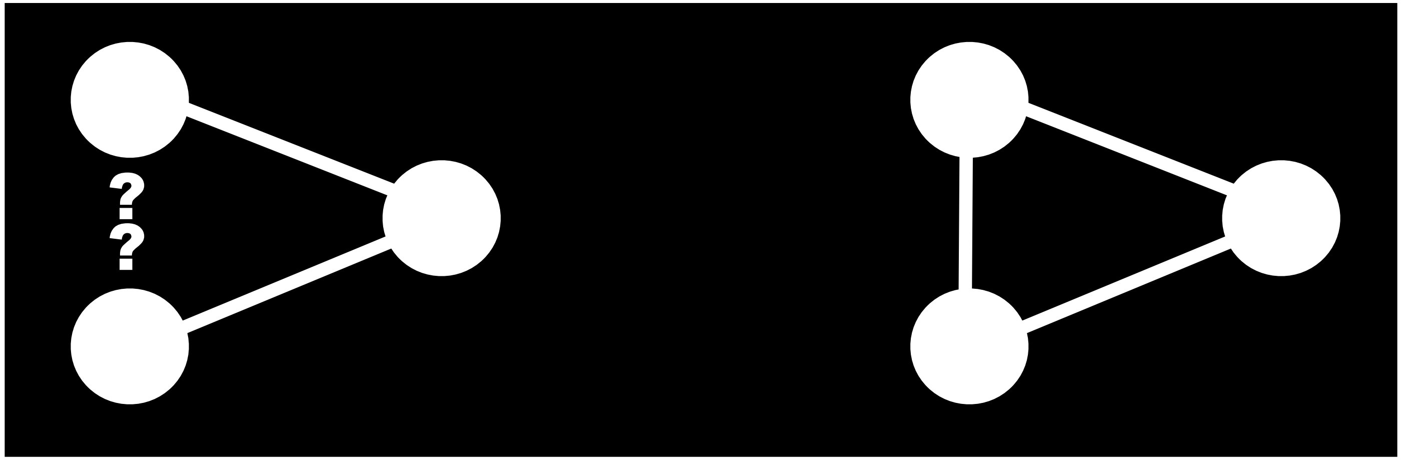
Inferring communities in large networks;

the blessing of transitivity.

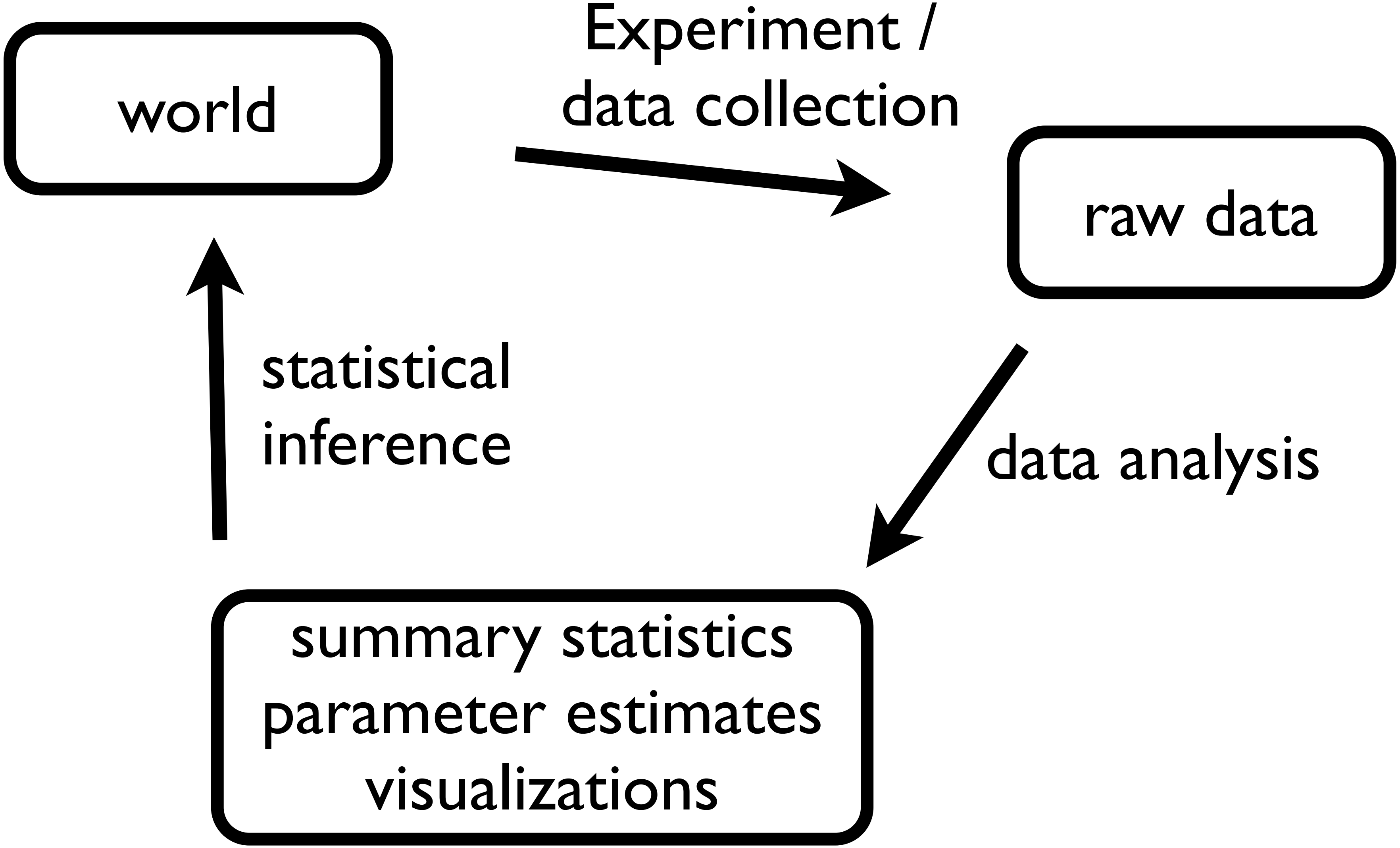


Karl Rohe
@stat.wisc.edu

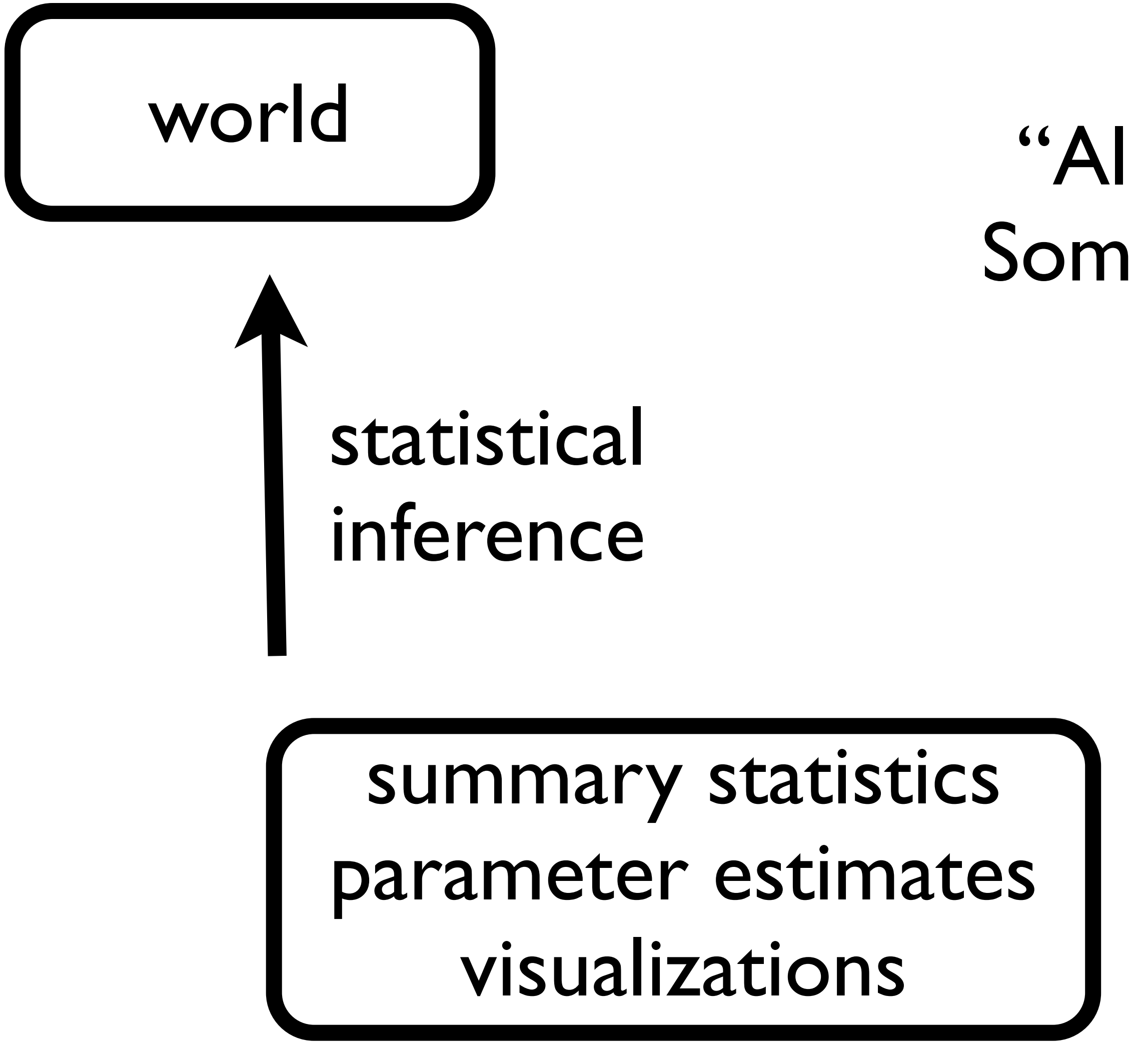
Join work with Tai Qin.

Research supported by DMS-1309998.

Historically, the overriding academic achievement of our field (stat) has been to develop a framework for making quantitatively rigorous inference.



Inference depends on a statistical model



world

statistical inference

summary statistics
parameter estimates
visualizations

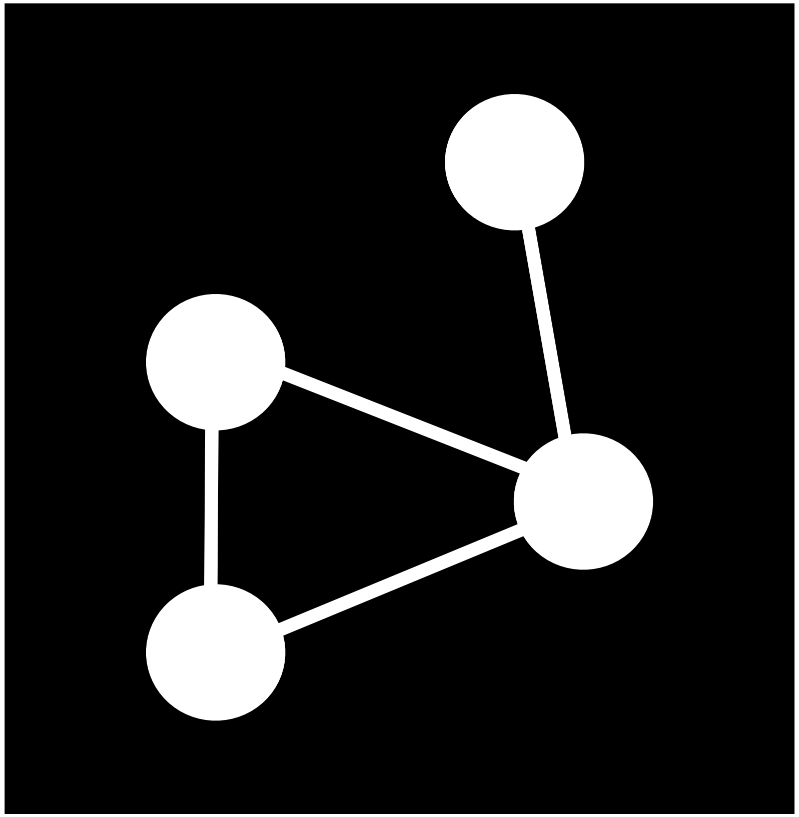
“All models are wrong.
Some models are useful.”

This talk will discuss network data

Networks or graphs are useful as a way of simplifying a complex system of interactions.

Facebook: edges could represent posting / commenting / liking

Biology: edges could represent something causal

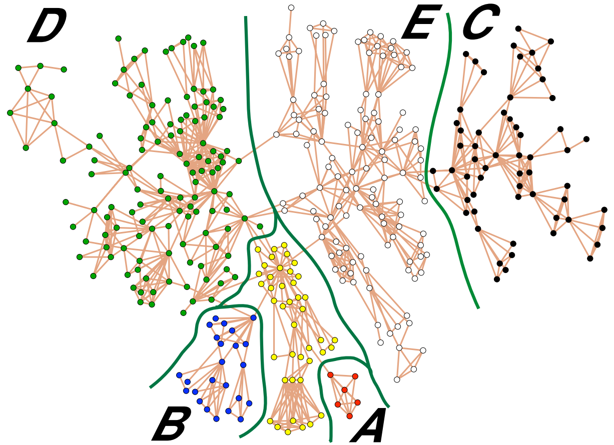


Graph = (node set, edge set)

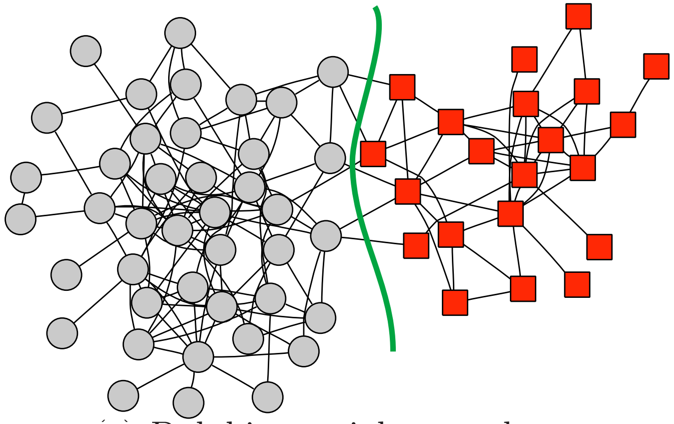
Partitions, clusters, modularities, communities represent latent structure

Algorithmic aim:
put "similar" nodes in the same set and "different" nodes in different sets.

- Edges simplify the underlying dyadic relationships.
- Communities suggest some latent structure in generating mechanism.



Network science network

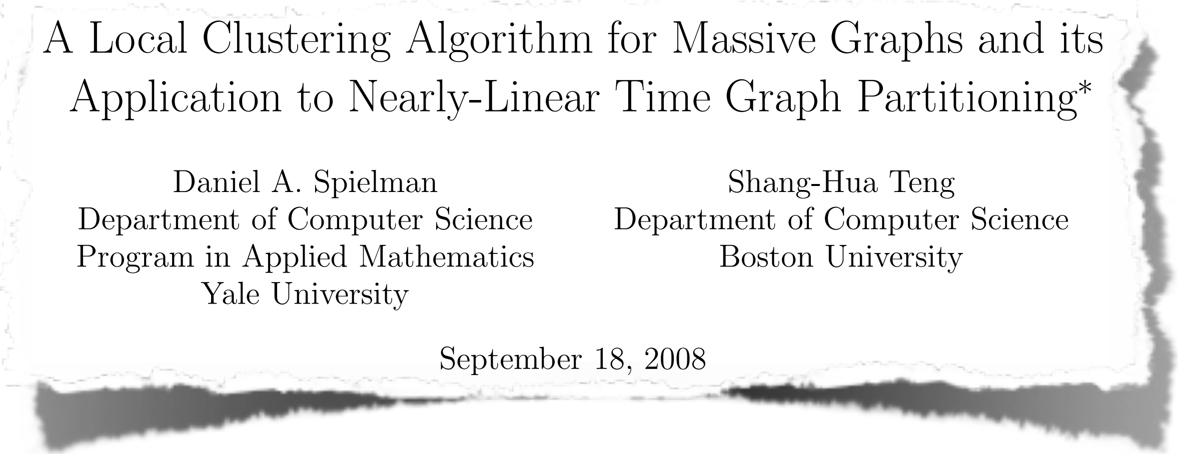


Dolphins social network

Leskovec et. al 2008

Local clustering algorithms were first proposed by Spielman and Teng.

- Spielman and Teng, 2008. “A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning”
    - really fast
    - empirical success
    - current theory:
        - perturbation bounds for graph conductance
        - bounds for running time.

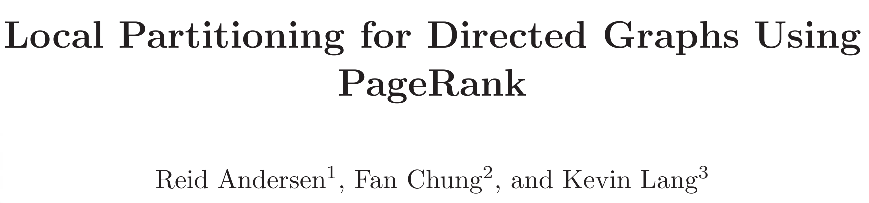


A Local Clustering Algorithm for Massive Graphs and its Application to Nearly-Linear Time Graph Partitioning*

Daniel A. Spielman
Department of Computer Science
Program in Applied Mathematics
Yale University

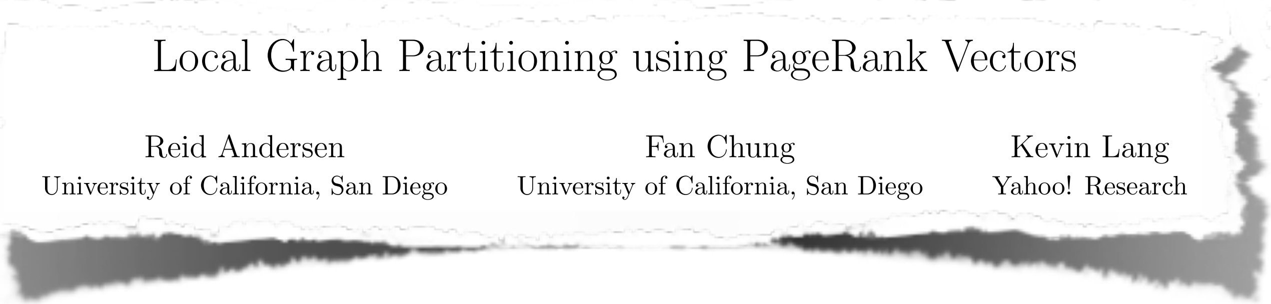
Shang-Hua Teng
Department of Computer Science
Boston University

September 18, 2008



Local Partitioning for Directed Graphs Using PageRank

Reid Andersen[1], Fan Chung[2], and Kevin Lang[3]

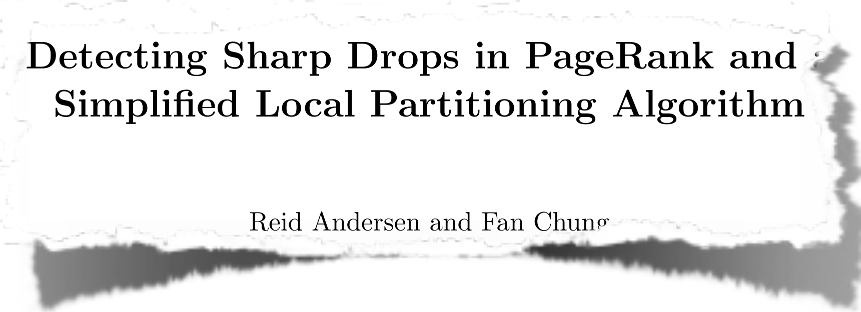


Local Graph Partitioning using PageRank Vectors

Reid Andersen
University of California, San Diego

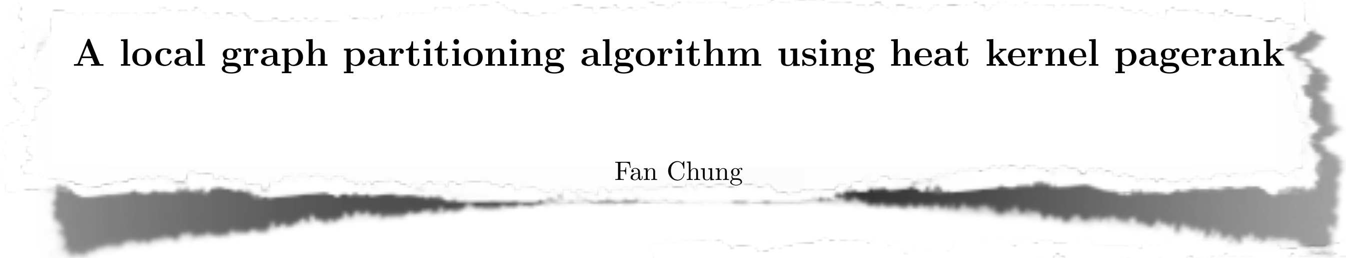
Fan Chung
University of California, San Diego

Kevin Lang
Yahoo! Research



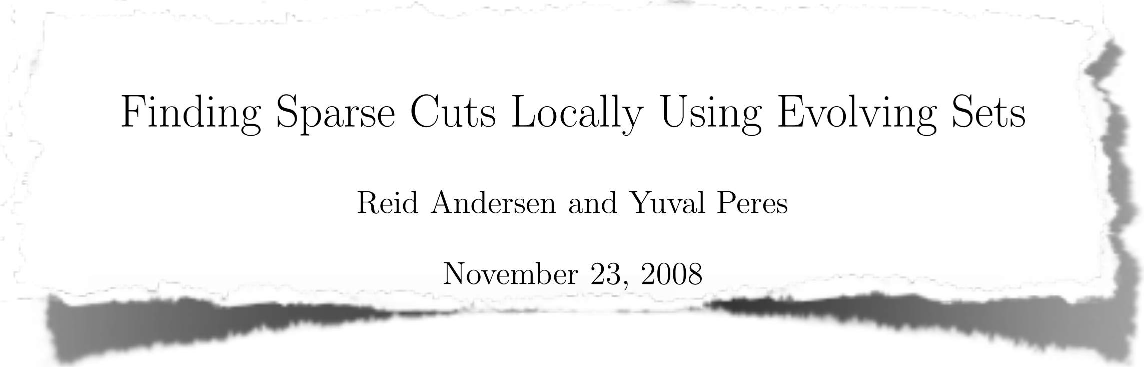
Detecting Sharp Drops in PageRank and Simplified Local Partitioning Algorithm

Reid Andersen and Fan Chung



A local graph partitioning algorithm using heat kernel pagerank

Fan Chung



Finding Sparse Cuts Locally Using Evolving Sets

Reid Andersen and Yuval Peres

November 23, 2008

This talk aims to provide a statistical framework for local clustering by

(1) showing that *sparse and transitive* Stochastic Blockmodels (aka planted partition models) naturally lead to local clustering.

(2) illustrating how the blessing of transitivity makes small clusters easy to estimate (both statistically and algorithmically).

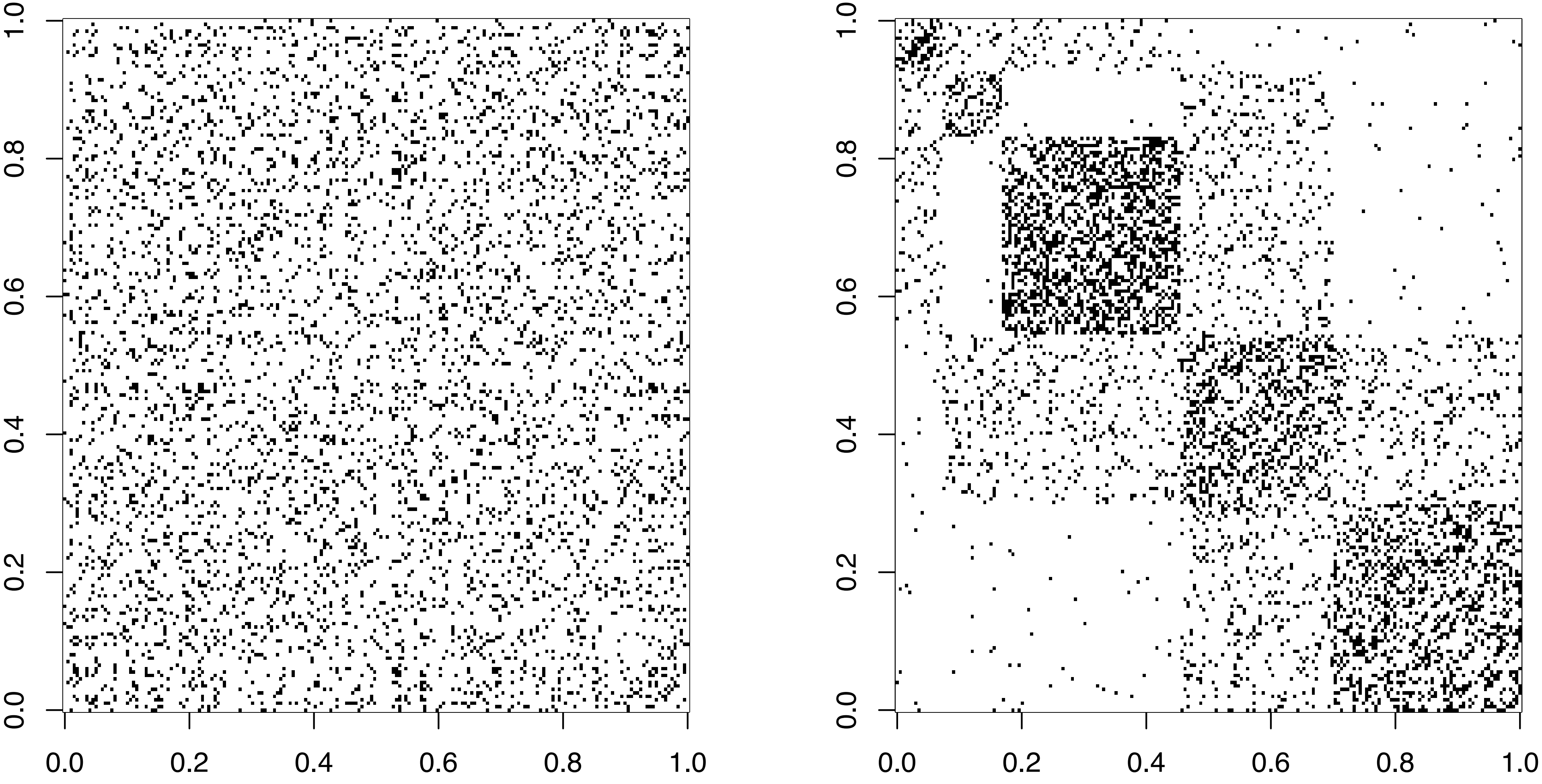
Consider the planted partition model (a simplified Stochastic Blockmodel)

- K = number of blocks
- s = population of each block. So, all blocks have equal population.
- r = probability of an out-of-block connection
- p = probability of an in-block connection

So, $n = Ks$

Assume r<p.

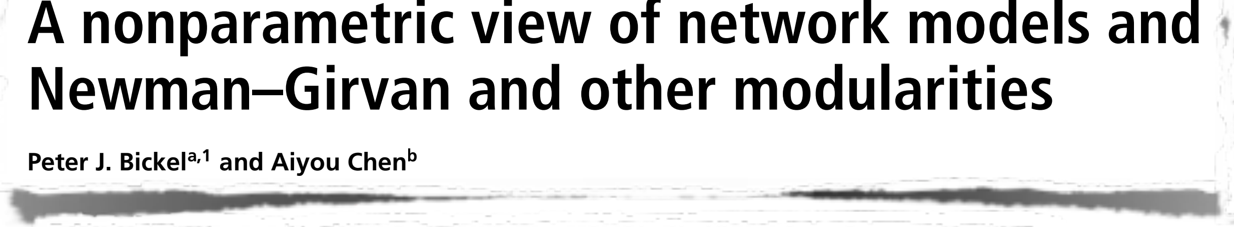
Adjacency matrix from the Stochastic Blockmodel



We want to estimate the partition of the nodes Z.

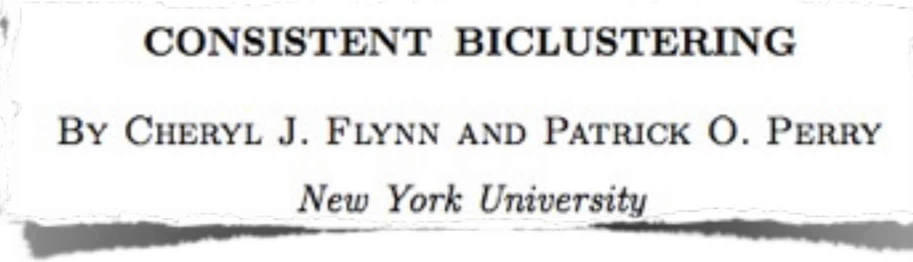
Extensive literature studies the estimation of Z

- Started in IEEE community
- McSherry. 2001. "Spectral partitioning of random graphs."
- Dasgupta, Hopcroft, and McSherry. 2004. "Spectral analysis of random graphs with skewed degree distributions."
- Great expansion in literature over past 4 years . . .



A nonparametric view of network models and Newman–Girvan and other modularities

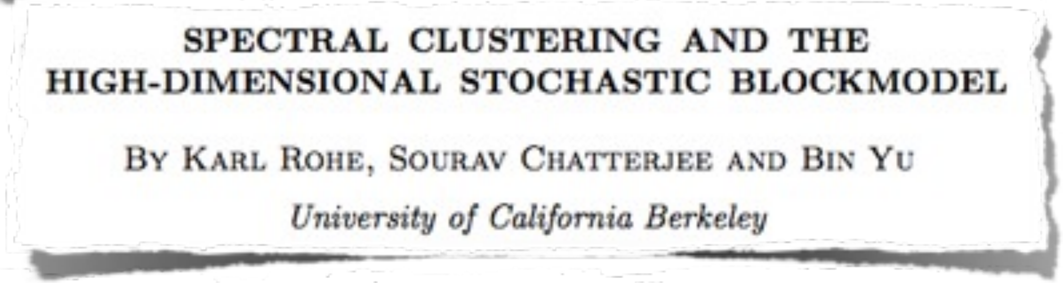
Peter J. Bickel[a,1] and Aiyou Chen[b]



CONSISTENT BICLUSTERING

BY CHERYL J. FLYNN AND PATRICK O. PERRY

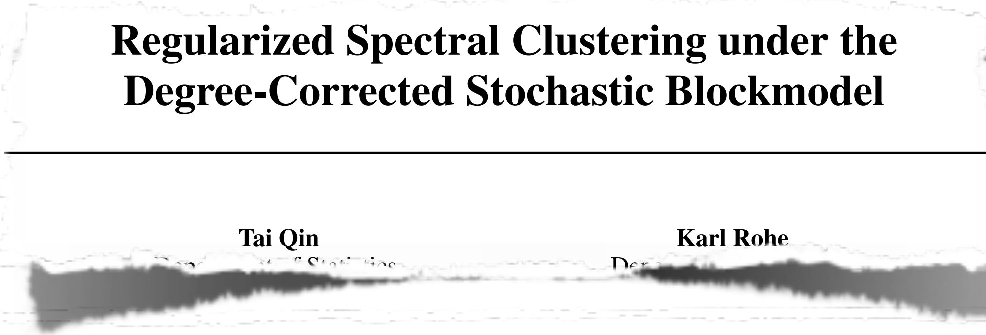
New York University



SPECTRAL CLUSTERING AND THE HIGH-DIMENSIONAL STOCHASTIC BLOCKMODEL

BY KARL ROHE, SOURAV CHATTERJEE AND BIN YU

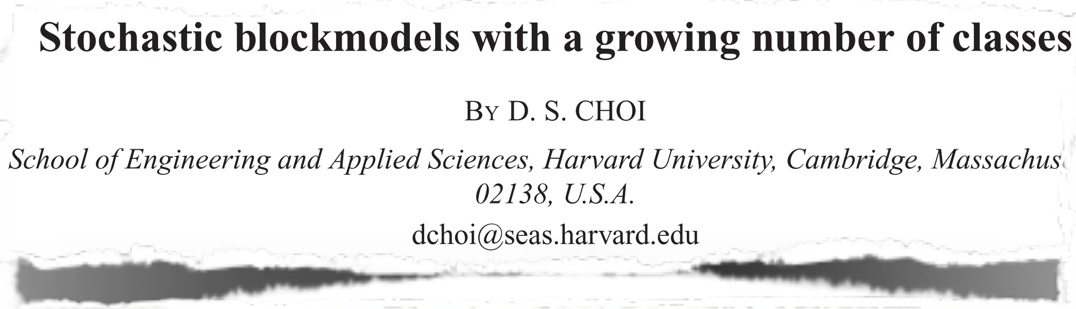
University of California Berkeley



Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel

Tai Qin

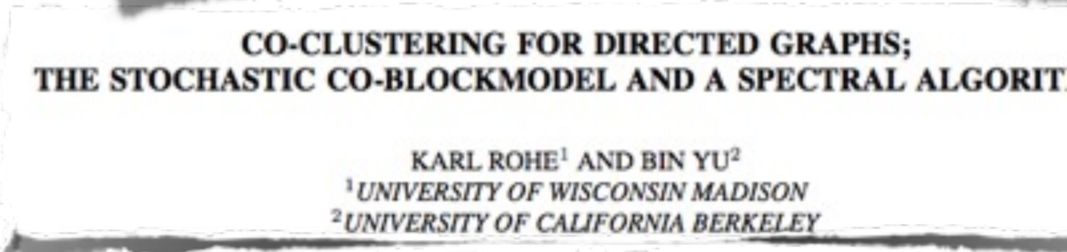
Karl Rohe



Stochastic blockmodels with a growing number of classes

BY D. S. CHOI

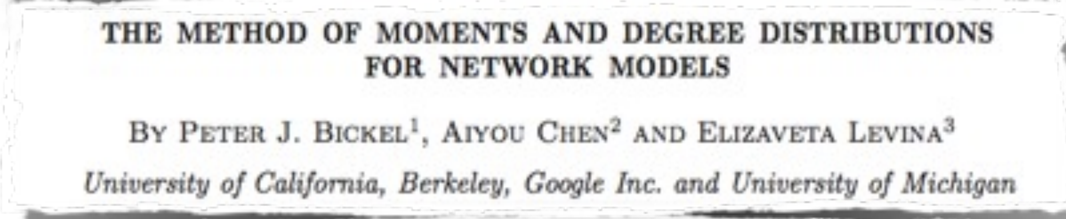
School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachus
02138, U.S.A.
dchoi@seas.harvard.edu



CO-CLUSTERING FOR DIRECTED GRAPHS; THE STOCHASTIC CO-BLOCKMODEL AND A SPECTRAL ALGORIT

KARL ROHE[1] AND BIN YU[2]

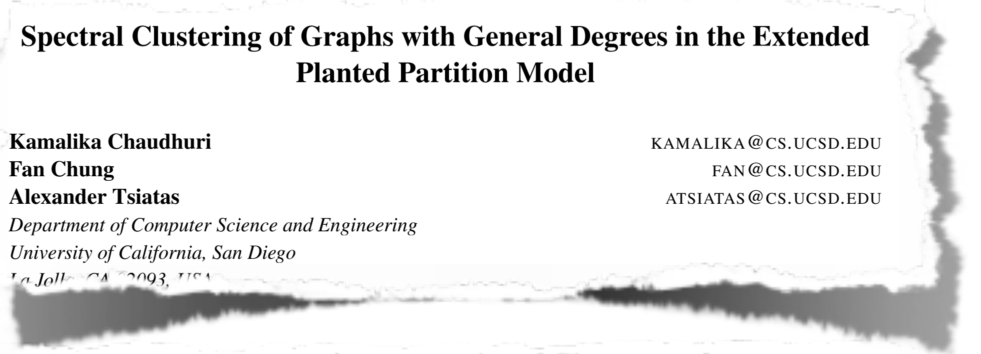
[1]*UNIVERSITY OF WISCONSIN MADISON*
[2]*UNIVERSITY OF CALIFORNIA BERKELEY*



THE METHOD OF MOMENTS AND DEGREE DISTRIBUTIONS FOR NETWORK MODELS

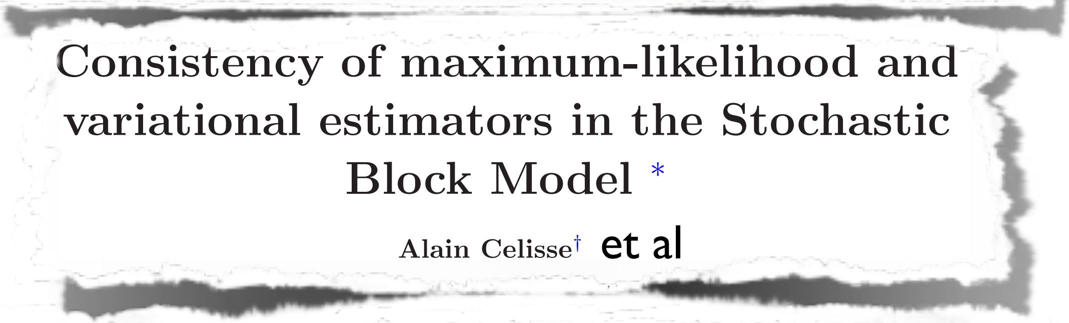
BY PETER J. BICKEL[1], AIYOU CHEN[2] AND ELIZAVETA LEVINA[3]

University of California, Berkeley, Google Inc. and University of Michigan



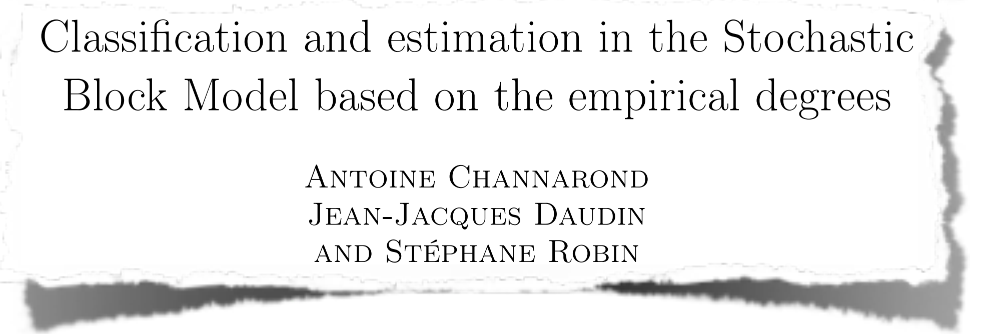
Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model

Kamalika Chaudhuri KAMALIKA@CS.UCSD.EDU
Fan Chung FAN@CS.UCSD.EDU
Alexander Tsiatas ATSIATAS@CS.UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA



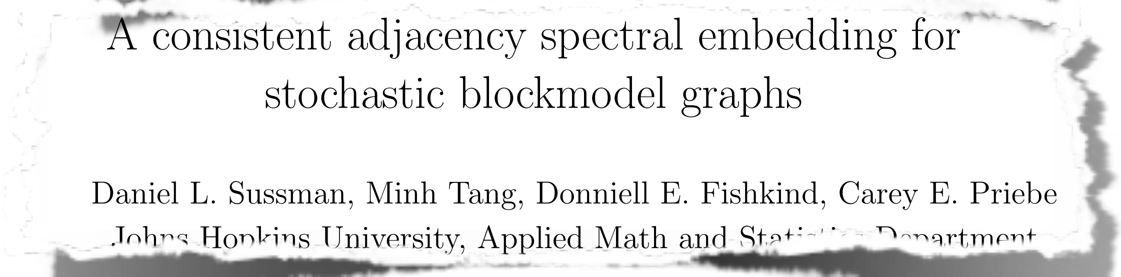
Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model *

Alain Celisse† et al



Classification and estimation in the Stochastic Block Model based on the empirical degrees

ANTOINE CHANNAROND
JEAN-JACQUES DAUDIN
AND STÉPHANE ROBIN



A consistent adjacency spectral embedding for stochastic blockmodel graphs

Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, Carey E. Priebe
Johns Hopkins University, Applied Math and Statistics Department

Current theory suggests that three quantities regulate the difficulty of estimating Z.

1. Edge sparsity. More edges are better.
2. Size of the smallest cluster. Bigger is better.
3. Difference between in-block and out-of-block probabilities (when they are equal the model is unidentifiable!)

A statistical framework for local clustering

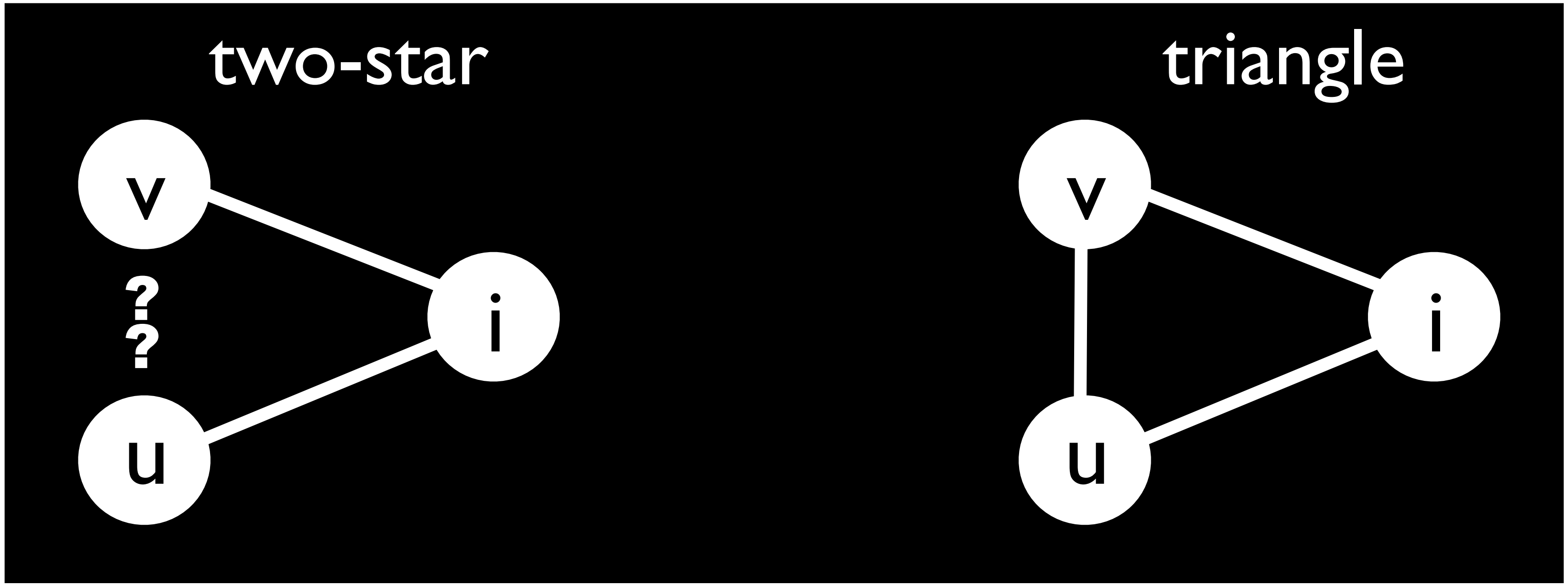
(1) *Sparse and transitive* models (stochastic block or planted partition) naturally lead to local clustering with large p, vanishing r.

(2) this blessing of transitivity makes small clusters easy to estimate (both statistically and algorithmically), even under "semi-parametric" models

Empirical networks are sparse and transitive.

- Empirically, the average node degree in most networks is between 10 and 100, even in "massive graphs."
- Moreover, most networks have many more triangles than we would expect under an Erdős–Rényi random graph.

Transitivity: Friends-of-friends are likely to be friends.



$$\text{transitivity ratio} \propto \frac{\text{number of triangles}}{\text{number of 2 stars}}$$

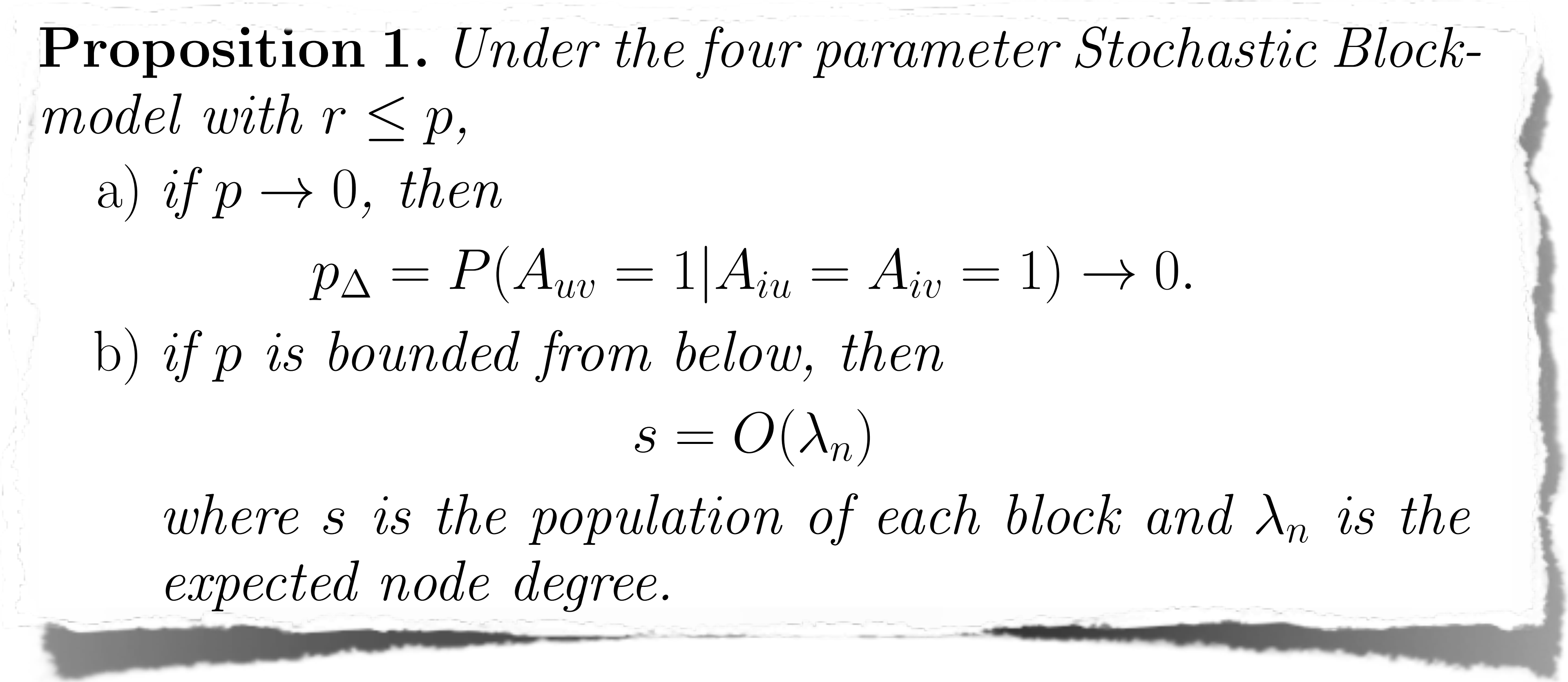
Another popular measure of transitivity is the clustering coefficient. (Watts and Strogatz 1998)

Under the planted partition model with $r < p$.

a) If p goes to zero, then you remove transitivity.

b) If p is bounded from below, then block size cannot grow faster than the expected degree.

Under the planted partition model with r < p.
a) If p goes to zero, then you remove transitivity.
b) If p is bounded from below, then block size cannot grow faster than the expected degree.



Proposition 1. *Under the four parameter Stochastic Blockmodel with $r \leq p$,*

a) *if $p \to 0$, then*

$$p_\Delta = P(A_{uv} = 1 | A_{iu} = A_{iv} = 1) \to 0.$$

b) *if p is bounded from below, then*

$$s = O(\lambda_n)$$

where s is the population of each block and λ_n is the expected node degree.

Similar results to part a) hold under the more general Exchangeable Random Graph Model.

The blessing of transitivity

- In planted partition model with bounded expected degree, transitivity implies:

$$(i)\ p > \epsilon > 0,\ (ii)\ r = O(1/n), \quad \text{and } (iii)\ s = O(1)$$

Previous results suggest estimation could be very difficult.

- Previous results require (1) *growing* degrees and (2) *growing* blocks
- We want (1) *bounded* degree and thus (2) *bounded* blocks
- Still possible because r --> 0 and p is fixed.

Blessing of transitivity doesn’t make “bad” edges disappear. It makes “bad” triangles disappear.

In sparse and transitive planted partition model

(1) Out-of-block edges can be common. O(n)

(2) Out-of-block triangles are unlikely. O(1)

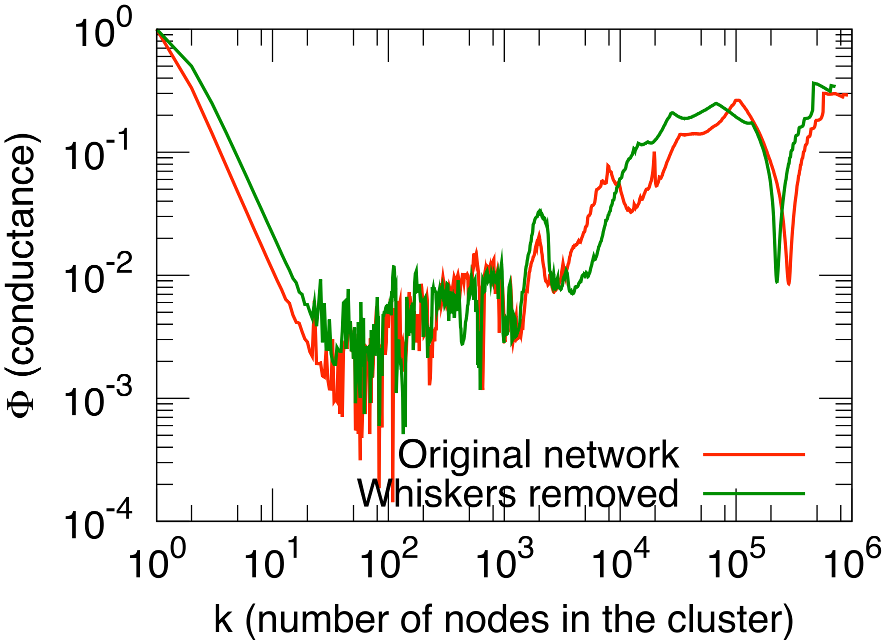
(3) In-block triangles are common. O(n)

Note: We look at triangles for computational purposes. Could look for 4-cliques, then (2) becomes O(1/n) and (3) gets a new constant.

Related research

- R, Qin, Fan 2012 “Highest dimensional Stochastic Blockmodel with a regularized estimator”
    - used the restricted MLE
- Verzelen and Arias-Castro, 2013. “Community Detection in Sparse Random Networks”
  - Hypothesis testing
    - Ho: massive, sparse Erdos–Rényi
    - Ha: hidden block, growing faster than log(N).
  - Number of triangles is a powerful test statistic.

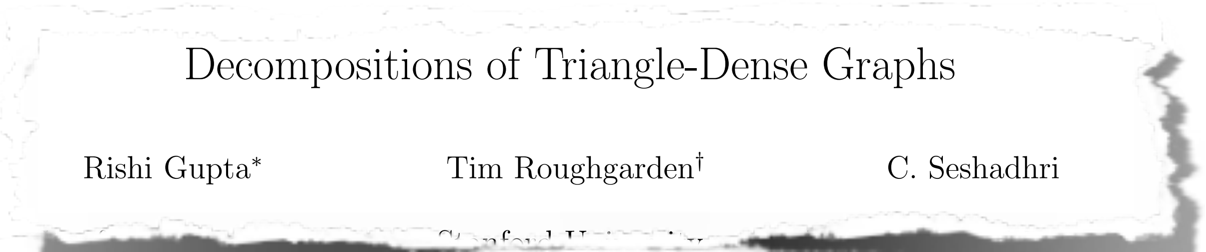
Empirical networks contain small communities.



(a) LiveJournal01

- Leskovec, Lang, Dasgupta, Mahoney 2008. In large empirical networks, communities with smallest conductance are no larger than 100 nodes.

- Transitivity plays a key role and we should use triangles to discover the local clusters.



Decompositions of Triangle-Dense Graphs

Rishi Gupta[*] Tim Roughgarden[†] C. Seshadhri

Stanford University

A statistical framework for local clustering

(1) *Sparse and transitive* models (stochastic block or planted partition) naturally lead to local clustering with large p, vanishing r.

(2) this blessing of transitivity makes small clusters easy to estimate (both statistically and algorithmically), even under "semi-parametric" models

Weaving together the two threads.

- Started off talking about local clustering algorithms.
- Then, we forgot that and used SBM + sparsity + transitivity to get a model with small blocks.
- Last section combines these two threads.
  - Propose a local clustering algorithm that looks for triangles.
  - Study the estimation performance of this algorithm under a <u>local</u> SBM. This is a *semi-parametric* network model.

There are global and local versions of our algorithm

Global:

Define $L^{\tau}_{ij} = [D_{\tau}^{-1/2} A D_{\tau}^{-1/2}]_{ij}$

where $[D_{\tau}]_{ii} = deg(i) + \tau$

Compute $[L^{\tau} L^{\tau}]_{ij} L^{\tau}_{ij}$

Run single linkage hierarchical clustering
(i.e. find maximum spanning tree)

There are global and local versions of our algorithm

Global:

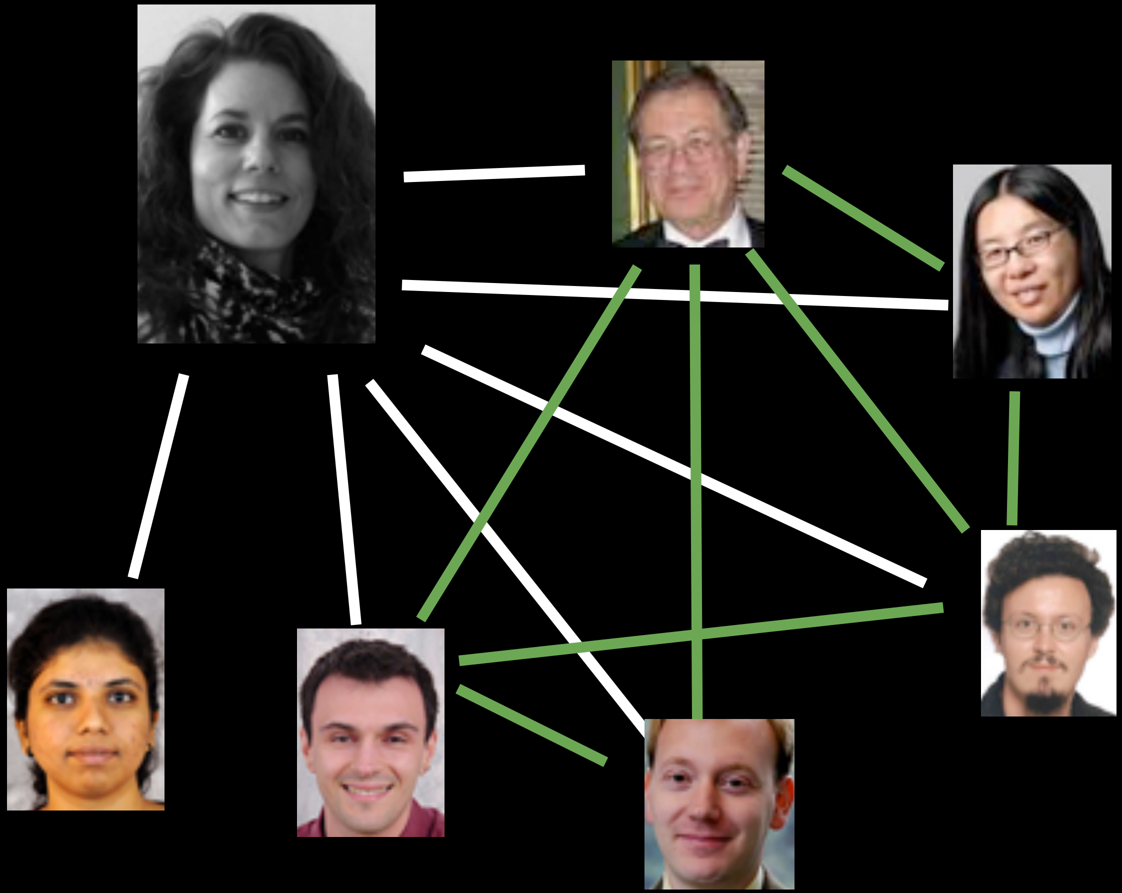
Define $L^{\tau}_{ij} = [D_{\tau}^{-1/2} A D_{\tau}^{-1/2}]_{ij}$

where $[D_{\tau}]_{ii} = deg(i) + \tau$

Compute $[L^{\tau} L^{\tau}]_{ij} L^{\tau}_{ij}$

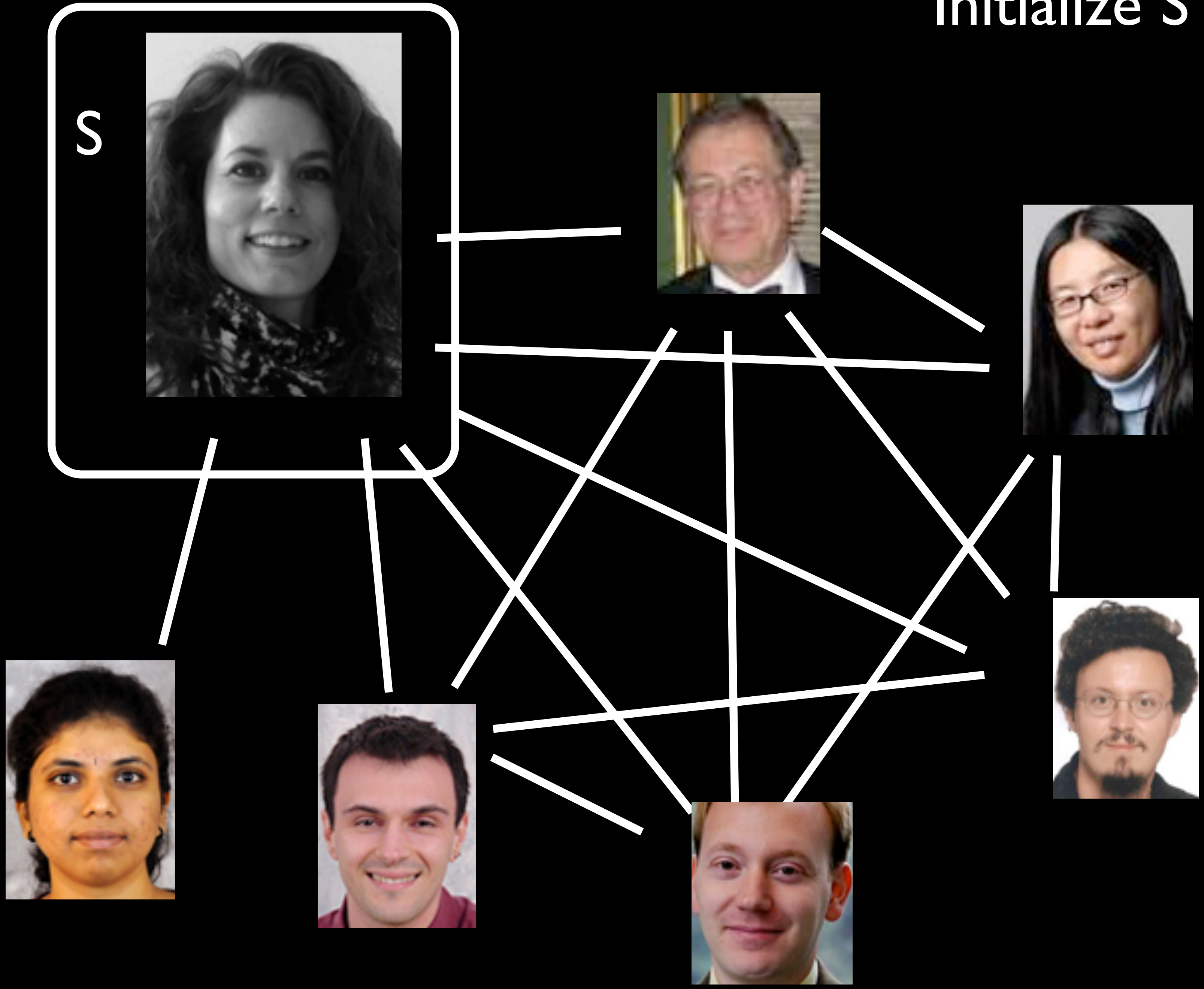
“regularized graph Laplacian” proposed by Chaudhuri, Chung, and Tsiatas (2012) and Amini, Chen, Bickel, and Levina (2012).

tau = average degree is a reasonable choice (Qin, R 2013)

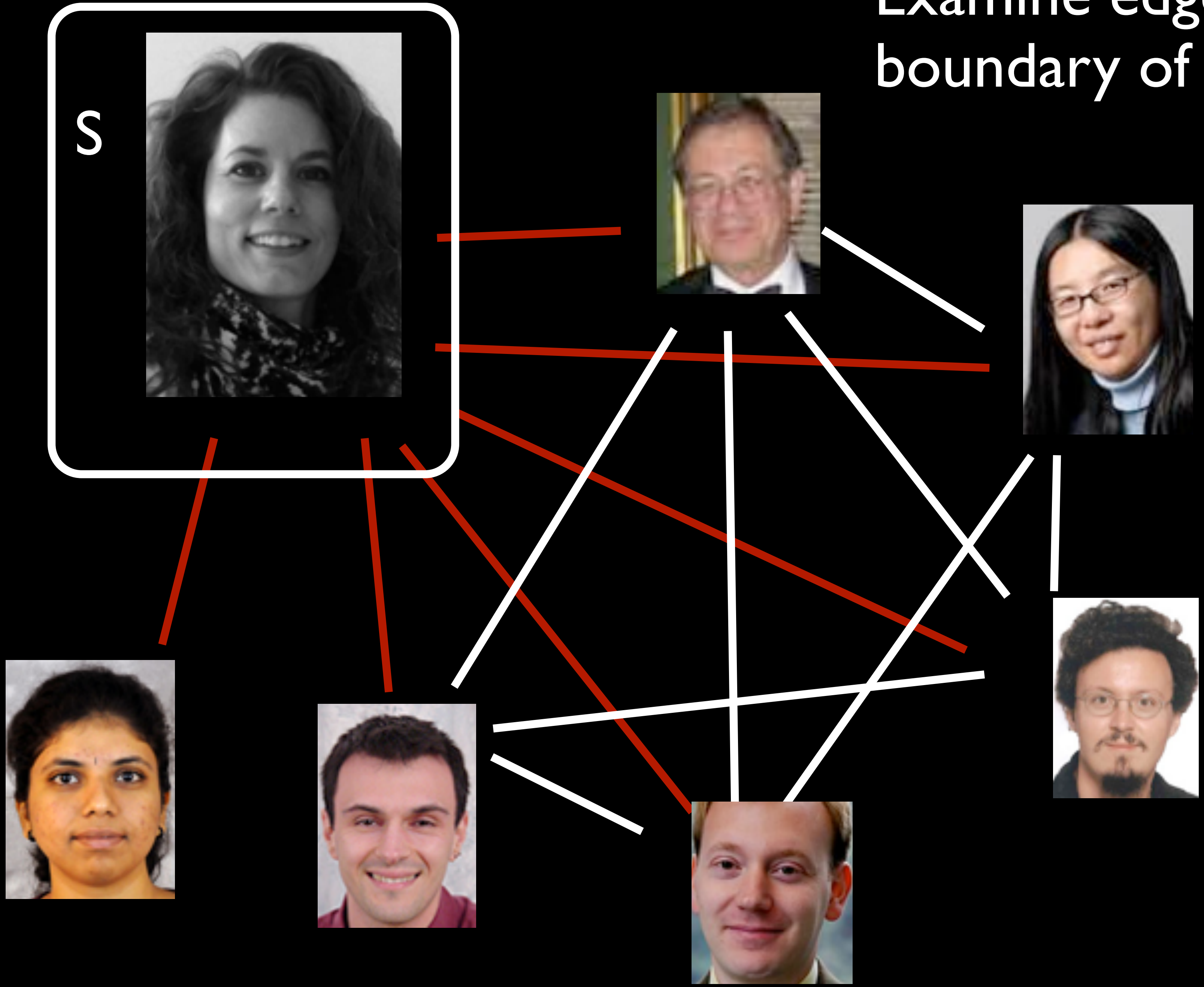


Our local algorithm searches for edges in several triangles...

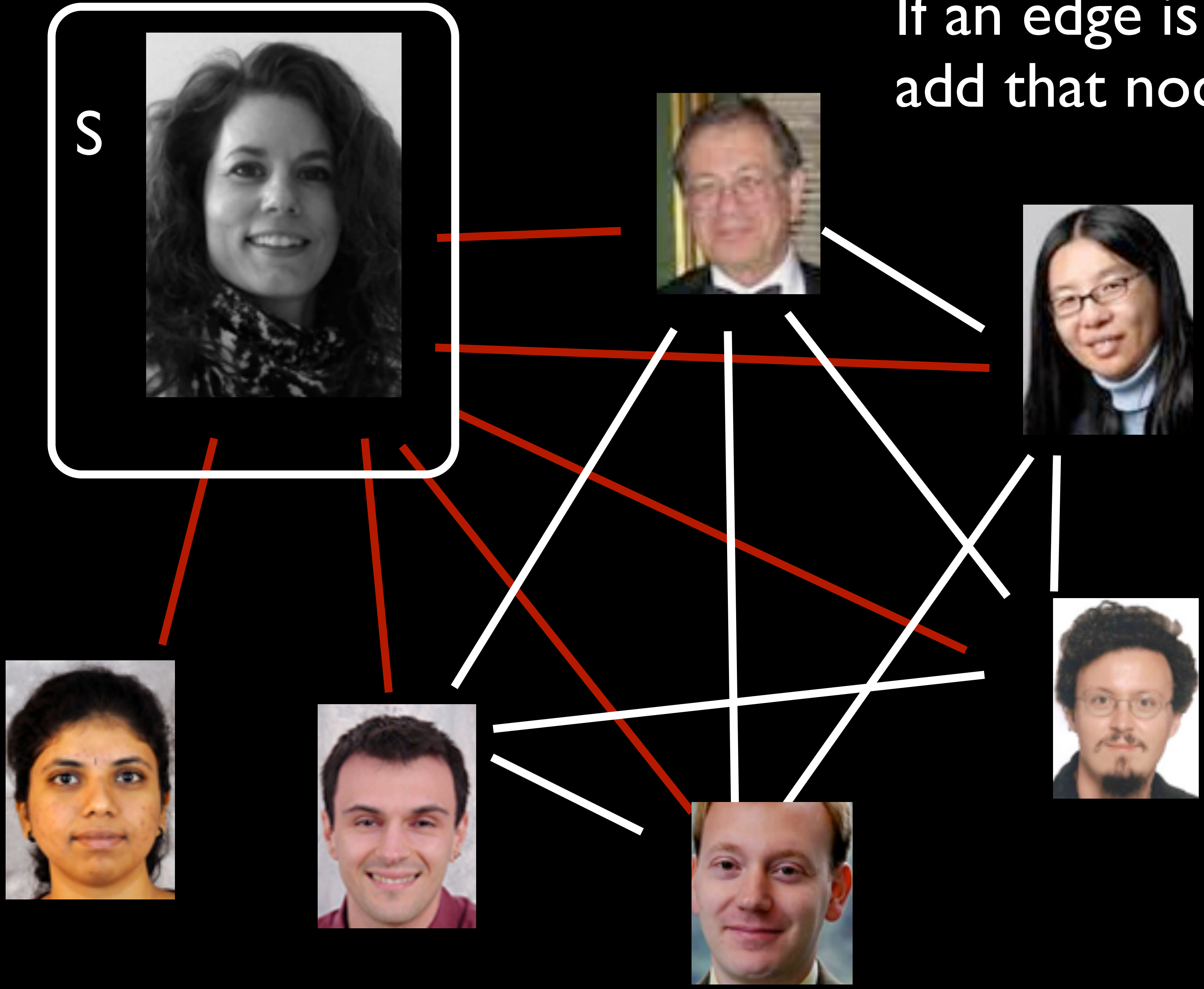
Initialize S = {Jennifer}



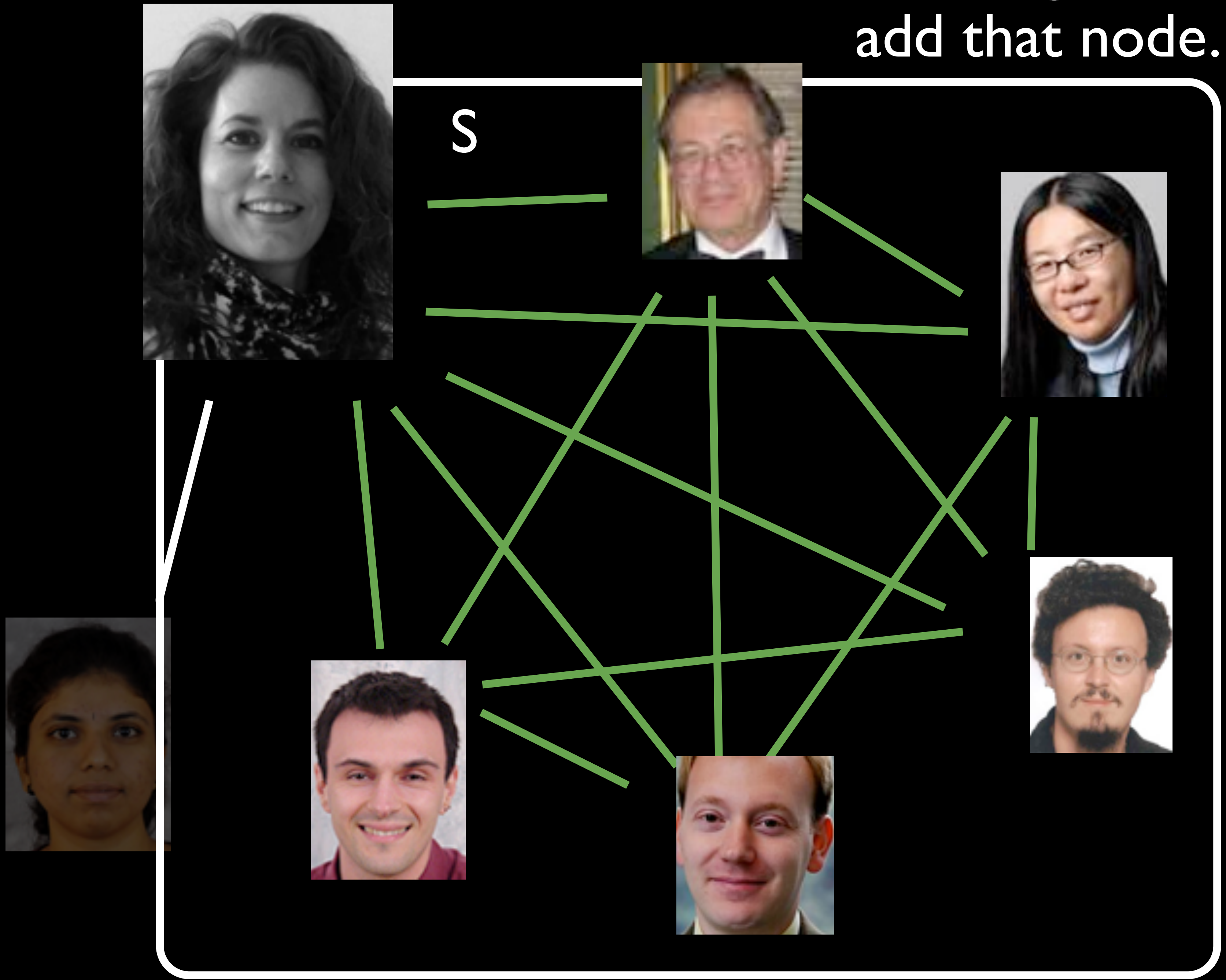
Examine edges crossing boundary of S.



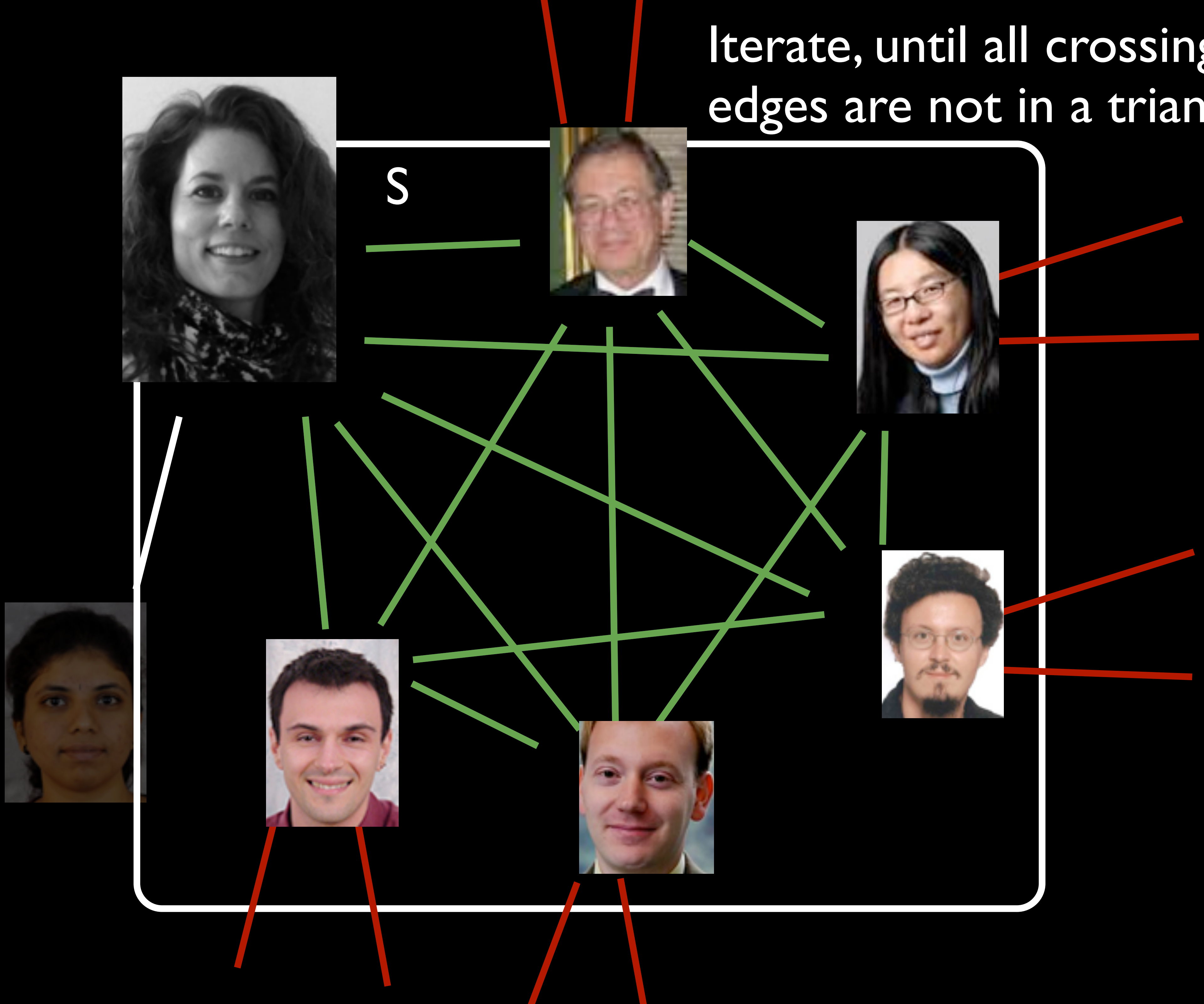
If an edge is in a triangle, add that node.



If an edge is in a triangle, add that node.



Iterate, until all crossing edges are not in a triangle.



Modification 1

- Tuning parameter: the required number of triangles.
    - regulates {size vs density} of cluster

Modification 1

Find the entire path of solutions, for every seed node:

$$O^3 \in R^{n\times n}. \;\; O^3_{ij} = [AA]_{ij} A_{ij}$$

O^3_{ij} Number of triangles that contain edge (i,j).

This is a similarity matrix.
Use it to perform single linkage hierarchical clustering.

Cut the dendrogram at the required number of triangles.
Returns ALL local clusters.

Modification 2.

- Edges from high degree nodes need to be down weighted.
- Replace the adjacency matrix with a row and column normalized version.

The regularized graph Laplacian

$$[D_\tau]_{ii} = deg(i) + \tau$$

$$L^\tau_{ij} = [D_\tau^{-1/2} A D_\tau^{-1/2}]_{ij} = \frac{A_{ij}}{\sqrt{[D_\tau]_{ii}[D_\tau]_{jj}}}$$

Proposed for spectral clustering by Chaudhuri, Chung, and Tsiatas (2012) and Amini, Chen, Bickel, and Levina (2012).

tau = average degree is a reasonable choice
(Qin, R 2013)

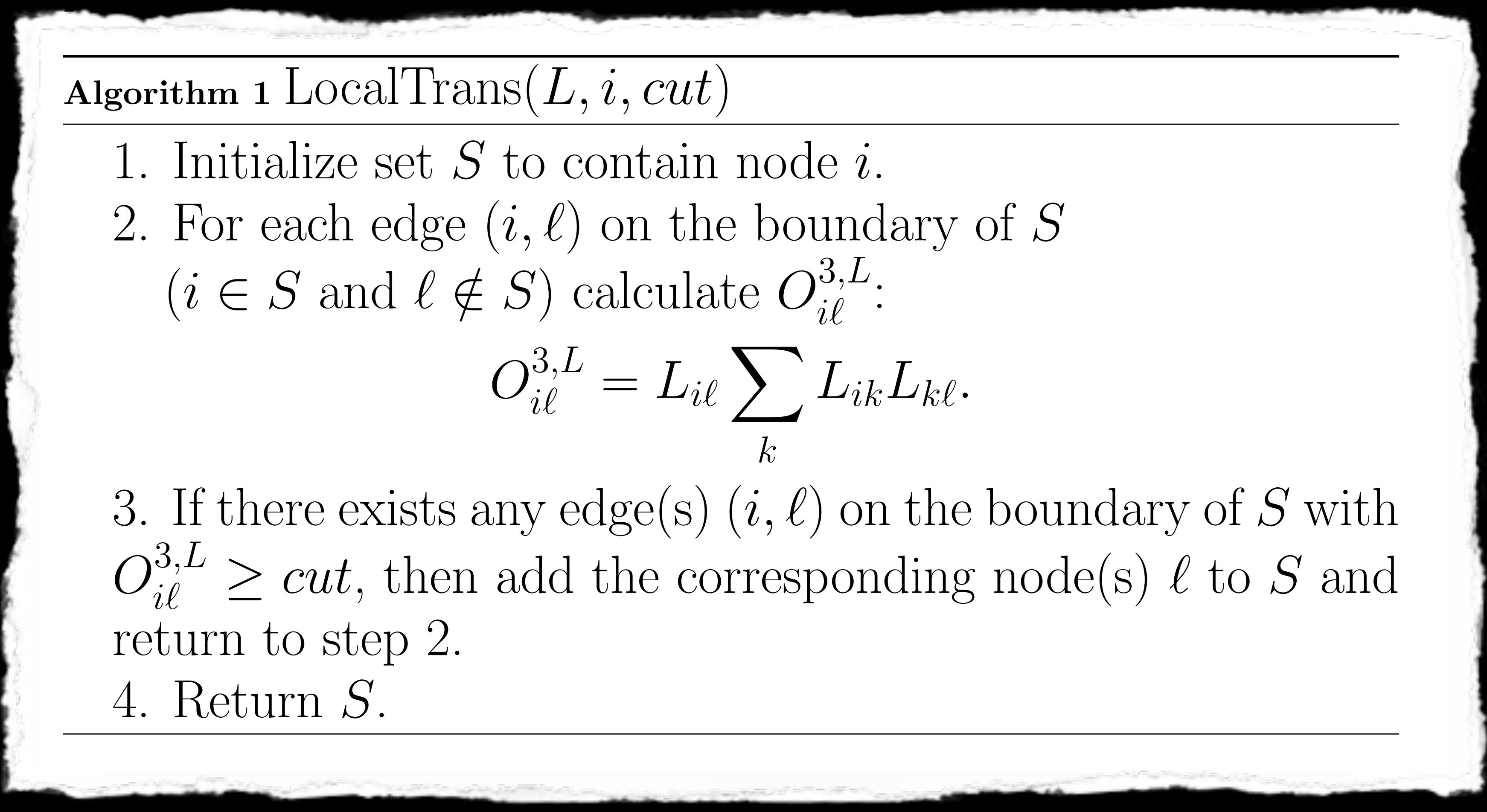
Modification 2

Replace $O^3_{ij} = [AA]_{ij} A_{ij}$

with $O^{3,L}_{ij} = [L^\tau L^\tau]_{ij} L^\tau_{ij}$

The local version of the algorithm only searches over edges connecting to the final cluster.

- To find one branch of the dendrogram . . .



Algorithm 1 LocalTrans(L, i, cut)

```
1. Initialize set $S$ to contain node $i$.
2. For each edge $(i, \ell)$ on the boundary of $S$
   ($i \in S$ and $\ell \notin S$) calculate $O_{i\ell}^{3,L}$:
       $$O_{i\ell}^{3,L} = L_{i\ell} \sum_k L_{ik} L_{k\ell}.$$
3. If there exists any edge(s) $(i, \ell)$ on the boundary of $S$ with
$O_{i\ell}^{3,L} \geq cut$, then add the corresponding node(s) $\ell$ to $S$ and
return to step 2.
4. Return $S$.
```

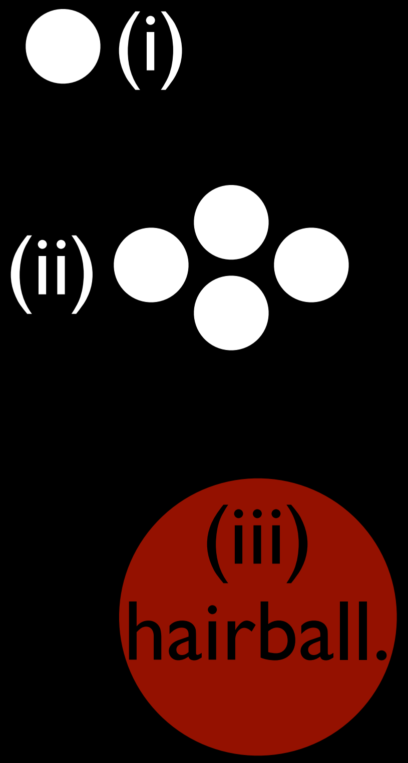
A local model to study a local algorithm.

Since the algorithm only searches locally, we should have a “local model.”

WAVE HANDS. SHOW MODEL IN TWO PARTS.

Local and semi-parametric Stochastic Blockmodel

- Network model contains three parts
  (i) the seed node,
  (ii) its local block,
  (iii) the rest of the network.
- Assume (iii) is sparse and is independent of edges from {(i) and (ii)} to (iii).
- No parametric / degree distribution / maximum degree assumptions on (iii).
- Asymptotically, (iii) grows and (ii) stays fixed.
- Node degrees in (ii) cannot grow too fast.



Local models

- First version is not degree corrected
    - good performance with unweighted triangle counting (i.e. use A)
    - This algorithm has not worked with real data.
- Second version is degree corrected.
    - use L.
    - This algorithm has worked with data.

Theorem

If:

1) correct seeding
2) ambient graph is sparse

then the local algorithm returns the correct clusters whp

Theorem

If:

1) correct seeding
2) ambient graph is sparse

then the local algorithm returns the correct clusters whp

Does **not** require:

1) growing degrees
2) growing s
3) specified model on the hairball
4) bounded degree in the hairball

Theorem

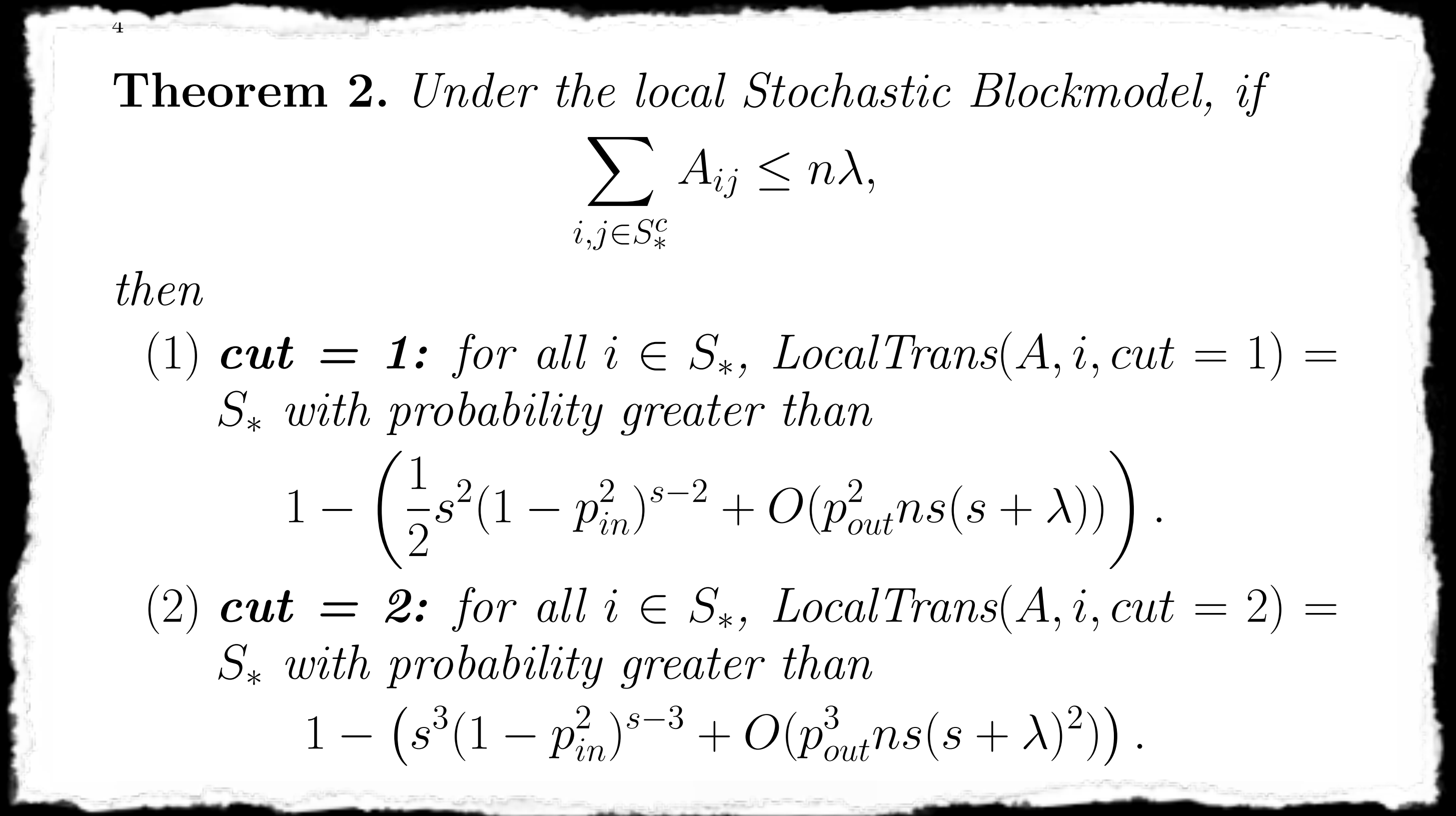
If:

1) correct seeding
2) ambient graph is sparse

then the local algorithm returns the correct clusters whp

Does **not** require:

1) growing degrees
2) growing s
3) specified model on the hairball
4) bounded degree in the hairball



Theorem 2. *Under the local Stochastic Blockmodel, if*

$$\sum_{i,j \in S_*^c} A_{ij} \leq n\lambda,$$

then

(1) ***cut = 1:*** *for all $i \in S_*$, LocalTrans$(A, i, cut = 1) = S_*$ with probability greater than*

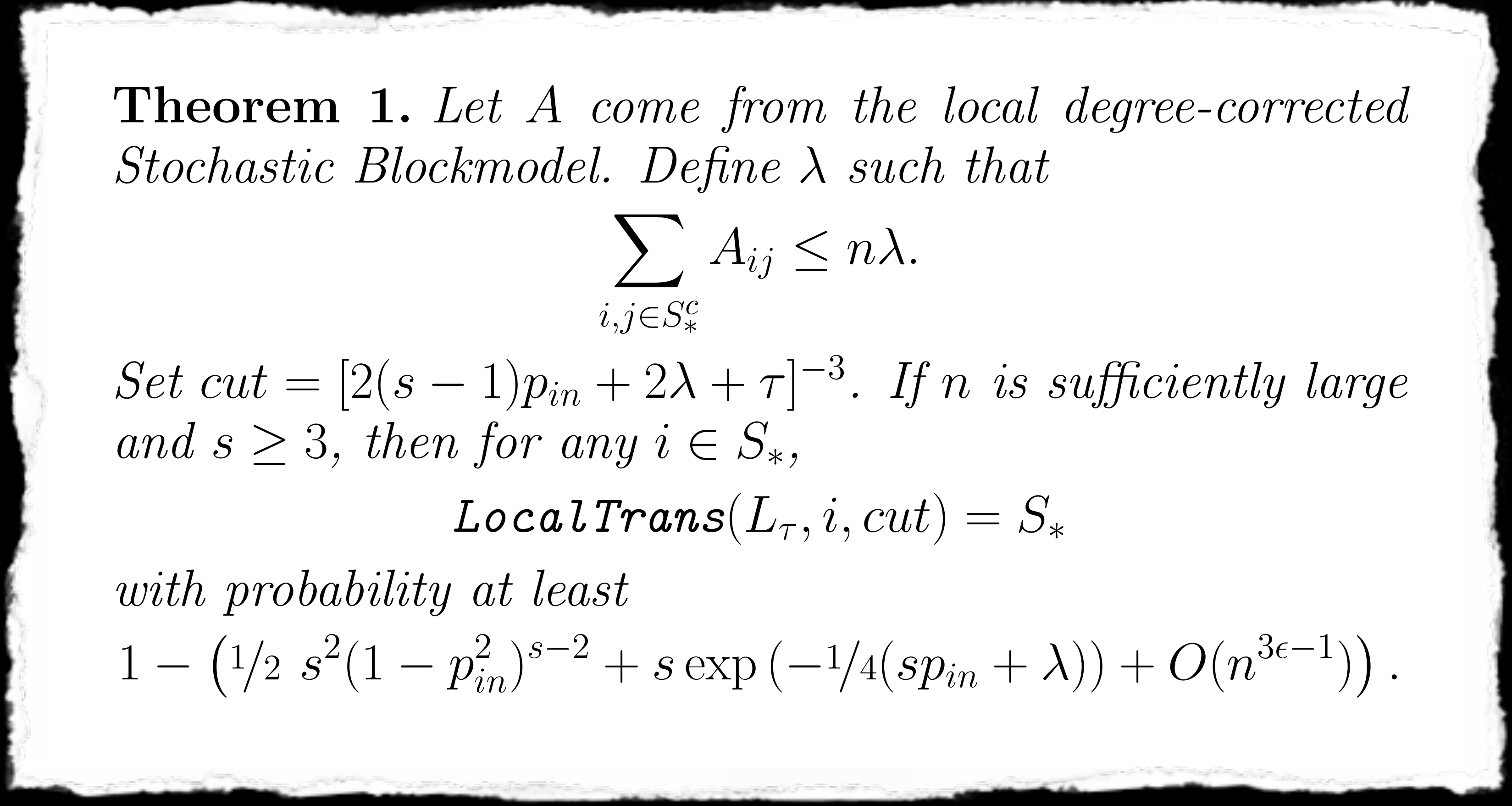
$$1 - \left(\frac{1}{2}s^2(1-p_{in}^2)^{s-2} + O(p_{out}^2 ns(s+\lambda))\right).$$

(2) ***cut = 2:*** *for all $i \in S_*$, LocalTrans$(A, i, cut = 2) = S_*$ with probability greater than*

$$1 - \left(s^3(1-p_{in}^2)^{s-3} + O(p_{out}^3 ns(s+\lambda)^2)\right).$$

Probability does not converge to one when s is fixed.

Similar theorem for the degree corrected model using Ltau



Theorem 1. *Let A come from the local degree-corrected Stochastic Blockmodel. Define λ such that*

$$\sum_{i,j \in S_*^c} A_{ij} \leq n\lambda.$$

Set $cut = [2(s-1)p_{in} + 2\lambda + \tau]^{-3}$. If n is sufficiently large and $s \geq 3$, then for any $i \in S_*$,

$$\texttt{LocalTrans}(L_\tau, i, cut) = S_*$$

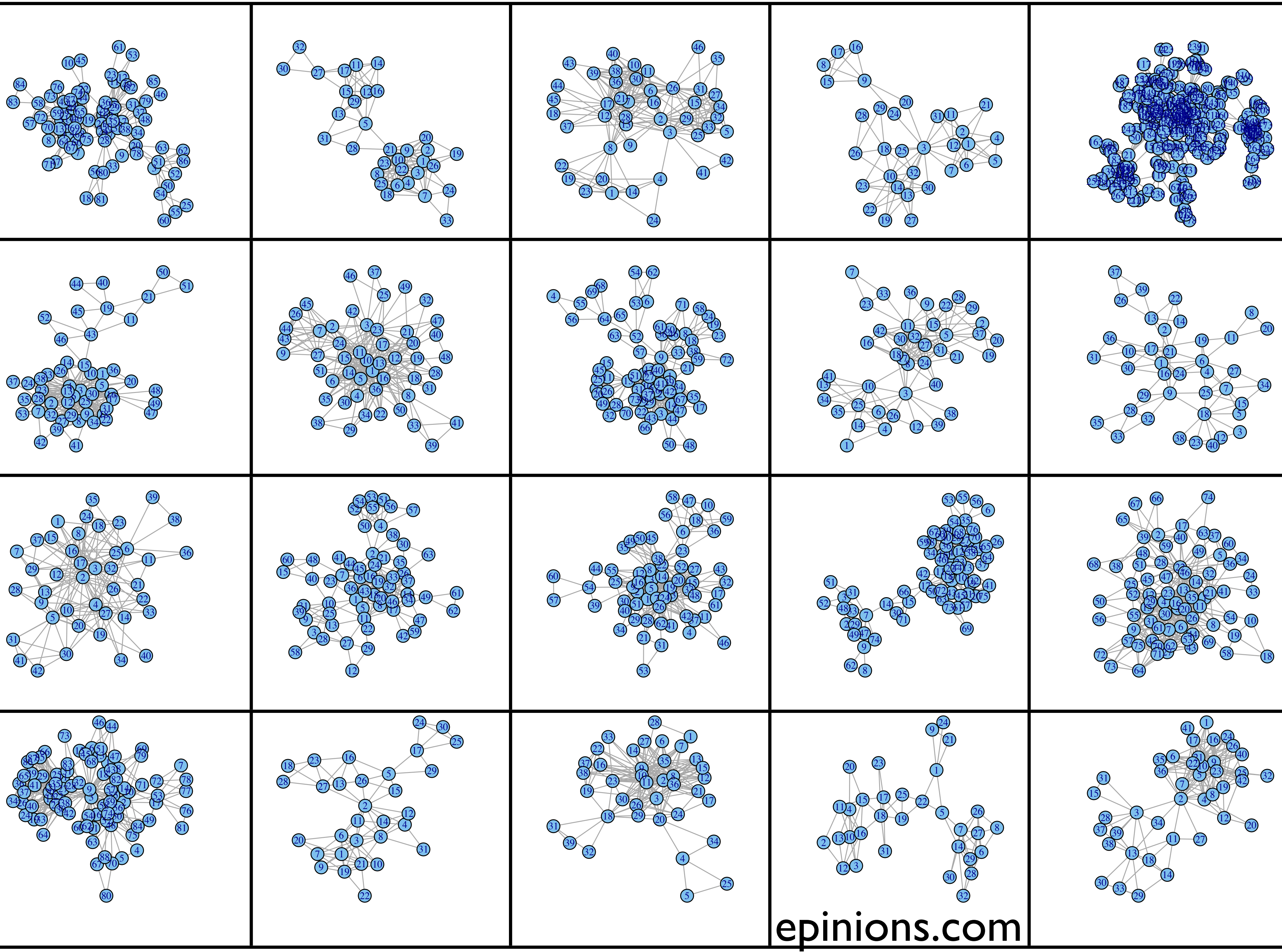
with probability at least

$$1 - \left({}^1\!/\!{}_2\ s^2(1-p_{in}^2)^{s-2} + s\exp\left(-{}^1\!/\!{}_4(sp_{in}+\lambda)\right) + O(n^{3\epsilon-1})\right).$$

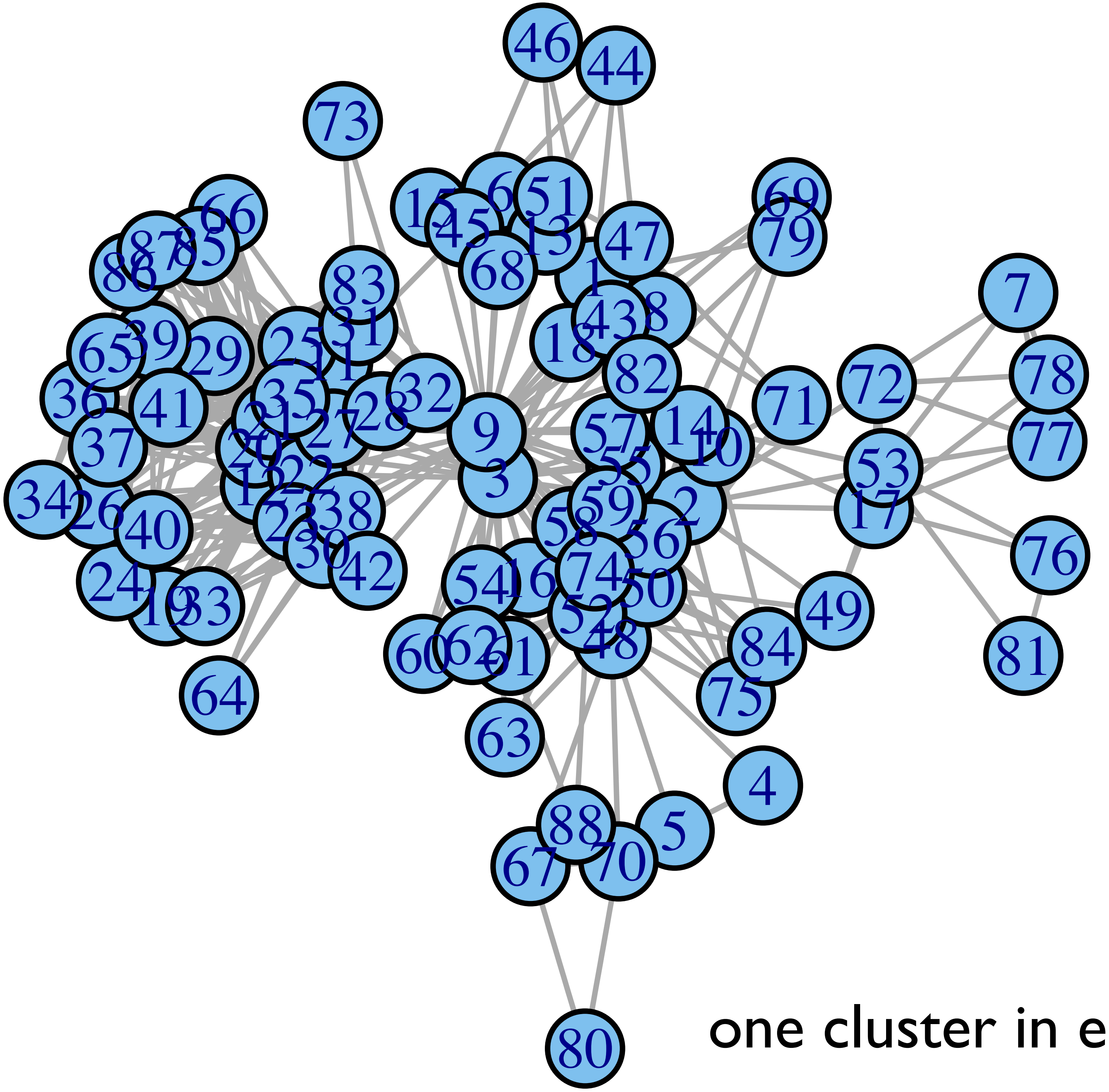
Probability does not converge to one when s is fixed.

The algorithm finds small communities in large networks.

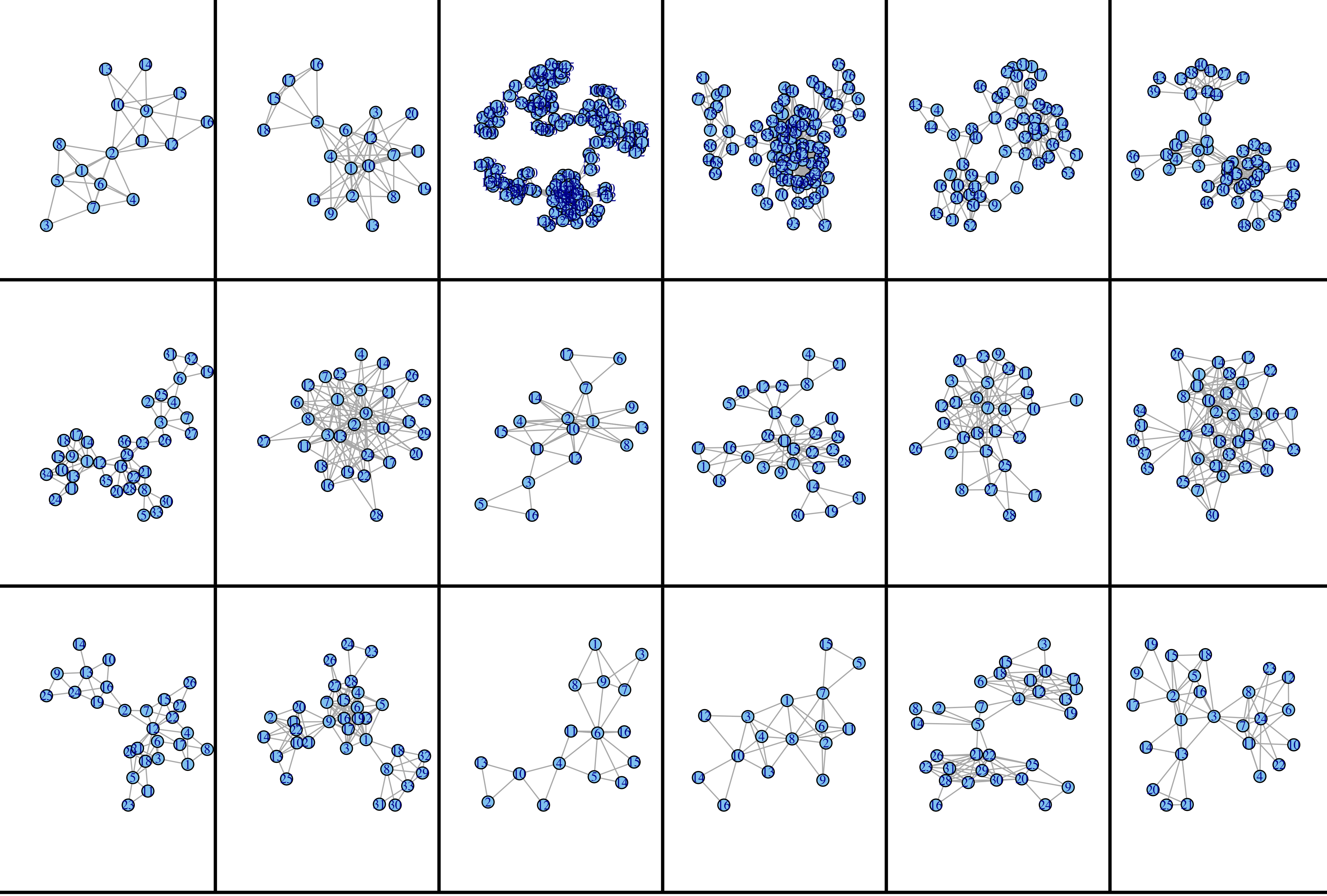
- Two online social networks.
  epinions and slashdot.
- Both 70,000+ nodes. Done on my wimpy laptop, in R!
- Data from http://snap.stanford.edu
- Next slides show the induced subgraphs of several local clusters



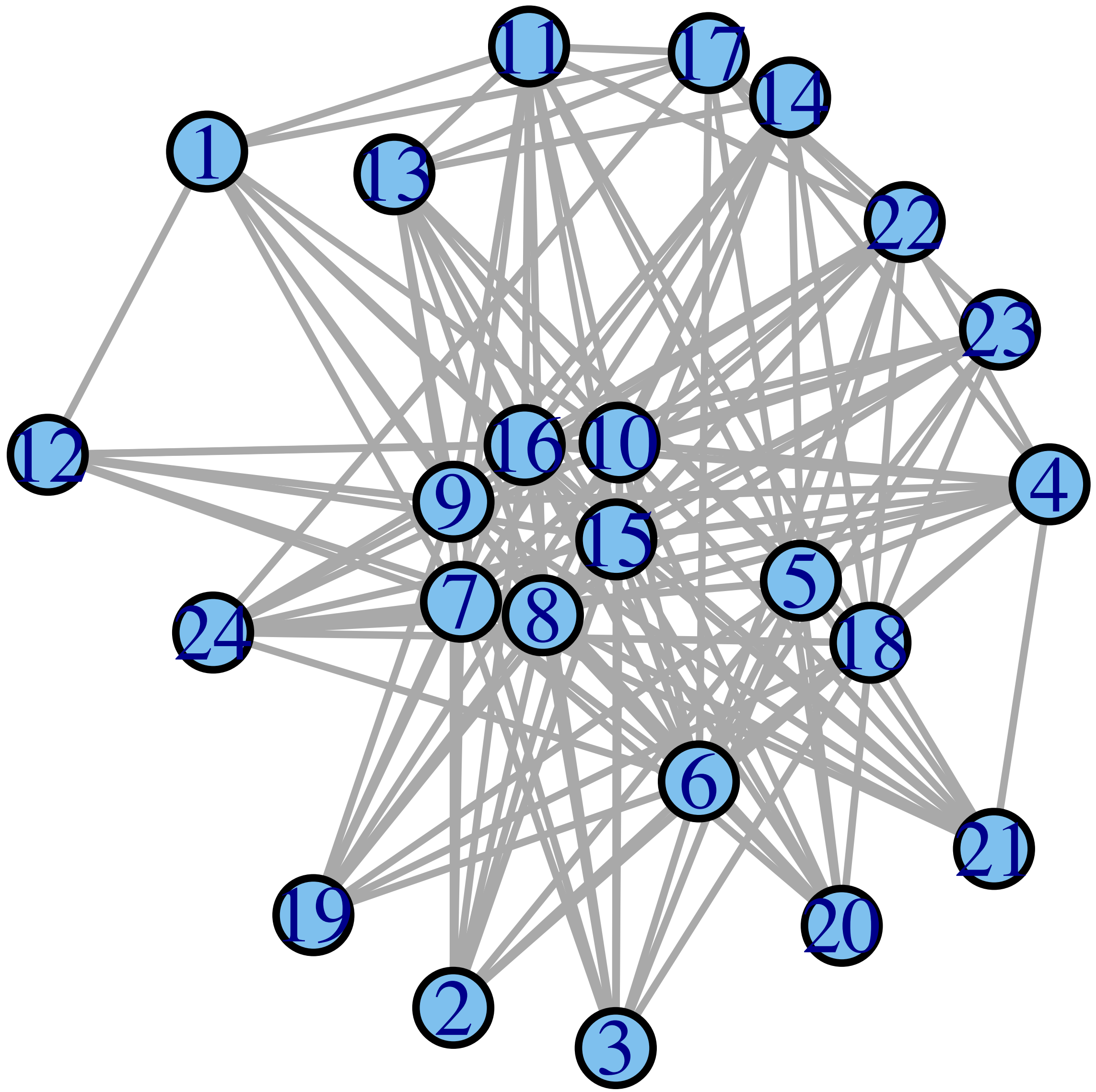
epinions.com



one cluster in epinions



slashdot.org



one cluster in slashdot.

Recap

1. Local clustering is justified for several reasons.
    a) new types of questions with massive networks, Dunbar’s number and other empirical evidence.
    b) computation, visualization, interpretation, diagnostics
2. Since we want to make inferences, we need a model with local clusters. Requires caution!
    a) Forgetting the motivation of local clusters...
    b) Sparse and transitive SBMs have small blocks.
3. Blessing of transitivity allows fast (local) algorithms and local inference with drastically reduced “global assumptions”

Recap   Thank you!

1. Local clustering is justified for several reasons.
   a) new types of questions with massive networks, Dunbar’s number and other empirical evidence.
   b) computation, visualization, interpretation, diagnostics
2. Since we want to make inferences, we need a model with local clusters. Requires caution!
   a) Forgetting the motivation of local clusters...
   b) Sparse and transitive SBMs have small blocks.
3. Blessing of transitivity allows fast (local) algorithms and local inference with drastically reduced “global assumptions”