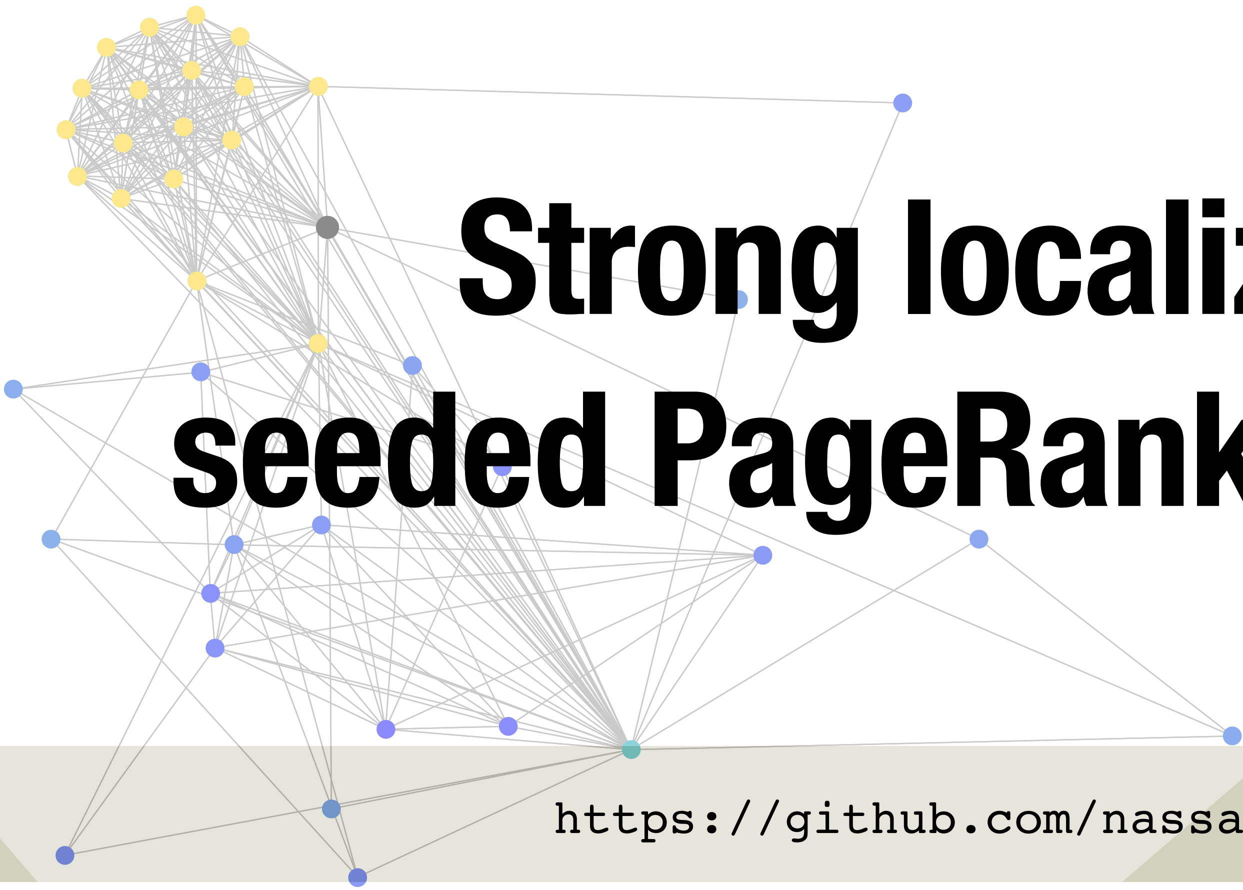
Strong localization in seeded PageRank vectors



https://github.com/nassarhuda/pprlocal

David F. Gleich
Computer Science

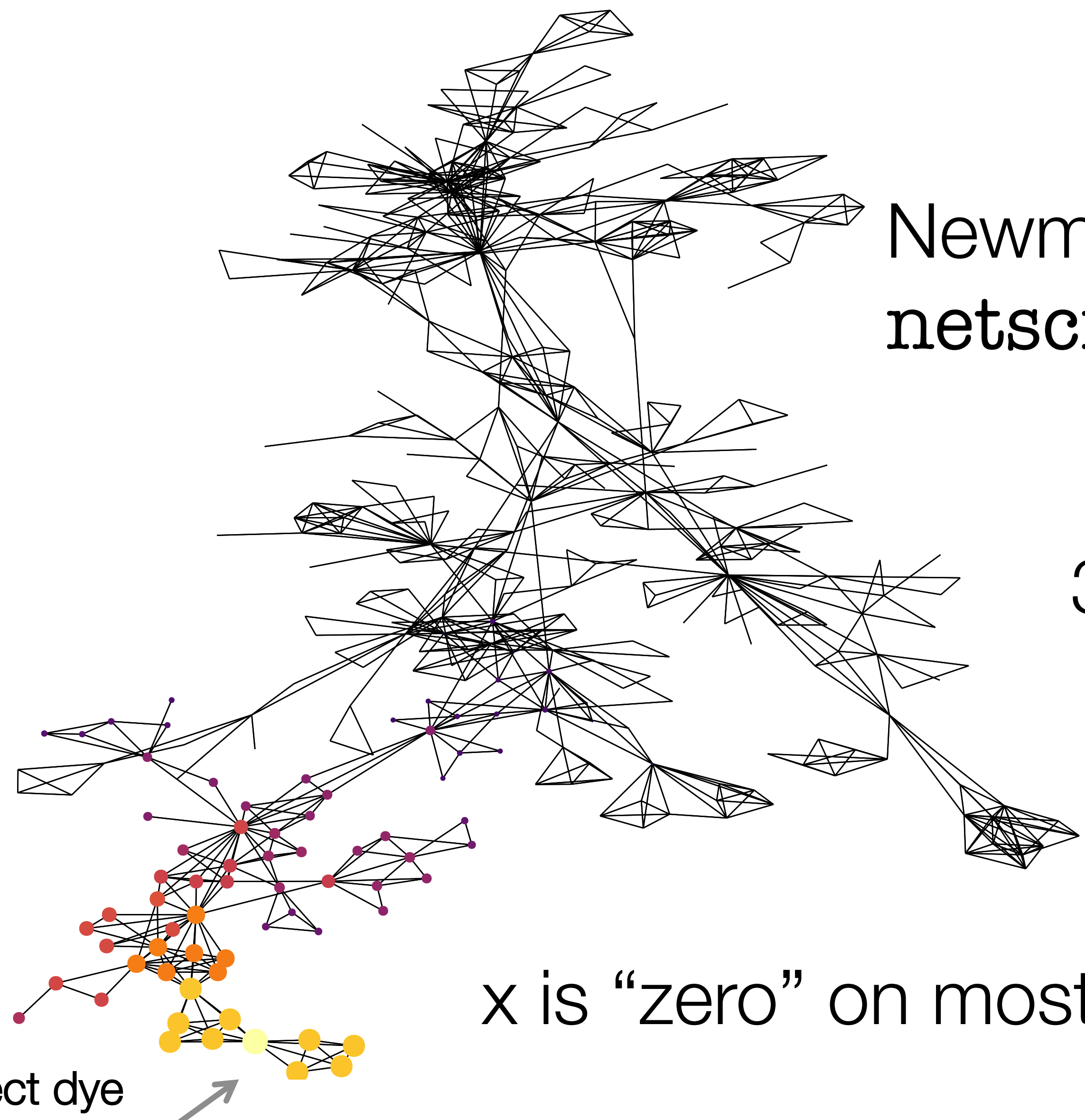
Huda Nassar
Computer Science

Kyle Kloster
Mathematics

supported by
NSF CAREER CCF-1149756,
IIS-1422918, DARPA SIMPLEX

PURDUE
UNIVERSITY

Localization in seeded PageRank



Newman's `netscience` graph

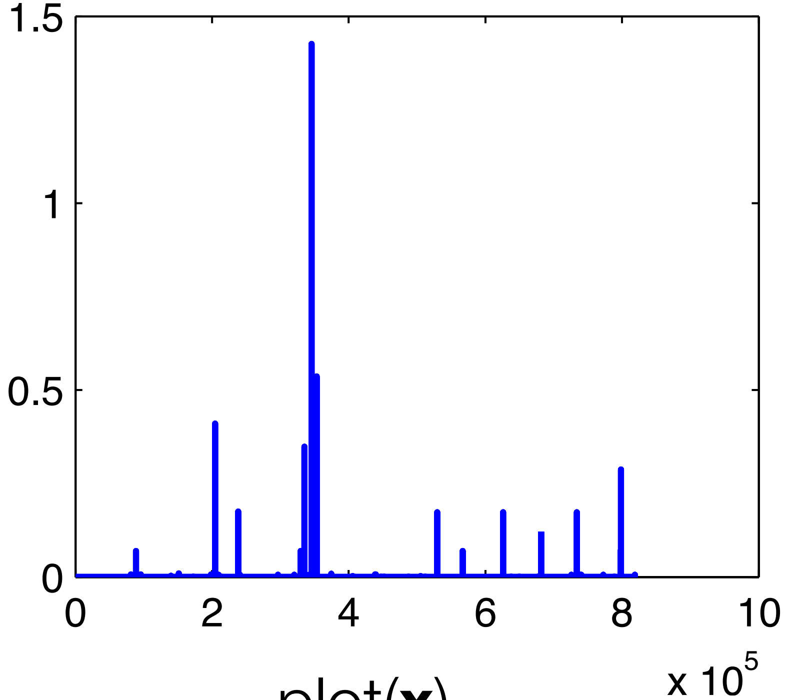
379 vertices
924 edges

x is "zero" on most of the nodes

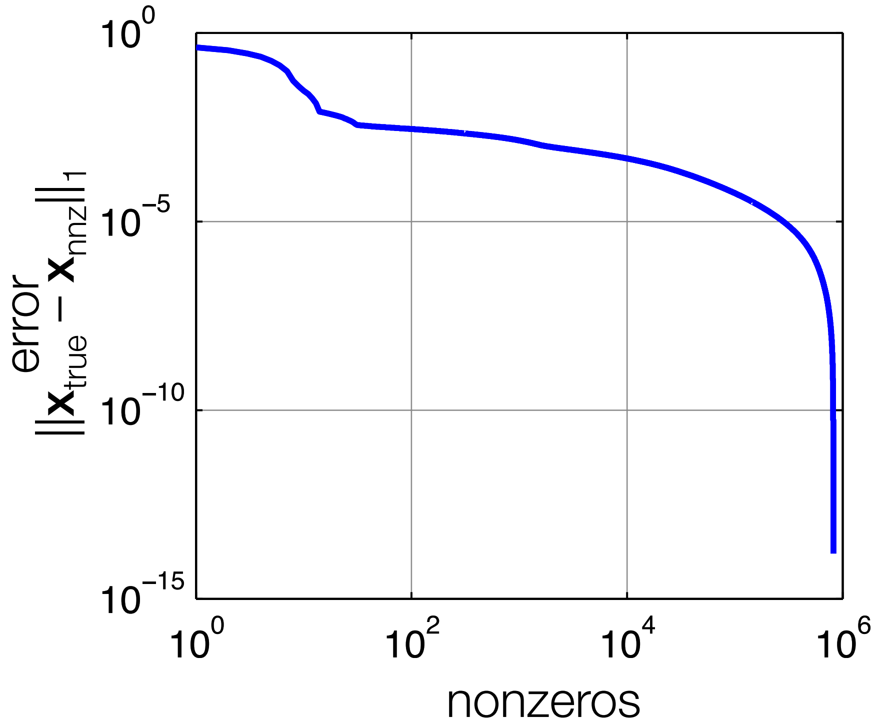
Inject dye here

An example on a bigger graph

Crawl of flickr from 2006: ~800K nodes, 6M edges, seeded PageRank with $\alpha = 0.5$



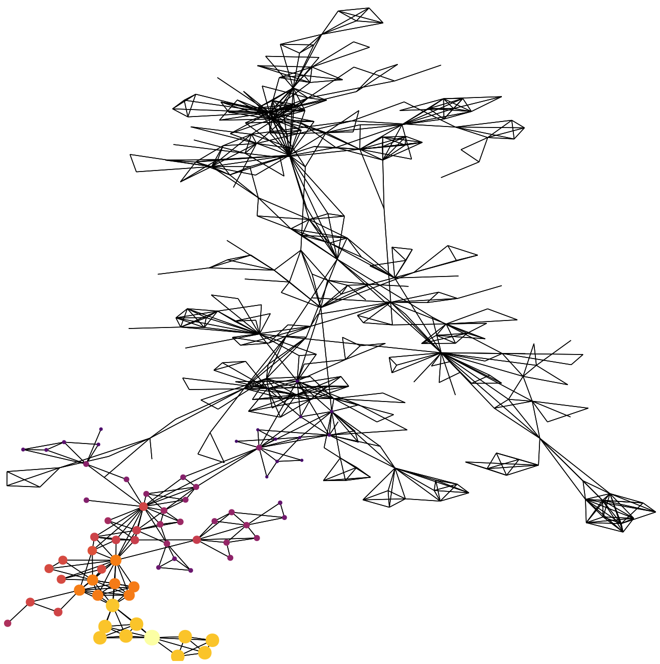
plot($\mathbf{x}$)



X-axis: node index

Y-axis: value at that index in true PageRank vector

Localization in seeded PageRank



Given a seed and a graph

$\mathbf{e}_s$ $\quad$ $\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{D}^{-1}$

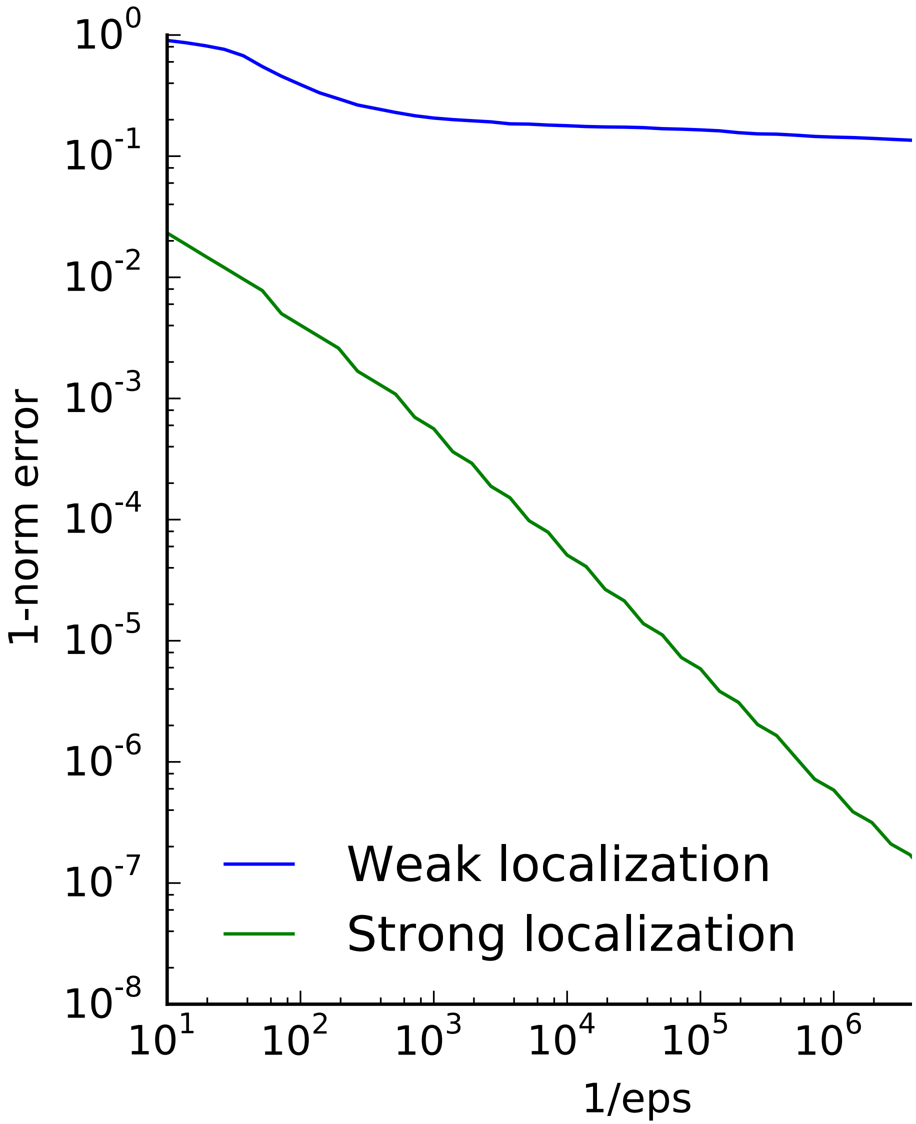
What can we say about localization in the seeded PageRank vector with parameter α ?

$$(\boldsymbol{I} - \alpha \boldsymbol{P})\mathbf{x} = (1 - \alpha)\mathbf{e}_s$$

THEOREM We show that if the graph has a type of skewed degree dist. then the solution **x** cannot have many big entries.

(Previously this was known only for const. degree or very slowly growing.)

Types of localization



Weak (entry-wise)

$$\|\mathbf{D}^{-1}(\mathbf{x} - \mathbf{x}^*)\|_\infty \le \varepsilon$$

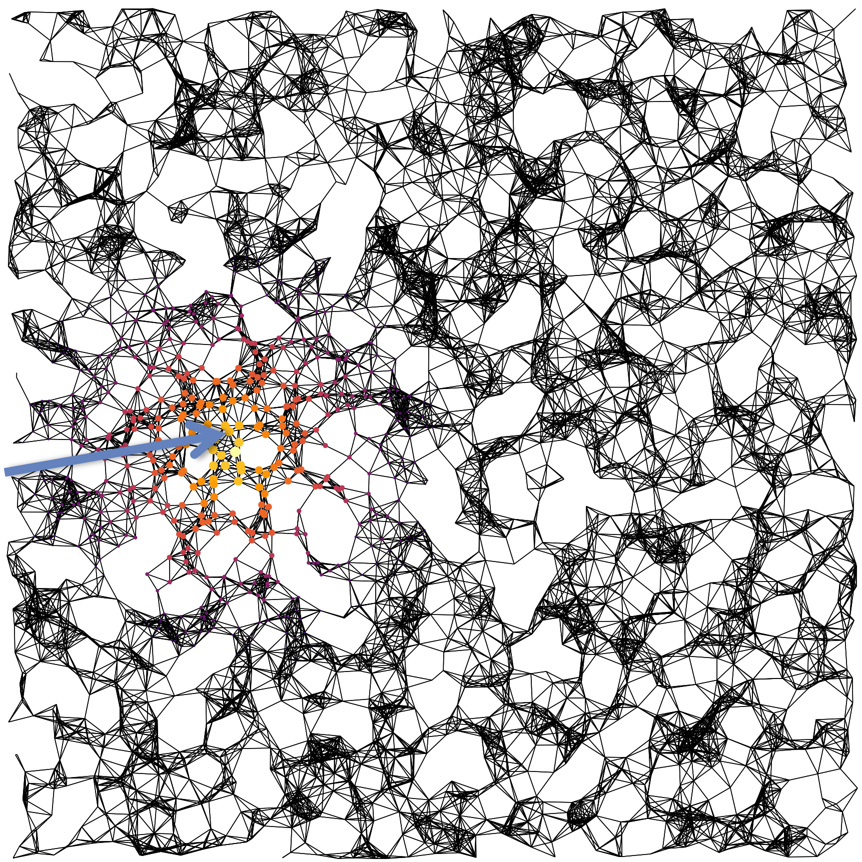
Andersen, Chung, and Lang proved that the PageRank vector is weakly localized in the famous 2006 “push” paper.

Strong (uniform)

$$\|\mathbf{x} - \mathbf{x}^*\|_1 \le \varepsilon$$

Strong localization

When is strong localization possible?



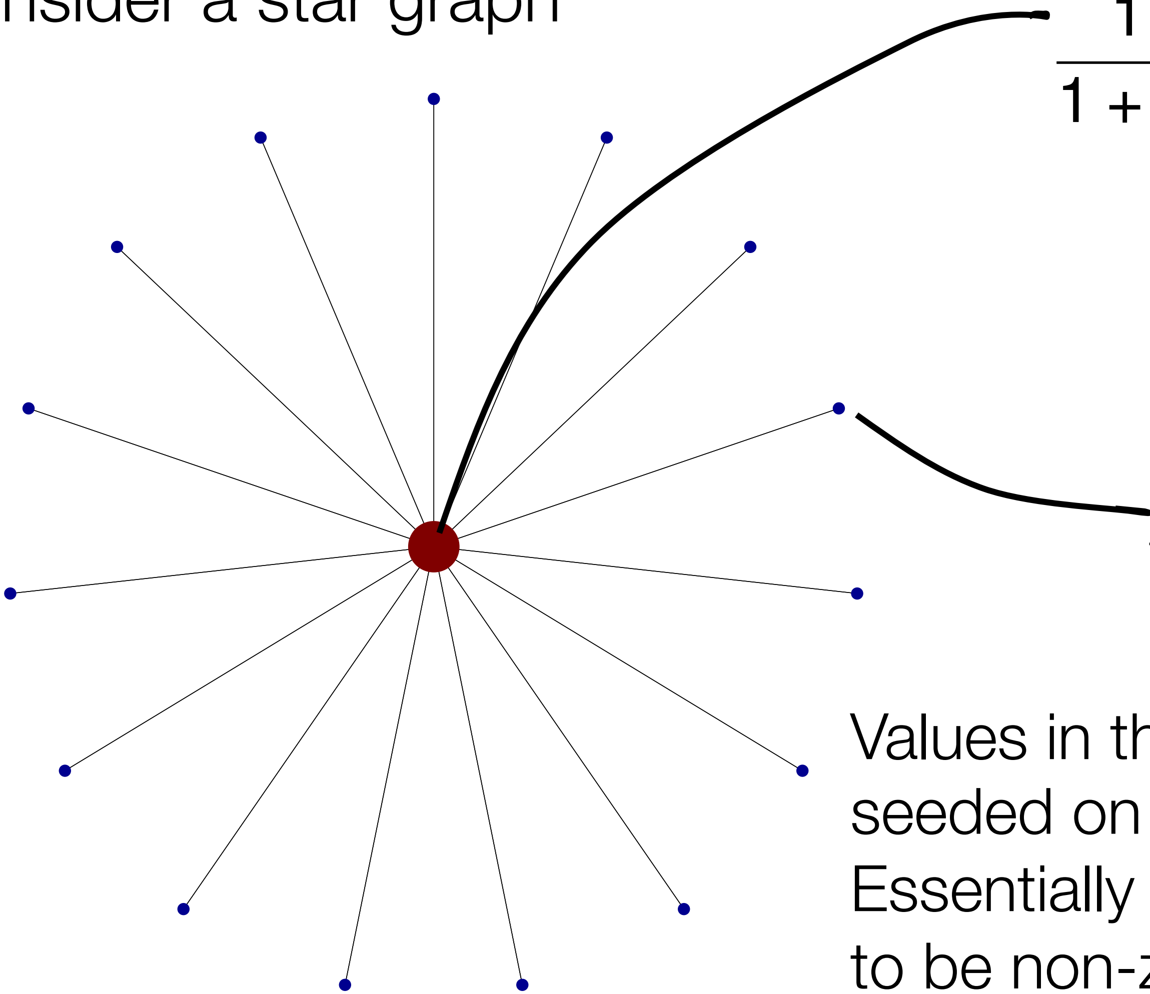
Consider graphs with very slowly growing degree or a constant degree.

An easy corollary of our subsequent theory. Also known from functions of sparse-matrix literature.

Handles cases like the Erdős-Réyni graphs and grid graphs.

Strong localization can be impossible

Consider a star graph



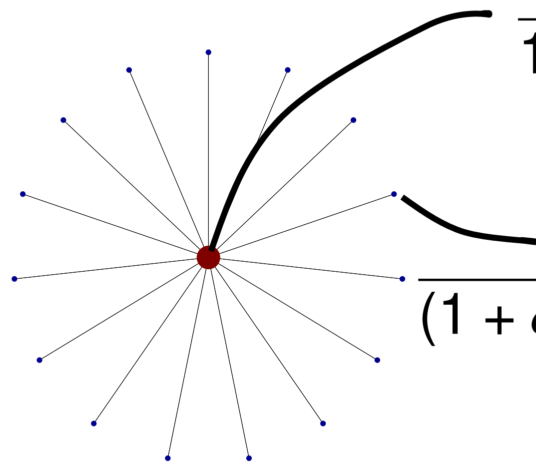
$$\frac{1}{1+\alpha}$$

$$\frac{\alpha}{(1+\alpha)(n-1)}$$

Values in the PageRank vector seeded on the center node. Essentially everything is needed to be non-zero to get a global error bound.

Strong localization can be impossible

Consider a star graph



If we round k entries to zero,

1-norm error is $k \cdot \frac{\alpha}{(1+\alpha)(n-1)}$

so…

this: $\|\mathbf{x} - \mathbf{x}^*\|_1 \le \varepsilon$

requires

$$\frac{(1+\alpha)\varepsilon}{\alpha} \cdot (n-1) \le k$$

Values in the PageRank vector seeded on the center node. Essentially everything is needed to be non-zero to get a global error bound.

Strong localization can be impossible

Seeded PageRank is also non-local on
any complete bipartite graphs (generalizing star graphs).

Why?
***Fact:* $\boldsymbol{P}$** is complete-bipartite iff eigenvalues = {-1,0,1}.

PageRank is really a matrix function, $f(x) = (1 - \alpha x)^{-1}$.

Strong localization can be impossible

Seeded PageRank is also non-local on
any complete bipartite graphs (generalizing star graphs).

Why?
Fact: $\boldsymbol{P}$ is complete-bipartite iff eigenvalues = {-1,0,1}.

PageRank is really a matrix function, $f(x) = (1 - \alpha x)^{-1}$.

Fact: a matrix function is equiv to interpolating polynomial

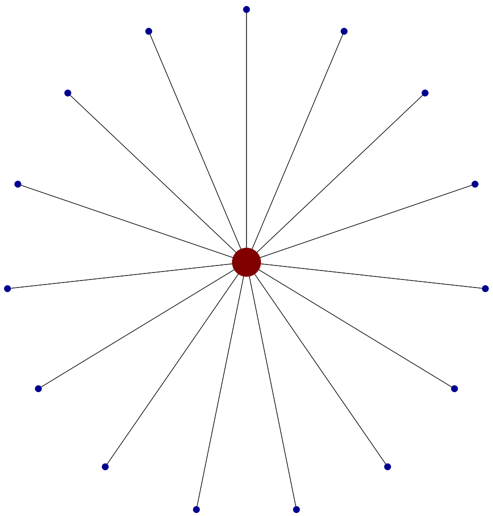
$$p(\lambda_i) = f(\lambda_i) \to p(\boldsymbol{P}) = f(\boldsymbol{P})$$

Only 3 eigenvalues ⟶ p(x) is degree 2 (!)

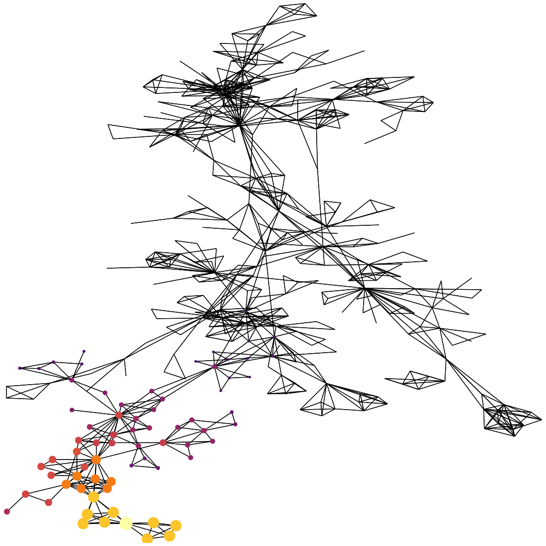
$$(\boldsymbol{I} - \alpha \boldsymbol{P})^{-1}\mathbf{e}_j \quad = \quad f(\boldsymbol{P})\mathbf{e}_j \quad = \quad (c_0\boldsymbol{I} + c_1\boldsymbol{P} + c_2\boldsymbol{P}^2)\mathbf{e}_j$$

When is localization possible?

Graphs exist where
seeded PageRank has
no local behavior (star graphs)



& graphs exist with
local behavior **everywhere**
(degree <= constant, or log log(n))



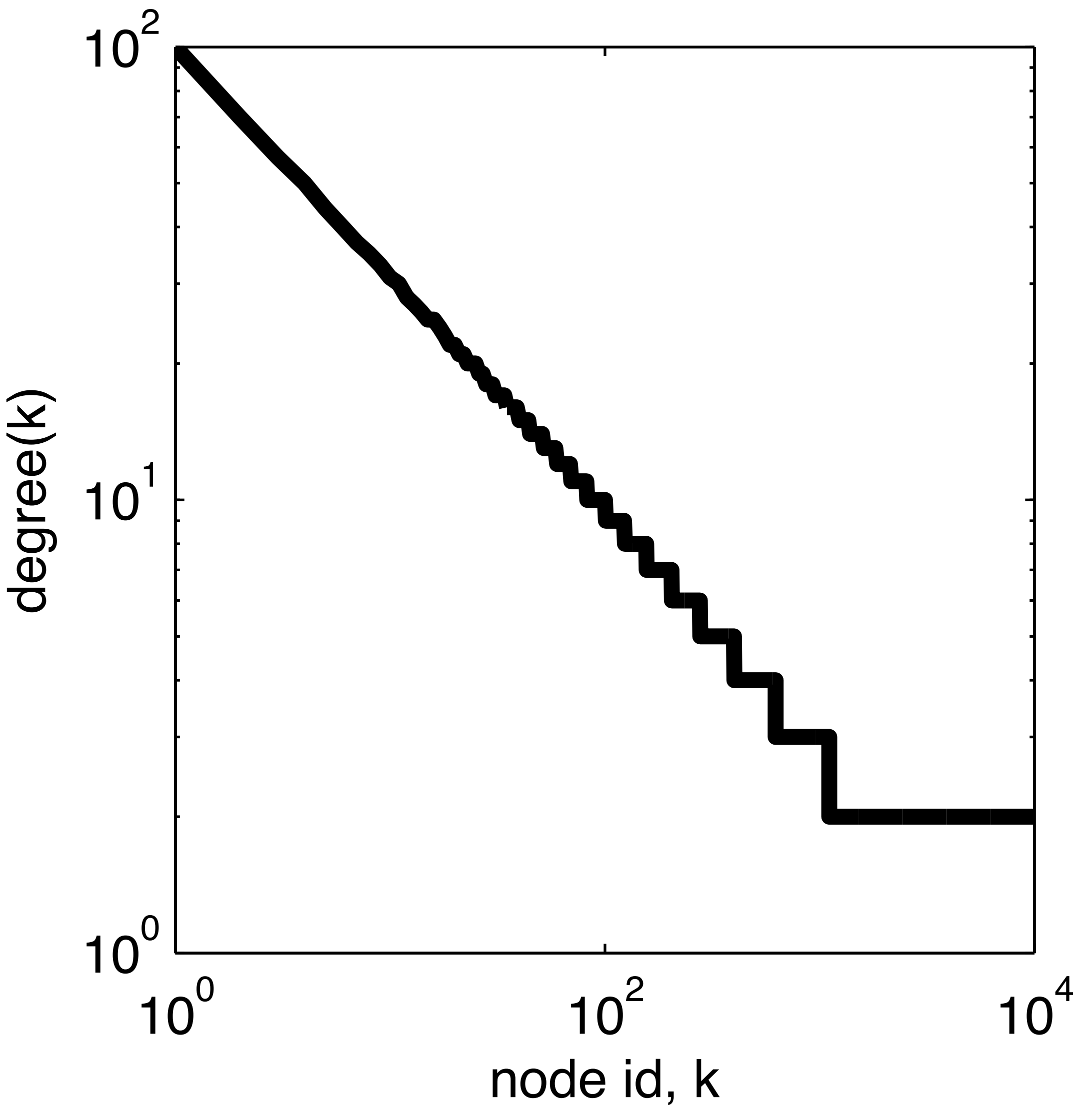
So what properties can

determine localization

in seeded PageRank?

Skewed degree distributions

The k-th largest degree $d(k) \leq \max(dk^{-p}, \delta)$



(δ is min degree,
p is decay exponent)

log log plot of the degree sequence for a synthetic example with

10,000 nodes
d = 100 (max degree)
δ = 2 (min degree)
p = 0.5 (decay exponent)

Distinct model from Pareto power law!

Strong localization in personalized PageRank Vectors

Theorem (Nassar, K., Gleich):
Let *d* be the max-degree, δ be the min-degree,
n be the number of nodes, *p* be the decay exponent.

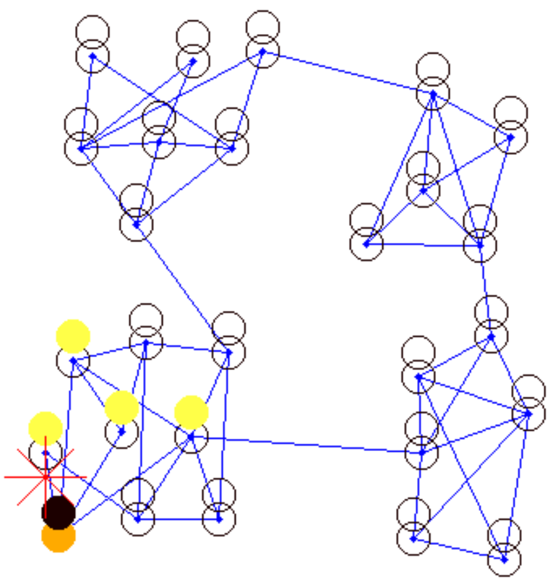
Then the number of non-zeros *N* needed for $\|\mathbf{x} - \mathbf{x}_\varepsilon\|_1 \le \varepsilon$

satisfies $N \le \min\left\{ n, \frac{1}{\delta} C_p (1/\varepsilon)^{\frac{\delta}{1-\alpha}} \right\}$

$$C_p = \begin{cases} d(1 + \log d) & p = 1 \\ d\left(1 + \frac{1}{1-p}(d^{(1/p)-1} - 1)\right) & \text{otherwise} \end{cases}$$

Due to the maximum degree d, this does not say anything about traditional power-law graphs (e.g. the Pareto case)

Strong localization in personalized PageRank Vectors (sketch)



We study the behavior of the *Gauss-Southwell or push algorithm* for computing PageRank

- residual = remaining rank/dye to assign
- solution = assigned rank/dye

Algorithm

1. pick node with most residual dye
2. assign dye to node
3. update residual dye on neighbors,
4. then repeat.

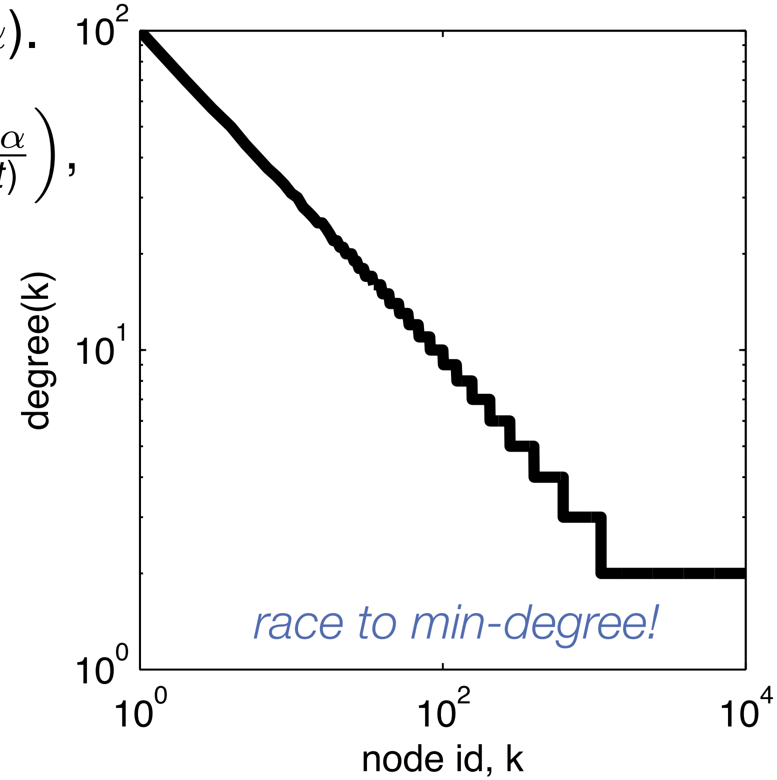
Strong localization in personalized PageRank Vectors (sketch)

Define the residual vector $\mathbf{r}$,
$\mathbf{r} = (1-\alpha)\mathbf{e}_s - (1-\alpha \boldsymbol{P})\hat{\mathbf{x}}$. Then
$\|\mathbf{x} - \hat{\mathbf{x}}\|_1 < \varepsilon$ is implied by $\|r\|_1 < \varepsilon(1-\alpha)$.

After k steps, $\|\mathbf{r}_k\|_1 \leq \|\mathbf{r}_0\|_1 \prod_{t=0}^{k} \left(1 - \frac{1-\alpha}{Z(t)}\right)$,
where $Z(t)$ denotes the number of non-zero entries in $\mathbf{r}_t$.

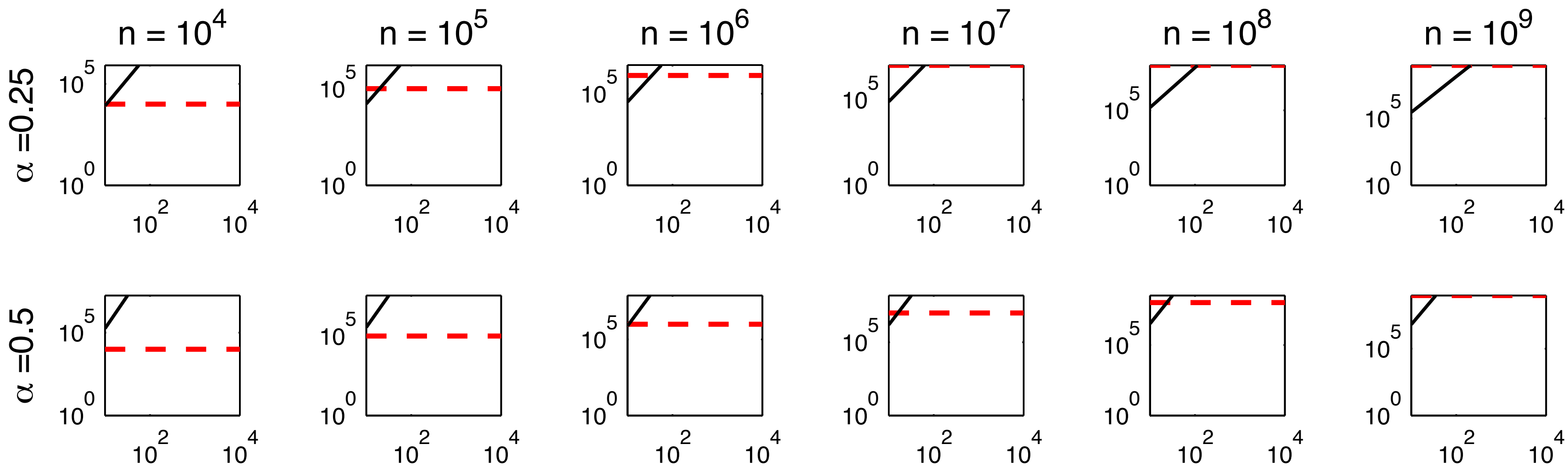
To guarantee $\|\mathbf{r}_k\|_1 < \varepsilon(1-\alpha)$,
it suffices to choose k so that,
$((\delta k + C_p)/C_p)(\alpha - 1)/\delta \leq \varepsilon$

The hard part is bounding $Z(t)$
we show $Z(t) \leq C_p + \delta t$



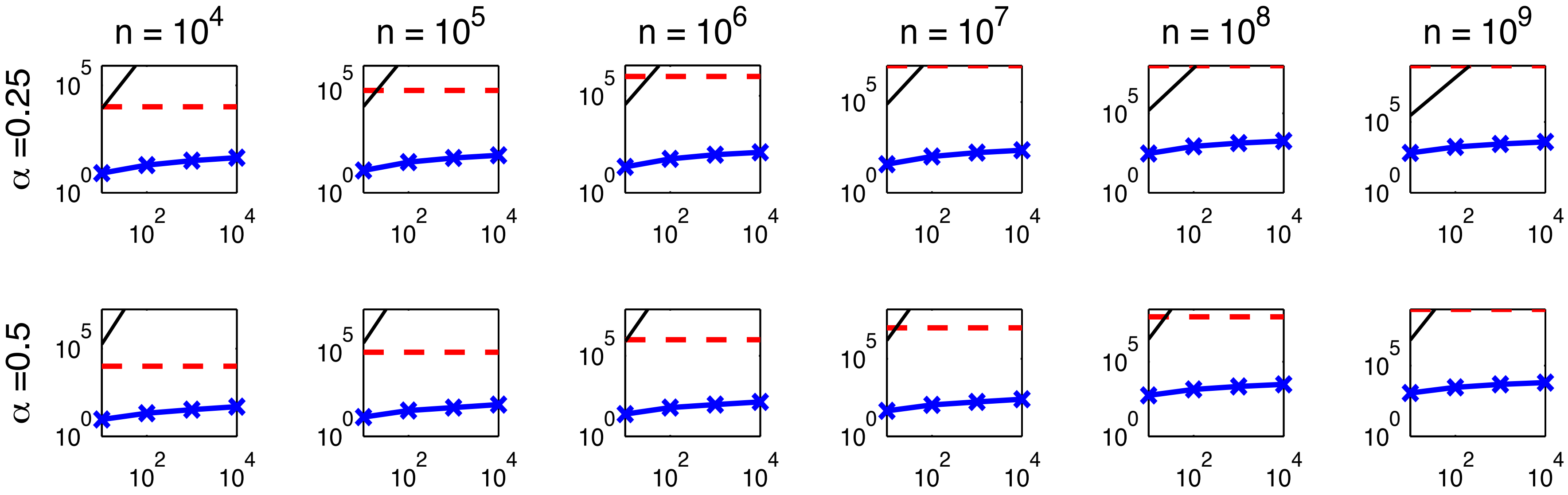
(The proof builds on techniques from [Gleich, K., Internet Math 2014])

Asymptotic theory prediction



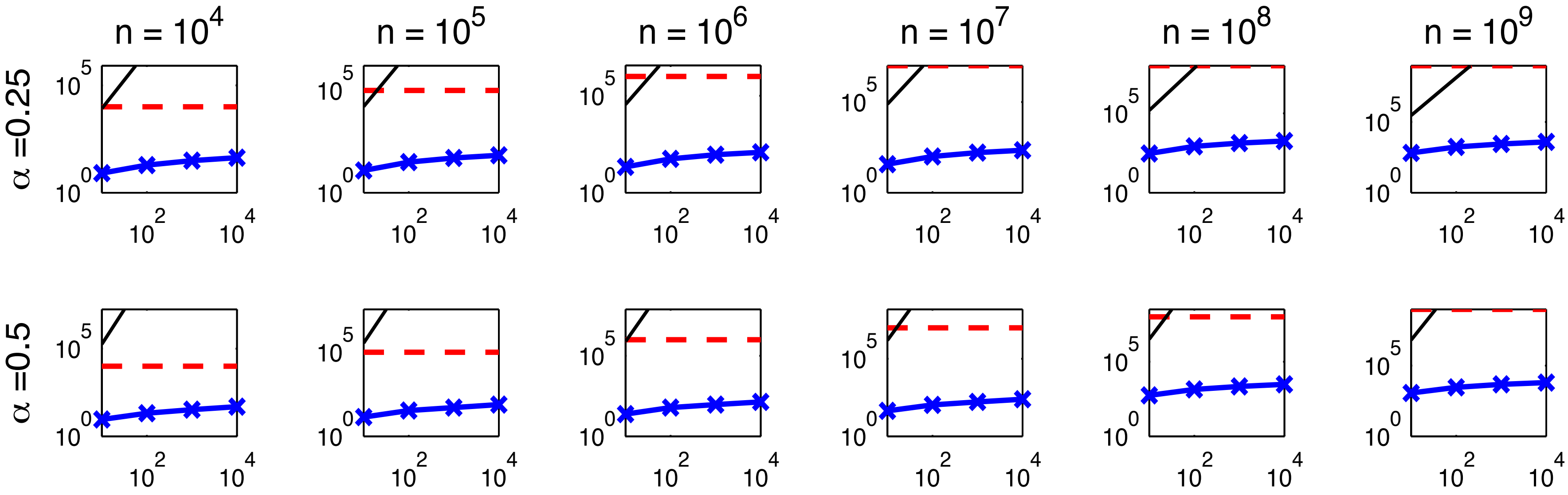
y-axis = number of non-zeros in approximate solution
x-axis = $1/\varepsilon$
red dashed line vector contains *all* non-zeros
black line bound on non-zeros predicted by theorem

Asymptotic theory prediction



red dashed line vector contains *all* non-zeros
black line bound on non-zeros predicted by theorem
blue line actual number of non-zeros in approximation

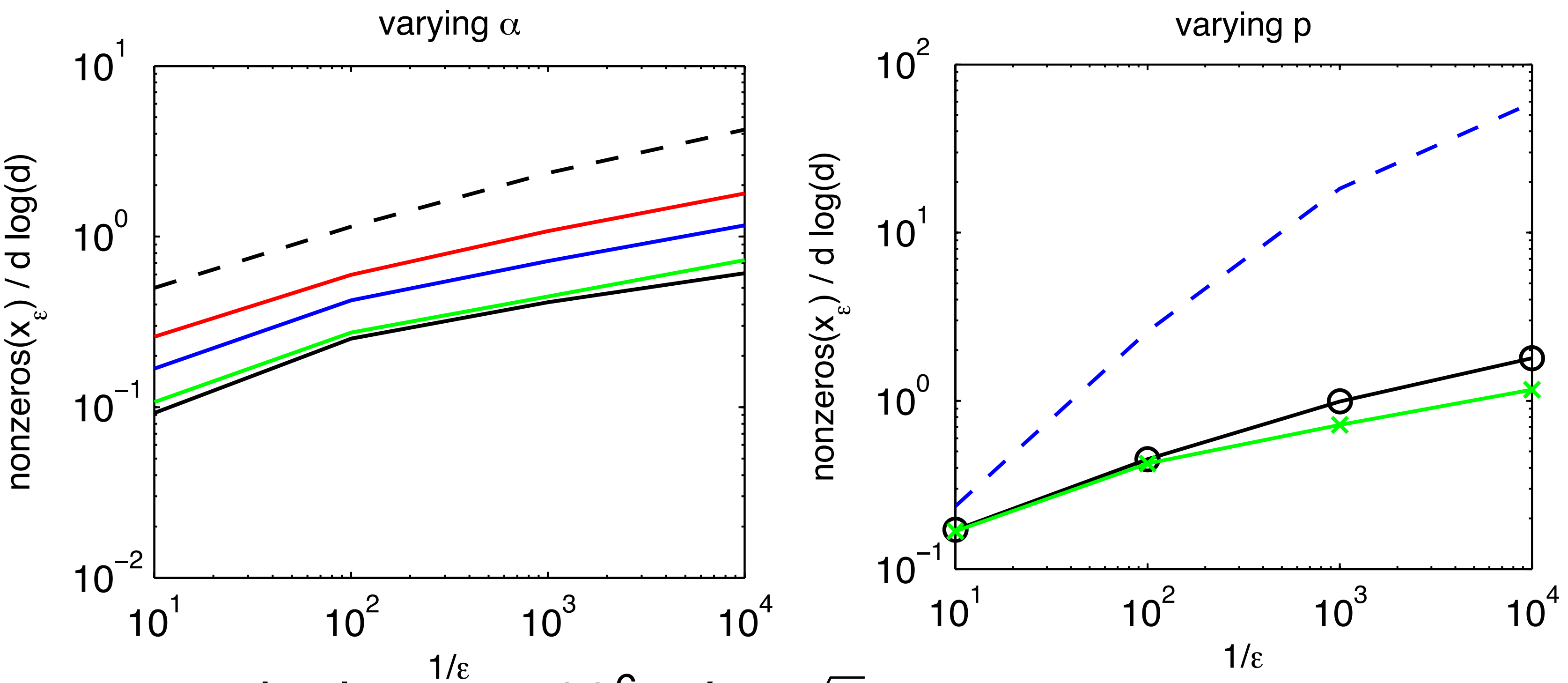
Asymptotic theory prediction



red dashed line vector contains *all* non-zeros
black line bound on non-zeros predicted by theorem
blue line actual number of non-zeros in approximation

☹ ➔ Need a better bound

Empirical scaling guides a new bound.

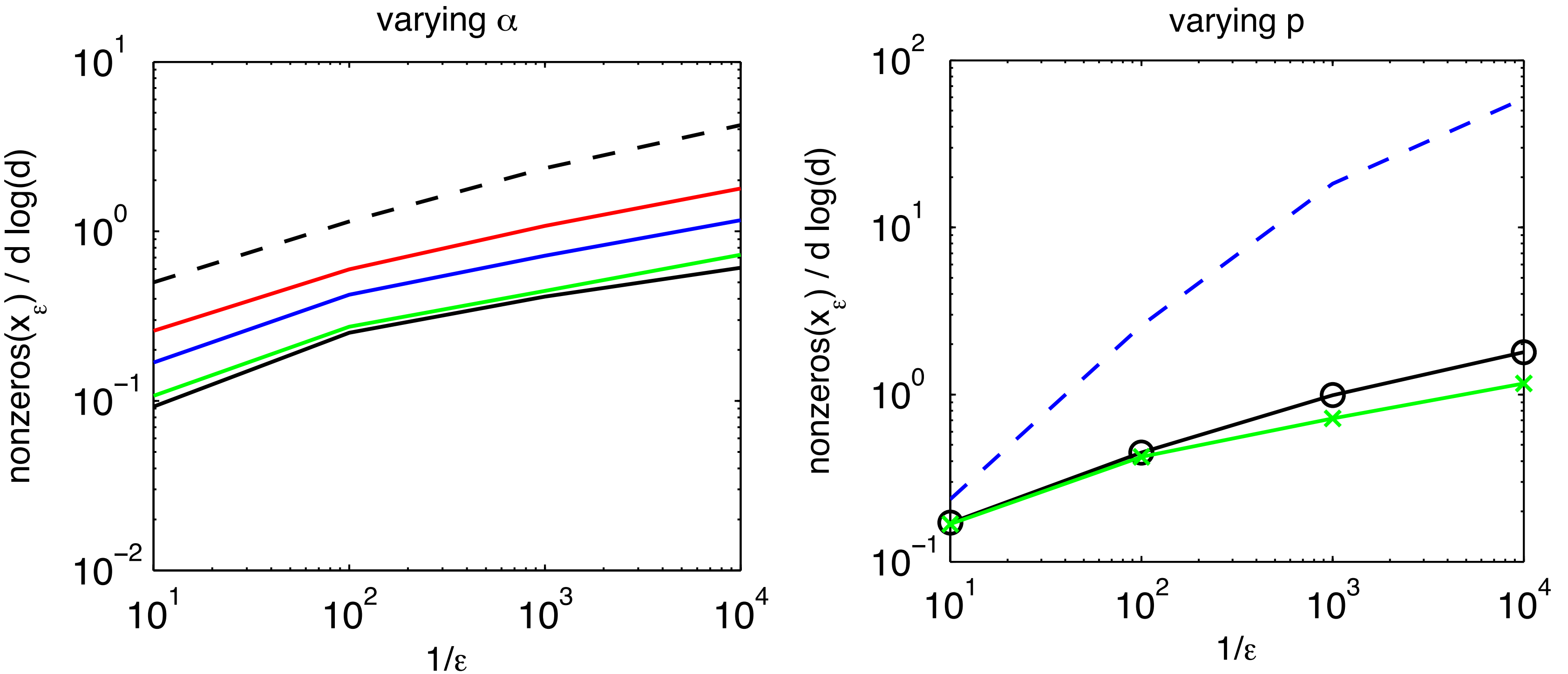


graph size, $n = 10^6$, $d = \sqrt{n}$

At left, $p = 0.95$, black, green, blue, red, represent $\alpha = \{0.25, 0.3, 0.5, 0.65, 0.85\}$

At right, $\alpha = 0.5$ and dashed blue, black and green represent $p = \{0.5, 0.75, 0.95\}$

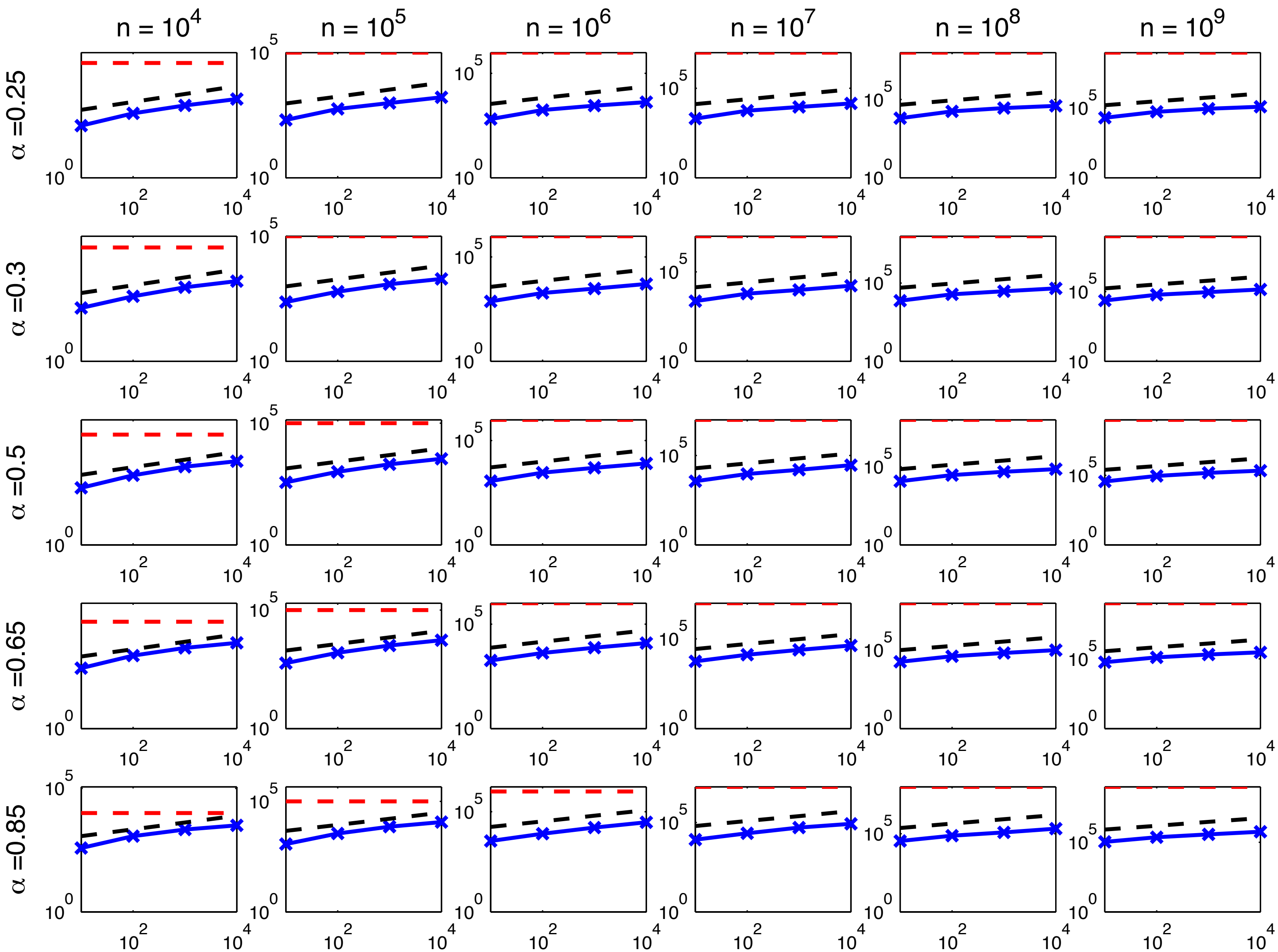
Empirical scaling guides a new bound.



We conjecture (bound')

$$\mathrm{nnz}(\mathbf{x}_\varepsilon) \leq d \log(d) \frac{0.2}{1-\alpha} \left(\frac{1}{\varepsilon}\right)^{(1/4p^2)}$$

Bound' accurately predicts localization



Conclusion and future work

- Examine broader classes of graphs empirically (like real-world networks)
- Improve the localization theory to apply to a wider range of degree distributions
- Explore other graphs without localization – more specifically, relationship between diameter and localization
- Get a theorem for the Pareto power-law case!