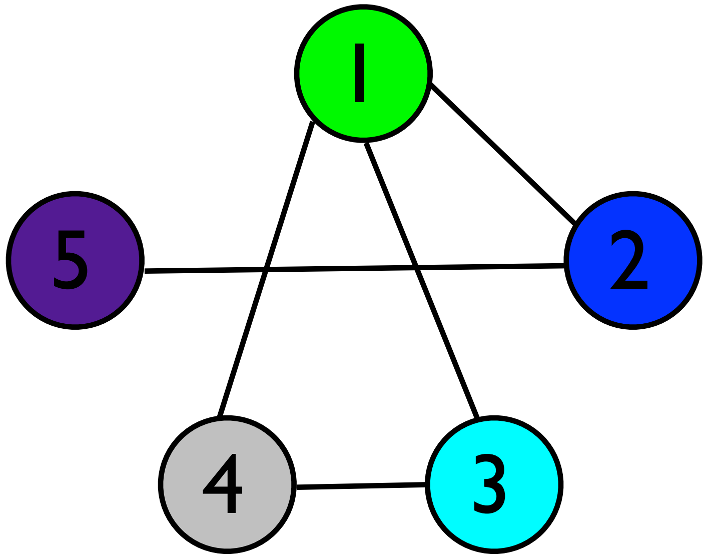
CS109/Stat121/AC209/E-109 Data Science

Network Models II

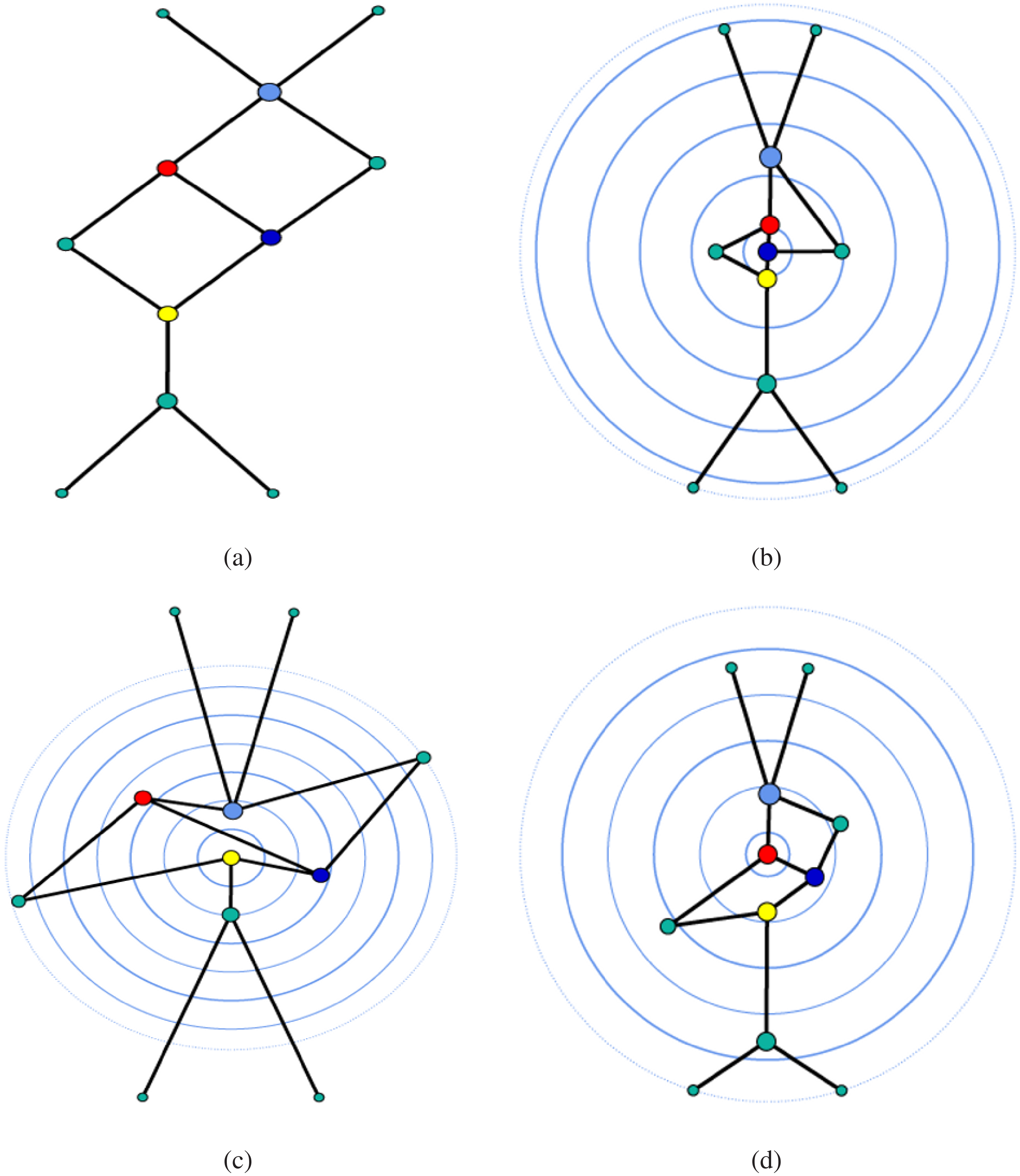
Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



This Week

- Project proposals due next Monday (Nov 11) http://cs109.org/projects/projects.php
- No late days or extensions are possible on project milestones or deadlines!
- HW5 due next Friday (Nov 15)
- Friday lab **10-11:30 am** in MD G115

Comparing centrality measures



(a)

(b)

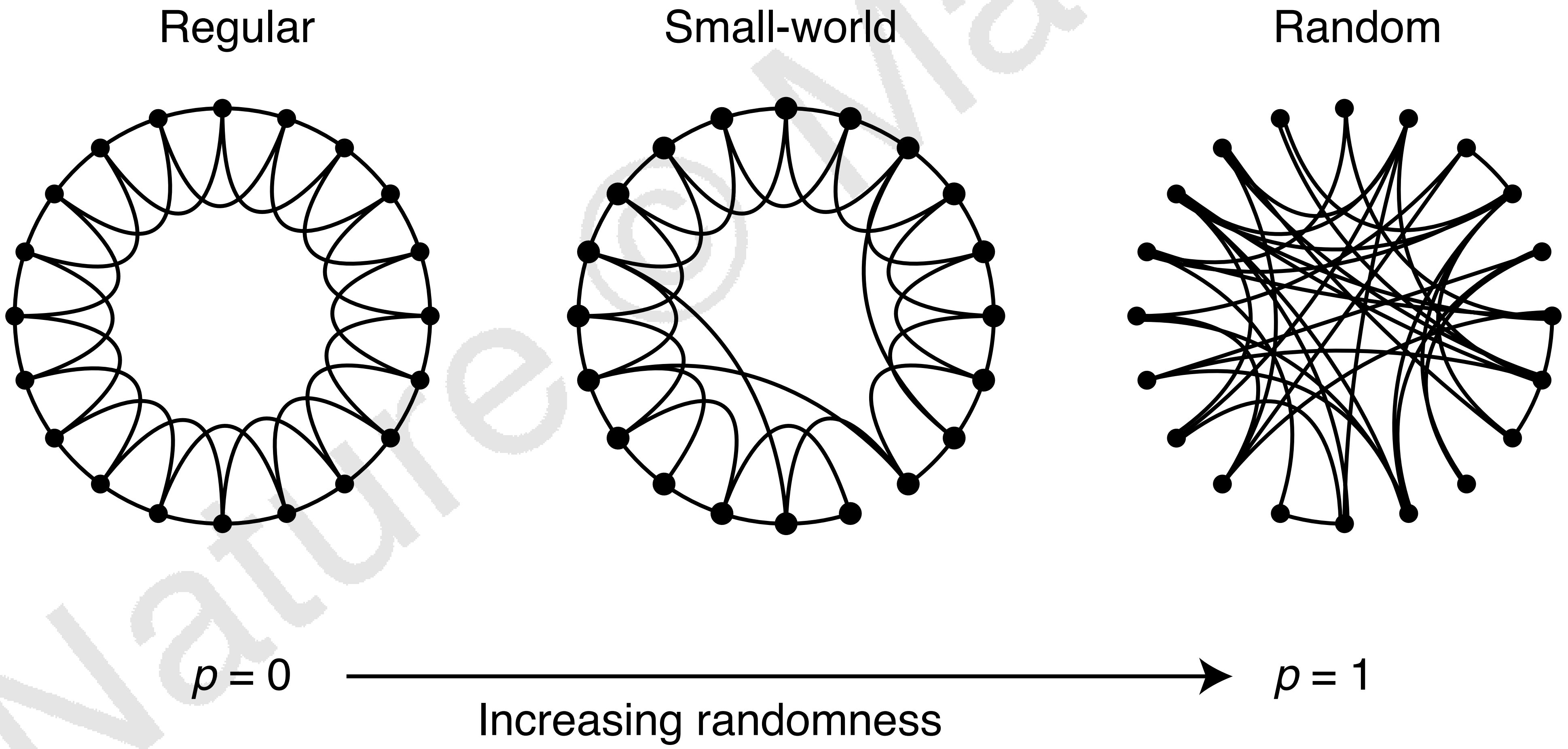
(c)

(d)

Fig. 4.4 Illustration of (b) closeness, (c) betweenness, and (d) eigenvector centrality measures on the graph in (a). Example and figures courtesy of Ulrik Brandes.

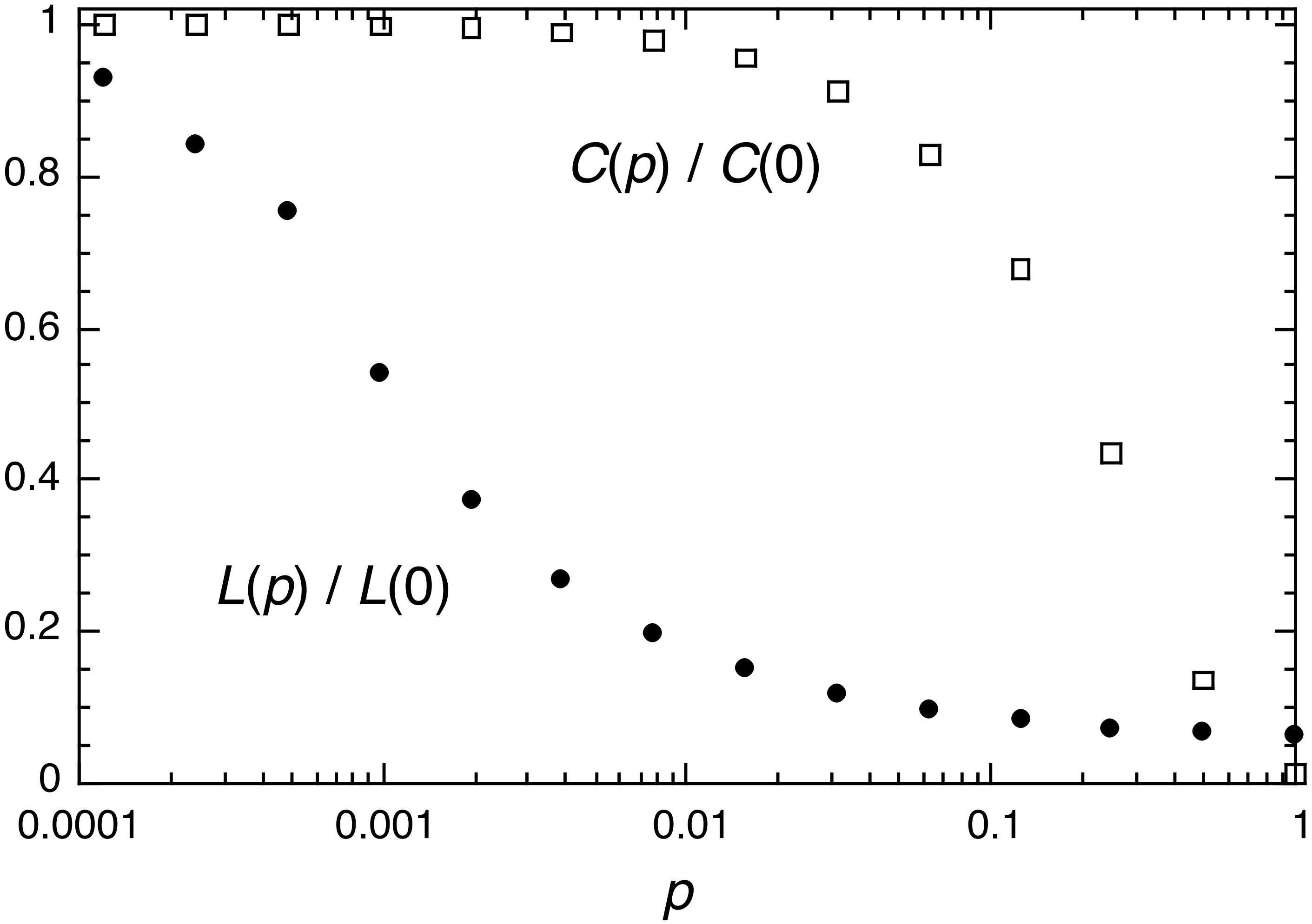
Kolaczyk (2009)

It’s a small world after all: Watts-Strogatz Model



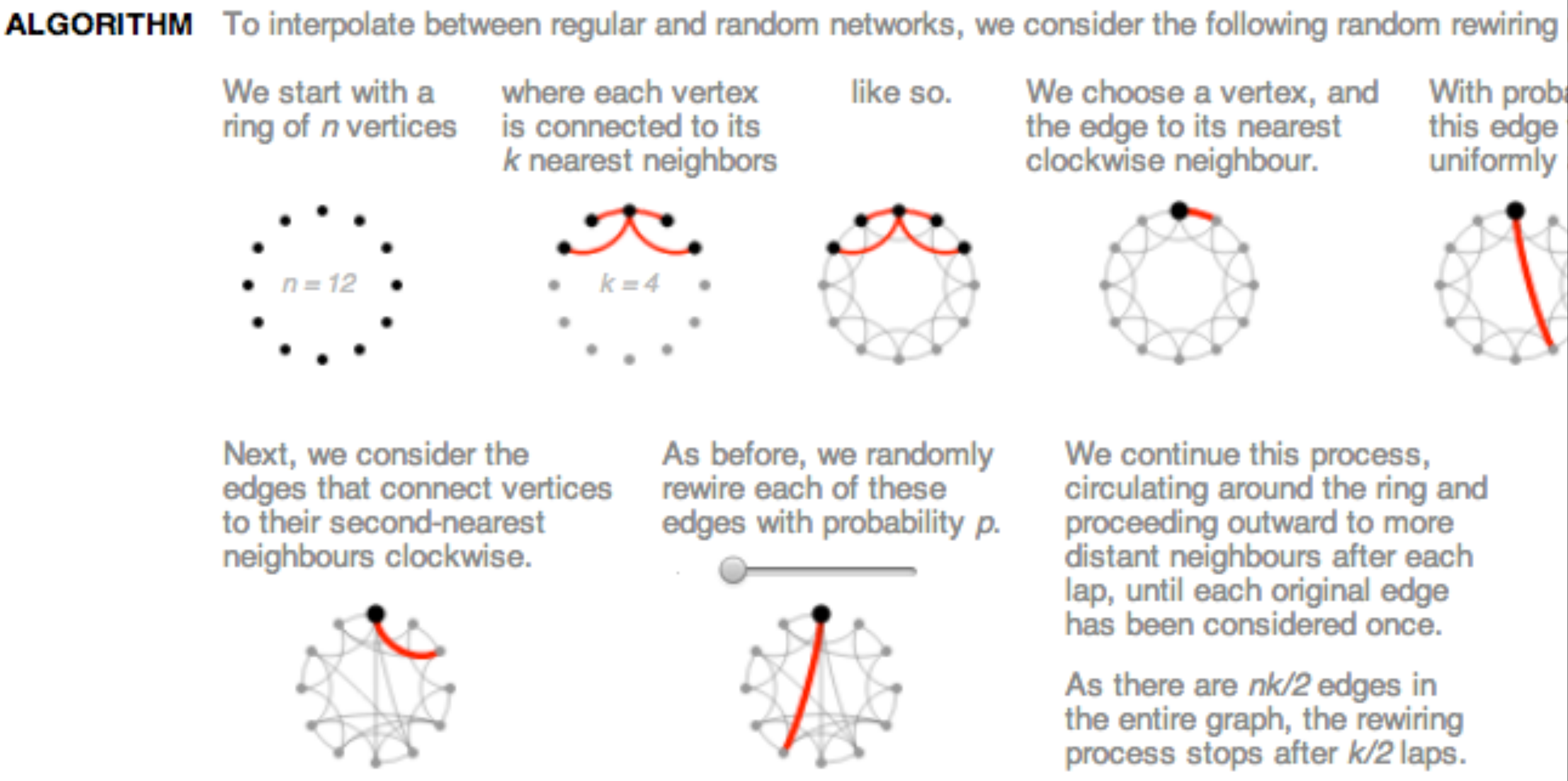
Watts-Strogatz (Nature, 1998)

Distances and clustering in Watts-Strogatz model



Watts-Strogatz (Nature, 1998)

Scientific Communication as Sequential Art (Bret Victor)



ALGORITHM To interpolate between regular and random networks, we consider the following random rewiring

We start with a ring of *n* vertices

where each vertex is connected to its *k* nearest neighbors

like so.

We choose a vertex, and the edge to its nearest clockwise neighbour.

With prob
this edge
uniformly

Next, we consider the edges that connect vertices to their second-nearest neighbours clockwise.

As before, we randomly rewire each of these edges with probability *p*.

We continue this process, circulating around the ring and proceeding outward to more distant neighbours after each lap, until each original edge has been considered once.

As there are *nk/2* edges in the entire graph, the rewiring process stops after *k/2* laps.

http://worrydream.com/ScientificCommunicationAsSequentialArt/

Class Size Paradox

Why do so many schools boast small average class size but then so many students end up in huge classes?

Simple example: each student takes one course; suppose there is one course with 100 students, fifty courses with 2 students.

Dean calculates: (100+50*2)/51 = 3.92

Students calculate: (100*100+100*2)/200 = 51

Class Size Paradox in Networks

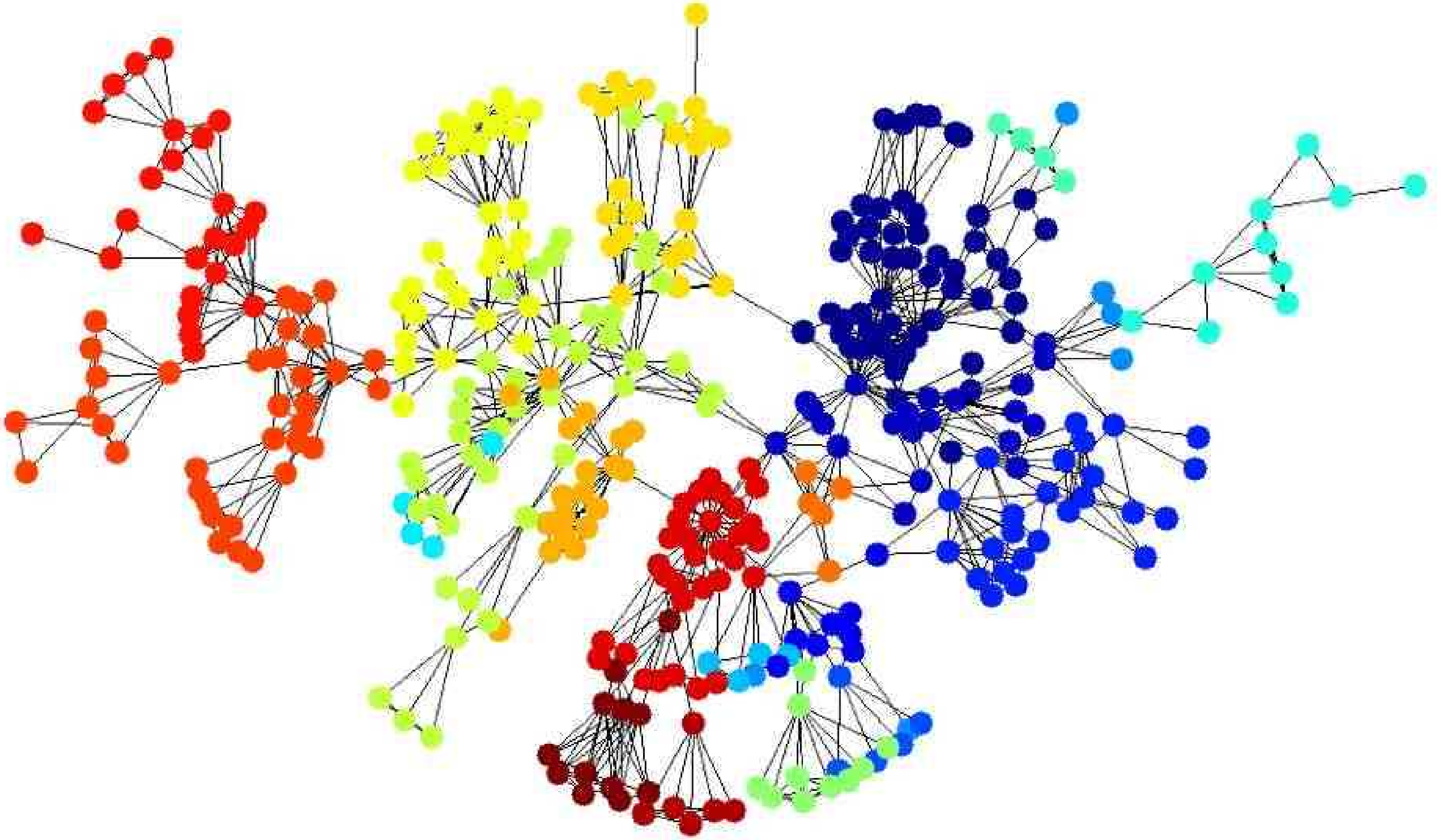
Average number of friends of a person’s friends is greater than average number of friends of a person!

Again a reminder of the importance of considering *sampling*.

Popular article on this phenomenon by Strogatz:

http://opinionator.blogs.nytimes.com/2012/09/17/friends-you-can-count-on/?_r=0

Community Detection



Porter et al survey: http://arxiv.org/pdf/0902.3788v2.pdf

Community Detection of Committees in Congress

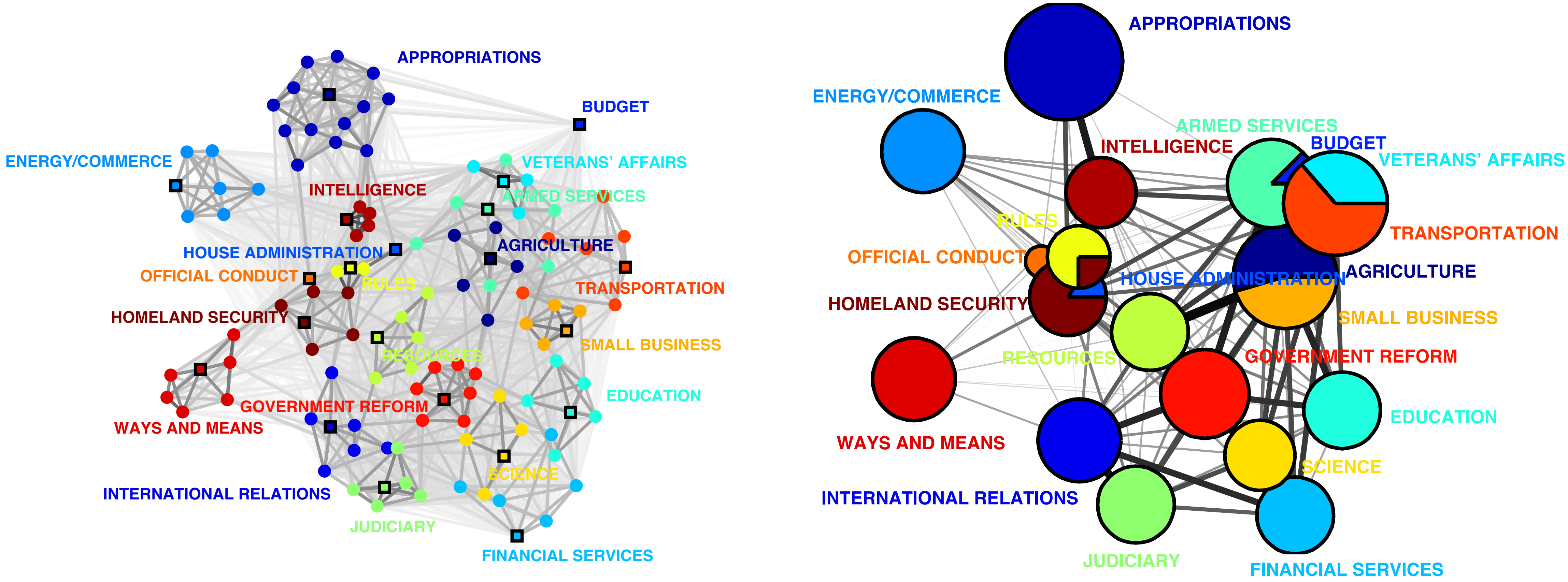


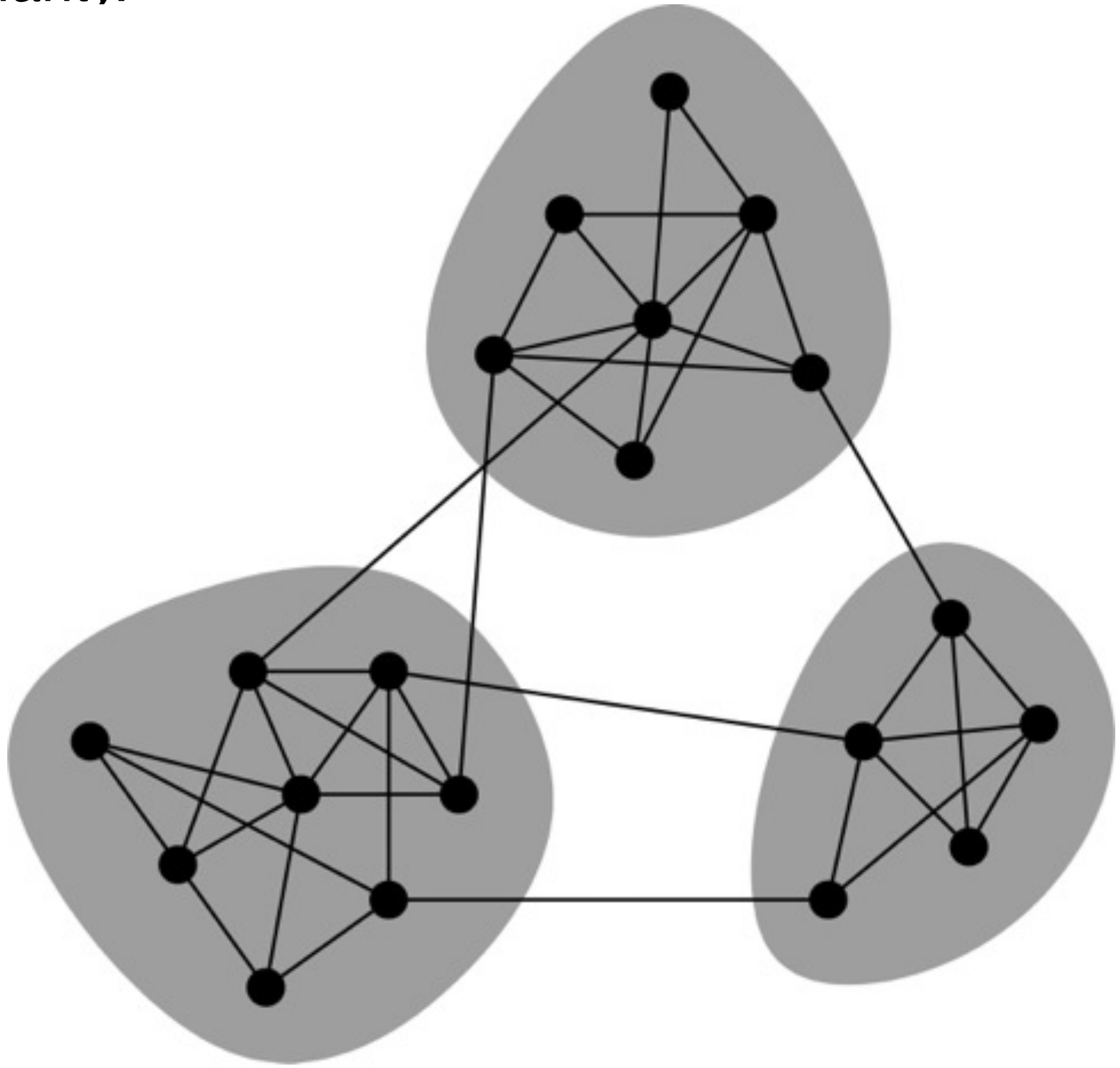
FIG. 0.4. *(Left) The network of committees (squares) and subcommittees (circles) in the 108th U.S. House of Representatives (2003-04), color-coded by the parent standing and select committees and visualized using the Kamada-Kawaii method [62]. The darkness of each weighted edge between committees indicates how strongly they are connected. Observe that subcommittees of the same parent committee are closely connected to each other. (Right) Coarse-grained plot of the communities in this network. Here one can see some close connections between different committees, such as Veterans Affairs/Transportation and Rules/Homeland Security.*

Porter et al survey: http://arxiv.org/pdf/0902.3788v2.pdf

Community Detection Algorithms

Girvan-Newman algorithm: iteratively remove edges by calculating betweennesses and removing the edge with maximum betweenness.

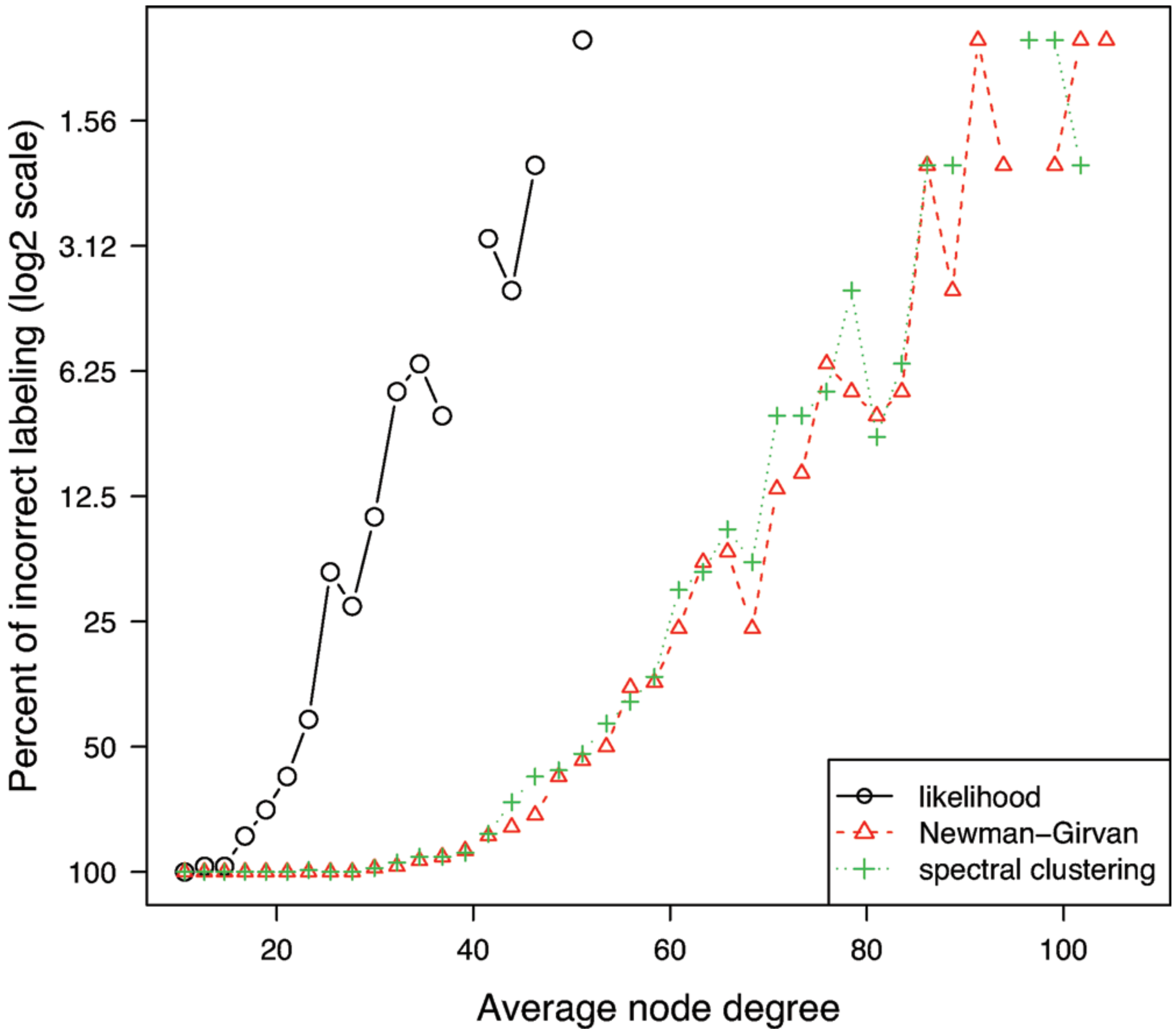
Metric called *modularity.*



http://www.pnas.org/content/103/23/8577/F1.expansion.html

Bickel-Chen on Community Detection

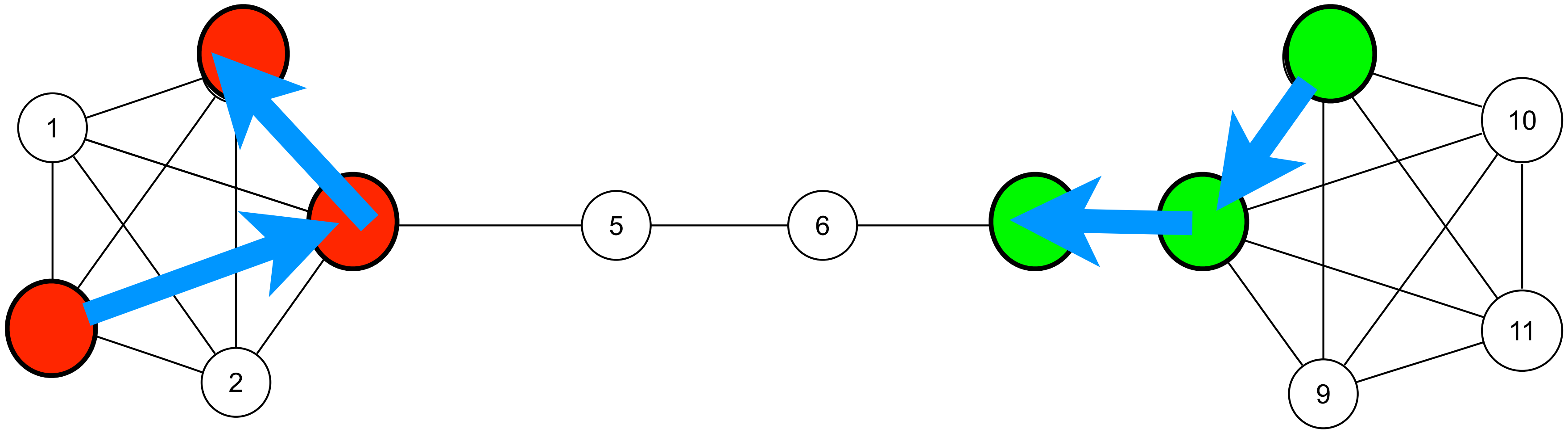
Inconsistency result for Newman-Girvan.



http://www.stat.berkeley.edu/~bickel/Bickel%20Chen%2021068.full.pdf

Respondent Driven Sampling (RDS)

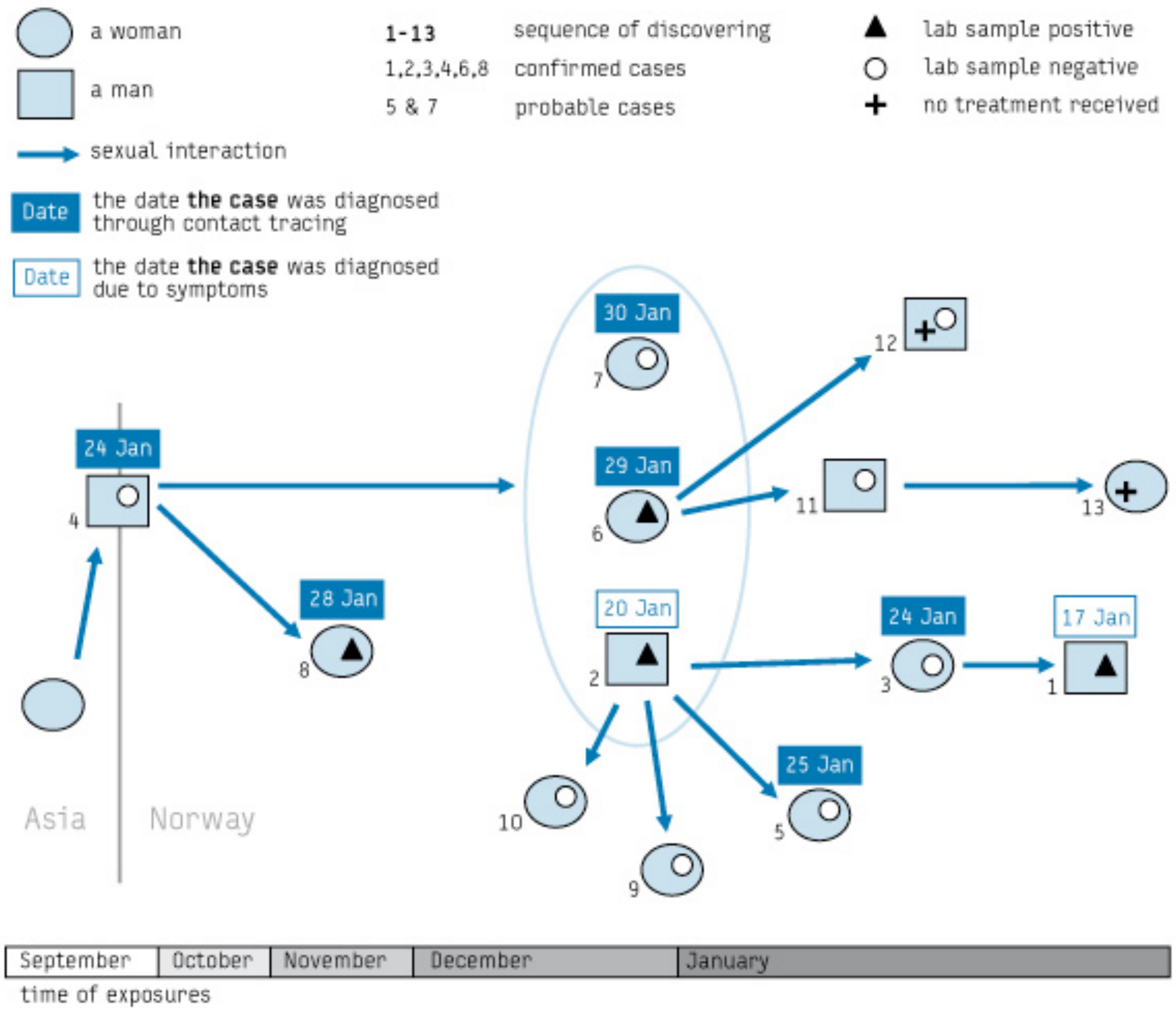
- sampling scheme for hard-to-reach populations, based on link-tracing across a social network with coupon incentives
- becoming extremely-widely used all over the world; hundreds of studies done or ongoing, e.g., CDC National HIV Behavioral Surveillance (NHBS) studies of injection drug users
- RDS as sampling vs. RDS estimation



Is RDS contact tracing?

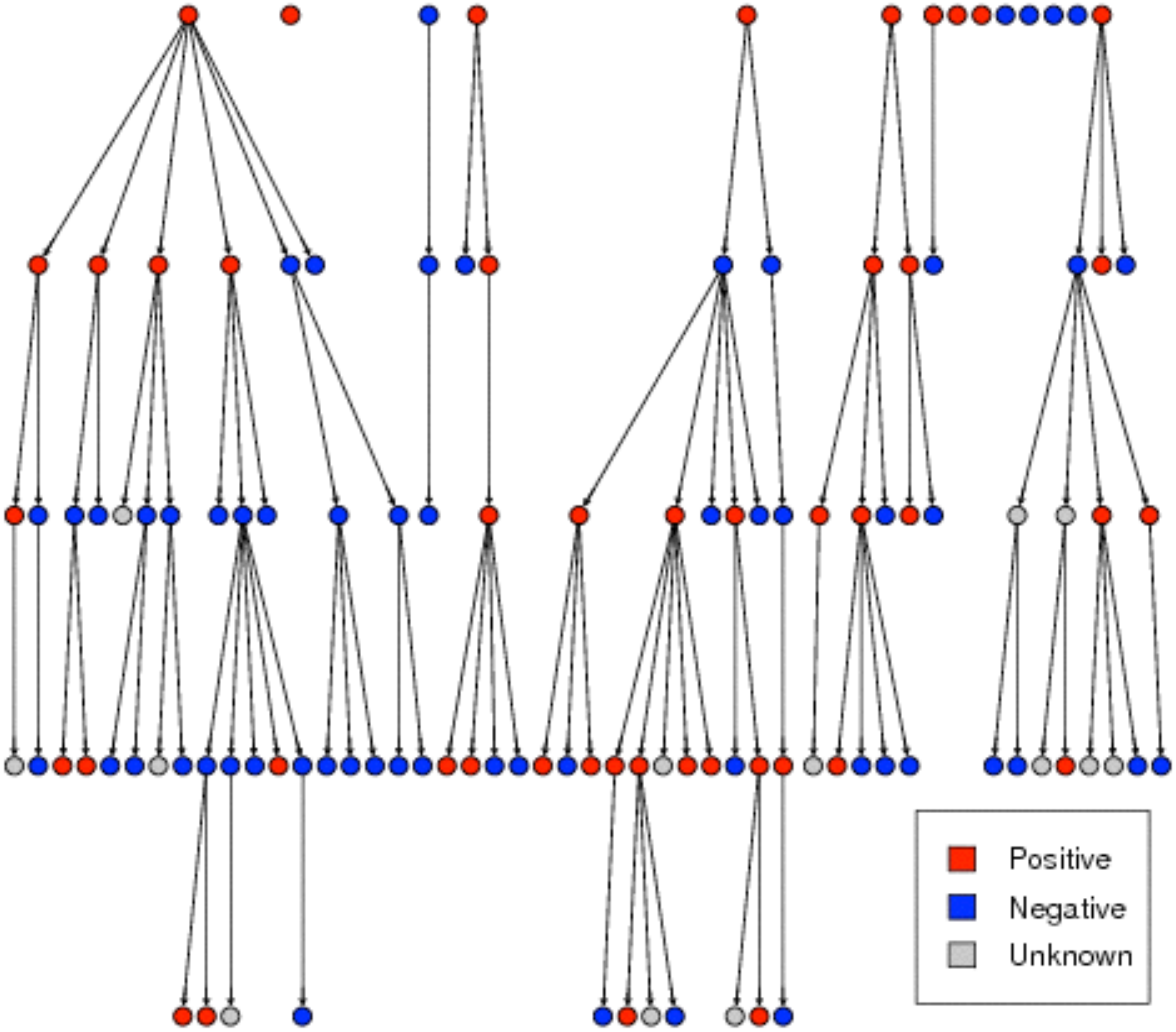
FIGURE

Sexual network of an outbreak of gonorrhoea in Norway, January 2008



Source: http://www.eurosurveillance.org/

Recruitment Tree Example



Volz-Heckathorn RDS Estimator

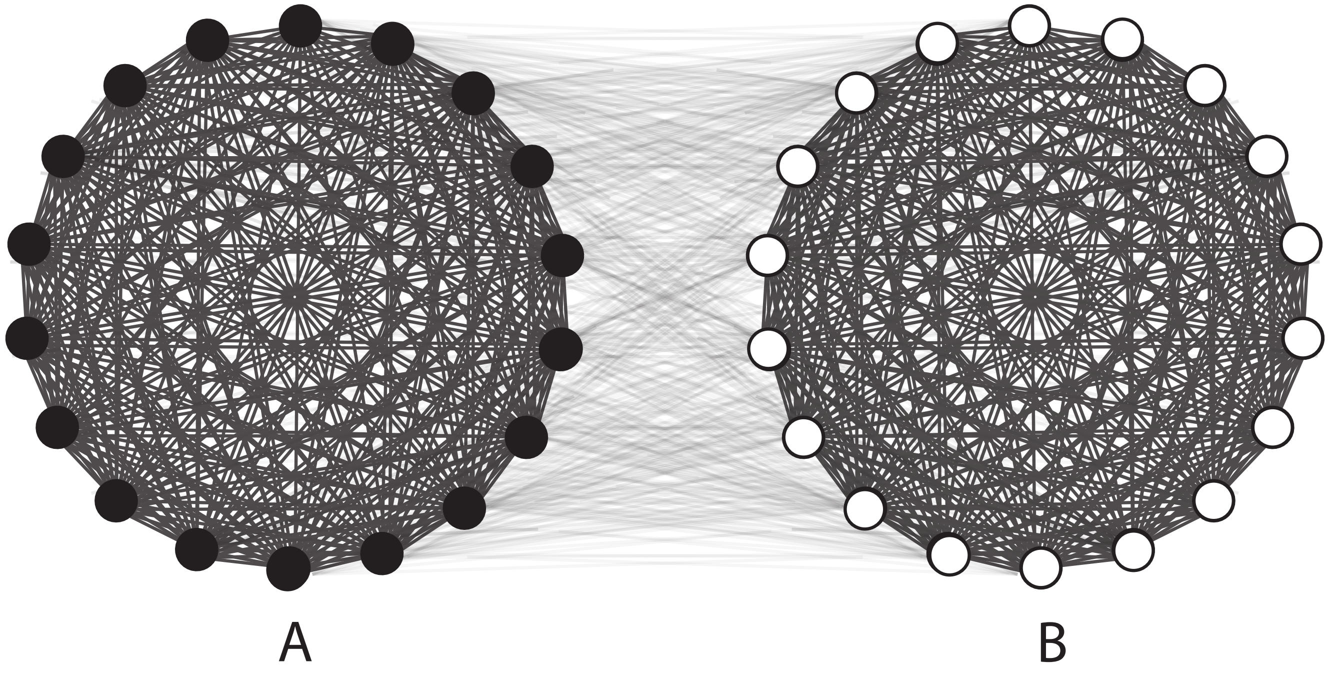
$$E(\hat{Y}) = \frac{\sum_{j=1}^{n} Y_j/d_j}{\sum_{j=1}^{n} 1/d_j}$$

This is a form of Horvitz-Thompson estimator, reweighting as in importance sampling.

Relies on a long list of strong assumptions; Handcock-Gile and Blitzstein-Nesterko perform sensitivity analyses under various conditions.

Goel-Salganik (Stats in Medicine 2009, PNAS 2010):

RDS variances can be extremely large, especially if there are bottlenecks in the network from modularity/communities, and from multiple recruitment. Typical design effects of 5-10, and coverage probabilities much lower than the nominal 95% values



What would Fisher say?

To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

-- R.A Fisher

To Model or Not To Model; Design-based vs. model-based

- Model the underlying network? What about unknown nodes?
- the recruitment process?
- coupon refusal?
- the outcome variables (such as HIV status)?