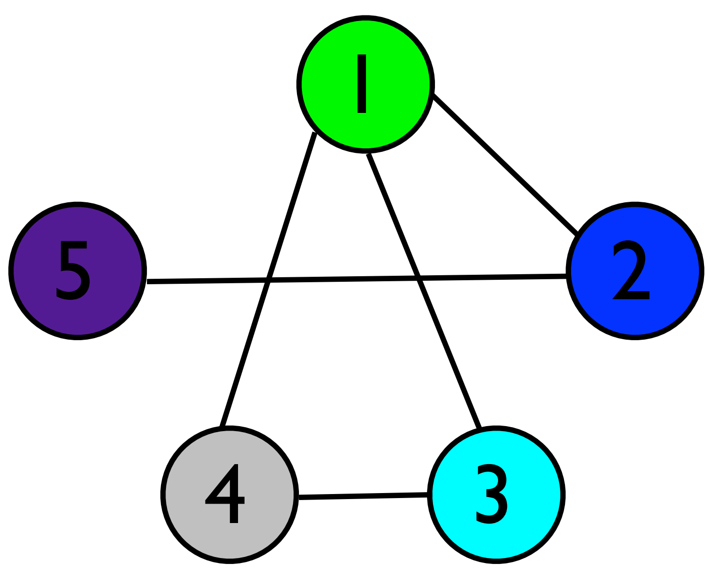
CS109/Stat121/AC209/E-109
Data Science
Network Models

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



This Week

- HW4 due tonight at 11:59 pm
- Friday lab **10-11:30 am** in MD G115

Examples from Newman (2003)

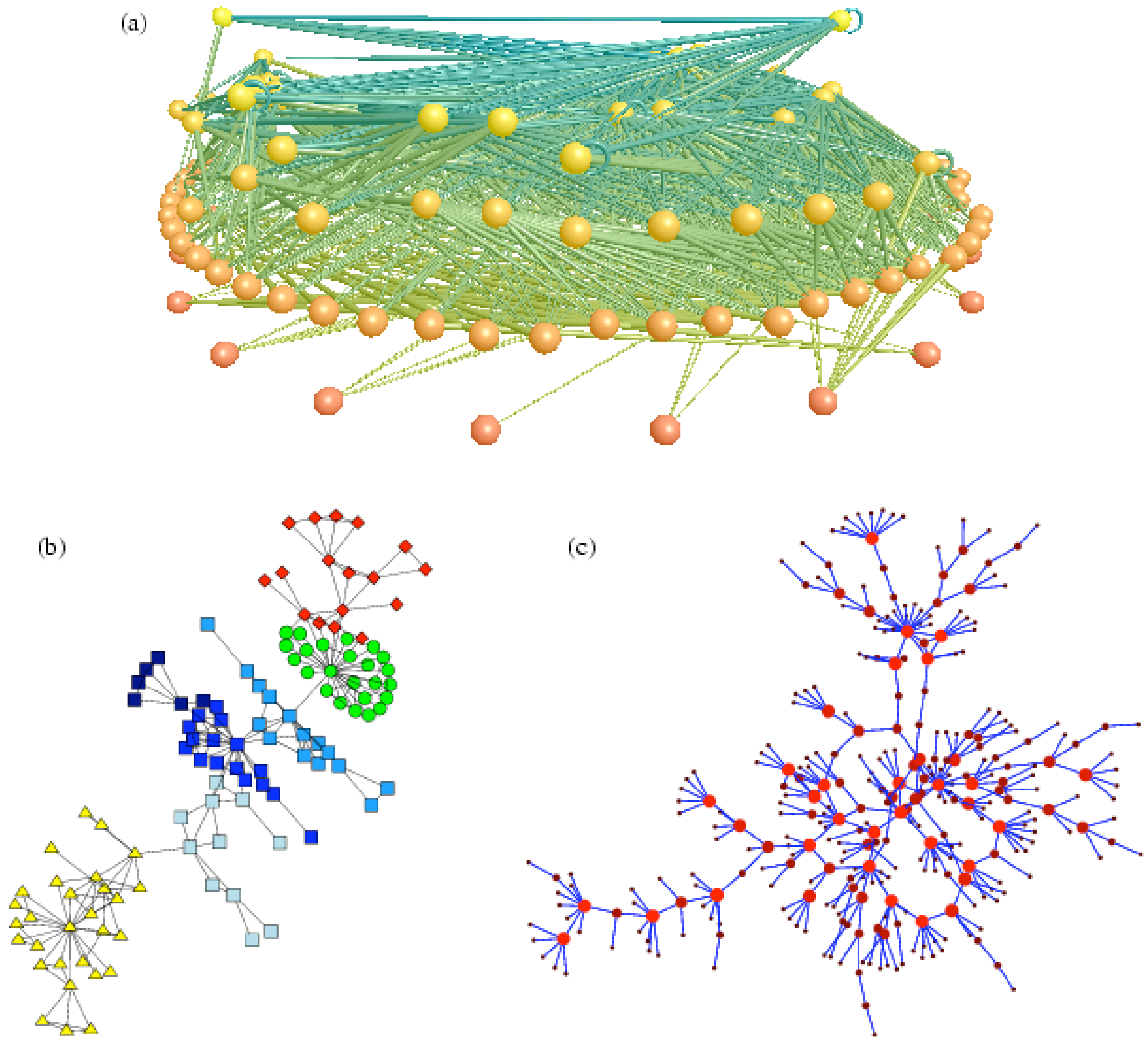
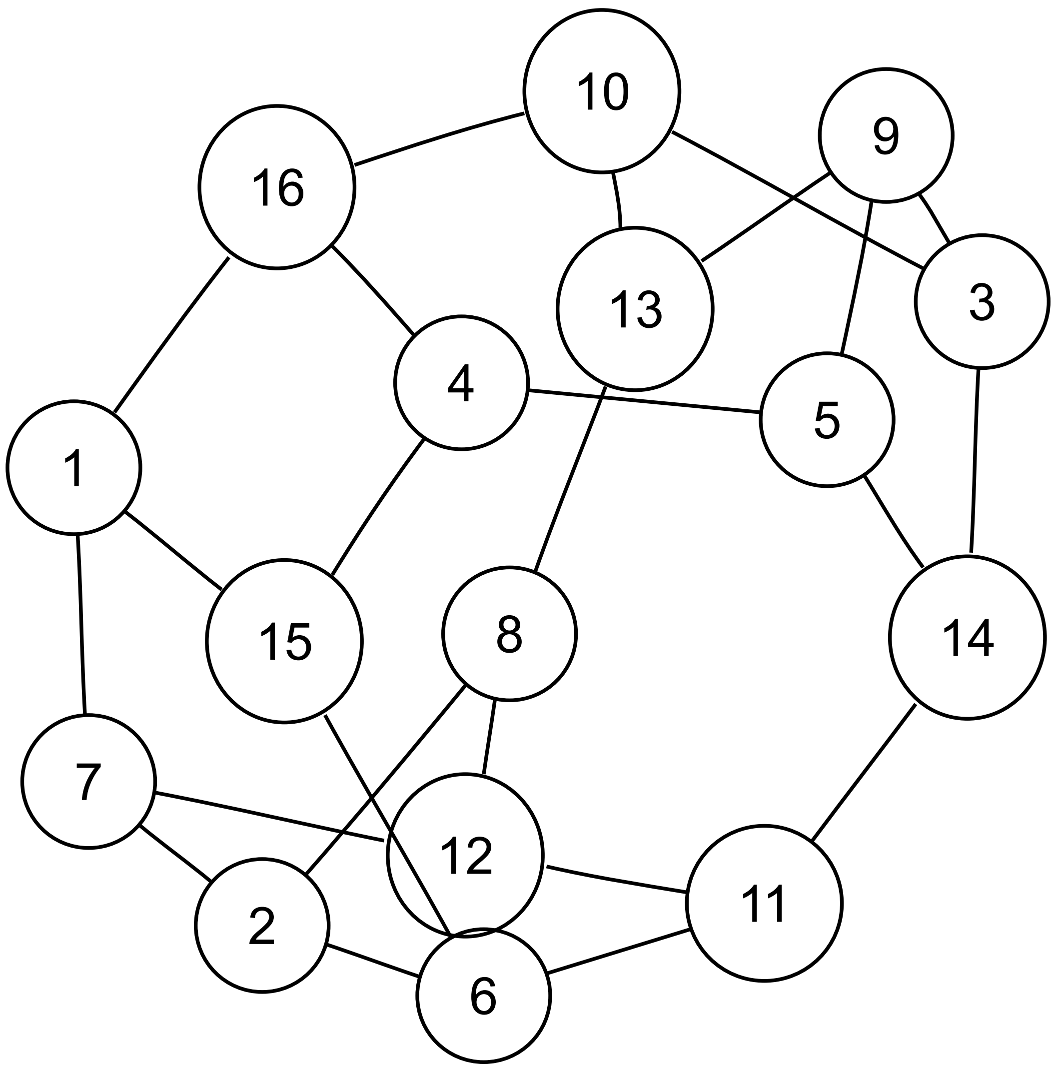


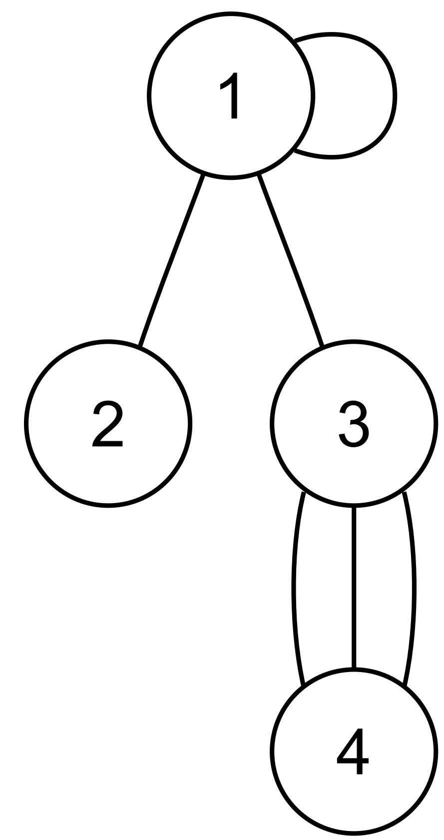
FIG. 2 Three examples of the kinds of networks that are the topic of this review. (a) A food web of predator-prey interactions between species in a freshwater lake [272]. Picture courtesy of Neo Martinez and Richard Williams. (b) The network of collaborations between scientists at a private research institution [171]. (c) A network of sexual contacts between individuals in the study by Potterat *et al.* [342].

Graphs

A graph G=(V,E) consists of a *vertex set* V and an *edge set* E containing unordered pairs {i,j} of vertices.



graph



multigraph

The degree of vertex v is the number of edges attached to it.

A Plea for Clarity: What is a Network?

- graph vs. multigraph (are loops, multiple edges ok? What is a "simple" graph?)
- directed vs. undirected
- weighted vs. unweighted
- dynamics of vs. dynamics on
- labeled vs. unlabeled
- network as quantity of interest vs. quantities of interest on networks

Why model networks?

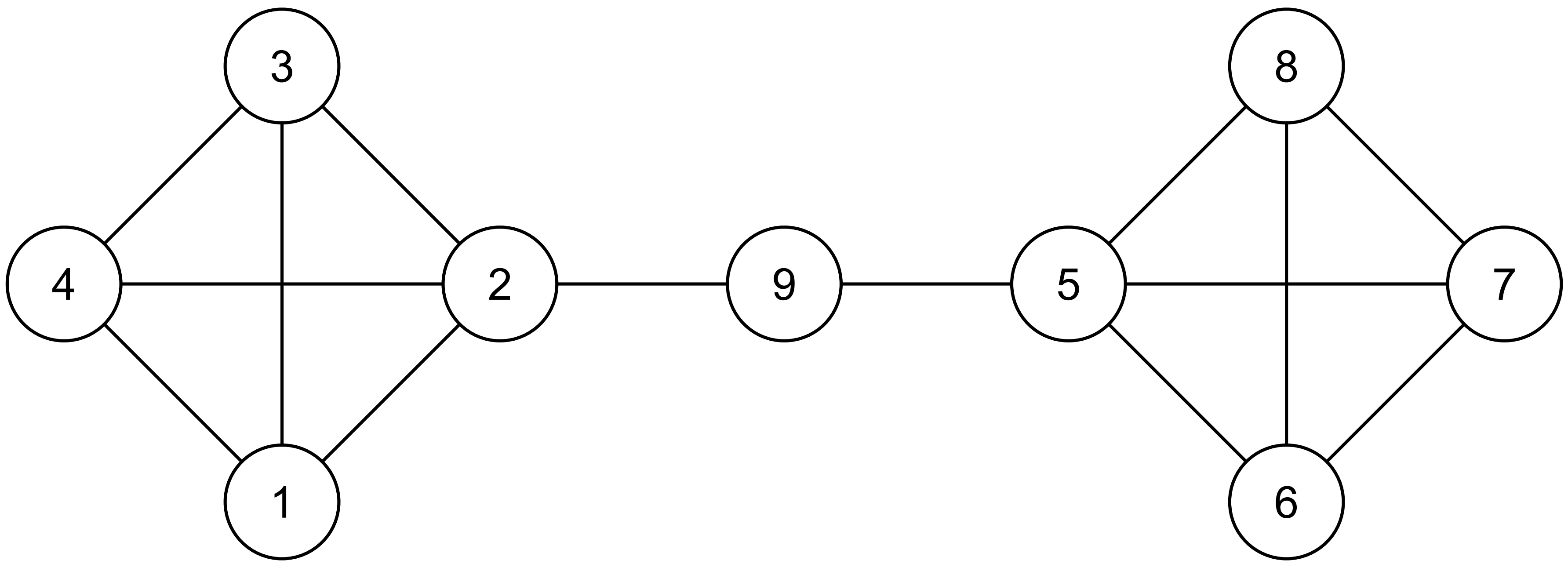
- Hard to interpret “hairballs”.
- We can define some interesting features (statistics) of a network, such as measures of clustering, and compare the observed values against those of a model
- Warning: much of the network literature carelessly ignores the way in which the network data were gathered (sampling) and whether there are missing/unknown nodes or edges!

Erdos-Renyi Random Graph Model

- Independently flip coins with prob. p of heads
- Let n get large and p get small, with the average degree $c = (n-1)p$ held constant.
- What happens for $c < 1$?
- What happens for $c > 1$?
- What happens for $c = 1$?

Degree Sequences

Take $V = \{1, \ldots, n\}$ and let d_i be the degree of vertex i. The degree sequence of G is $d = (d_1, \ldots, d_n)$.



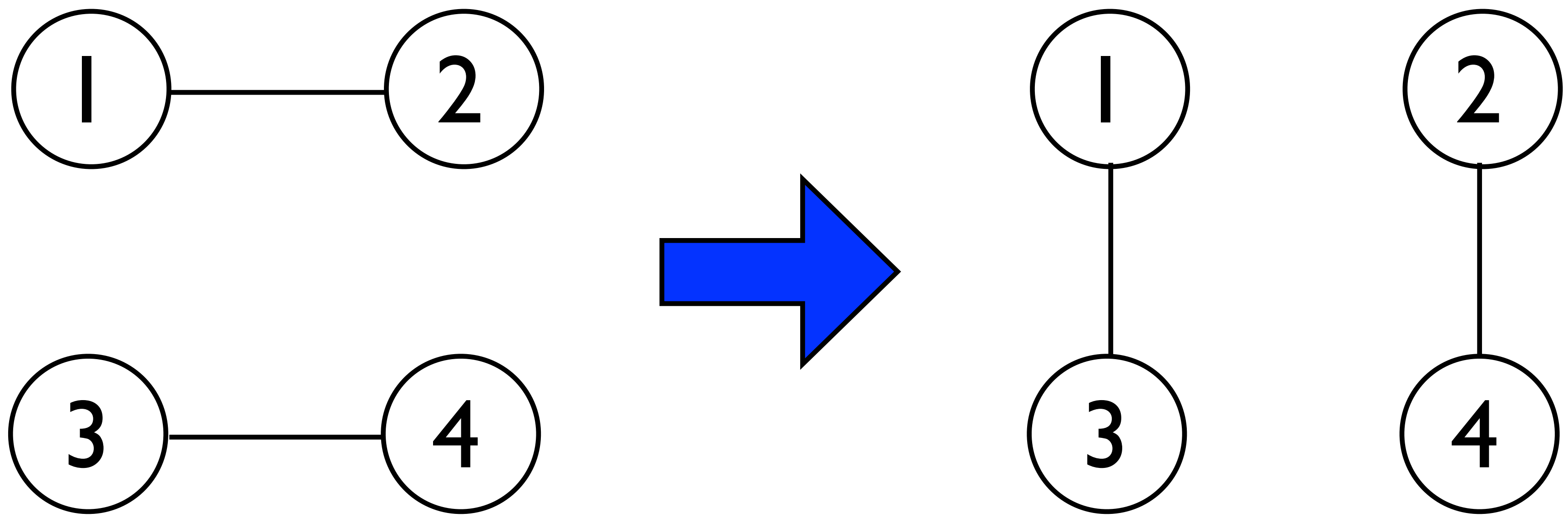
$$n = 9, d = (3, 4, 3, 3, 4, 3, 3, 3, 2)$$

A sequence d is *graphical* if there is a graph G with degree sequence d. G is a *realization* of d.

MCMC on Networks

mixing times, burn-in, bottlenecks, autocorrelation,...

Switchings Chain



Power Laws

- Power–law (a.k.a. scale–free) networks: the number of vertices of degree k is proportional to $k^{-\beta}$
- Stumpf et al (2005): Subnets of scale-free networks are not scale-free, especially for large β
- Their subnets are i.i.d. node-based.
- What about features other than degree distributions?

p1 Model (Holland-Leinhardt 1981)

- θ: a base rate for edge propagation,
- α_i (expansiveness): the effect of an outgoing edge from i,
- β_j (popularity): the effect of an incoming edge into j,
- ρ_{ij} (reciprocation/mutuality): the added effect of reciprocated edges.

ERGMs (Exponential Random Graph Models)

$$P_\beta(G) = Z^{-1} \exp\left(-\sum_{i=1}^{n} \beta_i d_i(G) \right)$$

How can we test and fit this model?
How can we use this model?

Pseudolikelihood (Strauss-Ikeda '80)

Fix a pair of nodes {i,j}, and consider the indicator r.v. of whether an edge {i,j} is present in G.

Conditioning on the rest of G yields great simplification:

$$\frac{P(\text{edge } \{i,j\}|\text{rest})}{P(\text{no edge } \{i,j\}|\text{rest})} = e^{\beta'(x(G^+)-x(G^-))}$$

So use logistic regression? Be careful of variance estimates!

MCMCMLE (Geyer-Thompson '92)

Write $$P_{\boldsymbol{\beta}}(G) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}(G))}{c(\boldsymbol{\beta})} = q_\beta(G)/c(\beta)$$

Fix some baseline β_0 and estimate log-likelihood ratio.

$$l(\beta) - l(\beta_0) = (\beta - \beta_0)' x(G) - \log \frac{c(\beta)}{c(\beta_0)}$$

Ratio of normalizing constants is: $$\frac{c(\beta)}{c(\beta_0)} = E_{\beta_0} \frac{q_\beta(G)}{q_{\beta_0}(G)}$$

So can approximate the MLE via MCMC.

What about the choice of β_0 though?

	i.i.d. node	i.i.d. edge	snowball	RDS	short paths
Erdos					
Dyad Indep.					
ERGM					
Fixed degree					
Geom					

Latent Space Models

Hoff et al (2002) model:

$$\eta_{i,j} = \log \text{odds}(y_{i,j} = 1 | z_i, z_j, x_{i,j}, \alpha, \beta)$$
$$= \alpha + \beta' x_{i,j} - |z_i - z_j|.$$

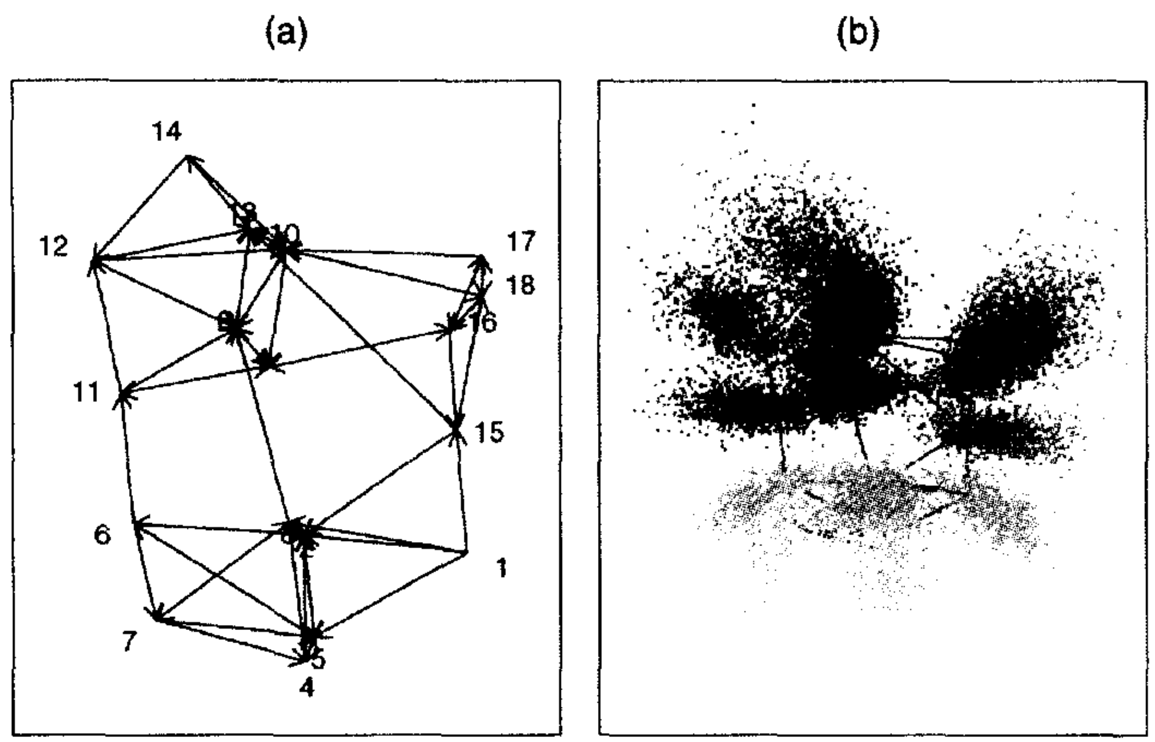
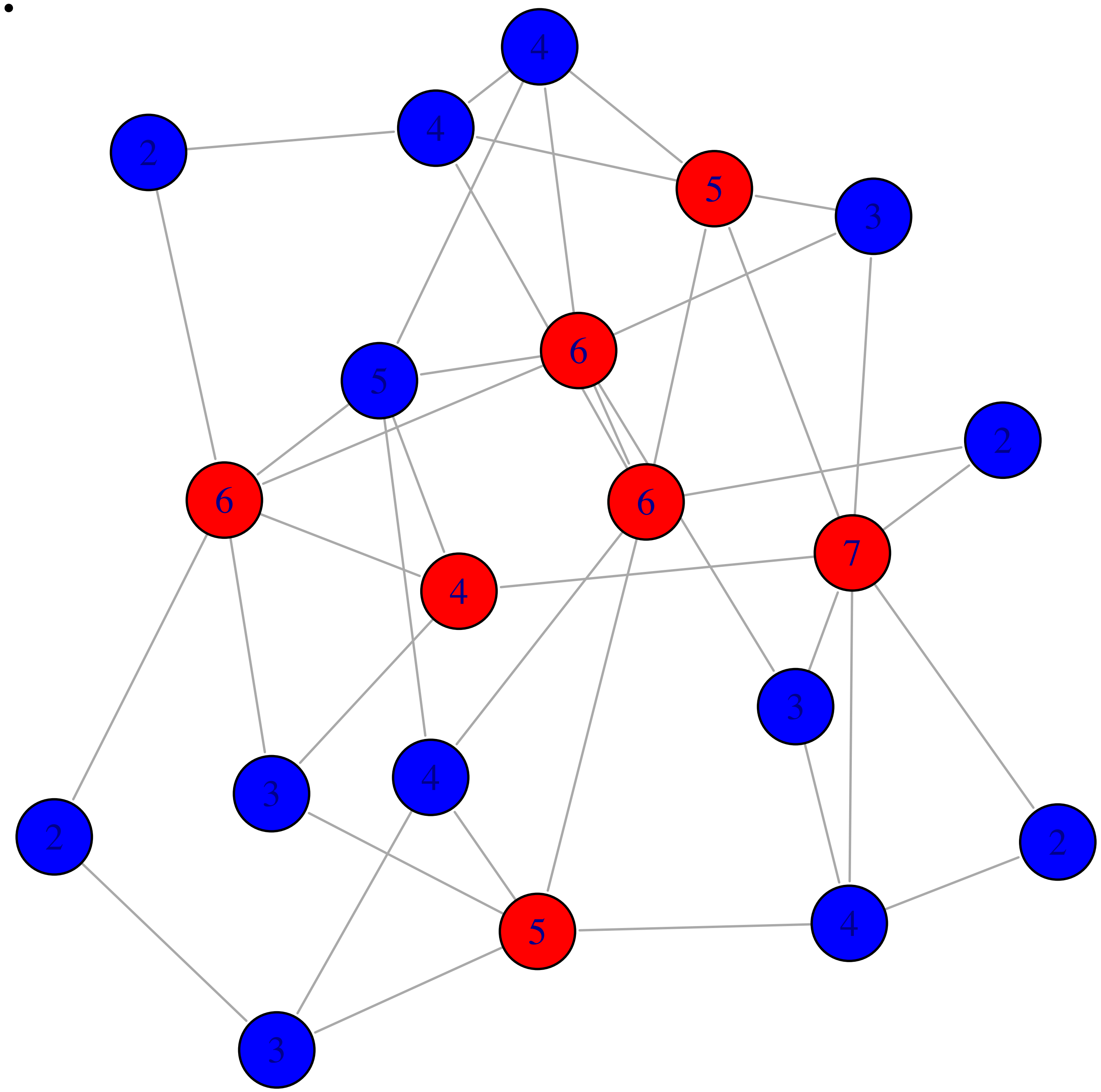


Figure 1. Maximum Likelihood Estimates (a) and Bayesian Marginal Posterior Distributions (b) for Monk Positions. The direction of a relation is indicated by an arrow.

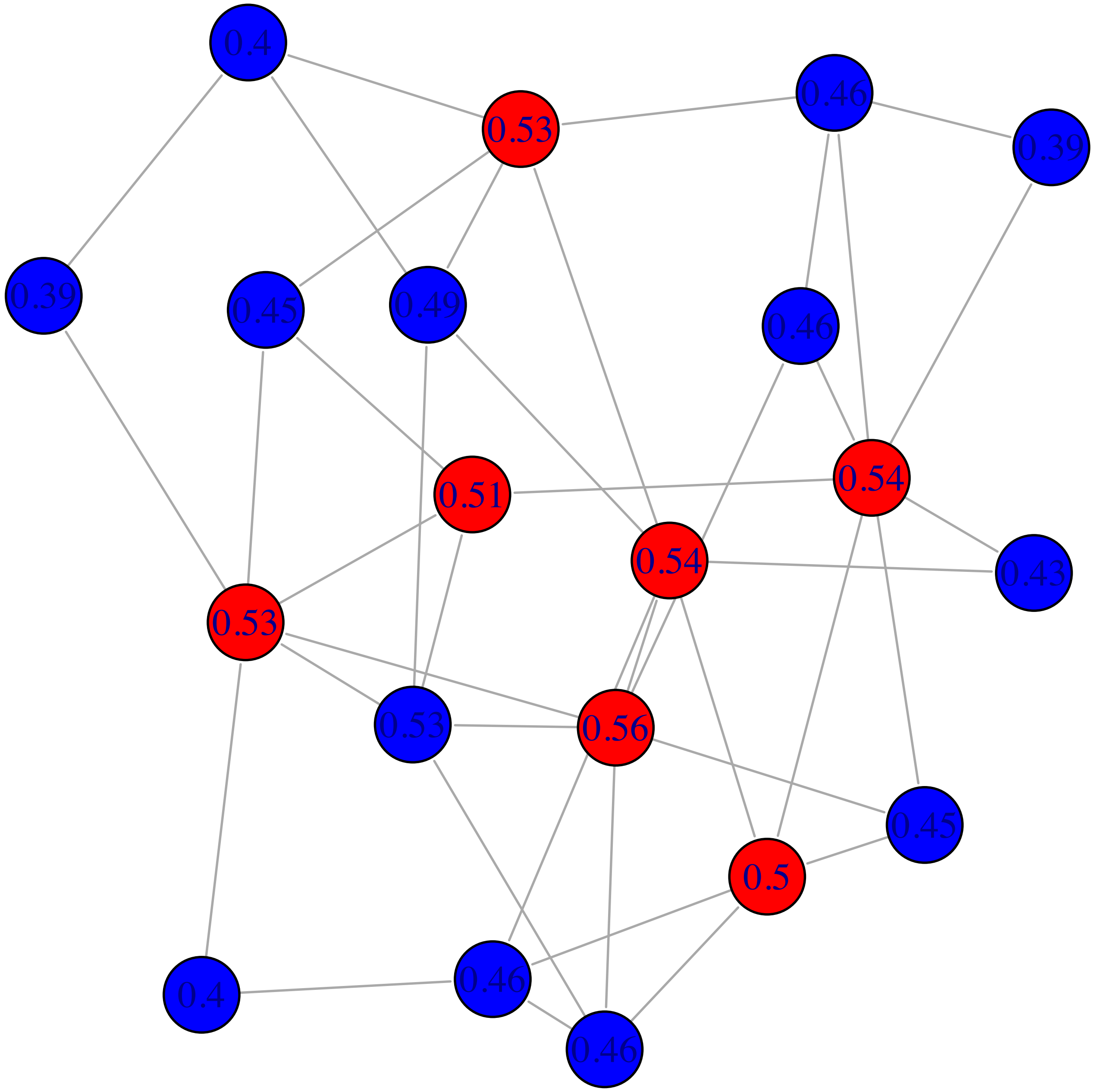
Degrees

Normalization?

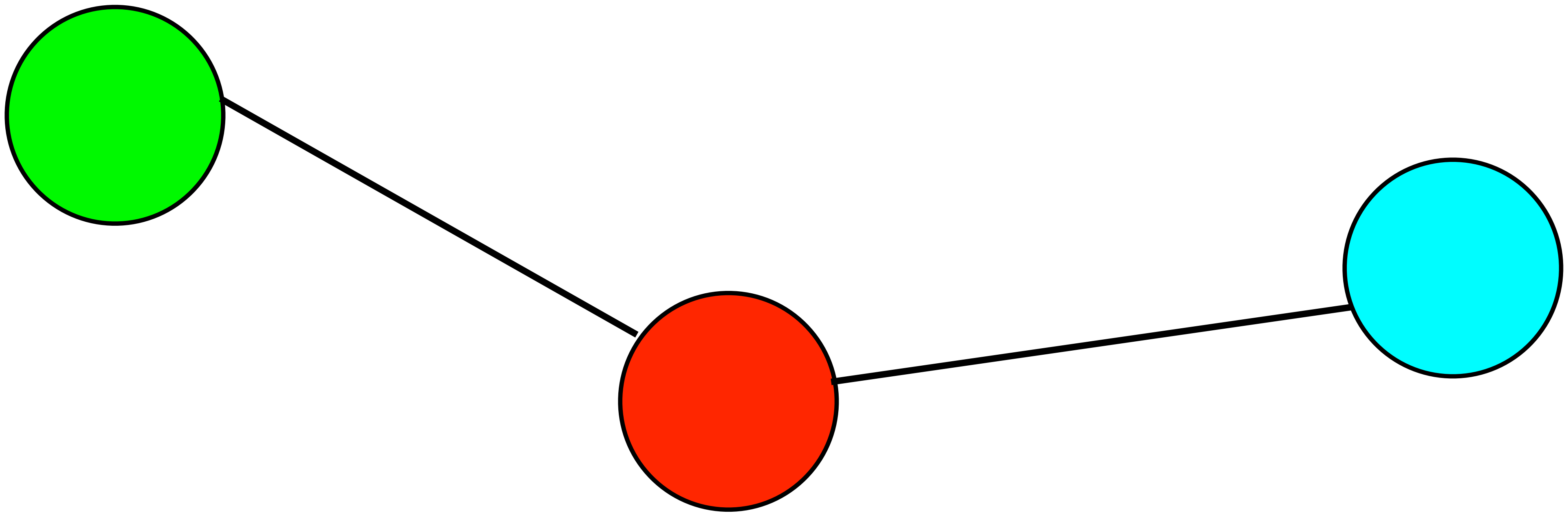


Closeness

uses the reciprocal of the average shortest distance to other nodes



Betweenness



many variations:
shortest paths
vs. flow
maximization
vs. all paths vs.
random paths

Eigenvector Centrality

use eigenvector of A corresponding to the largest eigenvalue (Bonacich); more generally, "power centrality"