal{S}_*| + \mathrm{vol}(\mathcal{S}_*)}{\sqrt{\mu}} \log\left(\frac{2}{\epsilon^2\rho^2\alpha^2 \min_j d_j}\right)\right)$$

LocalGraphClustering on GitHub

- Written in Python with C++ routines when required
- Graph analytics on 100 million edges graph on a 16GB RAM laptop
- Demonstrations on social and bioinformatics networks
- 8 Python notebooks with numerous examples and graph visualizations
- Video presentations
- 12 methods and pipelines

References

“An optimization approach to locally-biased graph algorithms”, K. Fountoulakis, D. Gleich, M. Mahoney Proceedings IEEE, 2016

“Variational perspective on local graph clustering”, K.Fountoulakis, F. Khorasani, J. Shun, X. Cheng, M. Mahoney, Math. Prog., 2017

“Capacity Releasing Diffusion for Speed and Locality”, D. Wang, K. Fountoulakis, M. Mahoney, S. Rao, ICML 2017

“Parallel Local Graph Clustering”, J. Shun, F. Khorasani, K. Fountoulakis, M. Mahoney, VLDB, 2016

Thank you!

Parallel local graph clustering methods in shared memory

- Why shared memory? Currently the largest publicly available graphs can be stored in computers with shared memory

- We parallelize 4 local spectral methods + rounding
  1. Approximate PageRank (as demonstrated in previous slides)
  2. Nibble
  3. Deterministic HeatKernel Approximate PageRank
  4. Randomized HeatKernel Approximate PageRank
  5. Sweep cut rounding algorithm

Based on

Parallel Local Graph Clustering, J. Shun, K. Fountoulakis, F. Khorasani, M. Mahoney, VLDB, 2016

Overview of results

-3-16x faster than serial version

-Parallelization allowed us to solve problems of billions of nodes and edges.

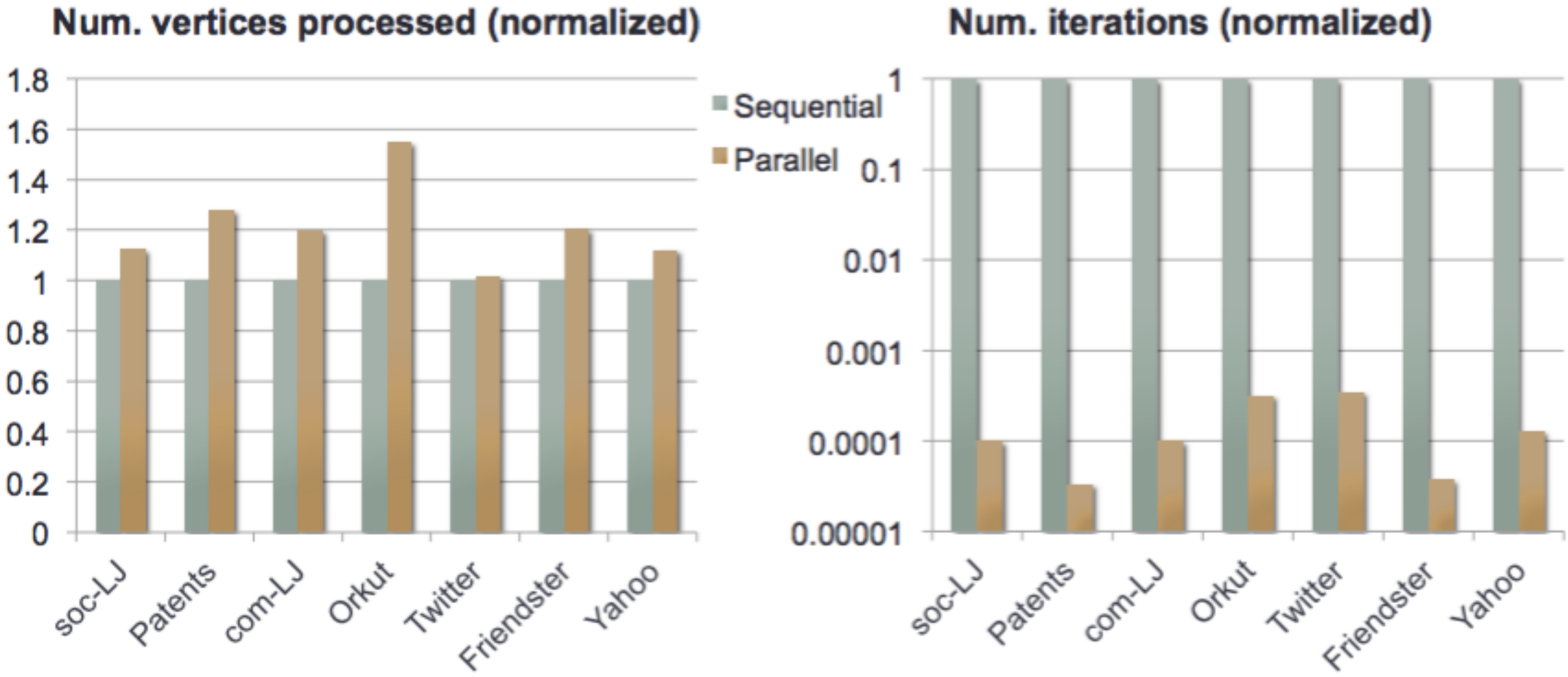
An example: Parallel Approximate Personalized PageRank

- Serial version: picks a node from candidates to distribute mass to its neighbors
- Parallel: pick **all** nodes from candidates and distribute mass simultaneously
- Asymptotic work (FLOPS) remains the same
  - But work is parallelized
  - We pay a small communication cost among cores

Data

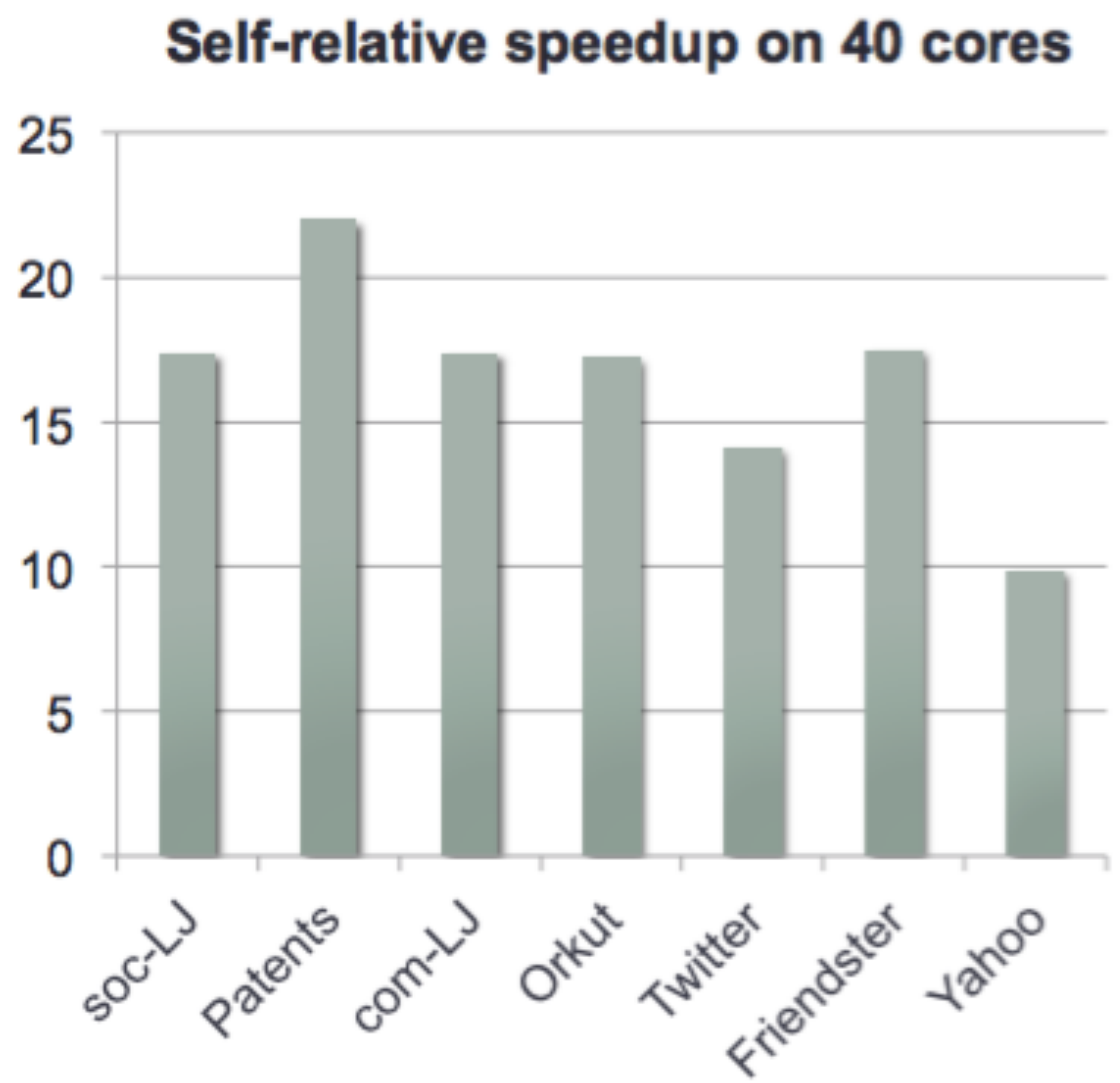
Input graph	Num. vertices	Num. edges
soc-JL	4,847,571	42,851,237
cit-Patents	6,009,555	16,518,947
com-LJ	4,036,538	34,681,189
com-Orkut	3,072,627	117,185,083
Twitter	41,652,231	1,202,513,046
Friendster	124,836,180	1,806,607,135
Yahoo	1,413,511,391	6,434,561,035

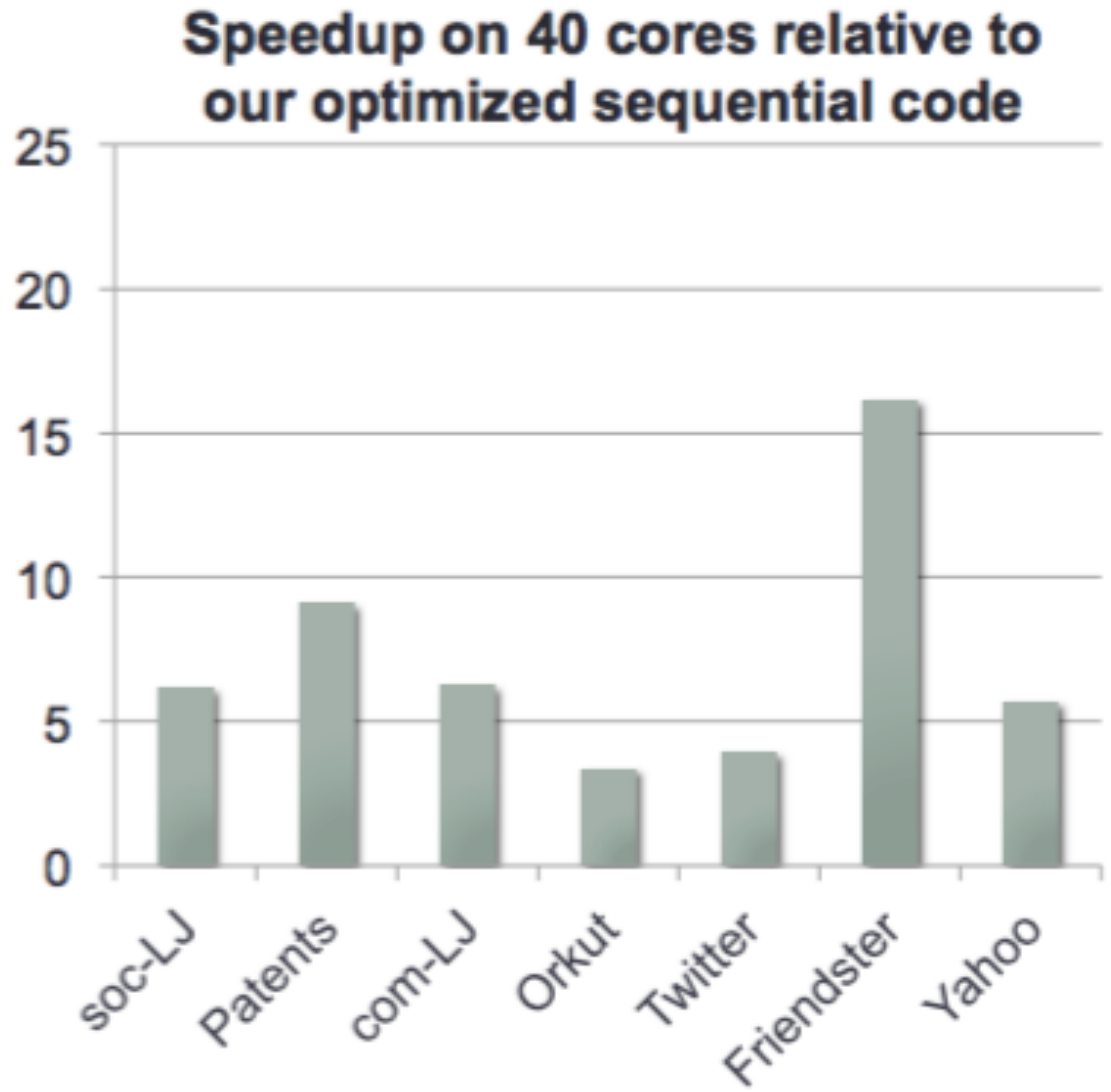
Performance



- Slightly more work for the parallel version
- Number of iterations is significantly less

Performance





- 3-16x speed up
- Speedup is limited by small active set in some iterations and memory effects