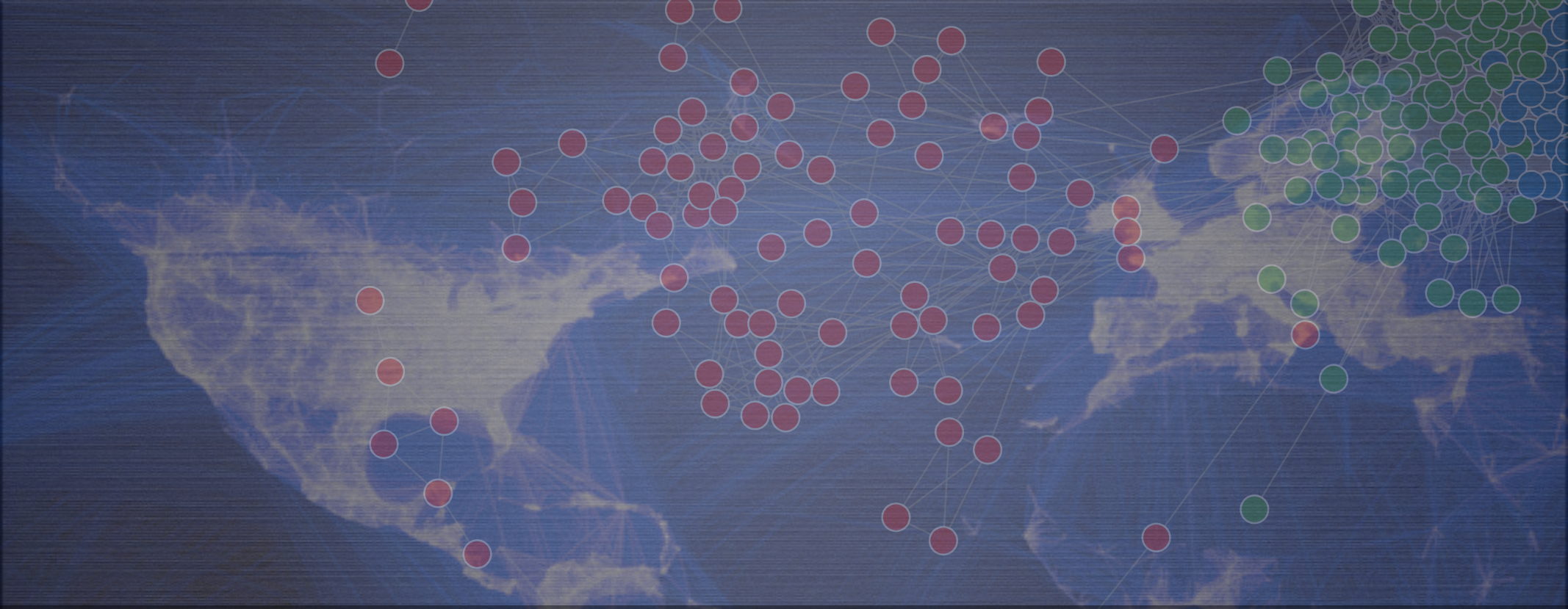
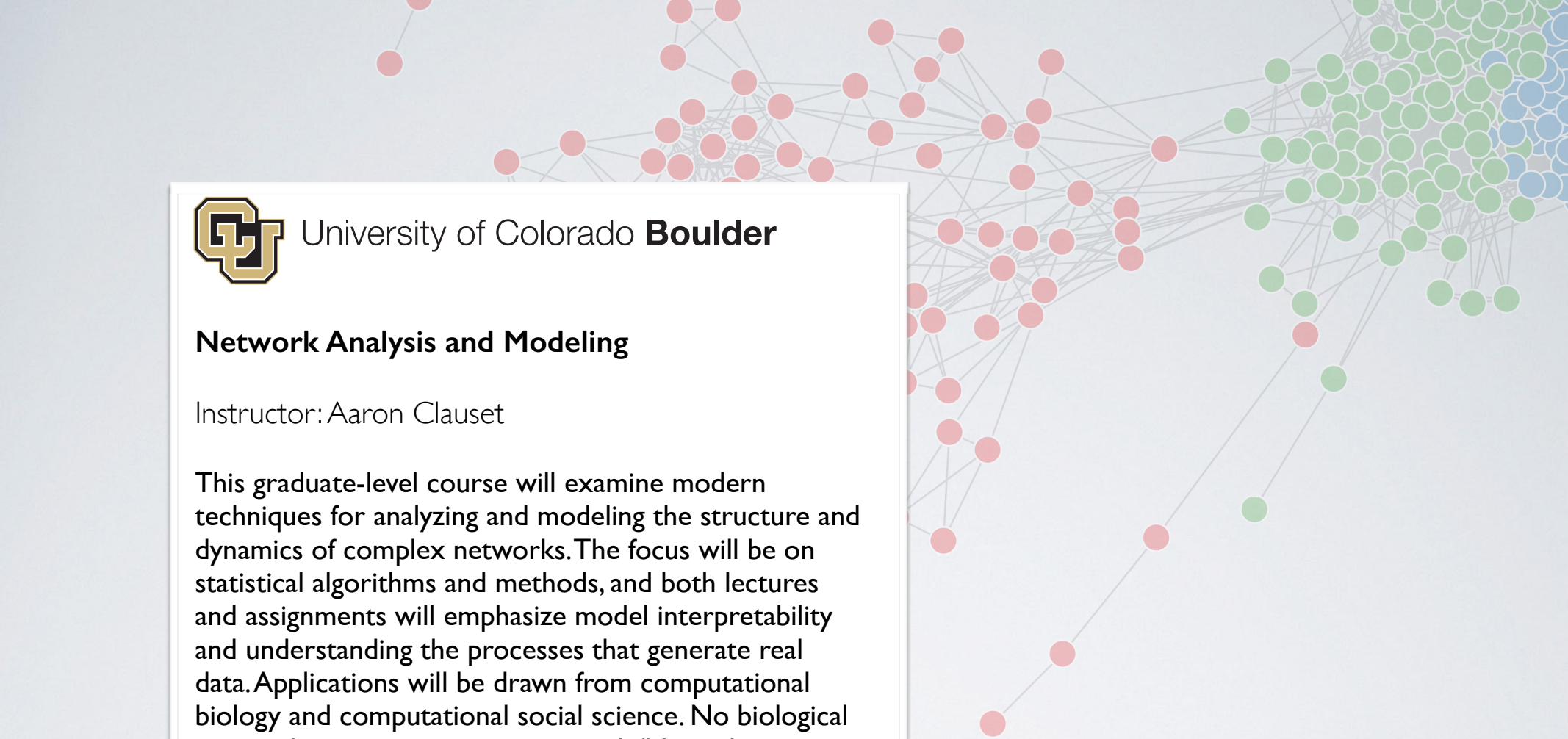


Five Lectures on Networks

Aaron Clauset
@aaronclauset
Assistant Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

lecture 4: dynamic and random graph models of structure

© 2014 Aaron Clauset



University of Colorado **Boulder**

Network Analysis and Modeling

Instructor: Aaron Clauset

This graduate-level course will examine modern techniques for analyzing and modeling the structure and dynamics of complex networks. The focus will be on statistical algorithms and methods, and both lectures and assignments will emphasize model interpretability and understanding the processes that generate real data. Applications will be drawn from computational biology and computational social science. No biological or social science training is required. (Note: this is not a scientific computing course, but there will be plenty of computing for science.)

Full lectures notes online (~150 pages in PDF)

http://santafe.edu/~aaronc/courses/5352/

Software

R
Python
Matlab
NetworkX [python]
graph-tool [python, c++]
GraphLab [python, c++]

Standalone editors

UCI-Net
NodeXL
Gephi
Pajek
Network Workbench
Cytoscape
yEd graph editor
Graphviz

Data sets

Mark Newman's network data sets
Stanford Network Analysis Project
Carnegie Mellon CASOS data sets
NCEAS food web data sets
UCI NET data sets
Pajek data sets
Linkgroup's list of network data sets
Barabasi lab data sets
Jake Hofman's online network data sets
Alex Arenas's data sets

1. defining a network
2. describing a network
3. **null models for networks**
4. statistical inference

citation networks

example of a dynamic network

ample data

pleasing narcissistic qualities

long history of study

generally well understood

citation networks

Networks of Scientific Papers

The pattern of bibliographic references indicates the nature of the scientific research front.

1965 Derek J. de Solla Price

Price's model:

- papers are published continually [growing network]
- each paper has bibliography of length c [mean out degree]
- new papers cite previously published only [directed acyclic graph]
- attachment mechanism:

so long as $\Pr(c)$ has finite variance, the model is well defined

citation networks

Networks of Scientific Papers

The pattern of bibliographic references indicates the nature of the scientific research front.

1965 Derek J. de Solla Price

Price's model:

- papers are published continually [growing network]
- each paper has bibliography of length c [mean out degree]
- new papers cite previously published only [directed acyclic graph]
- attachment mechanism:

$$p(j \text{ cites some paper } i) \propto k_i + a$$

preferential attachment **uniform attachment**

preferential attachment = cumulative advantage = the Matthew effect

so long as $\Pr(c)$ has finite variance, the model is well defined

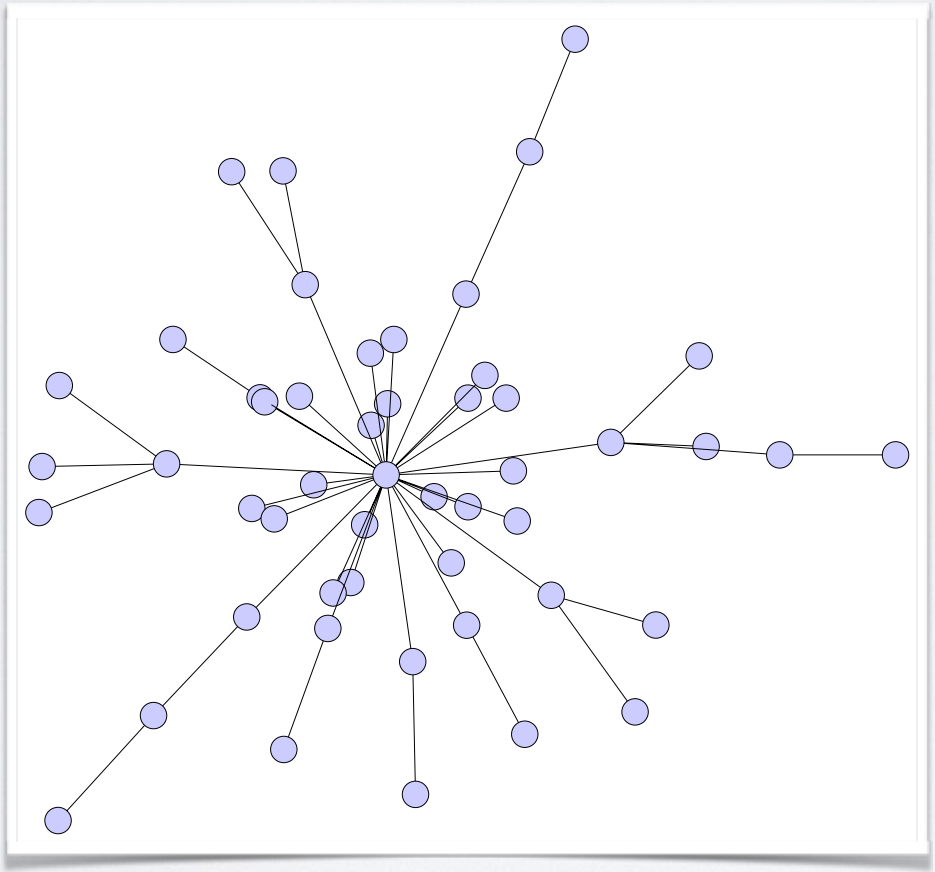
preferential attachment networks

"scale-free network"

$$n = 50$$

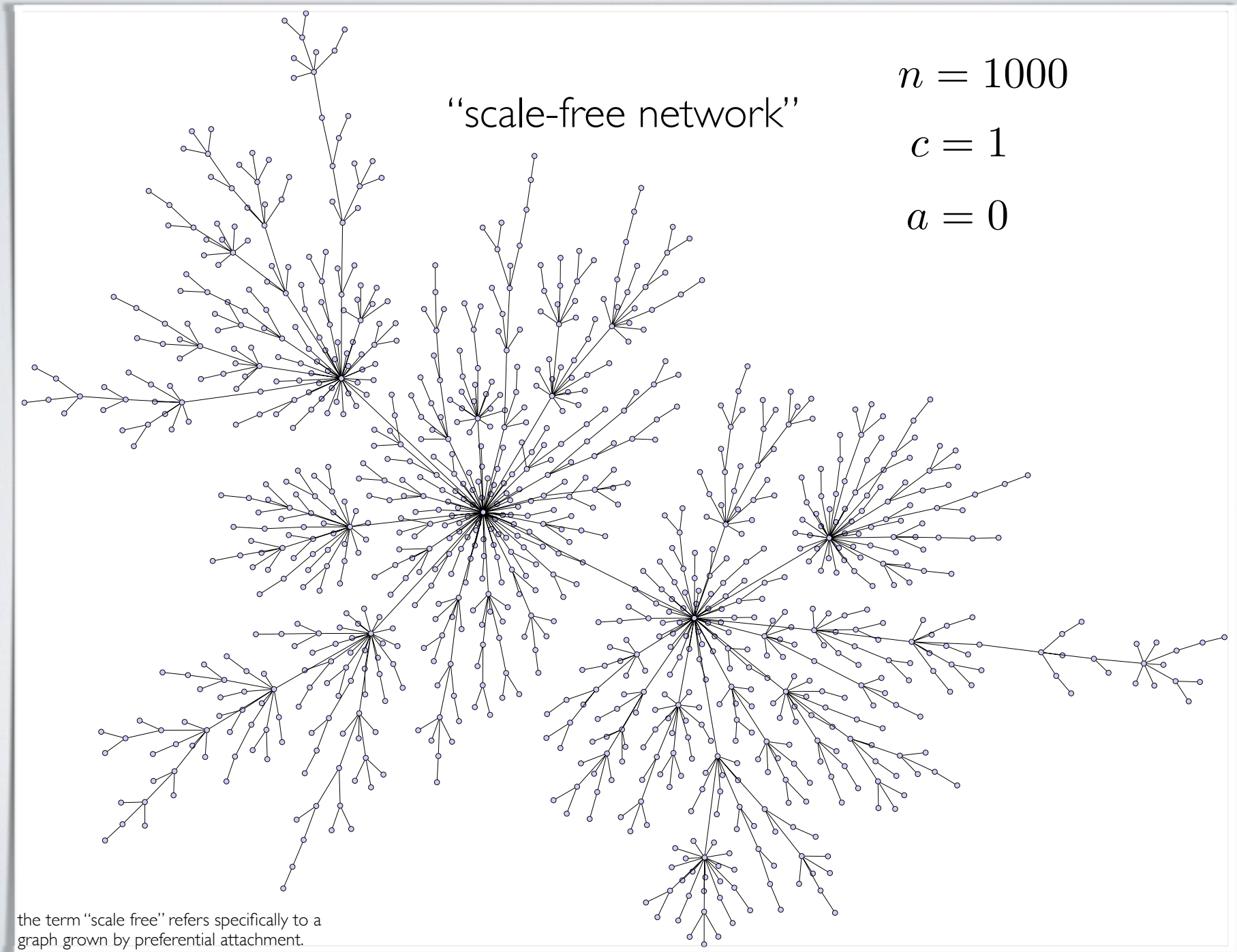
$$c = 1$$

$$a = 0$$



the term "scale free" refers specifically to a graph grown by preferential attachment.

preferential attachment networks



the term "scale free" refers specifically to a graph grown by preferential attachment.

degree distribution

exactly solvable in the limit [originally by Simon 1955]

$$p_k = \frac{\mathrm{B}(k+a, \alpha)}{\mathrm{B}(a, \alpha - 1)} \qquad \alpha = 2 + a/c$$

see Section 14.1 in *Networks: An Introduction*

degree distribution

exactly solvable in the limit [originally by Simon 1955]

$$p_k = \frac{\mathrm{B}(k+a, \alpha)}{\mathrm{B}(a, \alpha - 1)} \qquad \alpha = 2 + a/c$$

recall that

$$\mathrm{B}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$$

$$\mathrm{B}(a,b) \sim a^{-b} \qquad \text{(in the tail)}$$

see Section 14.1 in *Networks: An Introduction*

degree distribution

exactly solvable in the limit [originally by Simon 1955]

$$p_k = \frac{\mathrm{B}(k+a, \alpha)}{\mathrm{B}(a, \alpha - 1)} \qquad \alpha = 2 + a/c$$

recall that

$$\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$$

$$\mathrm{B}(a, b) \sim a^{-b} \qquad \text{(in the tail)}$$

thus, distribution of citations

$$p_k \approx (k+a)^{-\alpha}$$

see Section 14.1 in *Networks: An Introduction*

degree distribution

The first-mover advantage in scientific publication

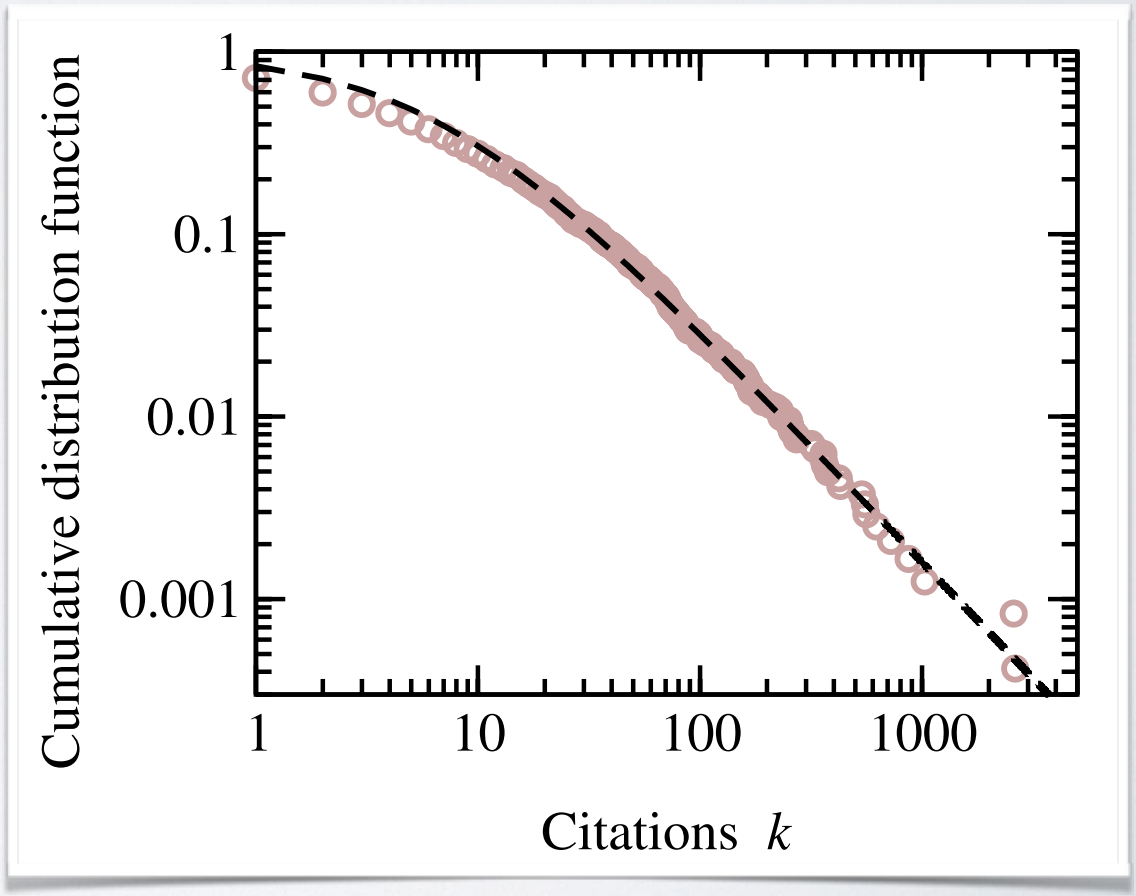
M. E. J. Newman[a] 2009

$$p_k \approx (k+a)^{-\alpha}$$

- 2407 network science papers
- from 1998-2008
- fitted parameters

$\alpha = 2.28$

$a = 6.38$



checking the model

Citation Statistics from 110 Years of *Physical Review* 2004

Sidney Redner

110 years of data (July 1893 - June 2003)

3.1 millions citations

330,000 papers with at least one citation

key question: is attachment function $\propto k_i$?

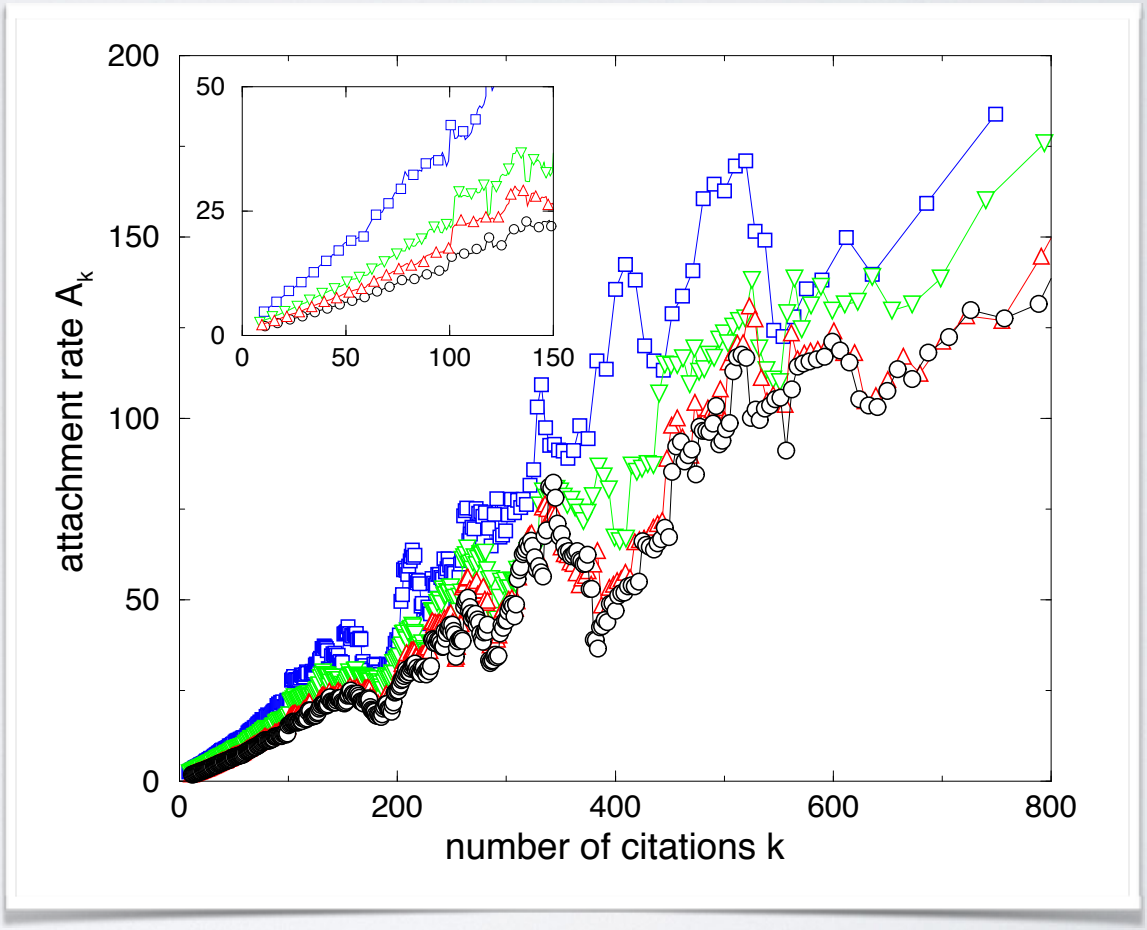
checking the model

key question: is attachment function $\propto k_i$?

pretty much.

caveat:

- ensemble only
  (*not* individual papers)



the first-mover effect

The first-mover advantage in scientific publication

M. E. J. Newman[a] 2009

- let t_i denote *time* that paper i was published
- new papers only cite older papers
- thus, first-mover effect: $k_i \propto 1/t_i$

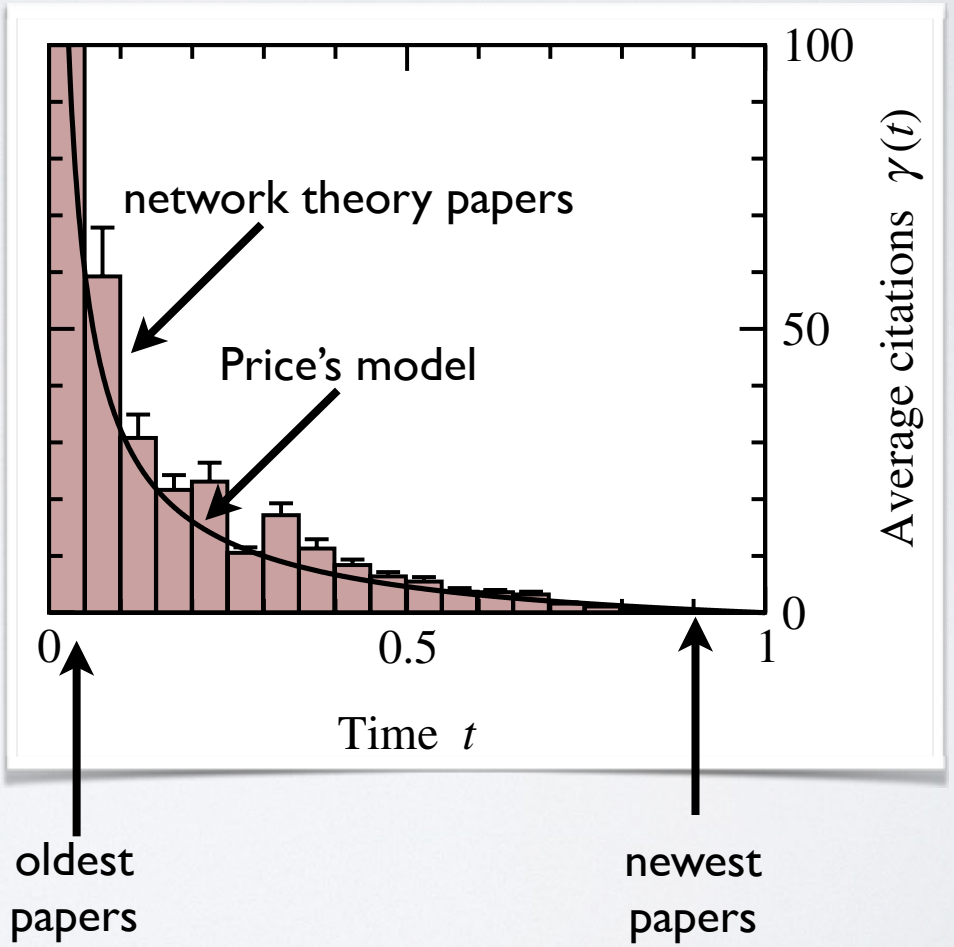
- Price's model fully specified by α and a
- idea:

1. estimate them from total citation distribution
2. derive predictions about citation counts vs. age of paper

the first-mover effect

average citations $\langle k \rangle$ vs. time of publication t

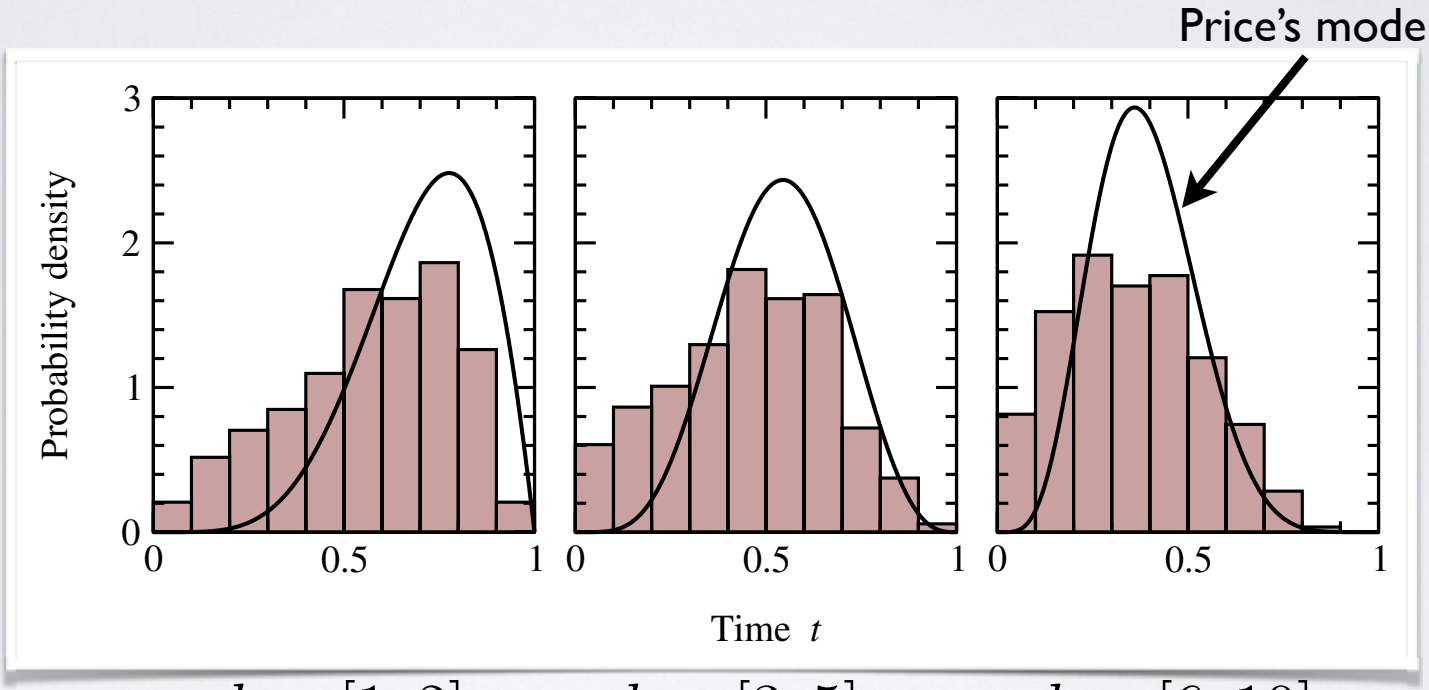
no free parameters



the first-mover effect

given k citations at time $t = 1$, probability of publication time t_i

no free parameters



citation networks

networks of scientific publications

summary of features

- Price's model: *preferential + uniform attachment*
  - excellent model of citation networks
  - also good model of WWW
  - a variation (duplication-mutation) good for gene networks
- not a great model of many other networks
  - especially social and spatial networks
  - ignores constraints (cost of edges)
- many additional mathematical, empirical results
  - see Redner's, Newman's, Fortunato's work

Erdos-Renyi model

denoted $G(n, p)$
where each edge (u, v) exists with probability p

defines a distribution over all networks:

$$P(G) = p^m (1-p)^{\binom{n}{2} - m}$$

where m is the number of edges in the graph

Erdos-Renyi model

denoted $G(n, p)$
where each edge (u, v) exists with probability p

defines a distribution over all networks:

$$P(G) = p^m (1-p)^{\binom{n}{2} - m}$$

where m is the number of edges in the graph

comments:

- highly unrealistic model (all edges iid)
- useful for building intuition
- the most well-studied random graph model
- warm up for more realistic models

degree distribution

mean degree: $\langle k \rangle = c = (n-1)p$

degree distribution:

$$p_k = \underbrace{\binom{n-1}{k}}_{\text{ways to choose those } k} \underbrace{p^k (1-p)^{n-1-k}}_{\text{probability of connecting to exactly } k \text{ vertices}}$$

degree distribution

mean degree: $\langle k \rangle = c = (n-1)p$

degree distribution:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

for $c = \text{const.}$ (or small p), we can show

$$(1-p)^{n-1-k} \approx \mathrm{e}^{-c}$$

expand log as Taylor series

$$\ln[(1-p)^{n-1-k}] = (n-1-k)\ln\left(1-\frac{c}{n-1}\right)$$
$$\approx -(n-1-k)\frac{c}{n-1} \approx -c$$

degree distribution

mean degree: $\langle k \rangle = c = (n-1)p$

degree distribution:

$$p_k = \binom{n-1}{k} p^k \mathrm{e}^{-c}$$

for $c = \mathrm{const.}$ (or small p), we can show

$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!k!} \approx \frac{(n-1)^k}{k!}$$

degree distribution

mean degree: $\langle k \rangle = c = (n-1)p$

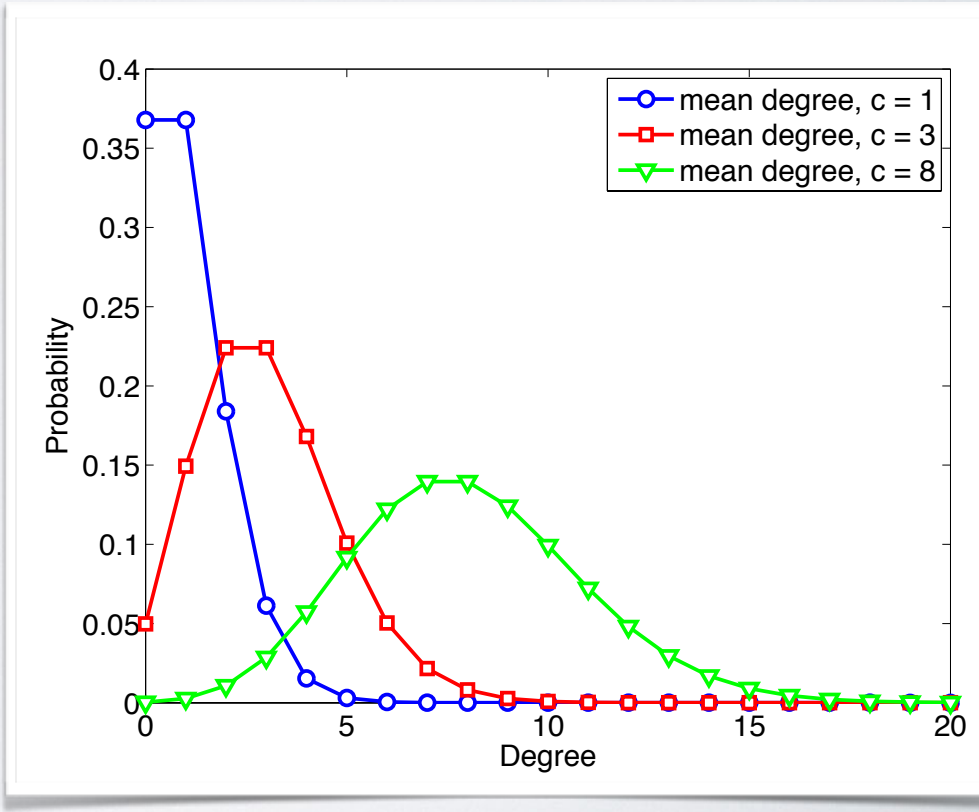
degree distribution:

$$p_k = \frac{(n-1)^k}{k!} p^k \mathrm{e}^{-c}$$

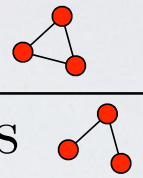
$$= \frac{(n-1)^k}{k!} \left(\frac{c}{n-1} \right)^k \mathrm{e}^{-c}$$

$$= \underbrace{\mathrm{e}^{-c} \frac{c^k}{k!}}$$

Poisson distribution



clustering coefficient

defined as $C = \dfrac{3 \times \#\text{triangles}}{\#\text{connected triples}}$ 

measures density of triangles in network

in social networks, $C \approx 0.2 - 0.4$

for $G(n,p)$,

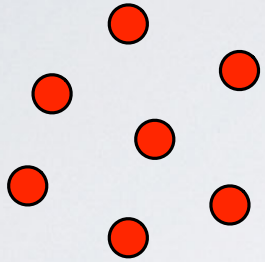
$$C = \frac{\binom{n}{3}p^3}{\binom{n}{3}p^2} = p = \frac{c}{n-1} = \underbrace{O(n^{-1})}$$

asymptotically zero clustering

giant component

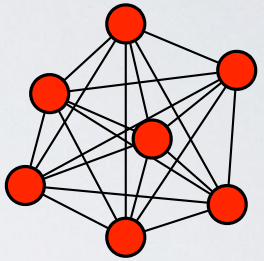
empty graph

n components of size 1



complete graph

1 component of size n



$p = 0$

$p = 1$

giant component

empty graph

n components of size 1

complete graph

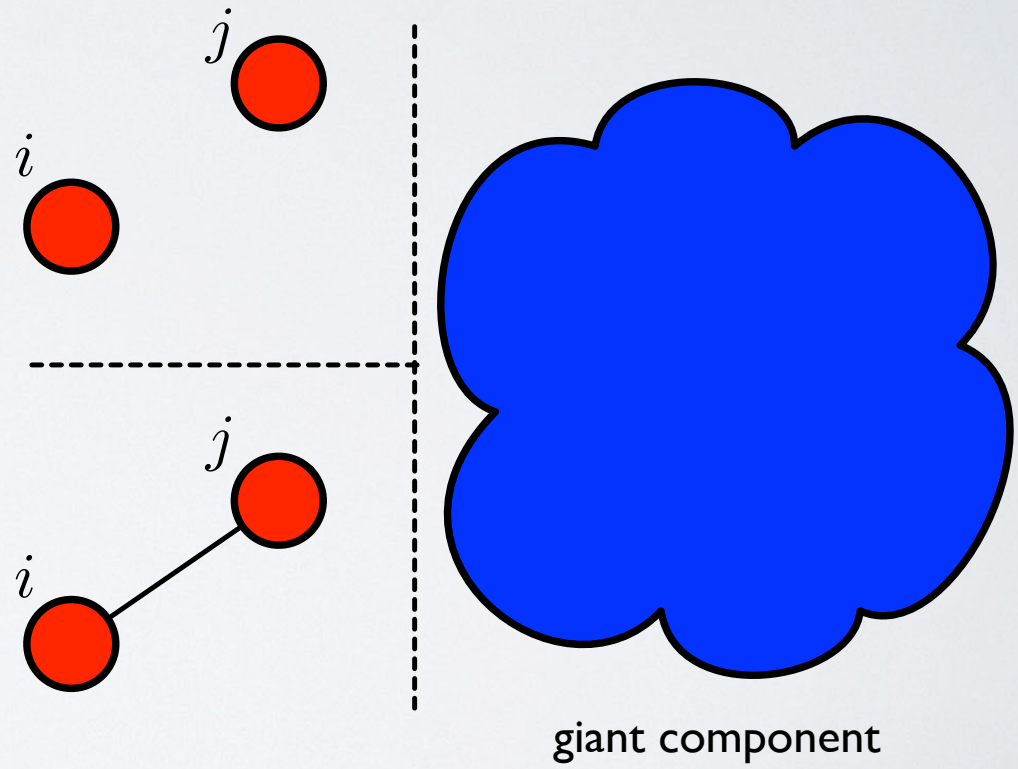
1 component of size n

giant component

let u be fraction of vertices *not* in **giant component**

for i not to be in the giant component, then for every j

1. i is not connected to j, or
2. i connects to j, and j is not part of the giant component



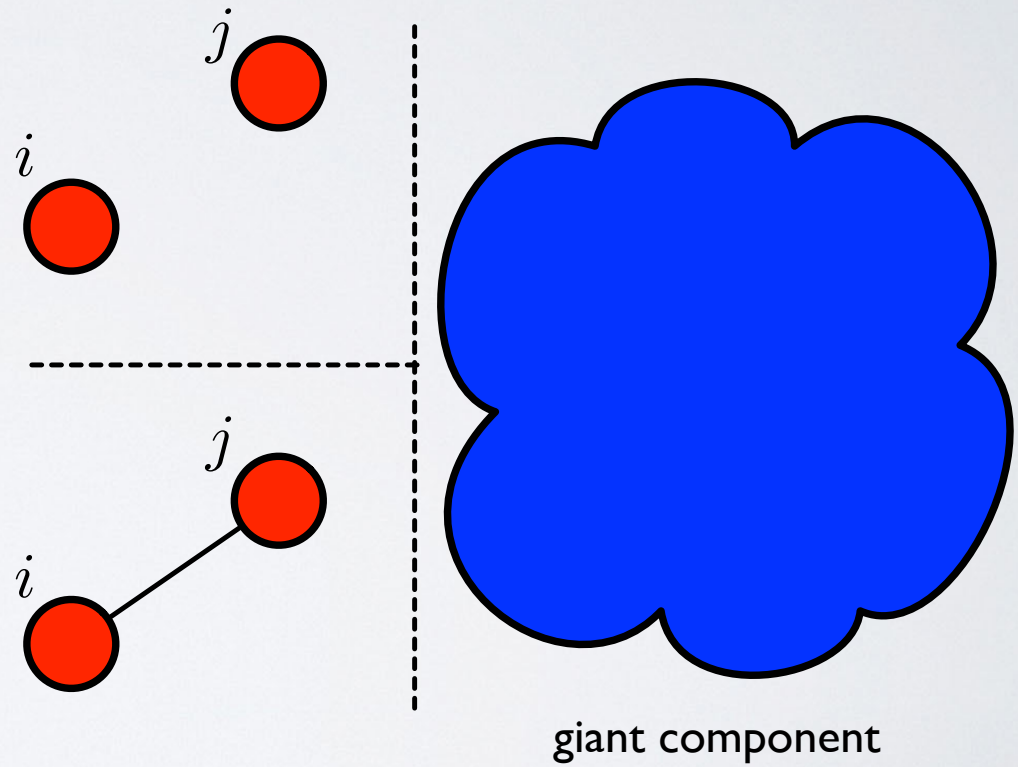
giant component

let u be fraction of vertices *not* in **giant component**

for i not to be in the giant component, then for every j

1. with probability $1 - p$

2. with probability pu



giant component

total probability that i **not** in giant component via any of the $n-1$ choices of j:

$$u = (1 - p + pu)^{n-1} = \left[1 - \frac{c}{n-1}(1-u)\right]^{n-1}$$

remember: $p = c/(n-1)$

giant component

total probability that i **not** in giant component via any of the $n-1$ choices of j:

$$u = (1 - p + pu)^{n-1} = \left[1 - \frac{c}{n-1}(1-u)\right]^{n-1}$$

taking logs of both sides, and approximating:

$$\begin{aligned}
\ln u &= (n-1)\ln\left[1 - \frac{c}{n-1}(1-u)\right] \\
&\approx -(n-1)\frac{c}{n-1}(1-u) \\
&= -c(1-u)
\end{aligned}$$

remember: $\ln(1+x) \approx x$ if $x \ll 1$

giant component

total probability that i **not** in giant component via any of the $n-1$ choices of j:

$$u = \mathrm{e}^{-c(1-u)}$$

and the fraction of vertices **in** the giant component is

$$S = 1 - u$$

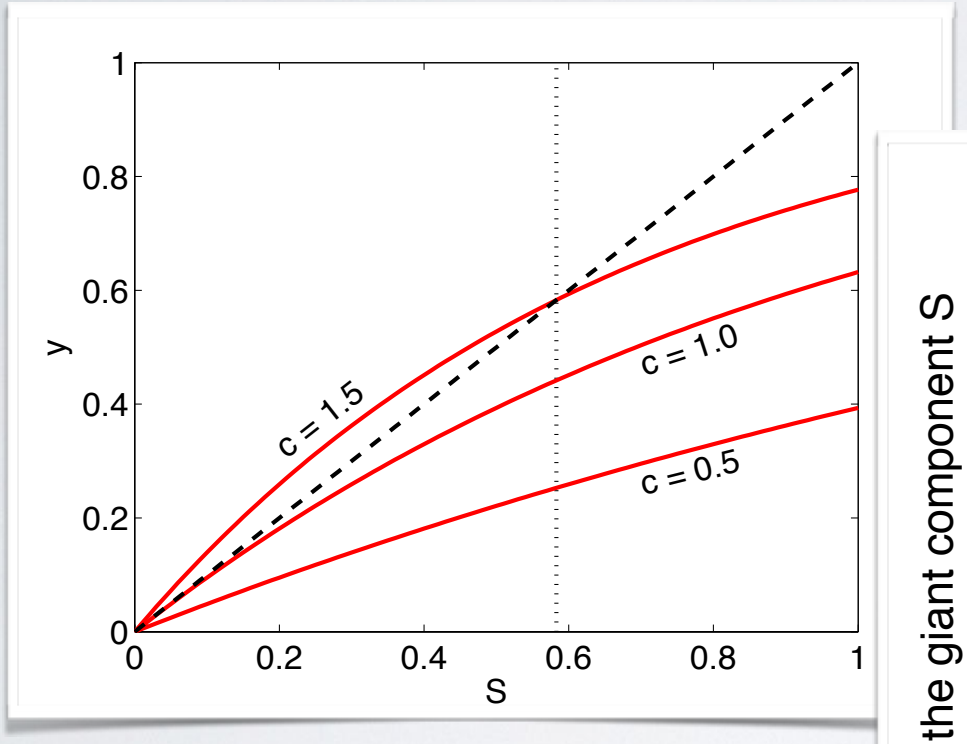
eliminating u for S yields the transcendental equation

$$S = 1 - \mathrm{e}^{-c\,S}$$

[first given by Erdos and Renyi in 1959]

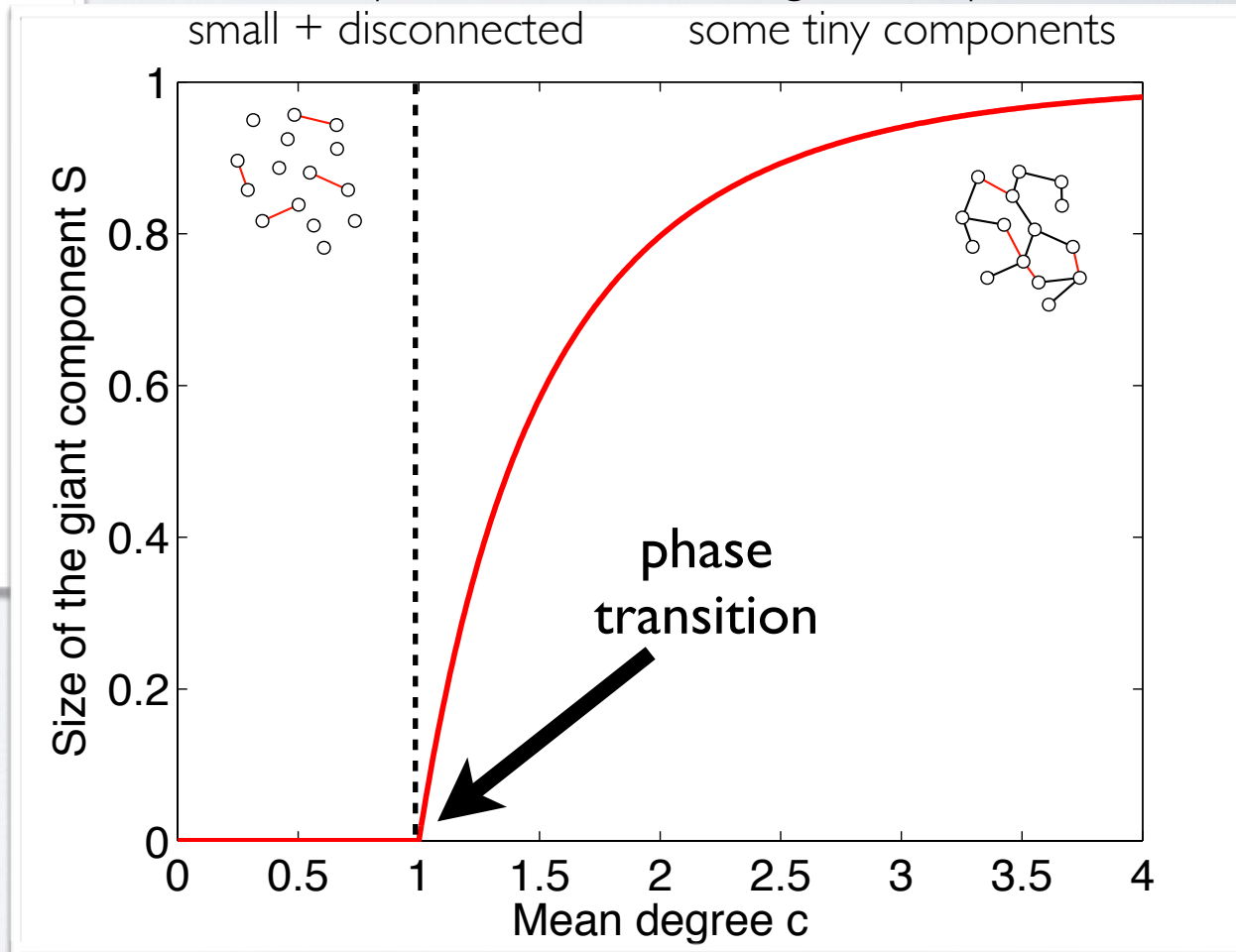
giant component

size of the giant component: $S = 1 - \mathrm{e}^{-c\,S}$

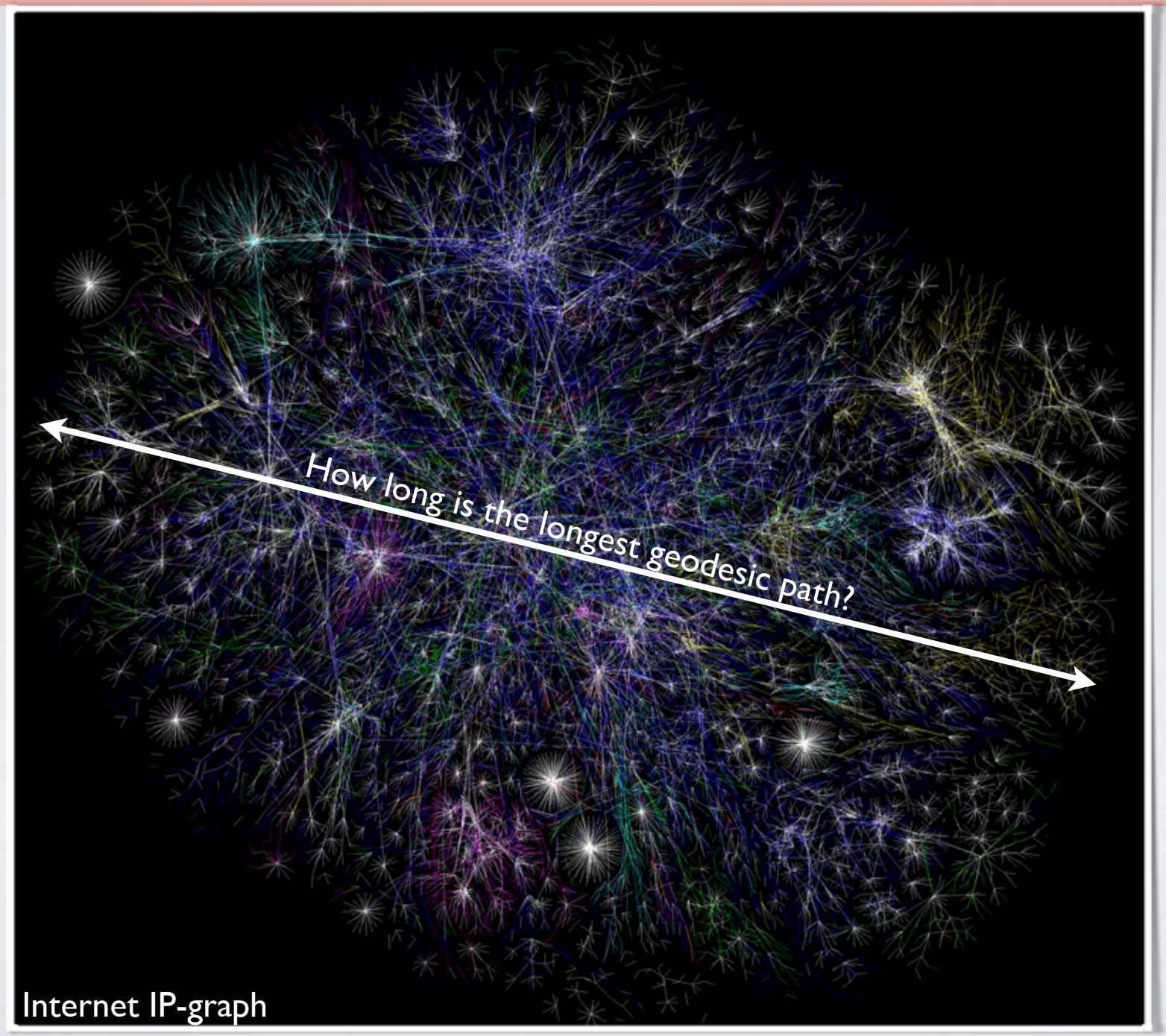


all components:
small + disconnected

one giant component,
some tiny components



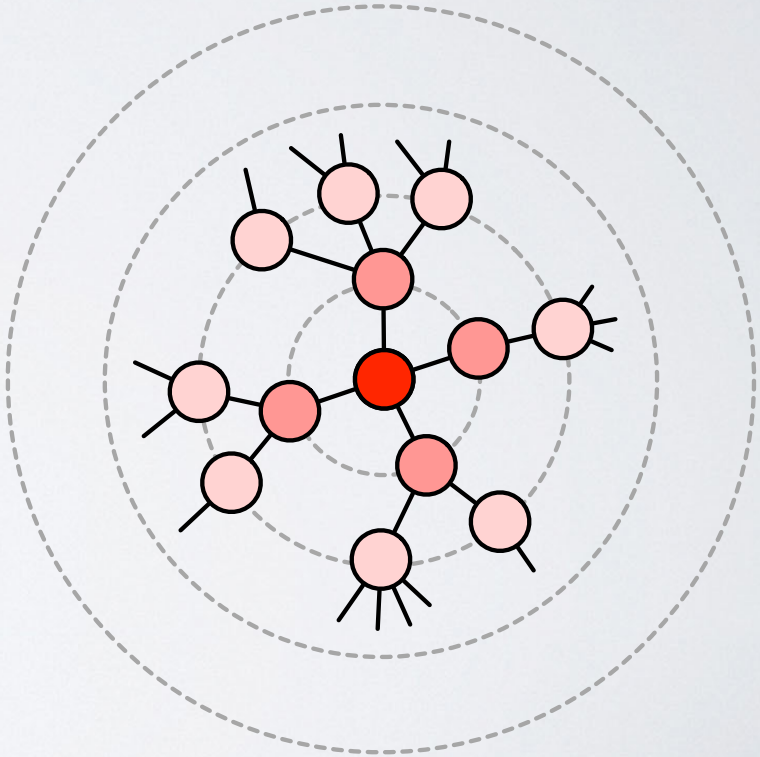
diameter



diameter

a rough argument:

- $G(n, p)$ is locally tree-like (no loops; low clustering coefficient)
- mean number of vertices within s steps is c^s
- all n vertices within ℓ steps
- thus, diameter is roughly $\ell \approx \ln n / \ln c$



this argument can be tightened up in several ways. however, these better versions yield the same asymptotic result

diameter

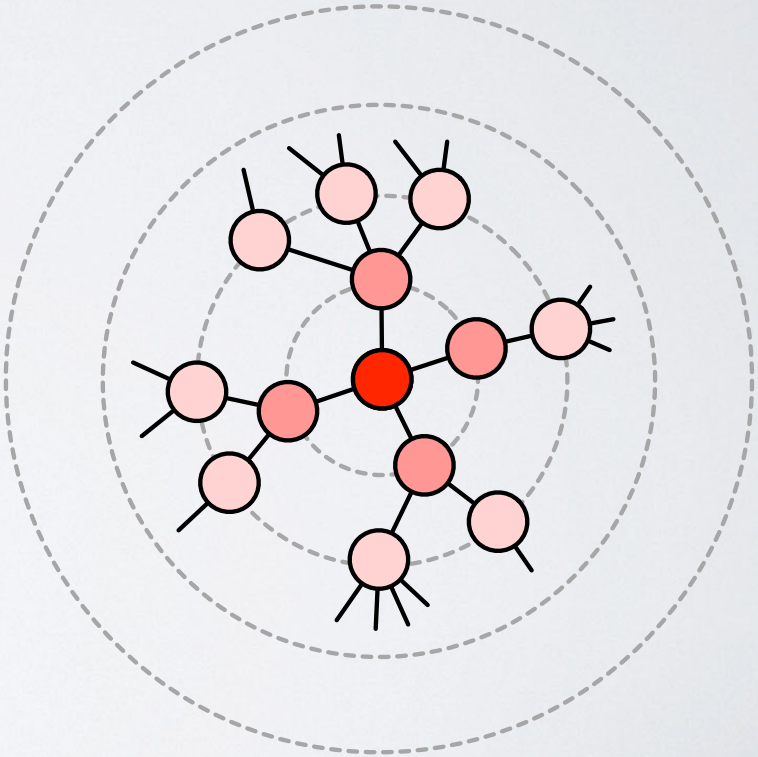
a rough argument:

- $G(n, p)$ is locally tree-like (no loops; low clustering coefficient)
- mean number of vertices within s steps is c^s
- all n vertices within ℓ steps
- thus, diameter is roughly $\ell \approx \ln n / \ln c$

---

it's a small world after all...

$$\ell \approx \frac{\ln n}{\ln c} = \frac{\ln 7 \times 10^9}{\ln 100} = 4.92\ldots$$



this argument can be tightened up in several ways. however, these better versions yield the same asymptotic result

Erdos-Renyi model

denoted $G(n, p)$
where each edge (u, v) exists with probability p

summary of features

- Poisson degree distribution
  - nearly every degree close to its average
- $O(n^{-1})$ loops of length ℓ
  - asymptotically zero clustering coefficient
  - graph is locally tree-like
- giant component phase transition
- diameter $\sim O(\ln n)$
  - small world property

how are we doing?

feature	$G(n,p)$	real networks
degree distribution	Poisson	heavy tailed
clustering coefficient	$O(n^{-1})$	social: high non-social: low
diameter	$O(\ln n)$	small
large-scale structure	none	communities, dense core, hierarchies, etc.

configuration model

a random graph conditioned on having the specified degree sequence $k_1, k_2, \ldots, k_n$

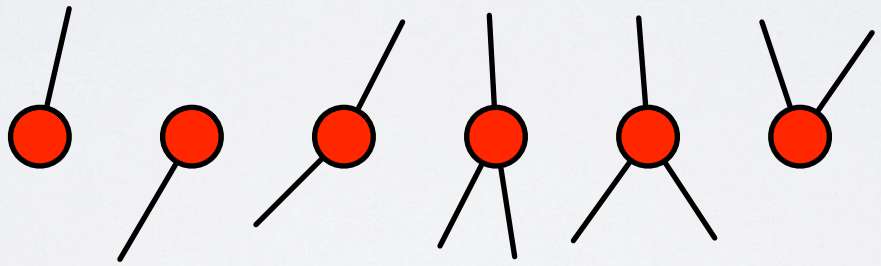
number of edges: $m = \frac{1}{2} \sum_i k_i$

each edge (i, j) exists with probability $p_{ij} = \frac{k_i k_j}{2m}$

configuration model

a random graph conditioned on having the specified degree sequence $k_1, k_2, \ldots, k_n$

how do we construct?



$1 \quad 1 \quad 2 \quad 3 \quad 3 \quad 2$

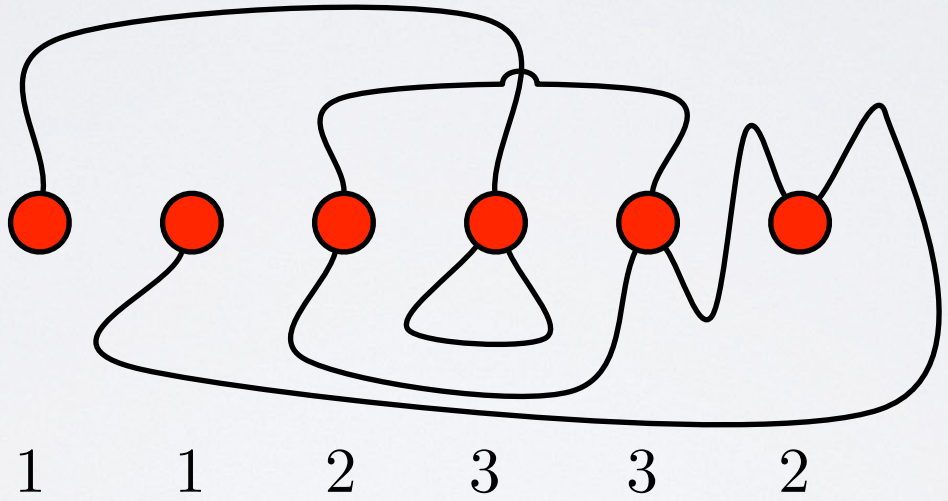
choose a *random matching* on the stubs

$$p_{ij} = \frac{k_i k_j}{2m}$$

configuration model

a random graph conditioned on having the specified degree sequence $k_1, k_2, \ldots, k_n$

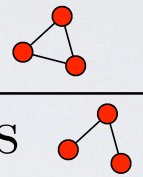
how do we construct?



choose a *random matching* on the stubs

$$p_{ij} = \frac{k_i k_j}{2m}$$

clustering coefficient

defined as $C = \dfrac{3 \times \#\text{triangles}}{\#\text{connected triples}}$ 

measures density of triangles in network

in social networks, $C \approx 0.2 - 0.4$

using generating functions:

$$C = \frac{1}{n} \underbrace{\frac{\left[\langle k^2 \rangle - \langle k \rangle\right]^2}{\langle k \rangle^3}} = O(n^{-1})$$

moments of degree sequence

see Section 13.5 of *Networks: An Introduction*

configuration model

a random graph conditioned on having the specified degree sequence $k_1, k_2, \ldots, k_n$

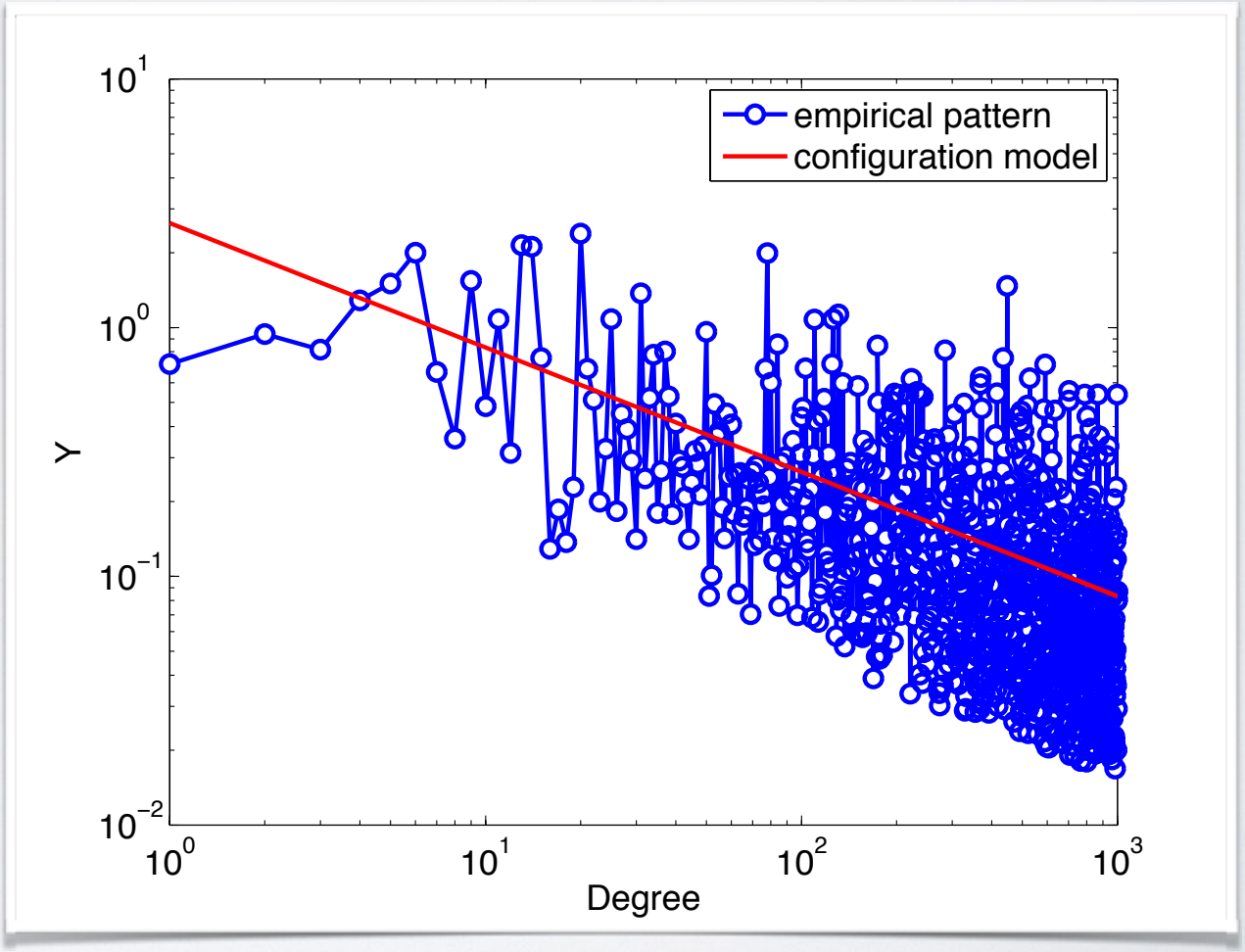
summary of features

- specified degree sequence
  - e.g., that of a specific empirical network
- $O(n^{-1})$ loops of length ℓ
  - graph is locally tree-like
  - but, depends on degree sequence
- giant component phase transition
- diameter $\sim O(\ln n)$
  - small world property
- *the standard null model for empirical patterns*

configuration model

the standard null model for empirical patterns

if your data and a random graph with the same degree sequence have the same pattern, is your pattern interesting?



how are we doing?

feature	$G(n,p)$	configuration	real networks
degree distribution	Poisson	specified	heavy tailed
clustering coefficient	$O(n^{-1})$	$O(n^{-1})$	social: high non-social: low
diameter	$O(\ln n)$	$O(\ln n)$	small
large-scale structure	none	none	communities, dense core, hierarchies, etc.

stochastic block model

classic SBM

- each vertex i has type $z_i \in \{1, \ldots, k\}$ (k vertex types or groups)
- stochastic block matrix M of group-level connection probabilities
- probability that i, j are connected = M_{z_i, z_j}

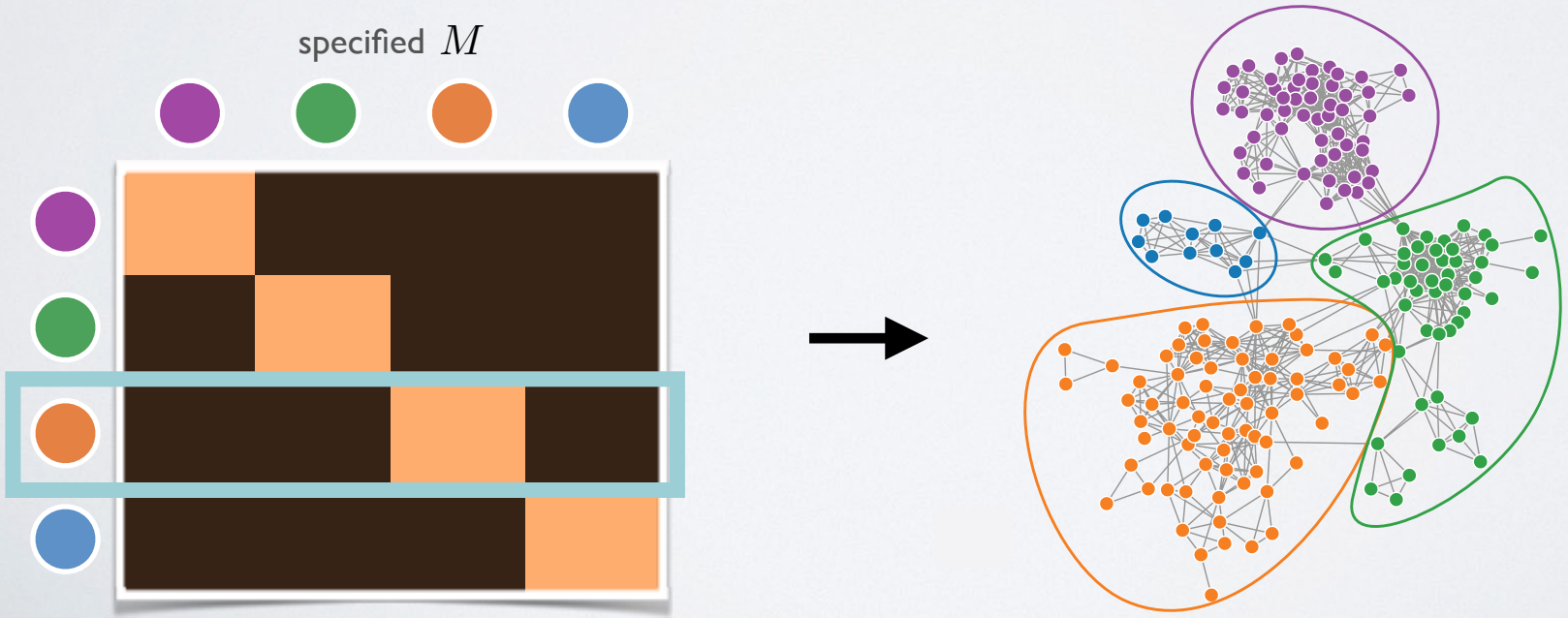
community = vertices with same pattern of inter-community connections

stochastic block model

classic SBM

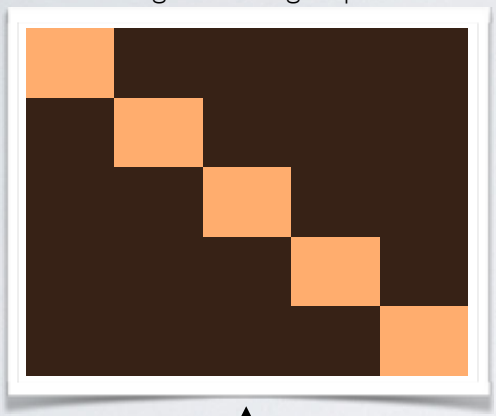
- each vertex i has type $z_i \in \{1, \ldots, k\}$ (k vertex types or groups)
- stochastic block matrix M of group-level connection probabilities
- probability that i, j are connected = M_{z_i, z_j}

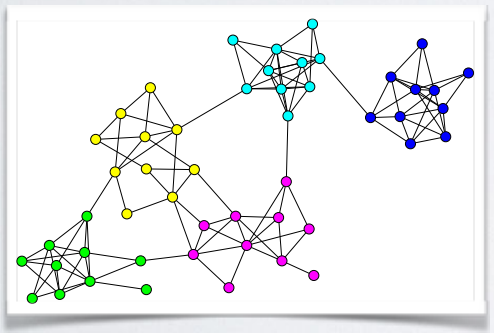
community = vertices with same pattern of inter-community connections



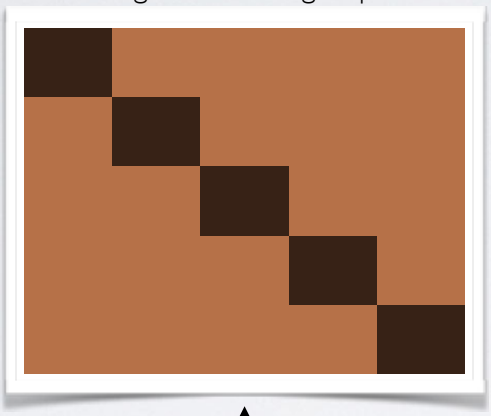
stochastic block model

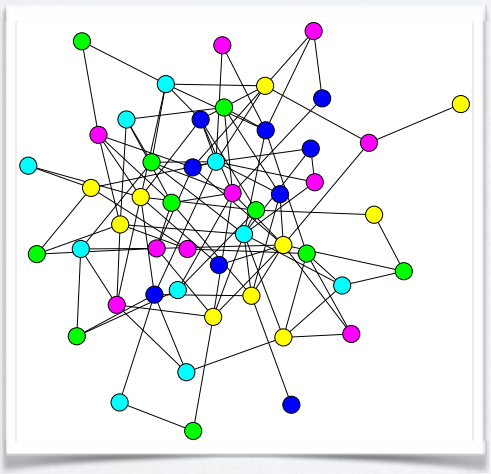
assortative
edges within groups



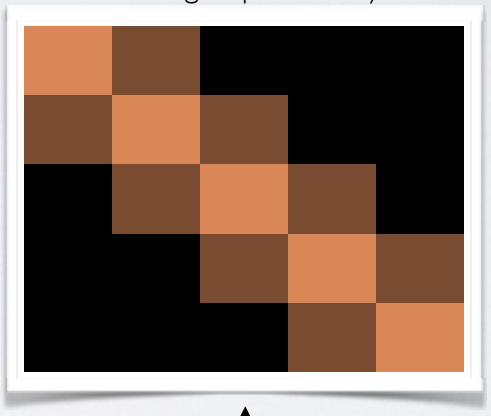


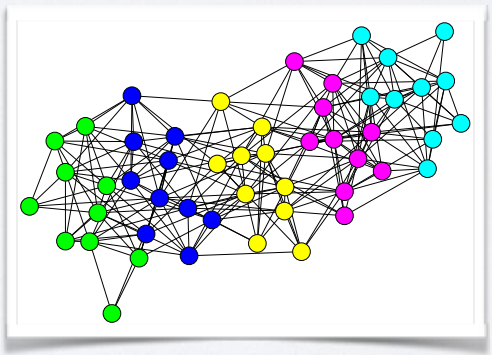
disassortative
edges between groups



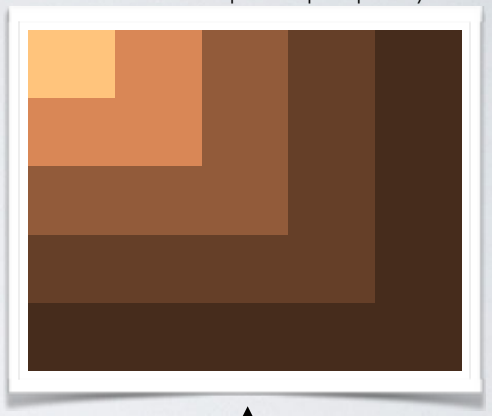


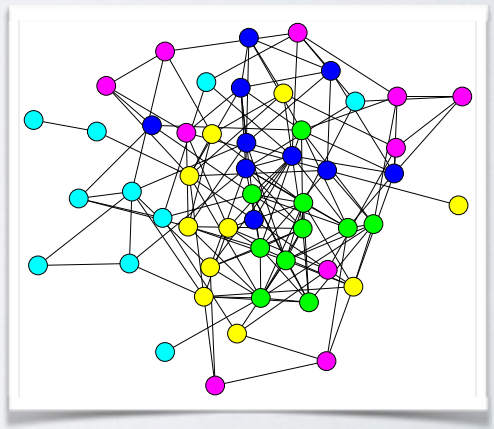
ordered
linear group hierarchy





core-periphery
dense core, sparse periphery





stochastic block model

SBM edge probability

given labeling z and block matrix M

$$\Pr(A_{ij} = 1 \,|\, M, z) = \text{Bernoulli}(M_{z_i, z_j})$$

(always produces *simple* networks)

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

stochastic block model

SBM edge probability

given labeling z and block matrix M

$$\Pr(A_{ij} = 1 \mid M, z) = \text{Bernoulli}(M_{z_i,z_j})$$

(always produces *simple* networks)

a similar model is then

$$\Pr(A_{ij} \mid M, z) = \text{Poisson}(M_{z_i,z_j})$$
$$= \frac{(M_{z_i,z_j})^{A_{ij}}}{A_{ij}!}\exp\left(-M_{z_i,z_j}\right)$$

(rarely produces *simple* networks)

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

stochastic block model

degree-corrected SBM

given labeling z, block matrix M, and vertex propensities θ

$$\begin{aligned}\Pr(A_{ij} \mid M, z, \theta) &= \text{Poisson}(\theta_i \theta_j M_{z_i, z_j}) \\ &= \frac{(\theta_i \theta_j M_{z_i, z_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\theta_i \theta_j M_{z_i, z_j}\right)\end{aligned}$$

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

stochastic block model

degree-corrected SBM

given labeling z, block matrix M, and vertex propensities θ

$$\Pr(A_{ij} \mid M, z, \theta) = \text{Poisson}(\theta_i \theta_j M_{z_i, z_j})$$
$$= \frac{(\theta_i \theta_j M_{z_i, z_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\theta_i \theta_j M_{z_i, z_j}\right)$$

in practice, we choose θ so that $\mathrm{E}(k_i) = \theta_i \, \kappa_{z_i}$

total degree of group containing i

and choose $M_{r,s}$ so that

$$M_{r,s} = M_{s,r} \text{ and } \sum_s M_{r,s} = \kappa_r$$

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

stochastic block model

degree-corrected SBM

given labeling z, block matrix M, and vertex propensities θ

$$\Pr(A_{ij} \mid M, z, \theta) = \text{Poisson}(\theta_i \theta_j M_{z_i, z_j})$$
$$= \frac{(\theta_i \theta_j M_{z_i, z_j})^{A_{ij}}}{A_{ij}!} \exp\left(-\theta_i \theta_j M_{z_i, z_j}\right)$$

in practice, we choose θ so that $\mathrm{E}(k_i) = \theta_i \, \kappa_{z_i}$

total degree of group containing i

and choose $M_{r,s}$ so that

$$M_{r,s} = M_{s,r} \text{ and } \sum_s M_{r,s} = \kappa_r$$

then, draw $\text{Poisson}(M_{r,s})$ number of edges for each pair r, s
and, for each edge bundle, assign an endpoint to vertex i with prob. θ_i

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

stochastic block model

SBM properties

k Erdos-Renyi random graphs

each with size n_r and internal density $M_{r,r}$

joined pairwise as random bipartite graph with density $M_{r,s}$

degree distribution: mixture of Poissons

diameter: $O(\ln n)$ or $O(\ln(kn))$

triangle density: low, except when $M_{r,s} \gg 0$

local structure: like a random graph

large-scale: mixtures of assortative & disassortative structure

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

stochastic block model

DC-SBM properties

k 'configuration model' random multi-graphs

each with size n_r, internal density $M_{r,r}$ and propensities $\{\theta_i\}_r$

joined pairwise as random bipartite graph with parameters $M_{r,s}$ and $\{\theta_i\}_{r,s}$

degree distribution: arbitrary ($\{\theta_i\}$)

diameter: $O(\ln n)$ or $O(\ln(kn))$

triangle density: low, except when $M_{r,s} \gg 0$

local structure: like a random multi-graph

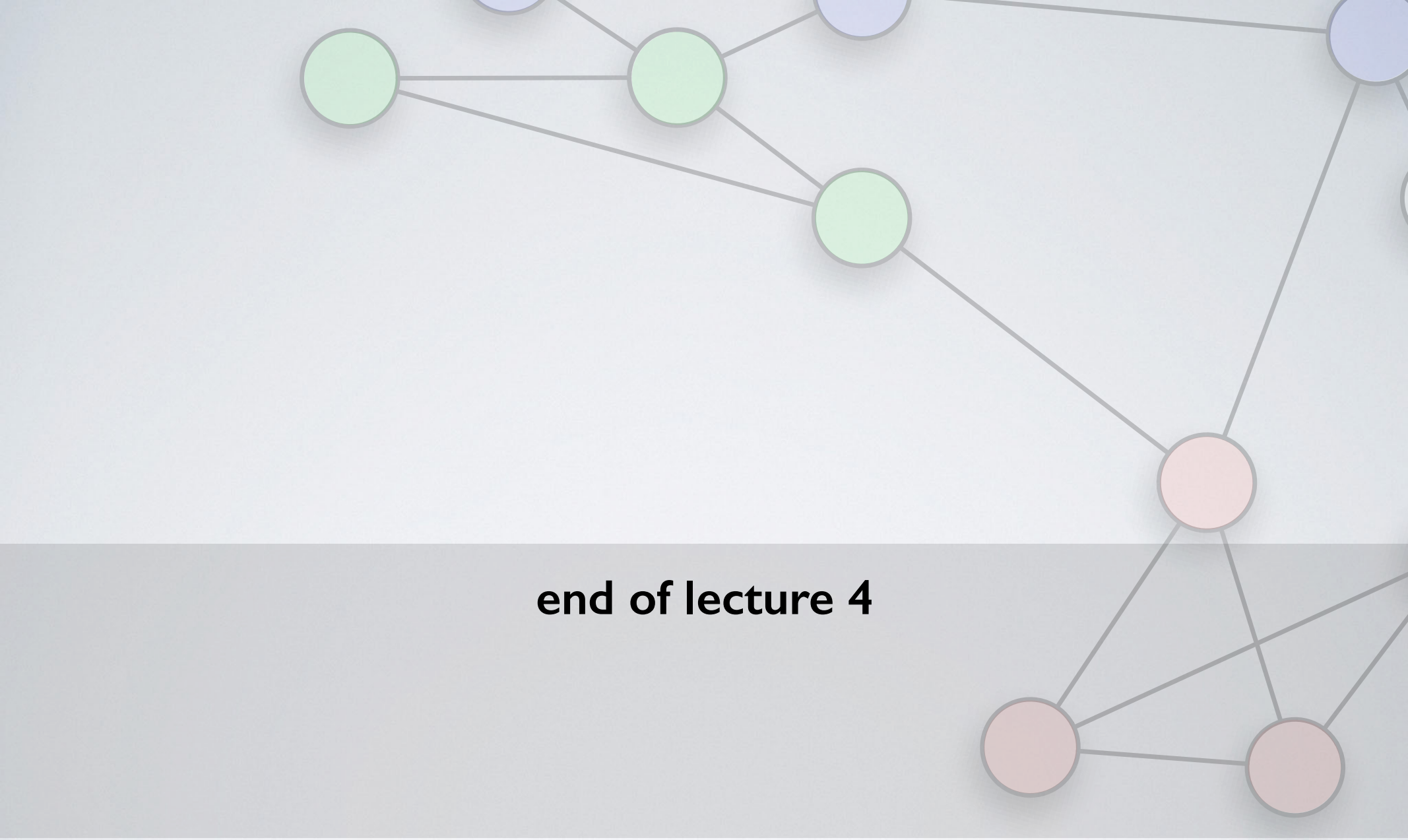
large-scale: mixtures of assortative & disassortative structure

Karrer & Newman, *Phys. Rev. E* **83**, 016107 (2011).

how are we doing?

feature	$G(n,p)$	configuration	DC SBM	real networks
degree distribution	Poisson	specified	specified	heavy tailed
clustering coefficient	$O(n^{-1})$	$O(n^{-1})$	$O(n^{-1})$	social: high non-social: low
diameter	$O(\ln n)$	$O(\ln n)$	$O(\ln n)$	small
large-scale structure	none	none	specified: communities, hierarchies, etc.	communities, dense core, hierarchies, etc.

Karrer & Newman, *Phys. Rev. E* **82**, 066118 (2010)



end of lecture 4

selected references

- The structure and function of complex networks. M. E. J. Newman, *SIAM Review* **45**, 167–256 (2003).
- *The Structure and Dynamics of Networks*. M. E. J. Newman, A.-L. Barabási, and D. J. Watts, Princeton University Press (2006).
- Hierarchical structure and the prediction of missing links in networks. A. Clauset, C. Moore, and M. E. J. Newman, *Nature* **453**, 98–101 (2008).
- Modularity and community structure in networks. M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
- Why social networks are different from other types of networks. M. E. J. Newman and J. Park, *Phys. Rev. E* **68**, 036122 (2003)
- Random graphs with arbitrary degree distributions and their applications. M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).
- Comparing community structure identification. L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas. *J. Stat. Mech*. P09008 (2005).
- Characterization of Complex Networks: A Survey of measurements. L. daF. Costa, F. A. Rodrigues, G. Travieso and P. R. VillasBoas. arxiv:cond-mat/050585 (2005).
- Evolution in Networks. S.N. Dorogovtsev and J. F. F. Mendes. *Adv. Phys*. **51**, 1079 (2002).
- Revisting "scale-free" networks. E. F. Keller. *BioEssays* **27**, 1060-1068 (2005).
- Currency metabolites and network representations of metabolism. P. Holme and M. Huss. arxiv:0806.2763 (2008).
- Functional cartography of complex metabolic networks. R. Guimera and L. A. N. Amaral. *Nature* **433**, 895 (2005).
- Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. J. Leskovec, J. Kleinberg and C. Faloutsos. *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* 2005.
- The Structure of the Web. J. Kleinberg and S. Lawrence. *Science* 294, 1849 (2001).
- Navigation in a Small World. J. Kleinberg. *Nature* 406 (2000), 845.
- Towards a Theory of Scale-Free Graphs: Definitions, Properties and Implications. L. Li, D. Alderson, J. Doyle, and W. Willinger. *Internet Mathematics 2(4), 2006*.
- A First-Principles Approach to Understanding the Internet's Router-Level Topology. L. Li, D. Alderson, W. Willinger, and J. Doyle. *ACM SIGCOMM* 2004.
- Inferring network mechanisms: The Drosophila melanogaster protein interaction network. M. Middendorf, E. Ziv and C. H. Wiggins. *Proc. Natl. Acad. Sci. USA* 102, 3192 (2005).
- Robustness Can Evolve Gradually in Complex Regulatory Gene Networks with Varying Topology. S. Ciliberti, O. C. Martin and A. Wagner. *PLoS Comp. Bio*. 3, e15 (2007).
- Simple rules yield complex food webs. R. J. Williams and N. D. Martinez. *Nature* 404, 180 (2000).
- A network analysis of committees in the U.S. House of Representatives. M. A. Porter, P. J. Mucha, M. E. J. Newman and C. M. Warmbrand. *Proc. Natl. Acad. Sci. USA* 102, 7057 (2005).
- On the Robustness of Centrality Measures under Conditions of Imperfect Data. S. P. Borgatti, K. M. Carley and D. Krackhardt. *Social Networks* **28**, 124 (2006).