

Video Lectures

[Help Center](#)

Having trouble viewing lectures? Try changing players. Your current player format is html5. [Change to flash.](#)

> Week 1: Introduction

- ✓ [1A Why Social Network Analysis? \(13:54\)](#)
- ✓ [1B Software Tools \(13:13\)](#)
- ✓ [1C Degree and Connected Components \(20:32\)](#)
- ✓ [1D Gephi Demo \(9:20\)](#)



> Week 2: Random Graph Models

- ✓ [2P intro remarks for week 2](#)
- ✓ [2A Introduction to random graph models \(16:58\)](#)
- ✓ [2B random graphs and alternative models \(20:04\)](#)
- ✓ [2C Models of network growth \(25:28\)](#)



> Week 3: Centrality

- ✓ [3A degree, betweenness, closeness \(26:41\)](#)
- ✓ [3B eigenvector & directed \(16:49\)](#)
- ✓ [3C centrality applications \(19:44\) \(optional\)](#)
- ✓ [3D power laws \(20:15\) \(optional\)](#)
- ✓ [3E Cameron Marlow on Data Science \(3:25\)](#)



> Week 4: Community structure

- ✓ [4A Why detect communities? \(10:22\)](#)
- ✓ [4B Heuristics for finding communities \(13:51\)](#)
- ✓ [4C community finding \(22:16\)](#)
- ✓ [4D SNA @ LinkedIn \(24:05\)](#)



> Week 5: Small world networks

- ✓ [5A Small world experiments \(12:28\)](#)
- ✓ [5B clustering and motifs \(20:45\)](#)



✓ 5C small world models (20:27)     

✓ 5D origins of small worlds (9:19)     

✓ 5E Mathieu Bastian (Gephi + LinkedIn) (4:58)   

> Week 6: processes on networks

✓ 6A network topology and diffusion (12:34)     

✓ 6B complex contagion (22:43)   

✓ 6C innovation and coordination (14:54)   

✓ 6D (optional) Eric Sun of Facebook on geo + social (~20mins)   

> Week 7: cool and unusual applications

✓ 7A Cool and unusual applications   

✓ 7B Predicting recipe ratings using ingredient networks     

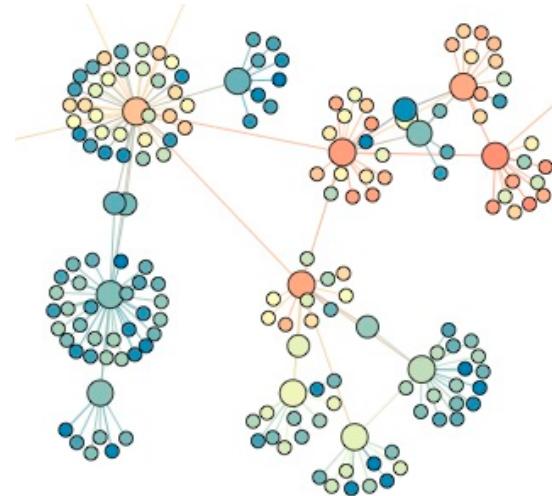
> Week 8: network resilience

✓ 8A network resilience     

✓ 8B resilience and assortativity   

✓ 8C resilience and the power grid   

✓ 8D concluding remarks   

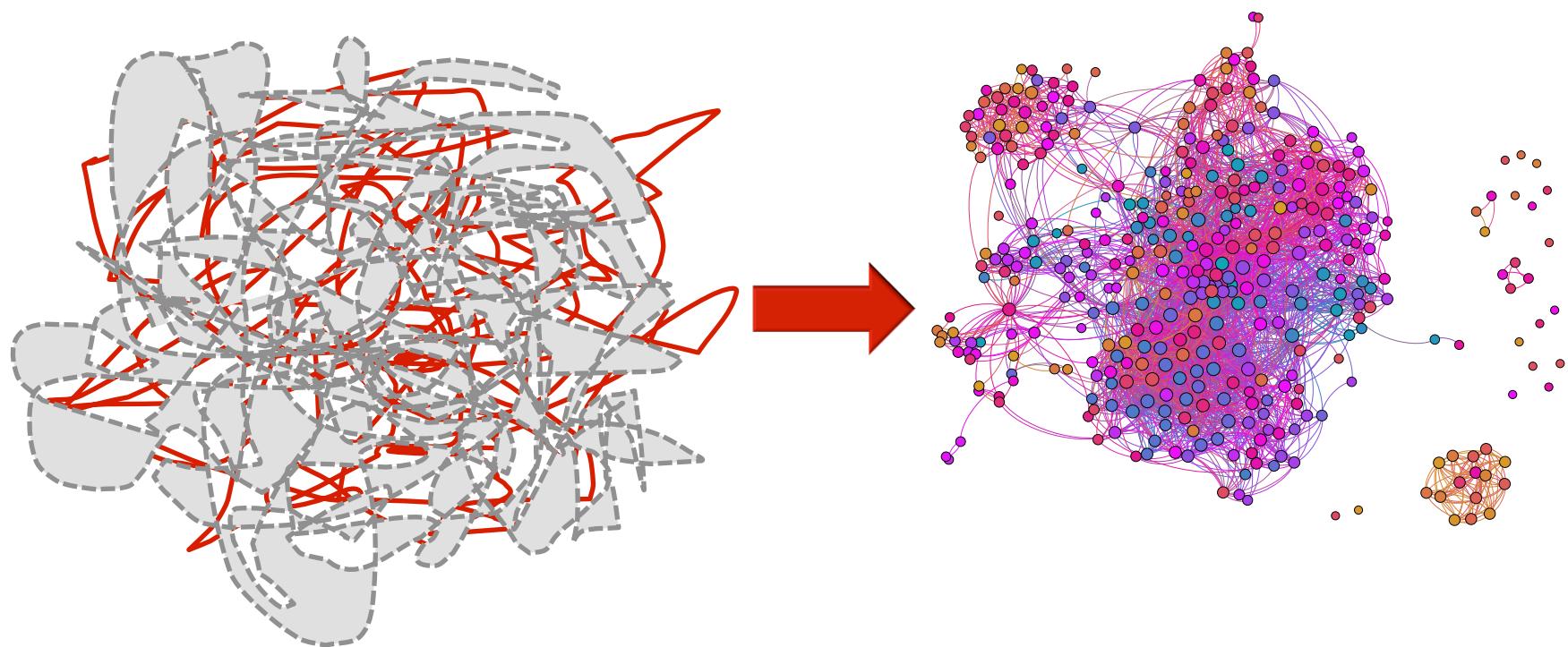


Social Network Analysis

Lada Adamic

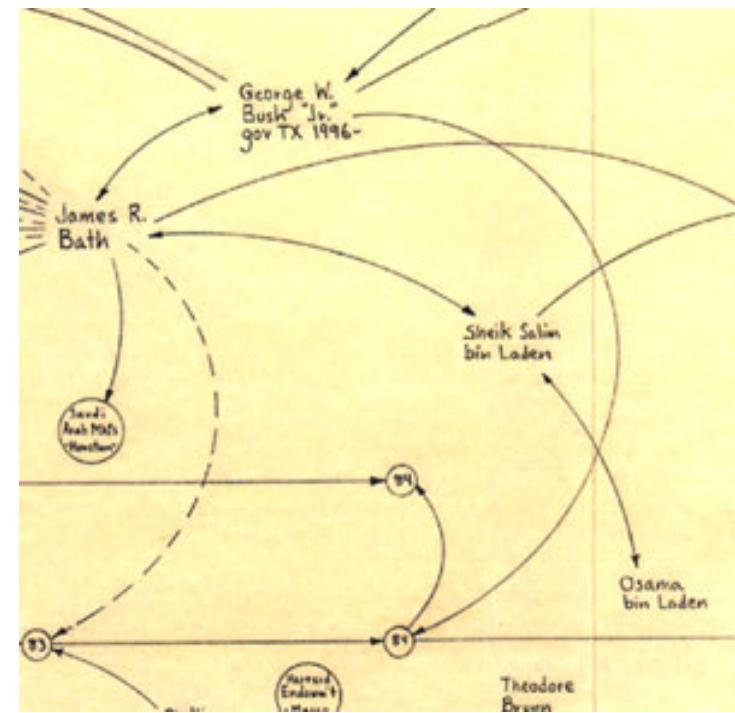
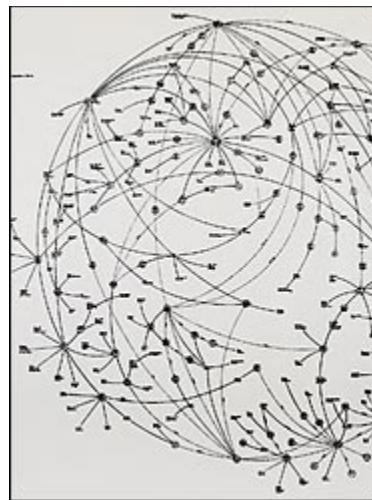


What do we get out of studying systems as networks?



examples: Political/Financial Networks

- Mark Lombardi: tracked and mapped global financial fiascos in the 1980s and 1990s from public sources such as news articles



Understanding through visualization

■ “I happened to be in the Drawing Center when the Lombardi show was being installed and several consultants to the Department of Homeland Security came in to take a look. They said they found the work revelatory, not because the financial and political connections he mapped were new to them, but because Lombardi showed them an elegant way to array disparate information and make sense of things, which they thought might be useful to their security efforts. I didn’t know whether to find that response comforting or alarming, but I saw exactly what they meant.”

Michael Kimmelman

Webs Connecting the Power Brokers, the Money and the World

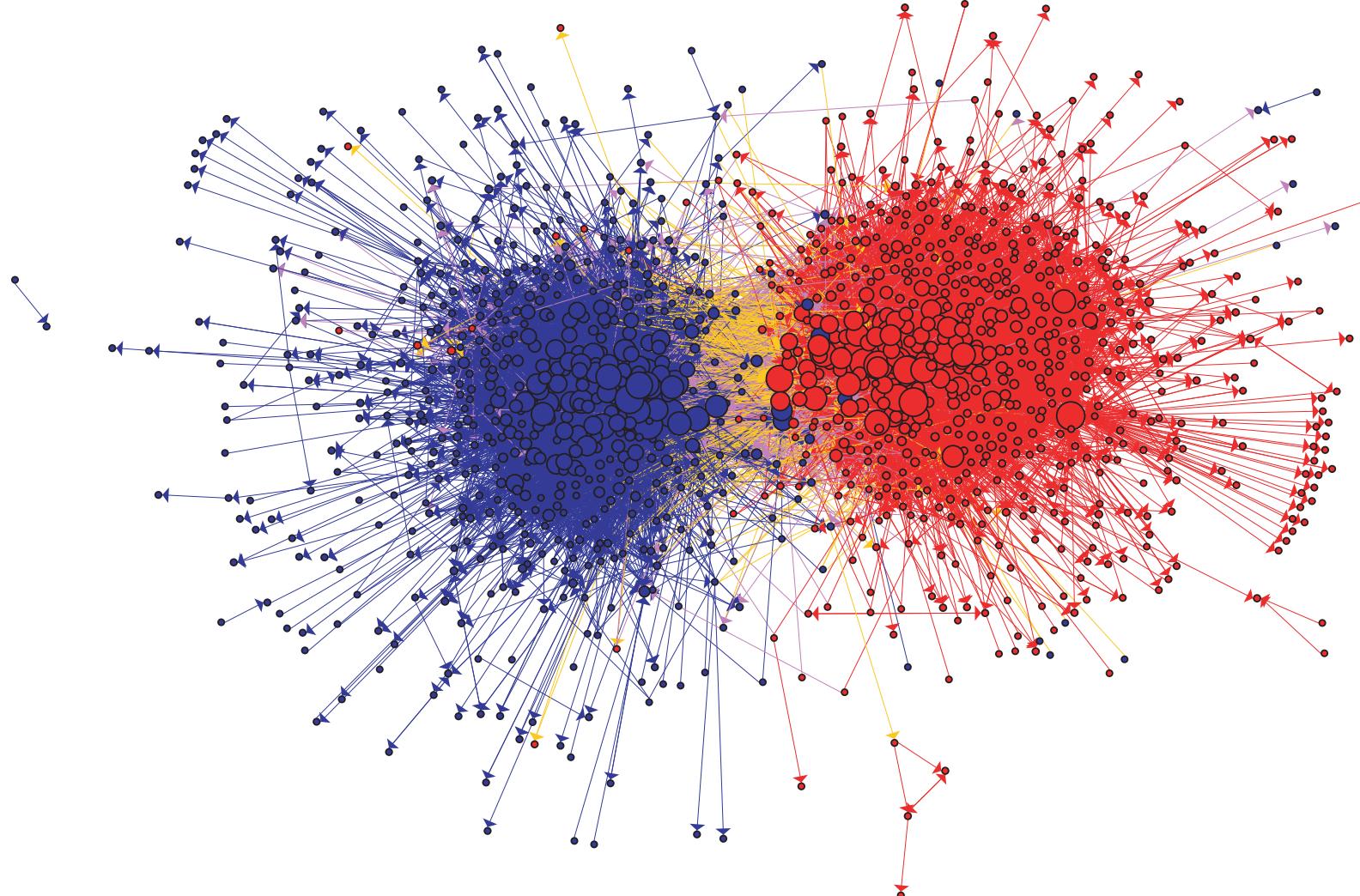
NY Times November 14, 2003

Internet

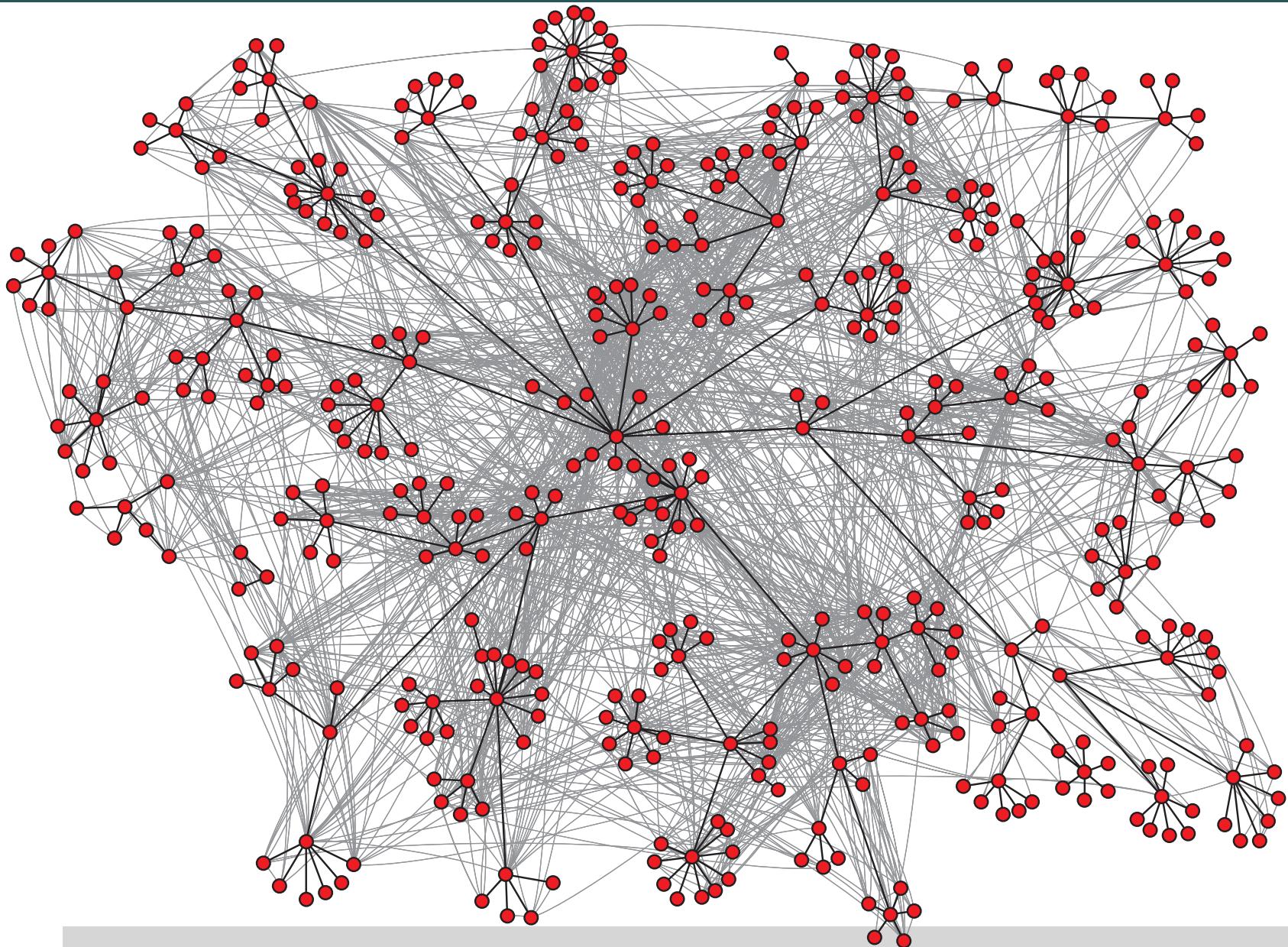


structure of the Internet at the level of autonomous systems. Data source: Mark Newman <http://www-personal.umich.edu/~mejn/netdata/>.

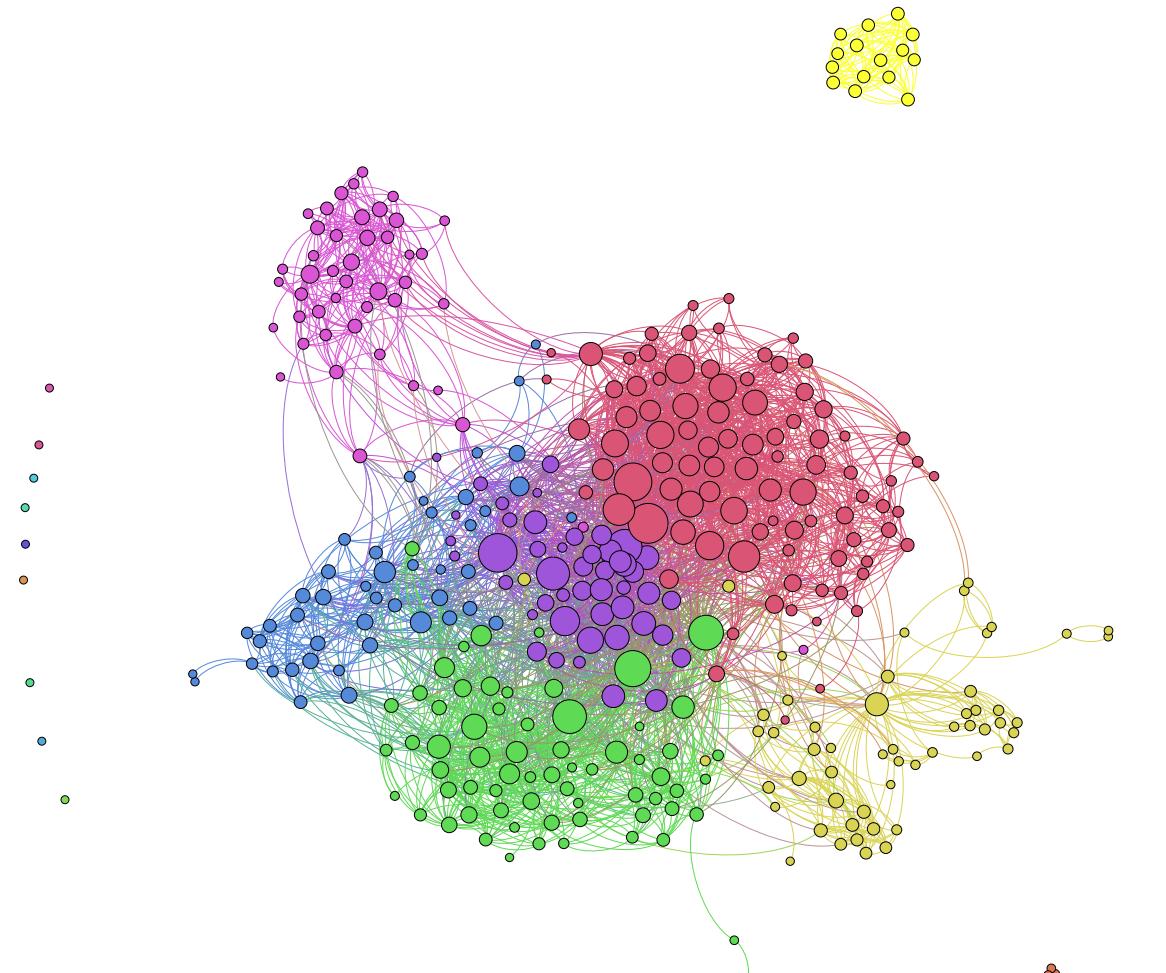
Political blogs



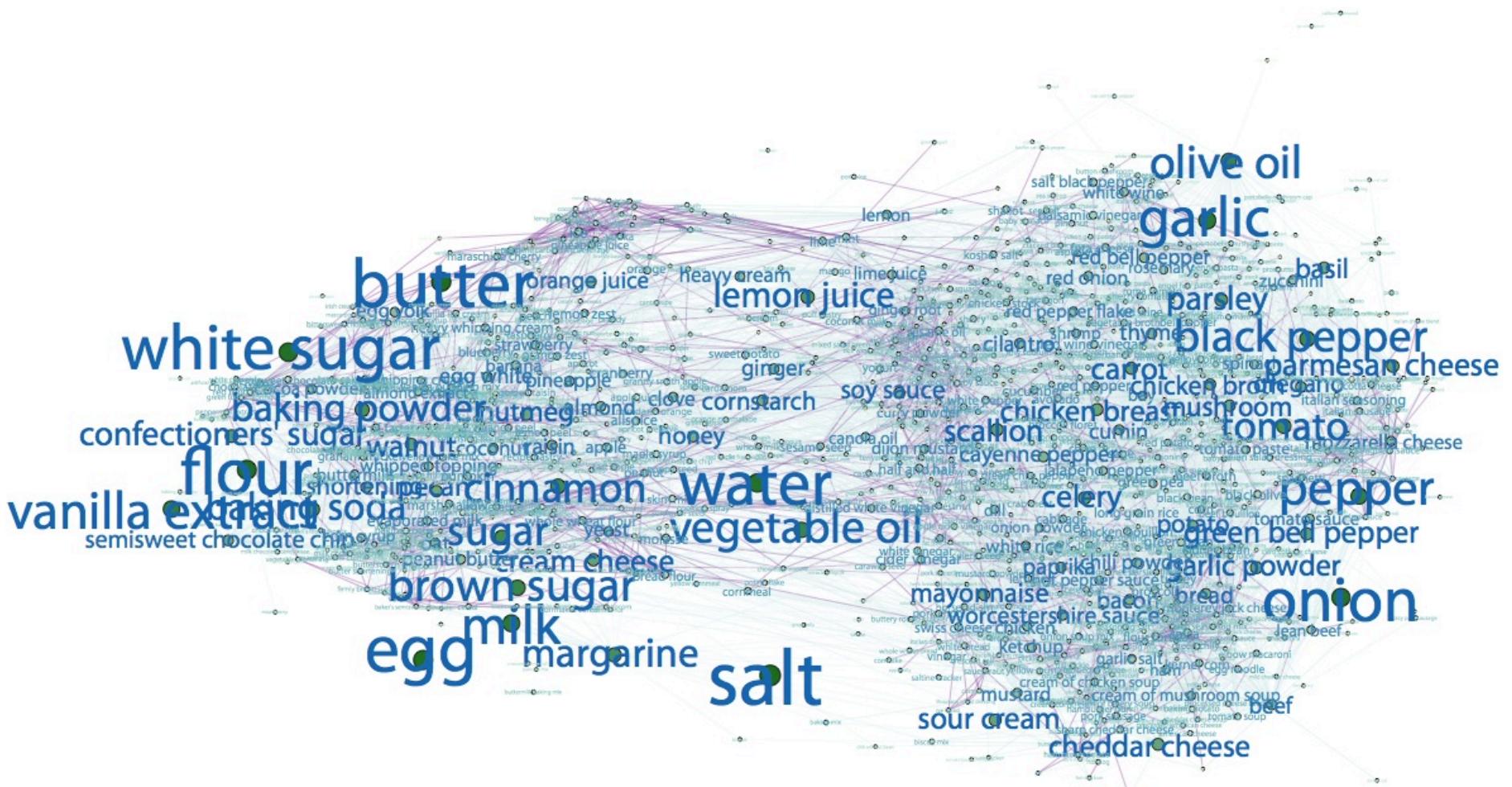
Organizations



Facebook networks

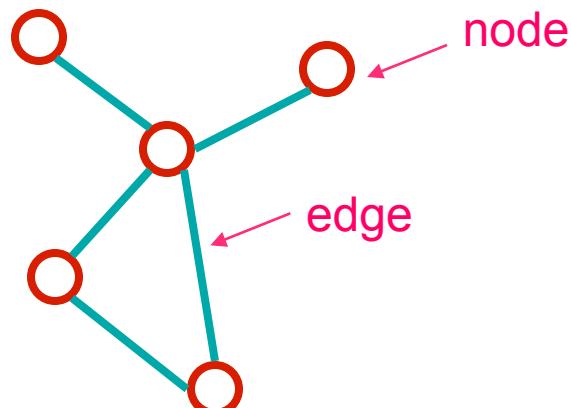


Ingredient networks



What are networks?

- Networks are sets of nodes connected by edges.



“Network” ≡ “Graph”

points	lines	
vertices	edges, arcs	math
nodes	links	computer science
sites	bonds	physics
actors	ties, relations	sociology

goal: characterize network structure

- ❑ Are nodes connected through the network? (week 1)
- ❑ How far apart are they? (week 1)
- ❑ Are some nodes more important due to their position in the network? (week 3)
- ❑ Is the network composed of communities? (week 4)

goal: model network formation

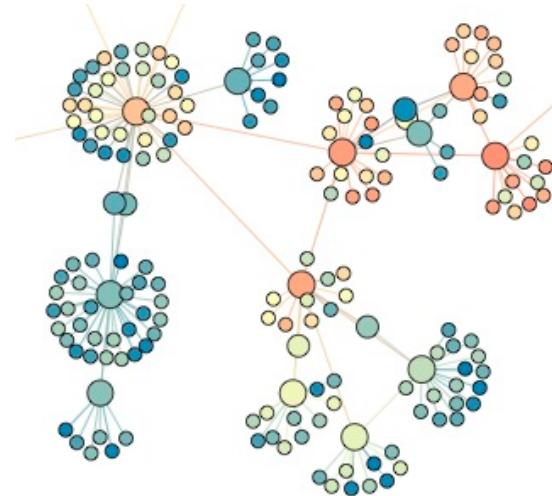
- ❑ Randomly generated networks (week 2)
- ❑ Preferential attachment (week 2)
- ❑ Small-world networks (week 5)
- ❑ Optimization, strategic network formation (week 5)

goal: understand how network structure affects processes

- ❑ information diffusion (weeks 2 & 6)
- ❑ opinion formation (week 6)
- ❑ coordination/cooperation (week 6)
- ❑ resilience to attack (week 2)

What about weeks 7 & 8?

- ❑ Week 7: cool and unusual applications of SNA
- ❑ Week 8: SNA and online social networks



SNA L1B: software tools

Lada Adamic



In this class

- ❑ Gephi (visualization and basic network metrics)
- ❑ NetLogo (modeling network dynamics)
- ❑ iGraph (for programming assignments)

Alternatives (User friendly)

- ❑ Pajek

- ❑ <http://pajek.imfm.si/doku.php>
- ❑ very extensive functionality via drop-down menus
- ❑ free
- ❑ Windows-only



Alternatives (User friendly)

❑ UCINet

- ❑ extensive, sociology-focused functionality
- ❑ Windows-only
- ❑ costs \$\$

Alternatives (User-friendly)

- ❑ NodeXL <http://nodexl.codeplex.com/>
 - ❑ SNA integrated into Excel
 - ❑ Windows-only
 - ❑ free
 - ❑ beta



Alternatives (Python)

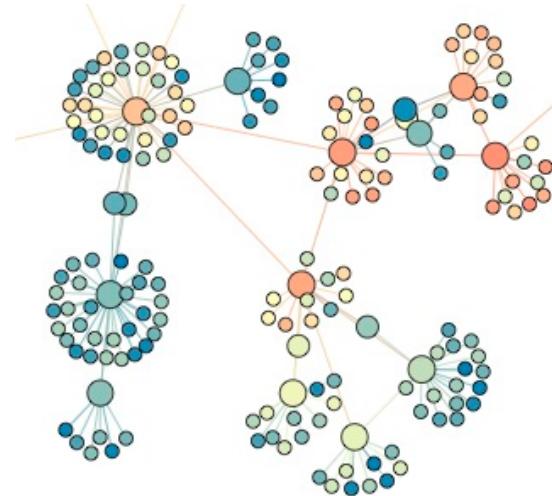
- ❑ NetworkX
 - ❑ extensive functionality
 - ❑ scales to large networks by taking advantage of existing C, Fortran libraries for large matrix computations
 - ❑ open source
 - ❑ <http://networkx.lanl.gov/>

Alternatives (Specialized)

- ❑ sna package for R
 - ❑ extensive, statistics-heavy functionality
 - ❑ <http://cran.r-project.org/web/packages/sna/index.html>
- ❑ SoNIA - Social Network Image Animator
 - ❑ <http://www.stanford.edu/group/sonia/>
 - ❑ specialized for longitudinal analysis of networks

Gephi

- ❑ Download from:
 - ❑ <http://gephi.org/>
 - ❑ download the datafile dining.gephi from Coursera
 - ❑ let's play



SNA L1C: degree, connected components

Lada Adamic



Network elements: edges

- ❑ Directed (also called arcs, links)
 - ❑ A -> B
 - ❑ A likes B, A gave a gift to B, A is B's child
- ❑ Undirected
 - ❑ A <-> B or A – B
 - ❑ A and B like each other
 - ❑ A and B are siblings
 - ❑ A and B are co-authors

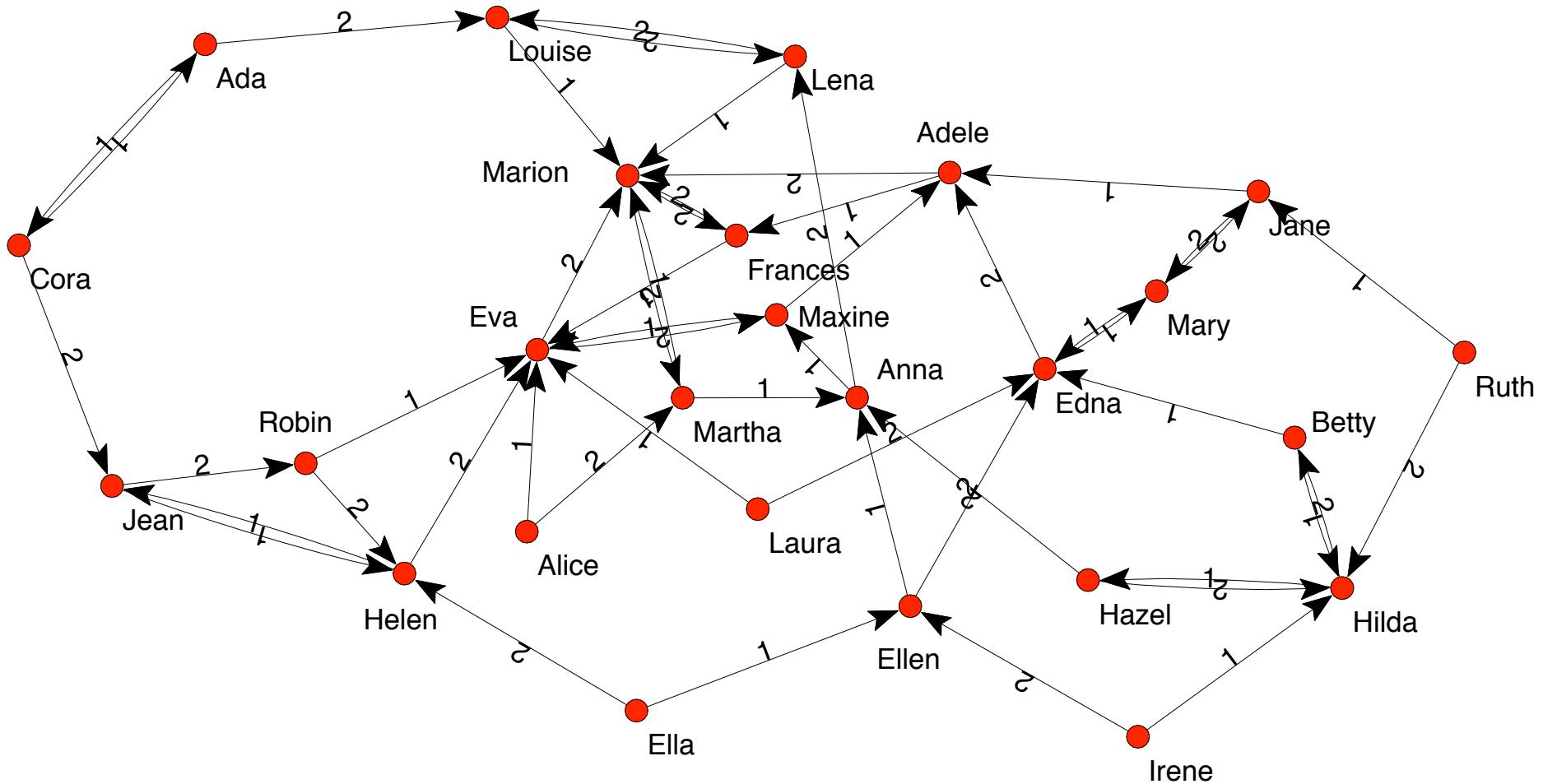
Edge attributes

■ Examples

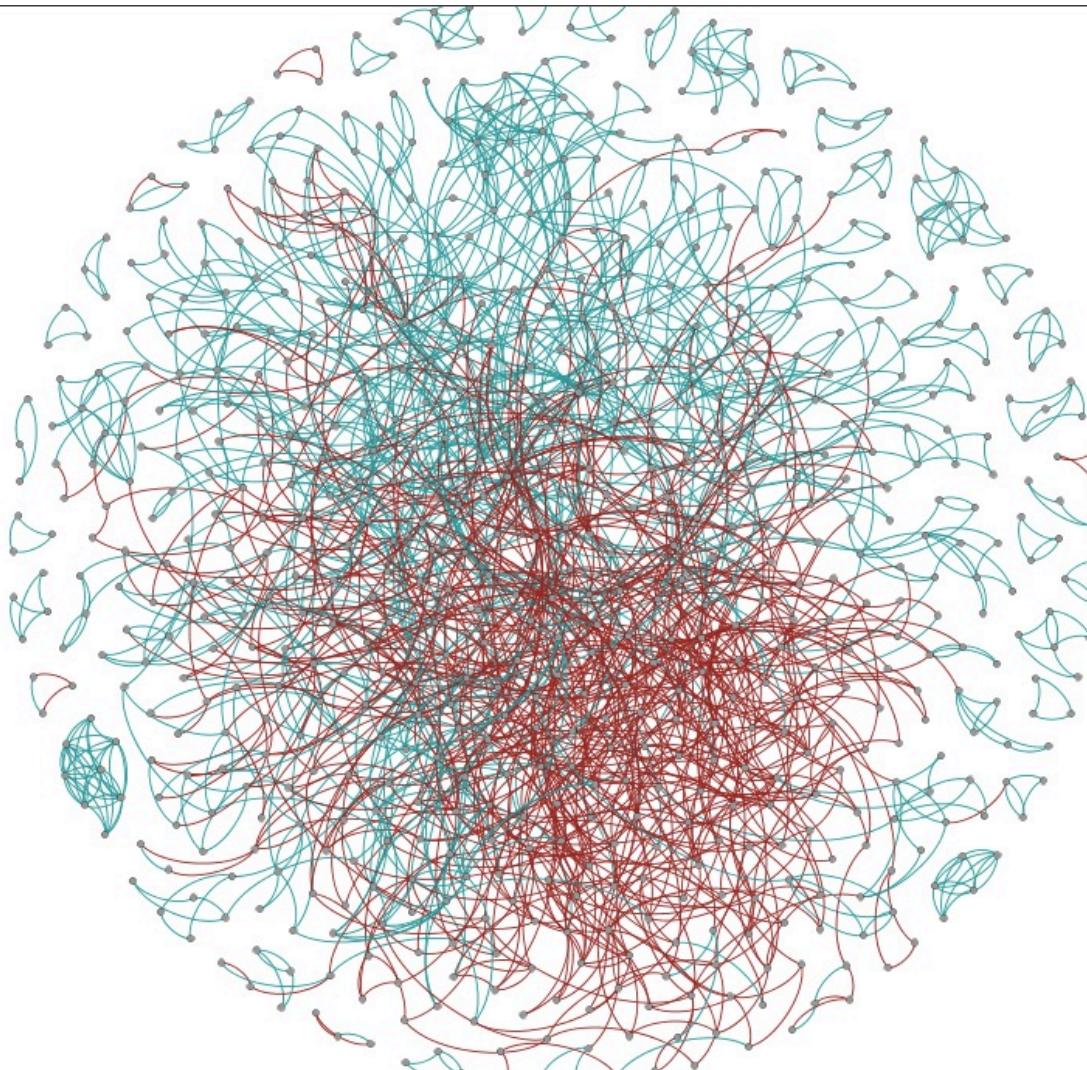
- weight (e.g. frequency of communication)
- ranking (best friend, second best friend...)
- type (friend, relative, co-worker)
- properties depending on the structure of the rest of the graph: e.g. betweenness

Directed networks

- girls' school dormitory dining-table partners, 1st and 2nd choices
(Moreno, *The sociometry reader*, 1960)



Positive and negative weights



sample of positive & negative ratings from *Epinions* network

- ❑ e.g. one person trusting/distrusting another
- ❑ Research challenge: How does one ‘propagate’ negative feelings in a social network? Is my enemy’s enemy my friend?

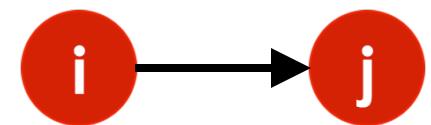
Data representation

- ❑ adjacency matrix
- ❑ edgelist
- ❑ adjacency list

Adjacency matrices

- Representing edges (who is adjacent to whom) as a matrix

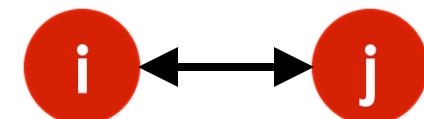
- $A_{ij} = 1$ if node j has an edge to node i
 $= 0$ if node j does not have an edge to i



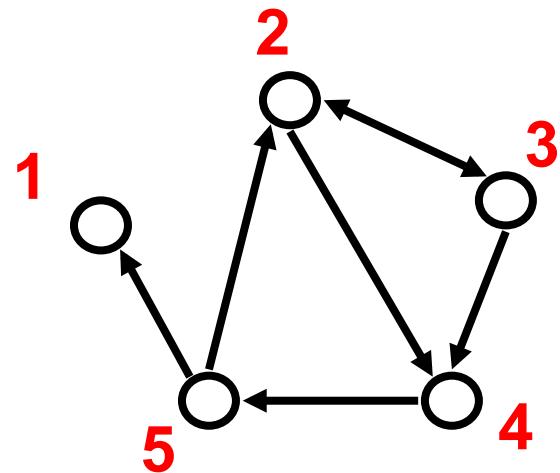
- $A_{ii} = 0$ unless the network has self-loops



- $A_{ij} = A_{ji}$ if the network is undirected,
or if i and j share a reciprocated edge



Example adjacency matrix

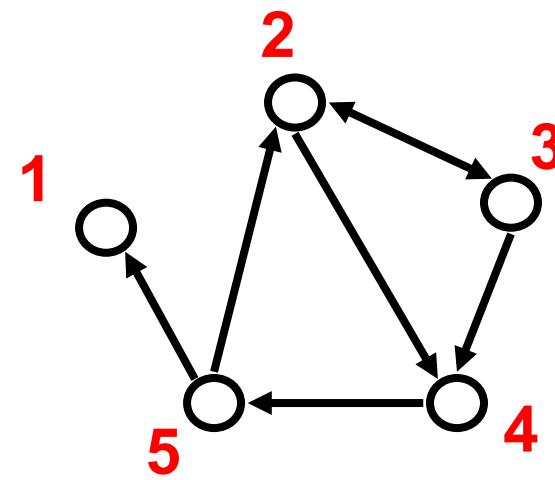


$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Edge list

- Edge list

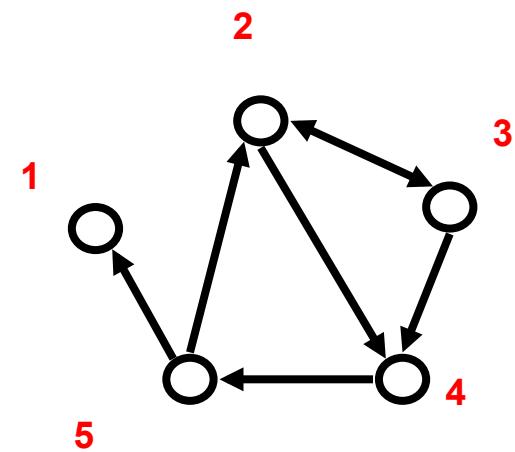
- 2, 3
- 2, 4
- 3, 2
- 3, 4
- 4, 5
- 5, 2
- 5, 1



Adjacency lists

❑ Adjacency list

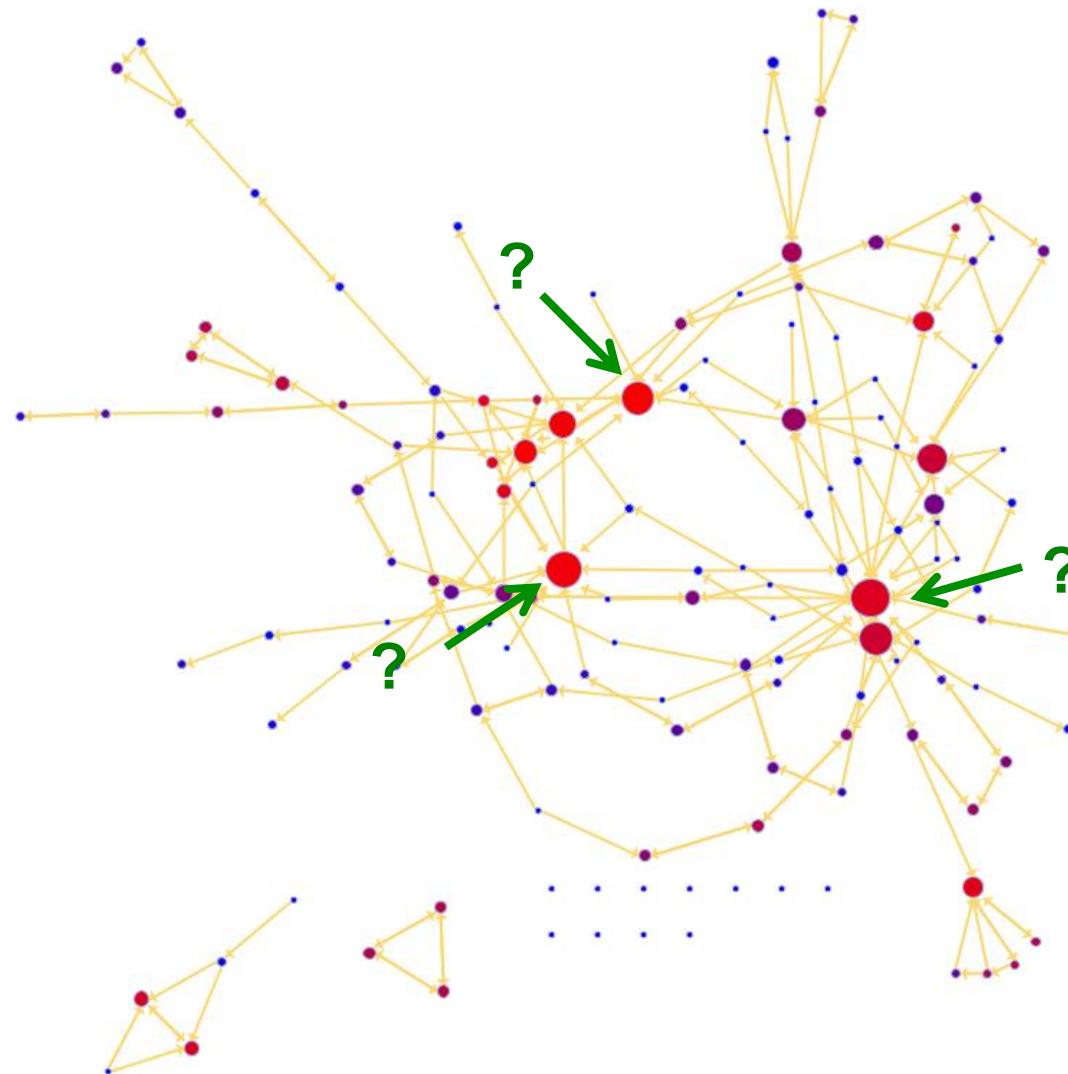
- ❑ is easier to work with if network is
 - ❑ large
 - ❑ sparse
- ❑ quickly retrieve all neighbors for a node
 - ❑ 1:
 - ❑ 2: 3 4
 - ❑ 3: 2 4
 - ❑ 4: 5
 - ❑ 5: 1 2



Computing metrics

- ❑ degree & degree distribution
- ❑ connected components

Degree: which node has the most edges?

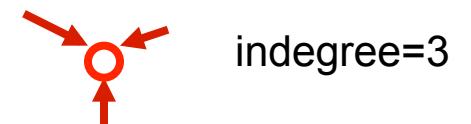


Nodes

- Node network properties
 - from immediate connections

- indegree

how many directed edges (arcs) are incident on a node



- outdegree

how many directed edges (arcs) originate at a node

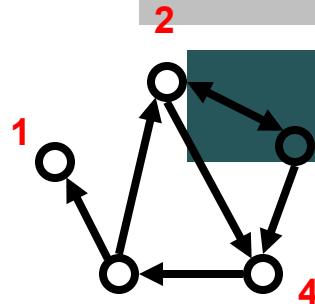


- degree (in or out)

number of edges incident on a node



- from the entire graph
 - centrality (betweenness, closeness)



Node degree from matrix values

■ Indegree = $\sum_{j=1}^n A_{ij}$

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

example: indegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd row

$$\sum_{j=1}^n A_{3j}$$

■ Outdegree = $\sum_{i=1}^n A_{ij}$

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

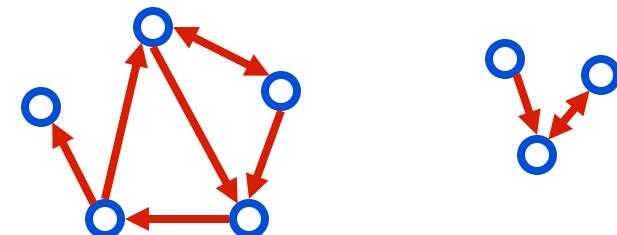
example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd column

$$\sum_{i=1}^n A_{i3}$$

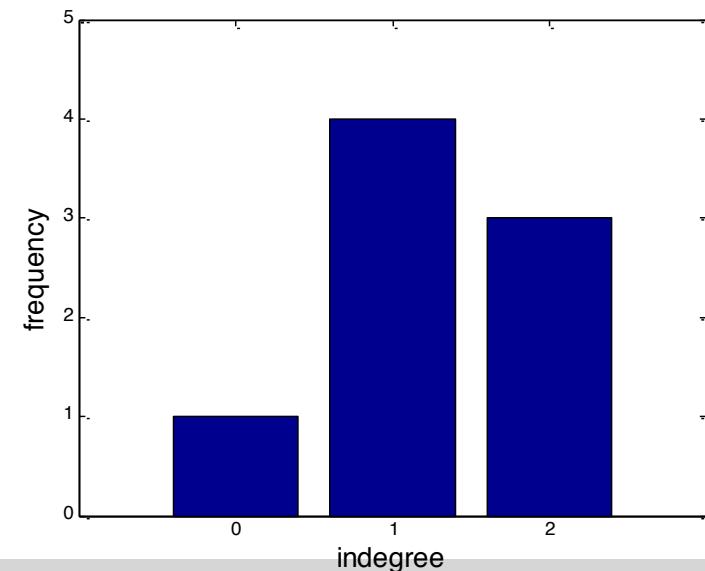
Network metrics: degree sequence and degree distribution

- Degree sequence: An ordered list of the (in,out) degree of each node

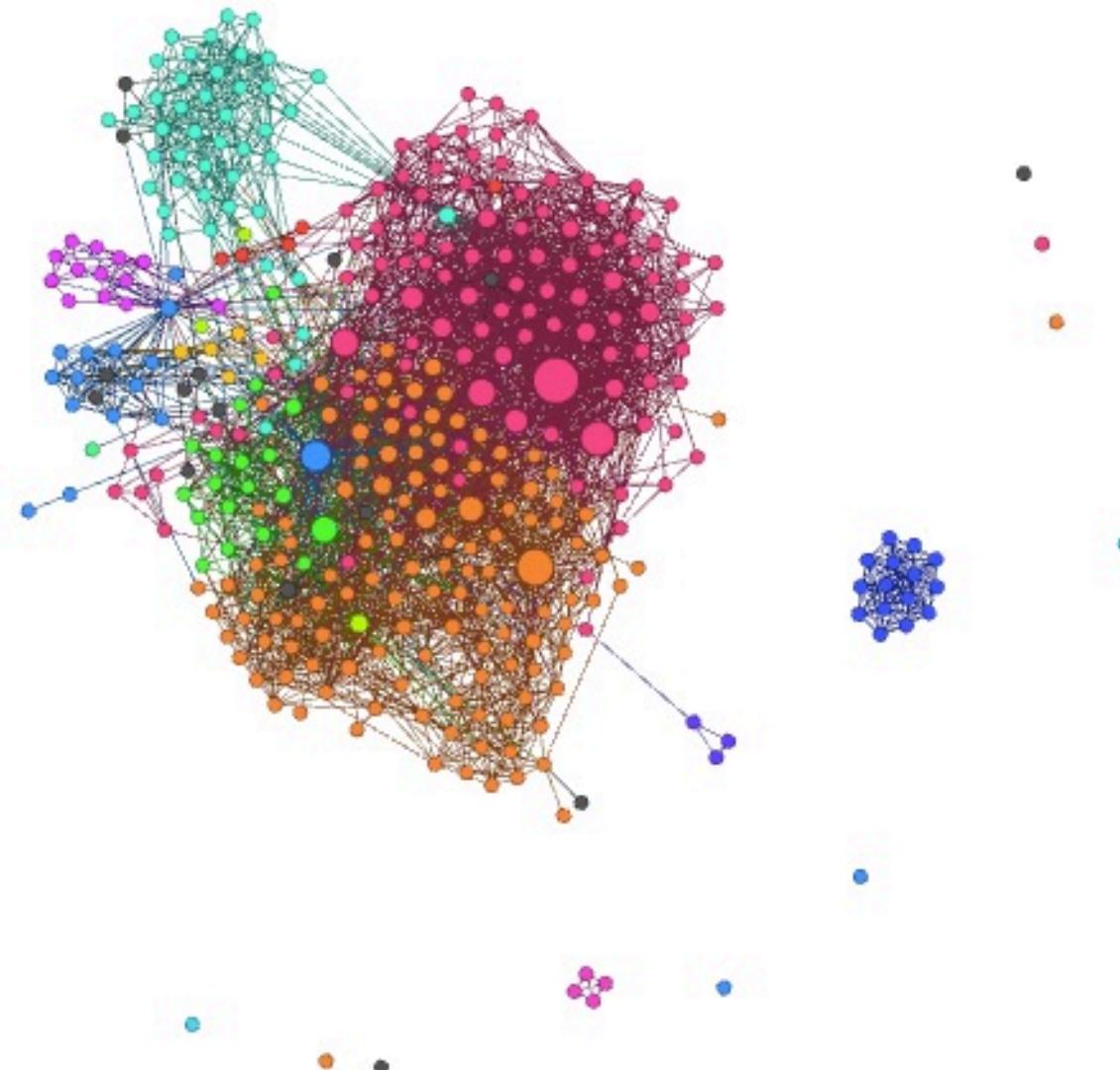
- In-degree sequence:
 - [2, 2, 2, 1, 1, 1, 1, 0]
- Out-degree sequence:
 - [2, 2, 2, 2, 1, 1, 1, 0]
- (undirected) degree sequence:
 - [3, 3, 3, 2, 2, 1, 1, 1]



- Degree distribution: A frequency count of the occurrence of each degree
 - In-degree distribution:
 - [(2,3) (1,4) (0,1)]
 - Out-degree distribution:
 - [(2,4) (1,3) (0,1)]
 - (undirected) distribution:
 - [(3,3) (2,2) (1,3)]



Is everything connected?



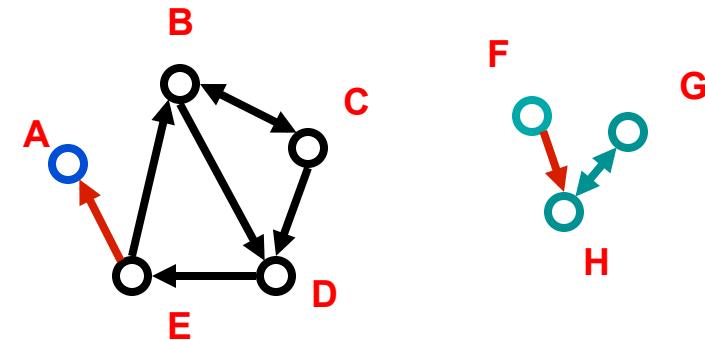
Connected components

- Strongly connected components

- Each node within the component can be reached from every other node in the component by following directed links

- Strongly connected components

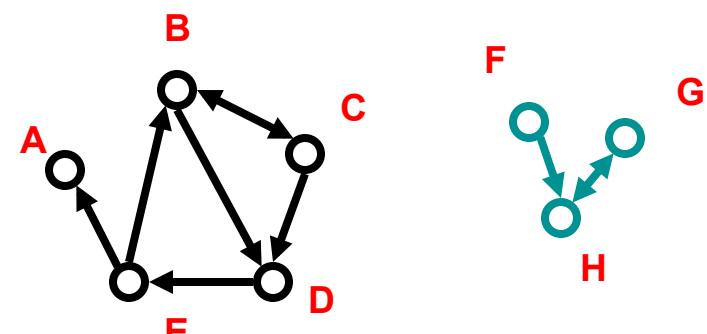
- B C D E
 - A
 - G H
 - F



- Weakly connected components: every node can be reached from every other node by following links in either direction

- Weakly connected components

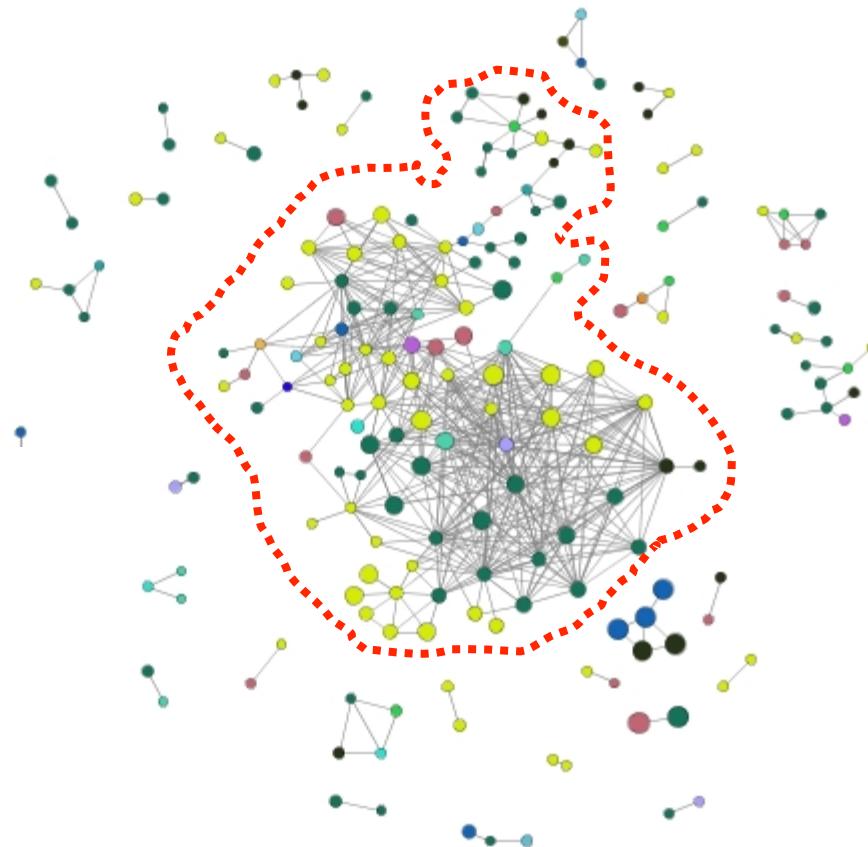
- A B C D E
 - G H F



- In undirected networks one talks simply about 'connected components'

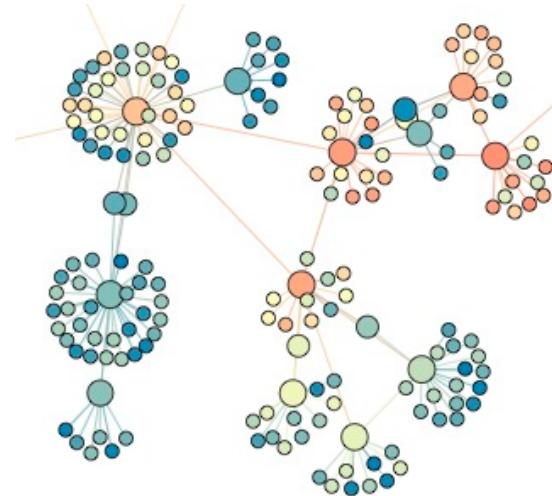
Giant component

- if the largest component encompasses a significant fraction of the graph, it is called the **giant component**



Recap

- ❑ Networks can be represented as matrices
- ❑ Useful network metrics:
 - ❑ degree and degree distribution
 - ❑ connected components
 - ❑ strong
 - ❑ weak
 - ❑ giant



SNA 2A: Intro to Random Graphs

Lada Adamic



Network models

- ❑ Why model?
 - ❑ simple representation of complex network
 - ❑ can derive properties mathematically
 - ❑ predict properties and outcomes

- ❑ Also: to have a strawman
 - ❑ In what ways is your real-world network different from hypothesized model?
 - ❑ What insights can be gleaned from this?

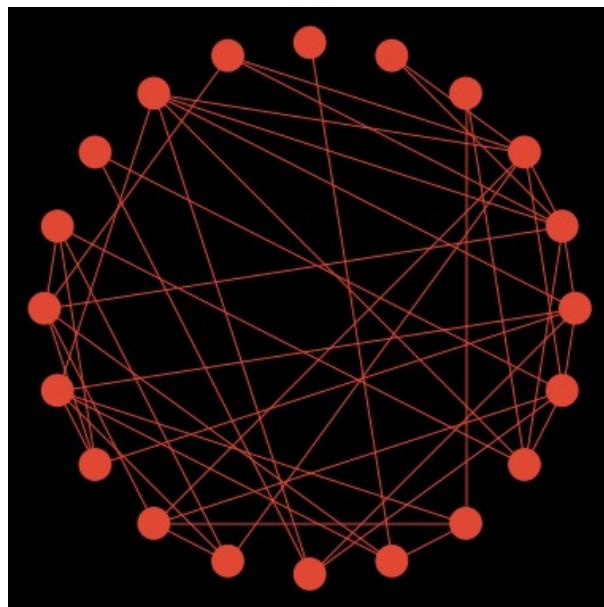
Erdős and Rényi



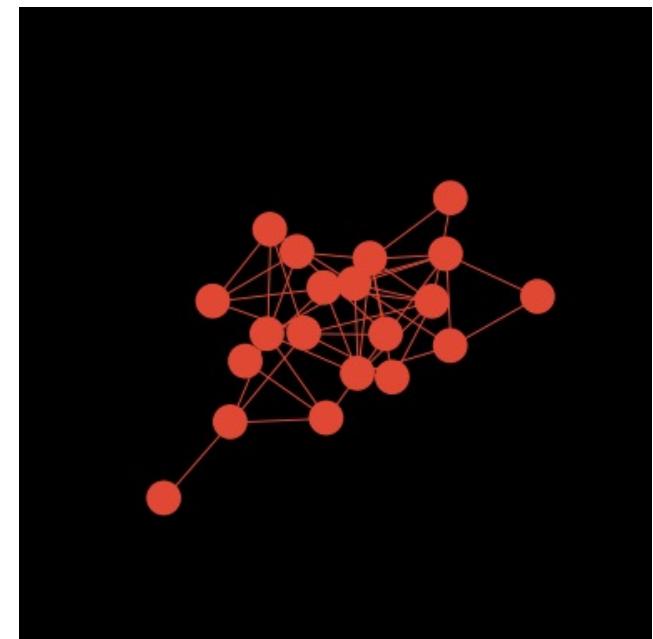
Erdös-Renyi: simplest network model

- Assumptions
 - nodes connect at random
 - network is undirected
- Key parameter (besides number of nodes N) : p or M
 - p = probability that any two nodes share an edge
 - M = total number of edges in the graph

what they look like



after spring
layout



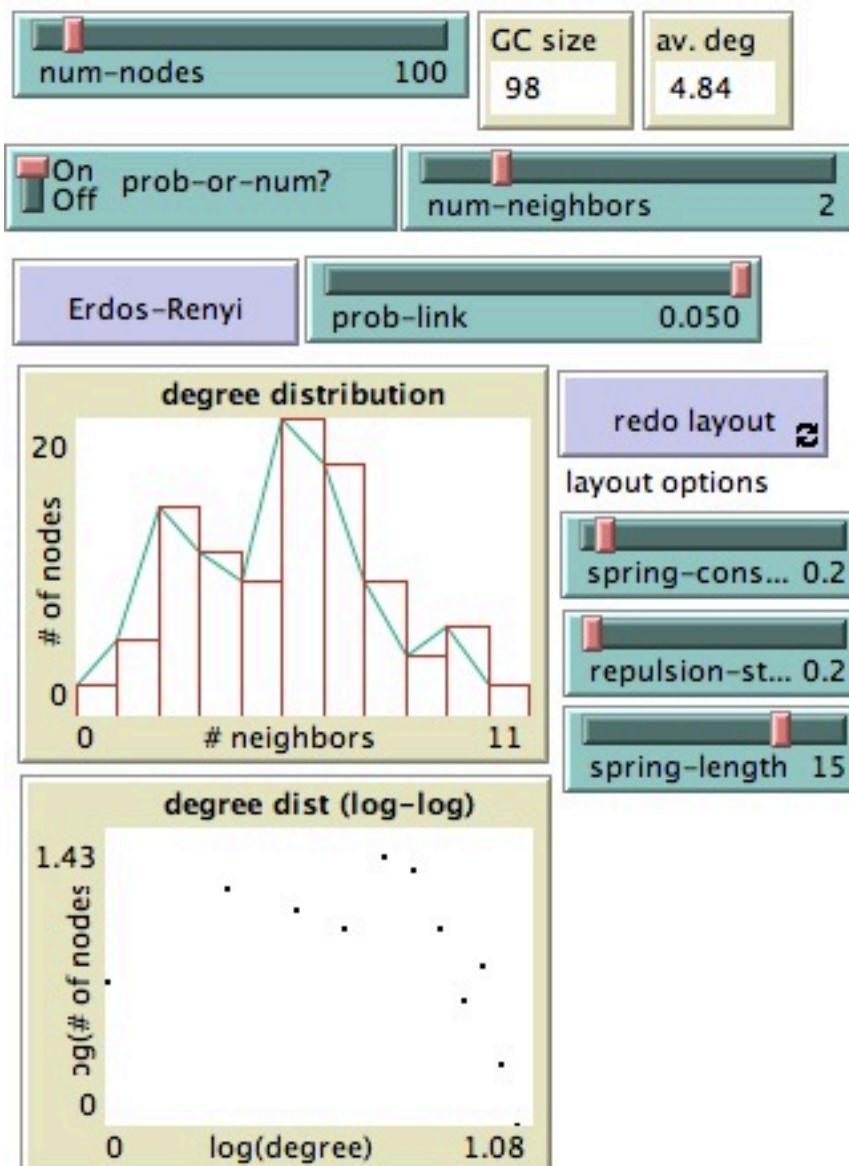
Degree distribution

- (N,p)-model: For each potential edge we flip a biased coin
 - with probability p we add the edge
 - with probability $(1-p)$ we don't

Quiz Q:

- ❑ As the size of the network increases, if you keep p , the probability of any two nodes being connected, the same, what happens to the average degree
 - ❑ a) stays the same
 - ❑ b) increases
 - ❑ c) decreases

<http://ladamic.com/netlearn/NetLogo501/ErdosRenyiDegDist.html>



<http://www.ladamic.com/netlearn/NetLogo501/ErdosRenyiDegDist.html>

Degree distribution

- ❑ What is the probability that a node has 0,1,2,3... edges?
- ❑ Probabilities sum to 1

How many edges per node?

- ❑ Each node has $(N - 1)$ tries to get edges
- ❑ Each try is a success with probability p
- ❑ The binomial distribution gives us the probability that a node has degree k :

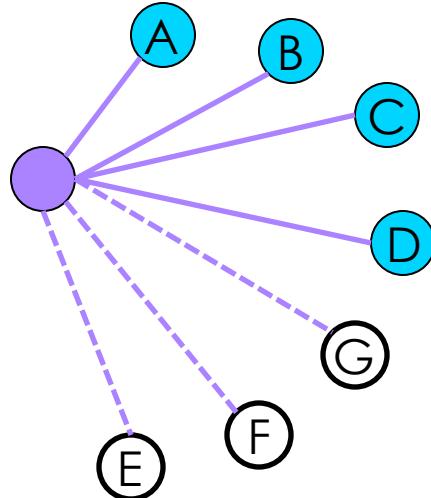
$$B(N - 1; k; p) = \binom{N - 1}{k} p^k (1 - p)^{N-1-k}$$

Quiz Q:

- ❑ The maximum degree of a node in a simple (no multiple edges between the same two nodes) N node graph is
 - ❑ a) N
 - ❑ b) $N - 1$
 - ❑ c) $N / 2$

Explaining the binomial distribution

- ❑ 8 node graph, probability p of any two nodes sharing an edge
- ❑ What is the probability that a given node has degree 4?



Binomial coefficient: choosing 4 out of 7

Suppose I have 7 blue and white nodes, each of them uniquely marked so that I can distinguish them. The blue nodes are ones I share an edge with, the white ones I don't.



How many different samples can I draw containing the same nodes but in a different order (the order could be e.g. the order in which the edges are added (or not)? e.g.



binomial coefficient explained

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

If order matters, there are **7!** different orderings:

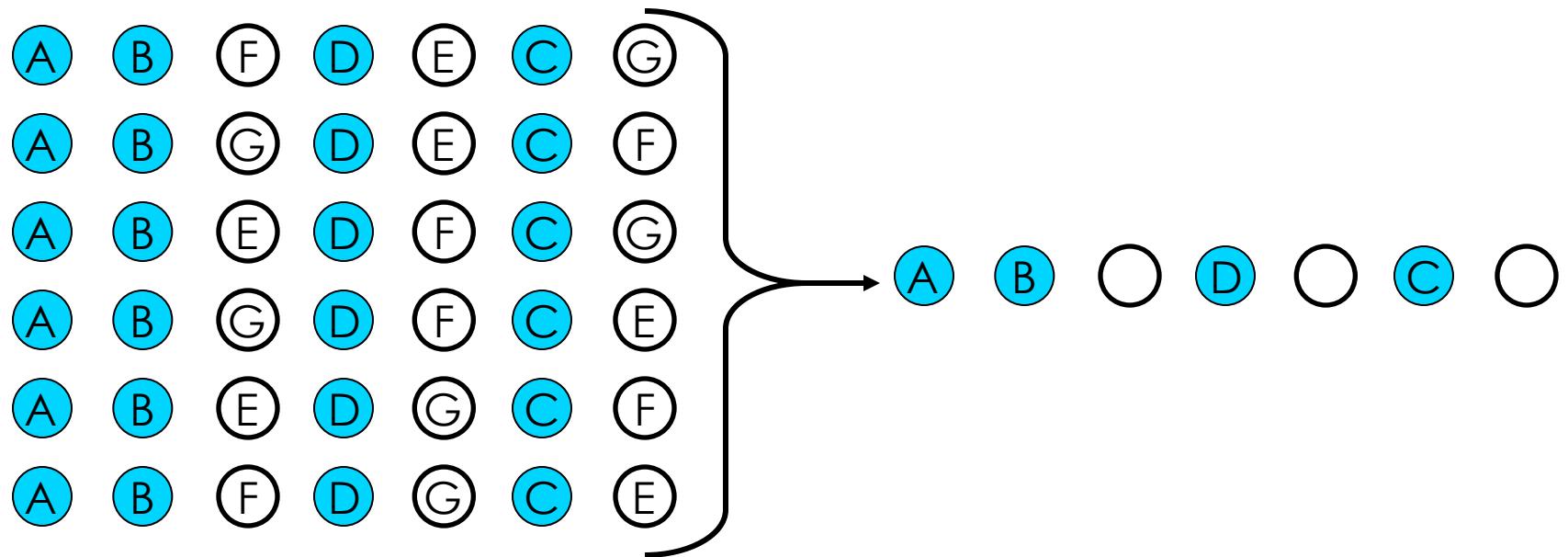
I have 7 choices for the first spot, 6 choices for the second (since I've picked 1 and now have only 6 to choose from),
5 choices for the third, etc.

$$7! = 7 * 6 * 5 * 4 * 3 * 2 * 1$$

binomial coefficient

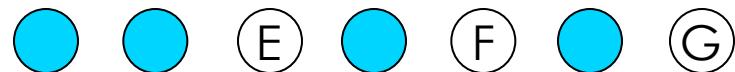
Suppose the order of the nodes I don't connect to (white) doesn't matter.

All possible arrangements ($3!$) of white nodes look the same to me.



Instead of $7!$ combinations, we have $7!/3!$ combinations

binomial coefficient explained



The same goes for the blue nodes, if we can't tell them apart, we lose a factor of 4!

binomial coefficient explained

number of ways of choosing k items out of $(n-1)$

$$= \frac{\text{number of ways of arranging } n-1 \text{ items}}{(\# \text{ of ways to arrange } k \text{ things}) * (\# \text{ ways to arrange } n-1-k \text{ things})}$$

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

Note that the binomial coefficient is symmetric – there are the same number of ways of choosing k or $n-1-k$ things out of $n-1$

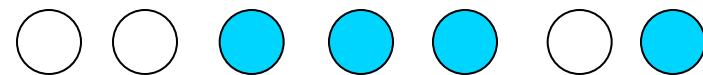
Quiz Q:

- ❑ What is the number of ways of choosing 2 items out of 5?
 - ❑ 10
 - ❑ 120
 - ❑ 6
 - ❑ 5

Now the distribution

- ❑ p = probability of having edge to node (blue)
- ❑ $(1-p)$ = probability of not having edge (white)
- ❑ The probability that you connect to 4 of the 7 nodes in some particular order (two white followed by 3 blues, followed by a white followed by a blue) is

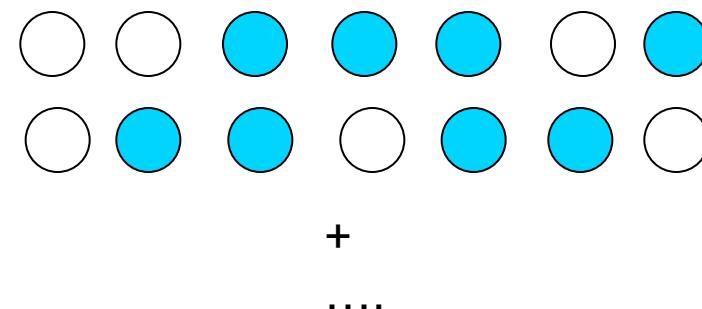
$$\begin{aligned} & P(\text{white}) * P(\text{white}) * P(\text{blue}) * P(\text{blue}) * P(\text{blue}) * P(\text{white}) * P(\text{blue}) \\ & = p^4 * (1-p)^3 \end{aligned}$$



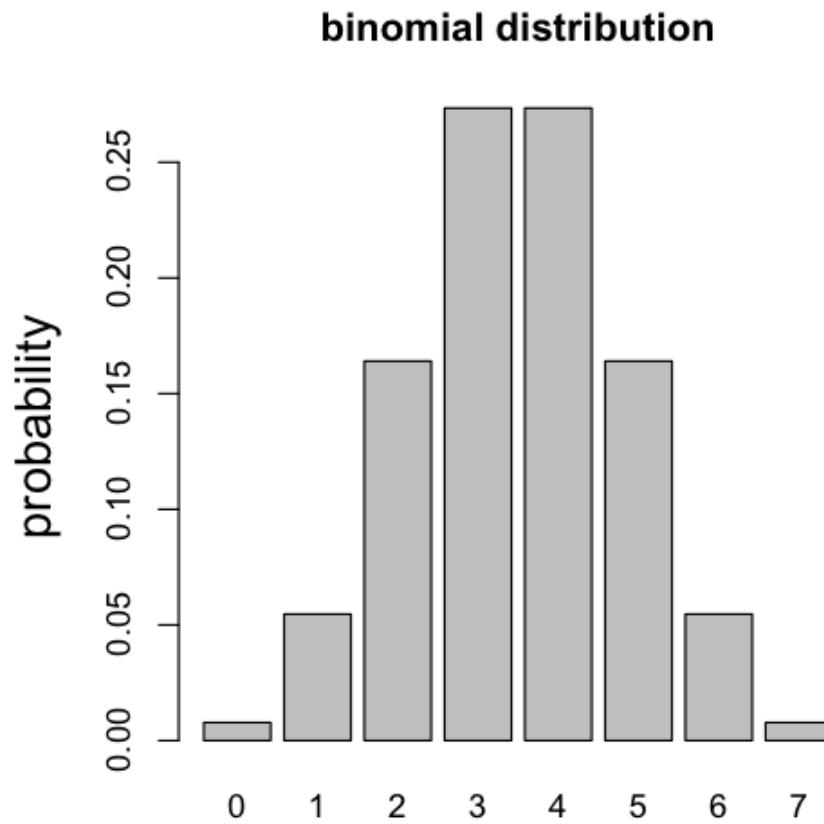
Binomial distribution

- If order doesn't matter, need to multiply probability of any given arrangement by number of such arrangements:

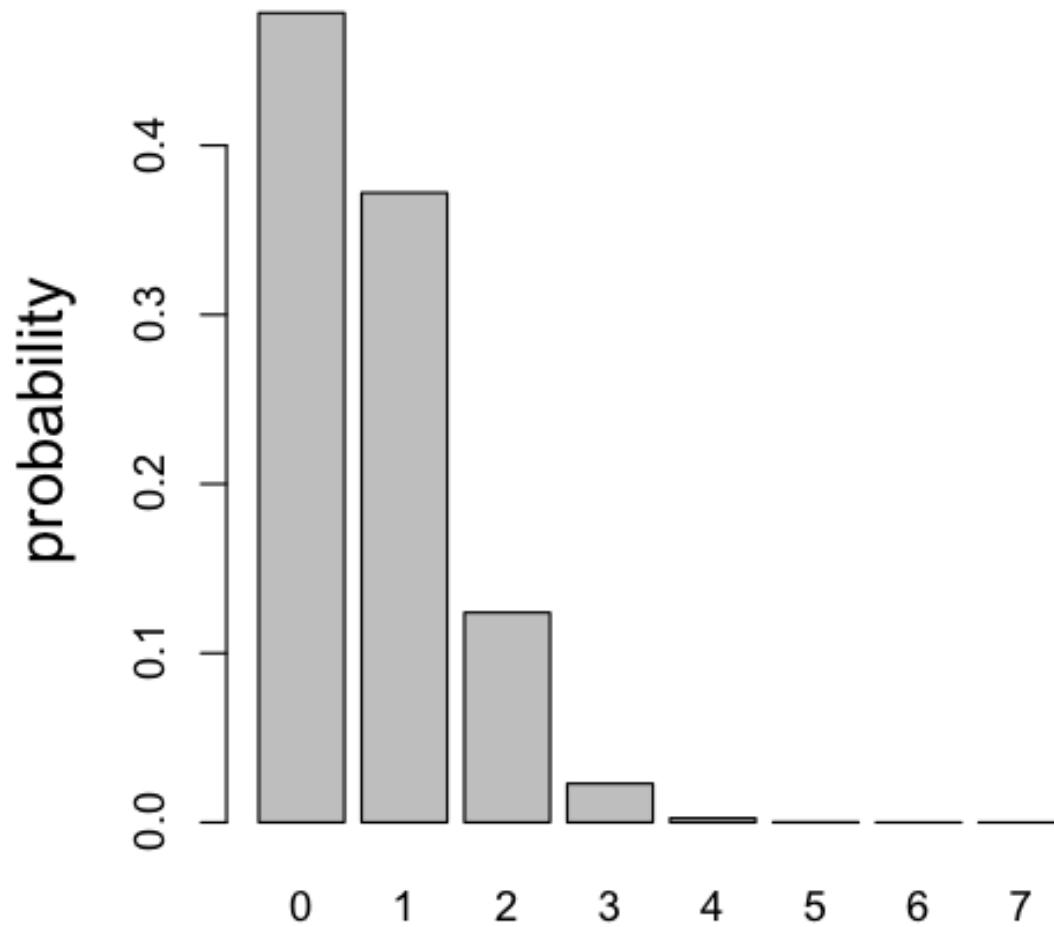
$$B(7; 4; p) = \binom{7}{4} p^4 (1-p)^3$$



if $p = 0.5$



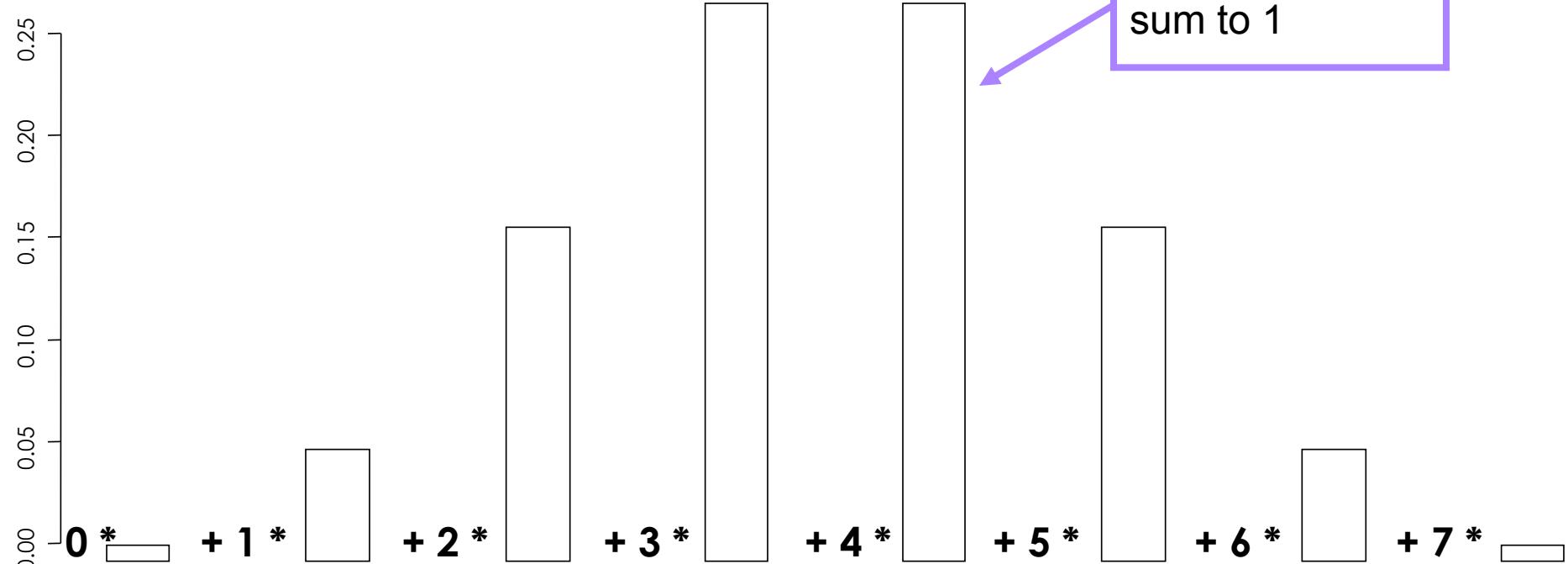
$$\overline{\sigma} = 0.1$$



What is the mean?

❑ Average degree $z = (n-1)*p$

❑ in general $\mu = E(X) = \sum x p(x)$



$$\mu = 3.5$$

Quiz Q:

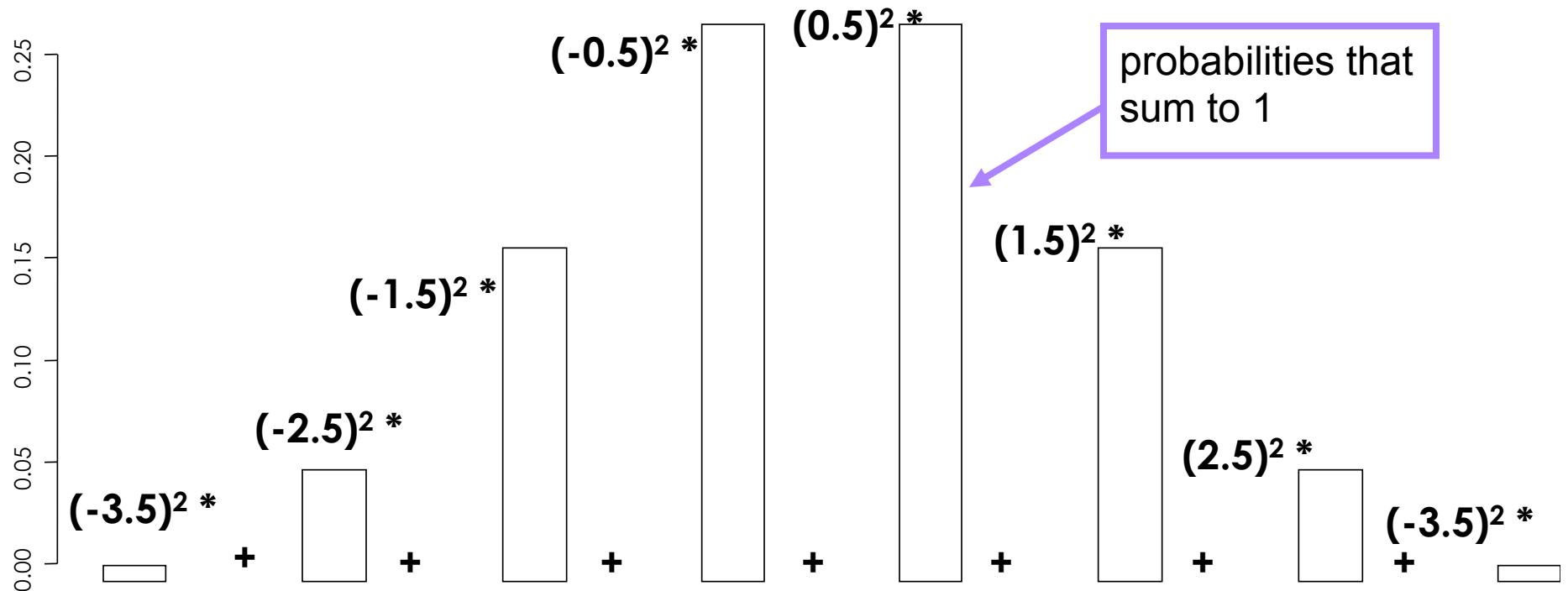
- ❑ What is the average degree of a graph with 10 nodes and probability $p = 1/3$ of an edge existing between any two nodes?
 - ❑ 1
 - ❑ 2
 - ❑ 3
 - ❑ 4

What is the variance?

- ❑ variance in degree

$$\sigma^2 = (n-1) * p * (1-p)$$

- ❑ in general $\sigma^2 = E[(X-\mu)^2] = \sum (x-\mu)^2 p(x)$



Approximations

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

$$p_k = \frac{z^k e^{-z}}{k!}$$

$$p_k = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-z)^2}{2\sigma^2}}$$

Binomial

limit p small

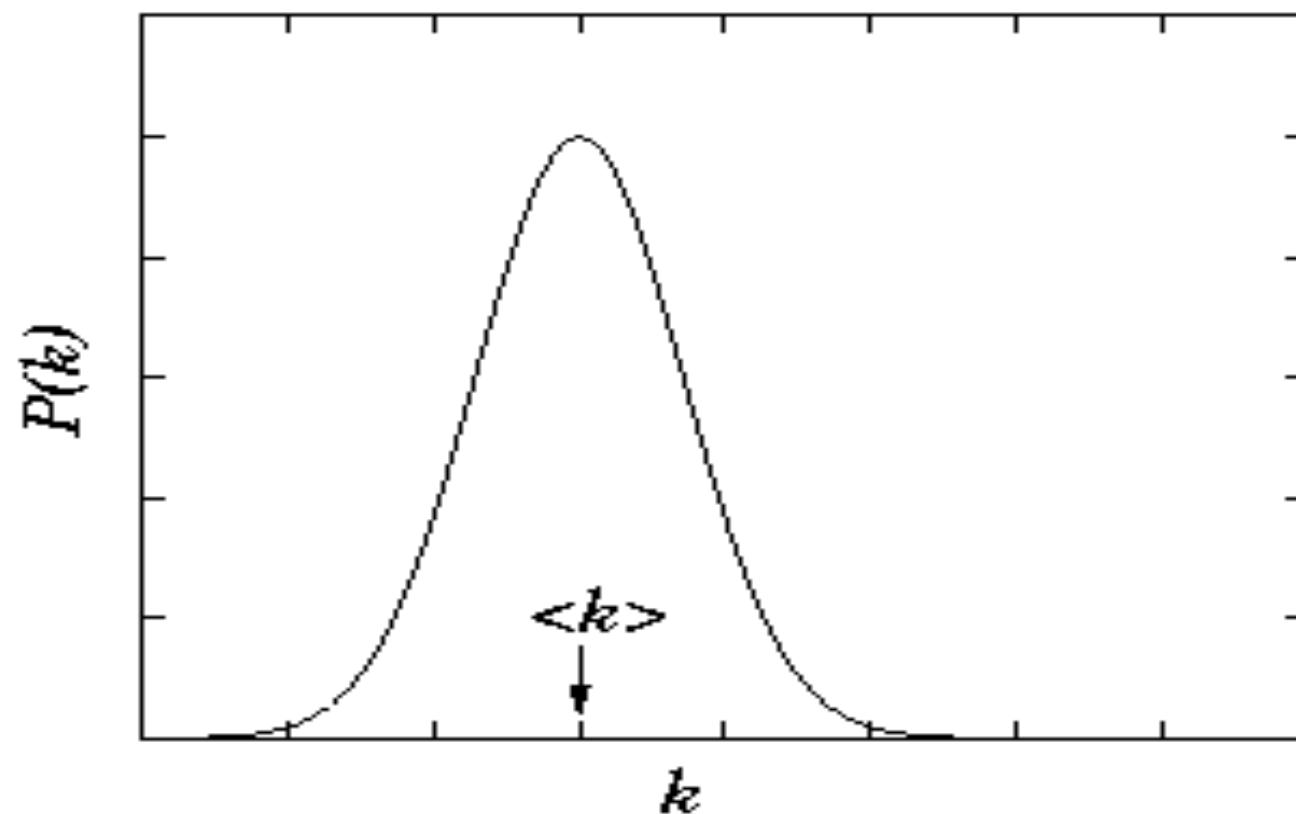
Poisson

limit large n

Normal

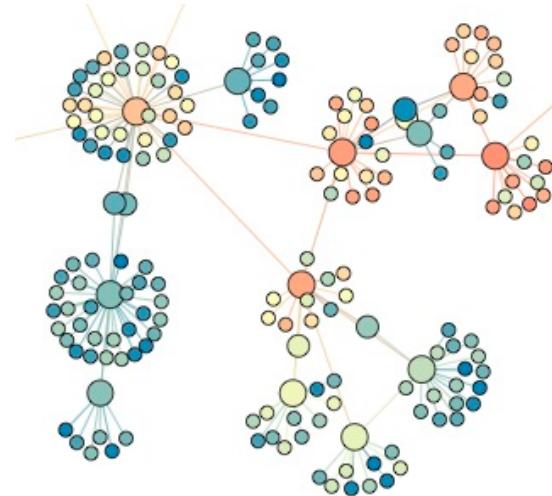
Poisson distribution

Poisson distribution



What insights does this yield? No hubs

- ❑ You don't expect large hubs in the network



SNA 2B: ER graphs: Insights and realism

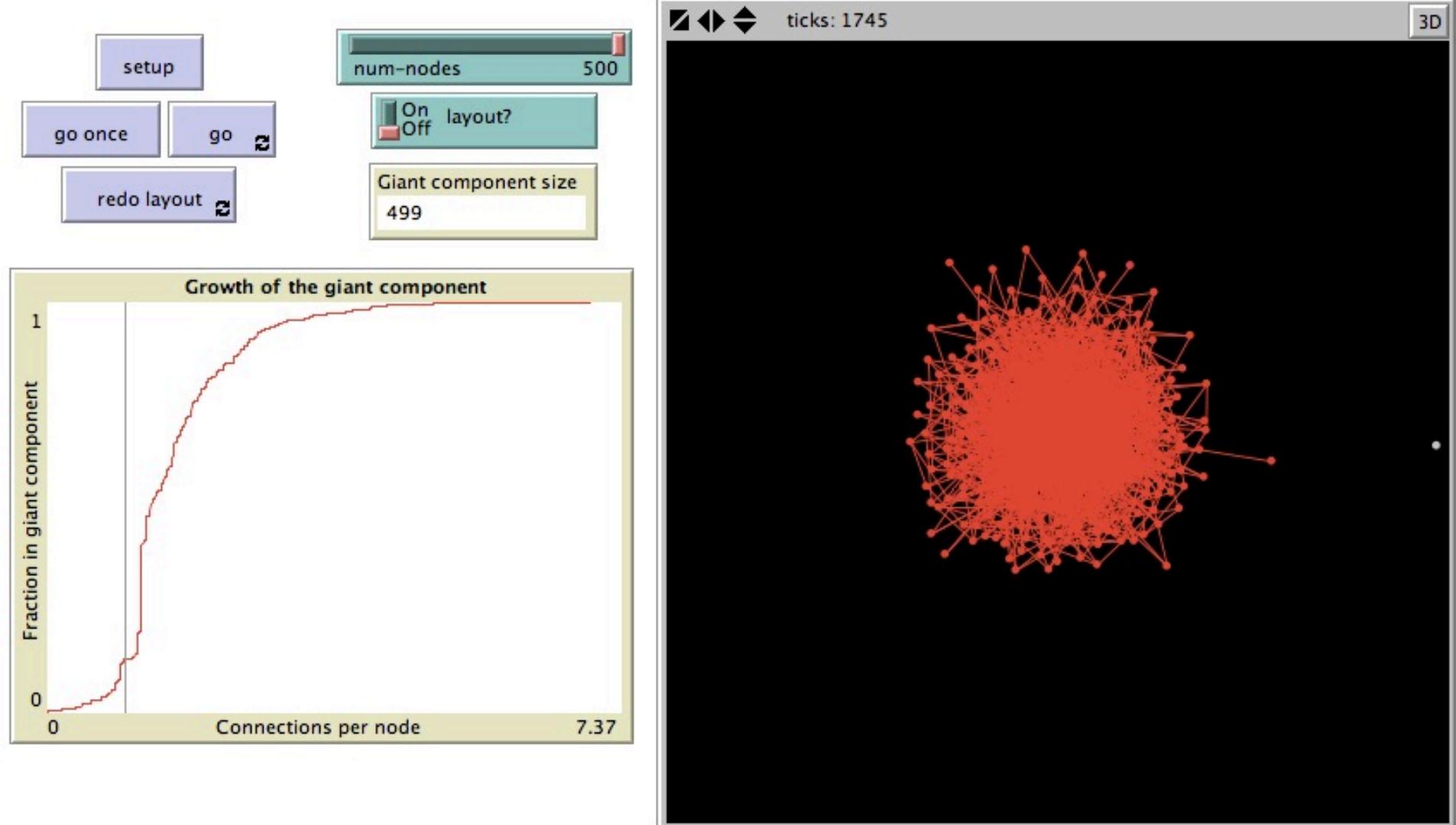
Lada Adamic



Insights

- ❑ Previously: degree distribution / absence of hubs
- ❑ Emergence of giant component
- ❑ Average shortest path

Emergence of the giant component

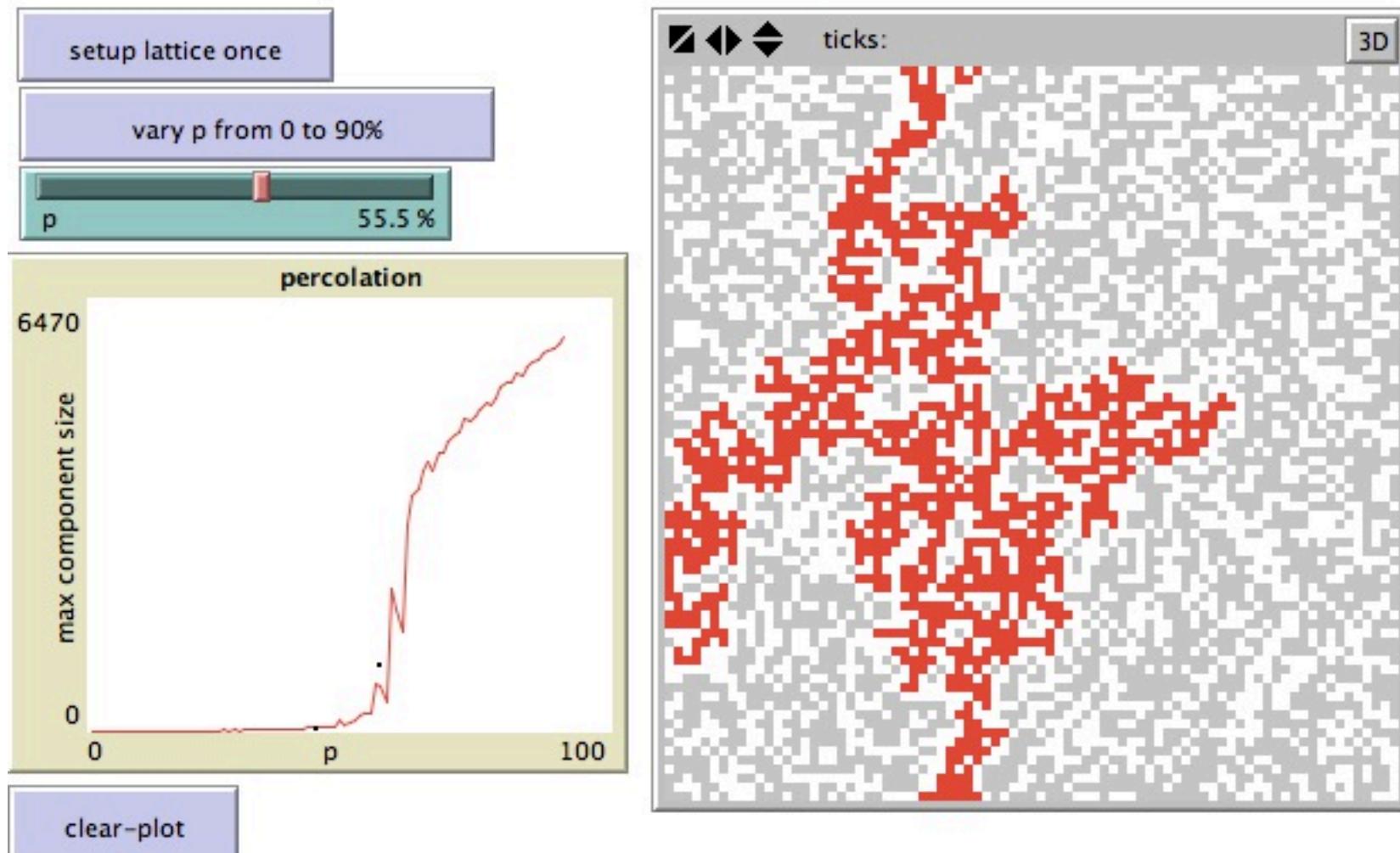


<http://ccl.northwestern.edu/netlogo/models/GiantComponent>

Quiz Q:

- ❑ What is the average degree z at which the giant component starts to emerge?
 - 0
 - 1
 - $3/2$
 - 3

Percolation on a 2D lattice

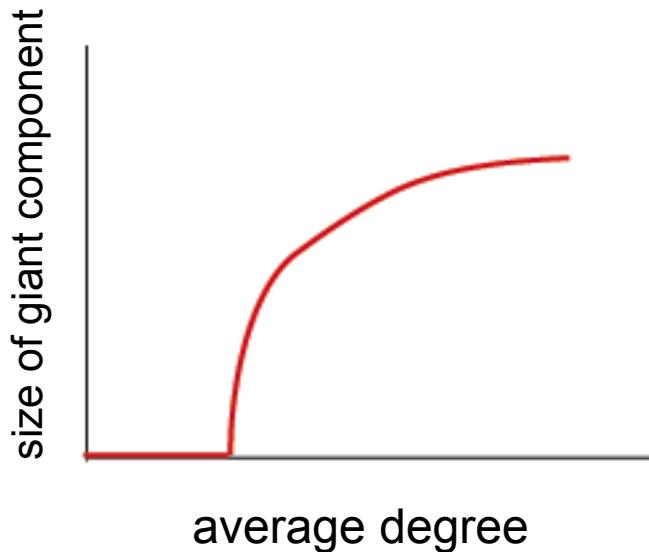


<http://www.ladamic.com/netlearn/NetLogo501/LatticePercolation.html>

Quiz Q:

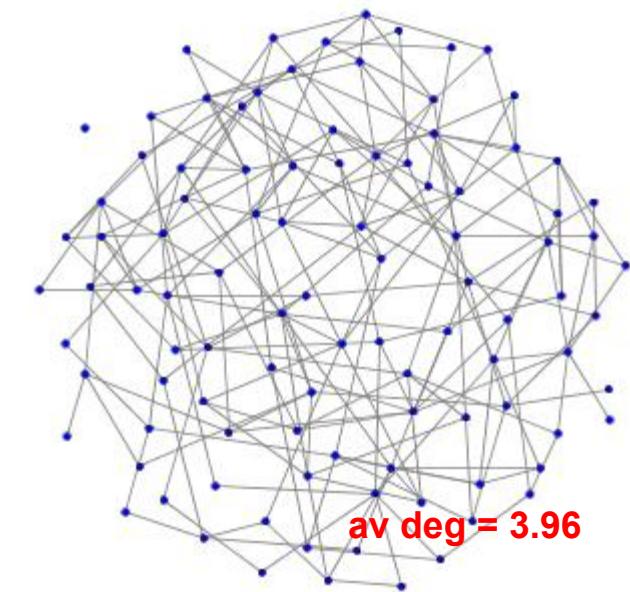
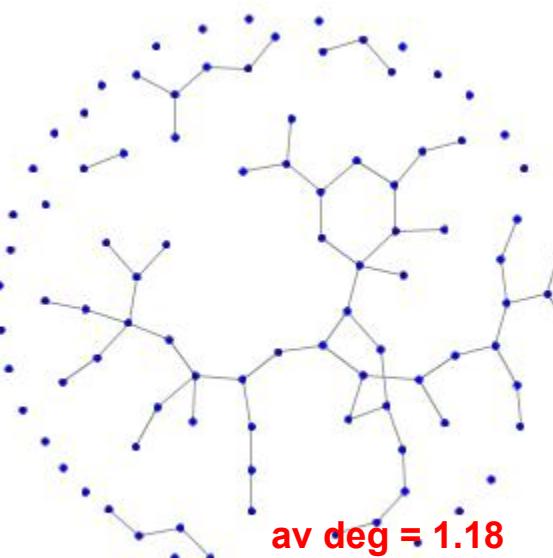
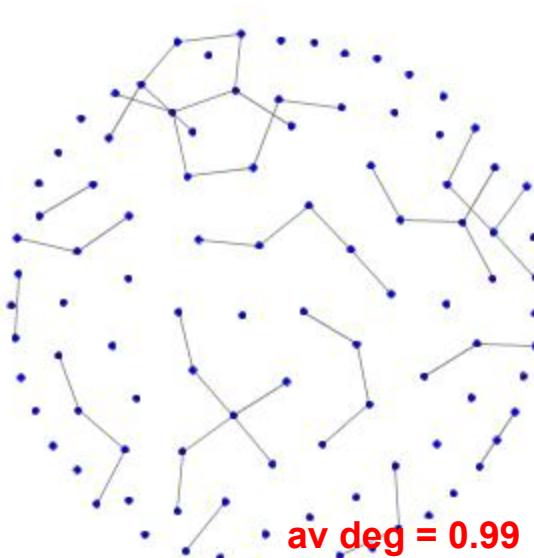
- ❑ What is the percolation threshold of a 2D lattice: fraction of sites that need to be occupied in order for a giant connected component to emerge?
 - 0
 - $\frac{1}{4}$
 - $\frac{1}{3}$
 - $\frac{1}{2}$

Percolation threshold



Percolation threshold: how many edges need to be added before the giant component appears?

As the average degree increases to $z = 1$, a giant component suddenly appears



Giant component – another angle

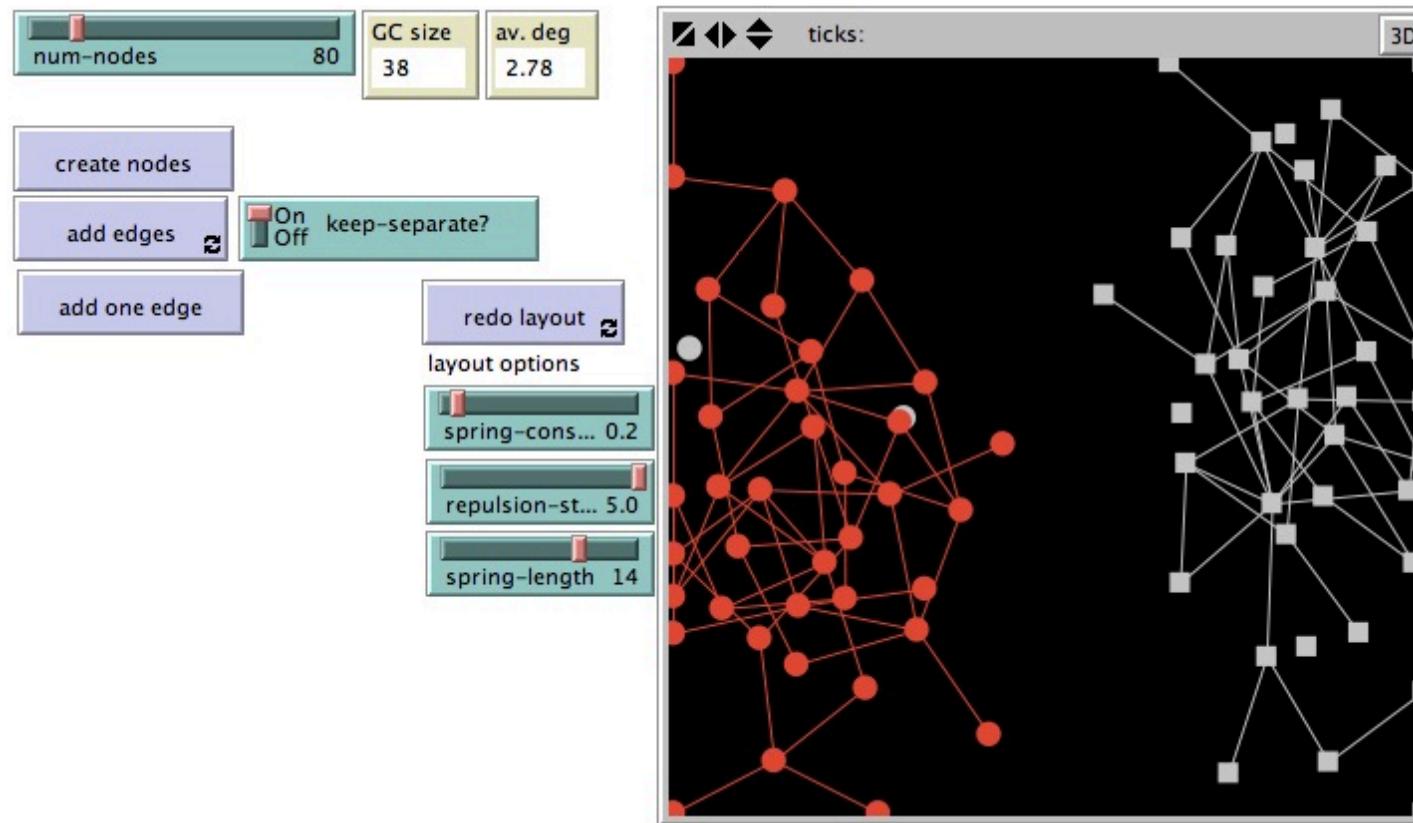
- ❑ How many other friends besides you does each of your friends have?
- ❑ By property of degree distribution
 - ❑ the average degree of your friends, you excluded, is z
 - ❑ so at $z = 1$, each of your friends is expected to have another friend, who in turn have another friend, etc.
 - ❑ the giant component emerges

Giant component illustrated



Why just one giant component?

- ❑ What if you had 2, how long could they be sustained as the network densifies?



<http://www.ladamic.com/netlearn/NetLogo501/ErdosRenyiTwoComponents.html>

Quiz Q:

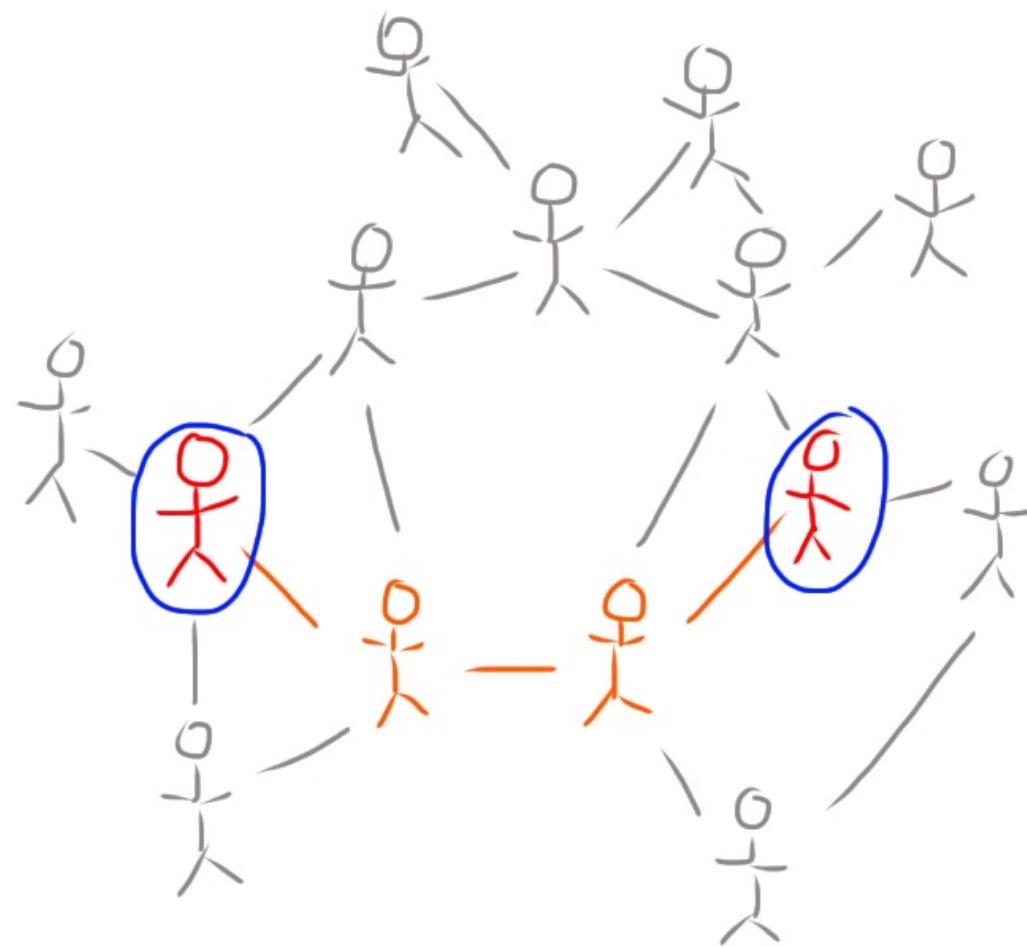
- ❑ If you have 2 large-components each occupying roughly 1/2 of the graph, how long does it typically take for the addition of random edges to join them into one giant component
 - ❑ 1-4 edge additions
 - ❑ 5-20 edge additions
 - ❑ over 20 edge additions

Average shortest path

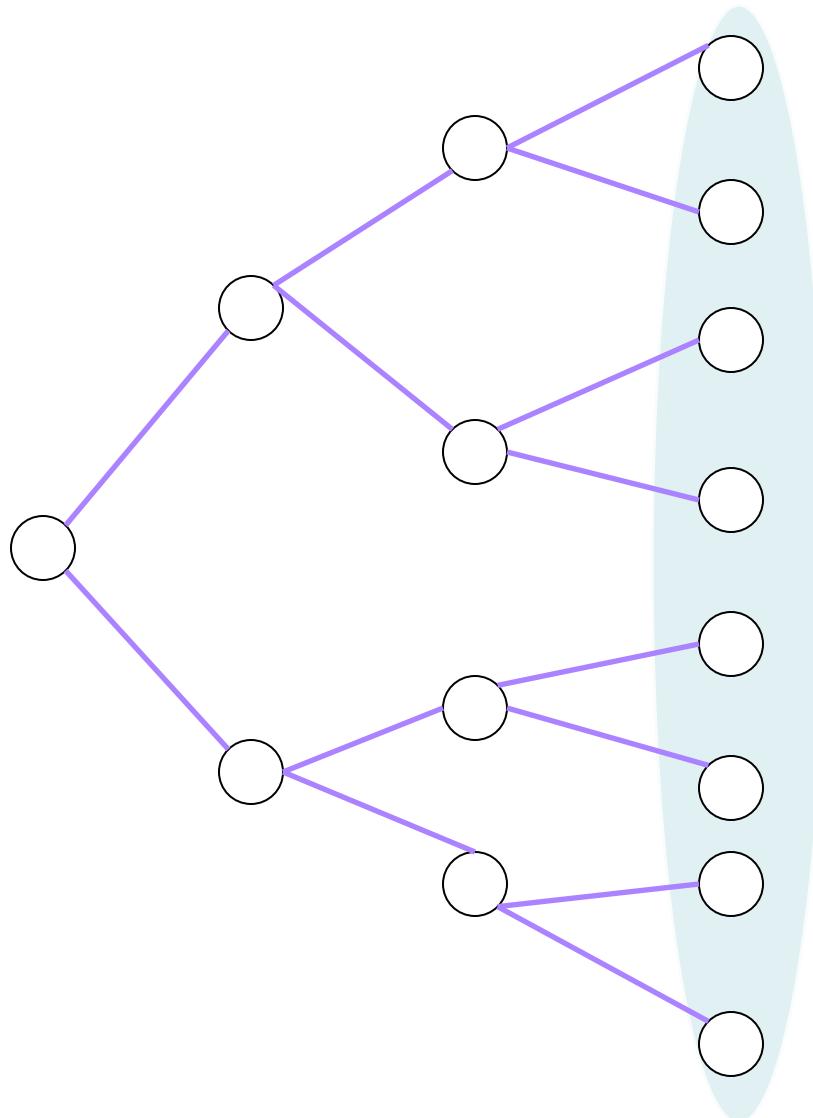
- ❑ How many hops on average between each pair of nodes?
- ❑ again, each of your friends has **$z = \text{avg. degree}$** friends besides you
- ❑ ignoring loops, the number of people you have at distance l is

$$z^l$$

Average shortest path



friends at distance l



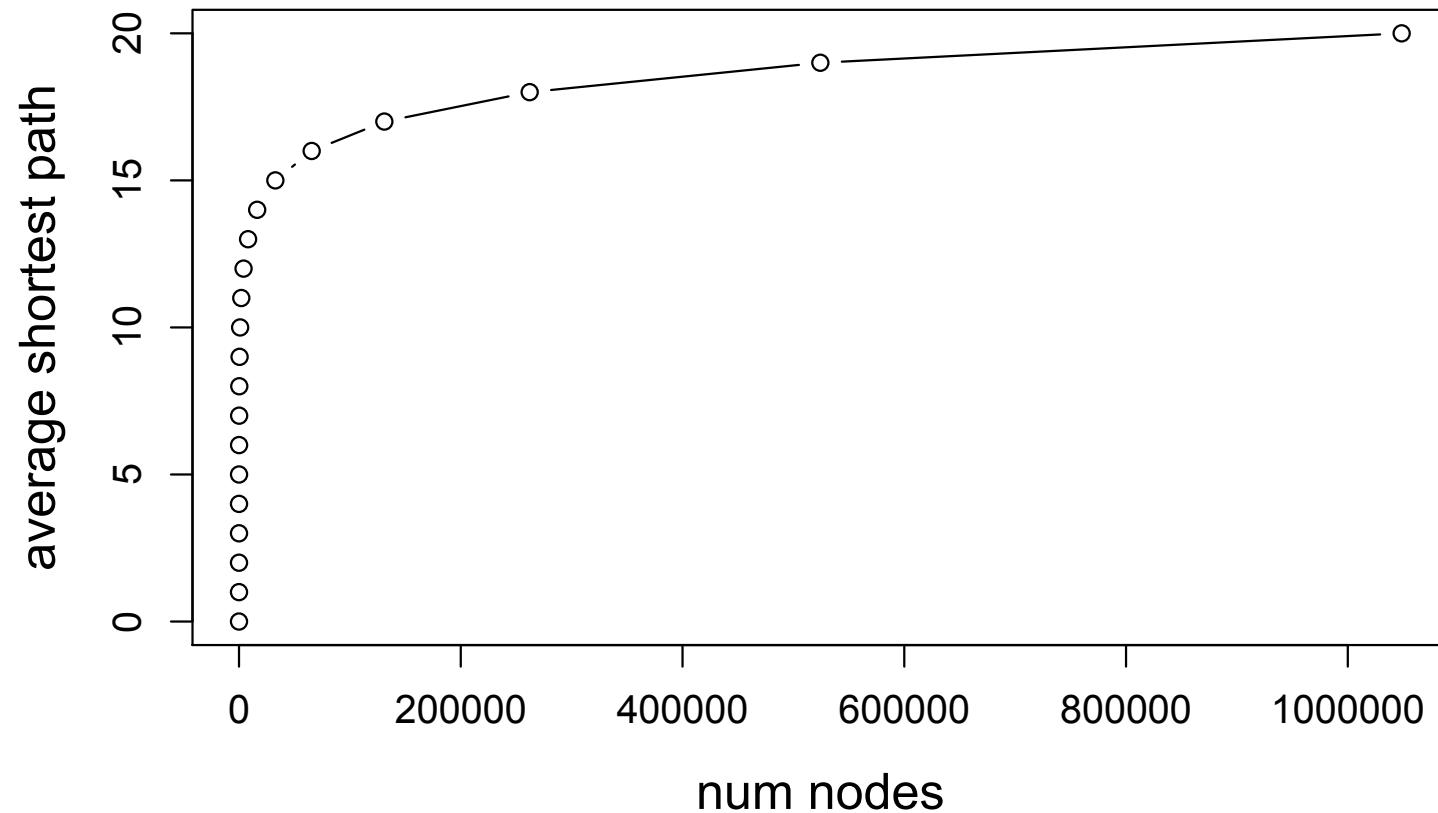
$$N_l = z^l$$

scaling:
average shortest path l_{av}

$$l_{av} \sim \frac{\log N}{\log z}$$

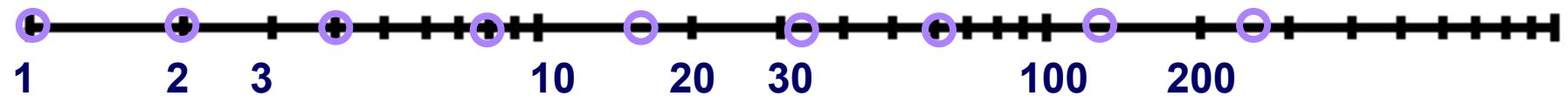
What this means in practice

- Erdös-Renyi networks can grow to be very large but nodes will be just a few hops apart



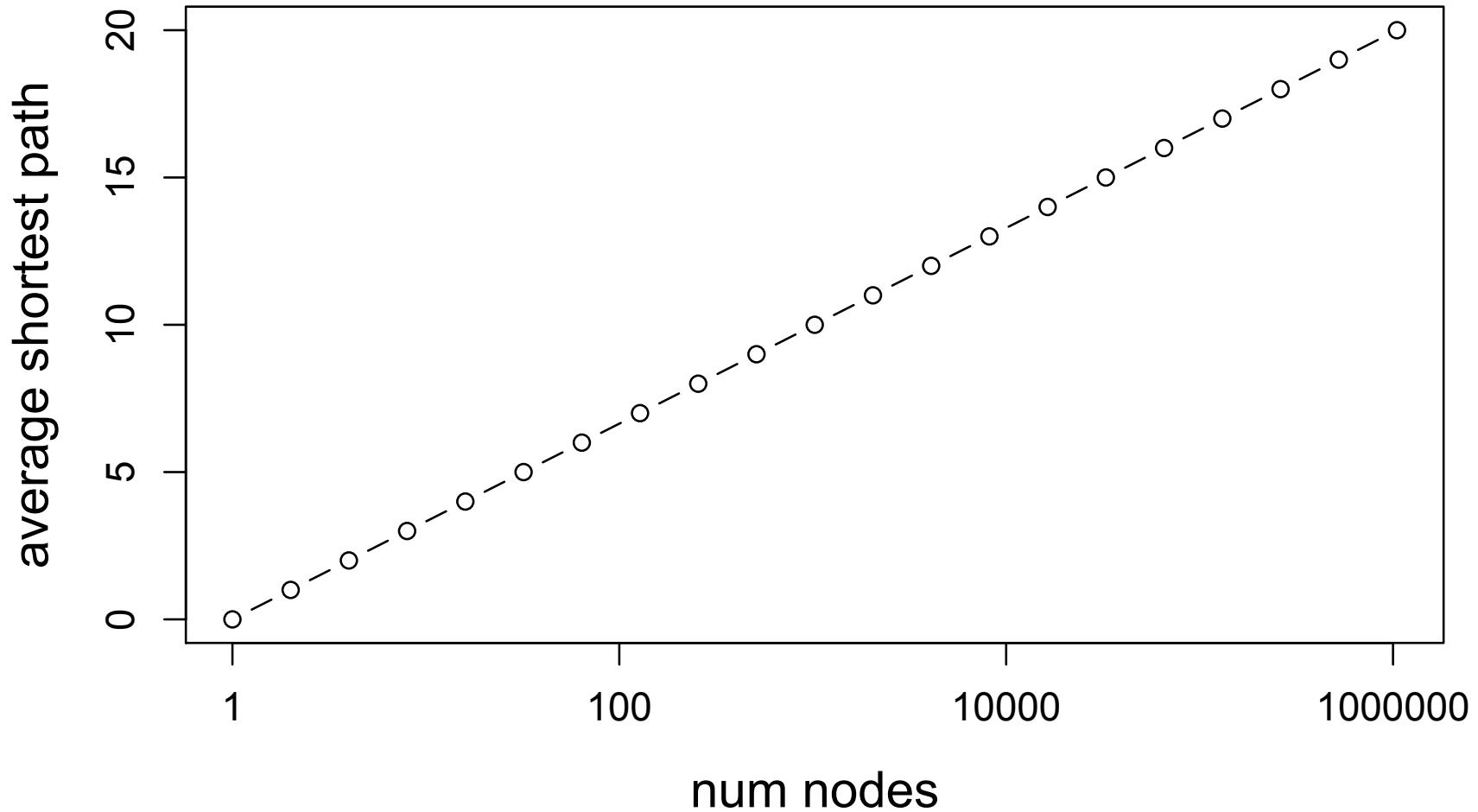
Logarithmic axes

- ☐ powers of a number will be uniformly spaced



- $2^0=1, 2^1=2, 2^2=4, 2^3=8, 2^4=16, 2^5=32, 2^6=64, \dots$

Erdös-Renyi avg. shortest path



Quiz Q:

- ❑ If the size of an Erdös-Renyi network increases 100 fold (e.g. from 100 to 10,000 nodes), how will the average shortest path change
 - ❑ it will be 100 times as long
 - ❑ it will be 10 times as long
 - ❑ it will be twice as long
 - ❑ it will be the same
 - ❑ it will be 1/2 as long

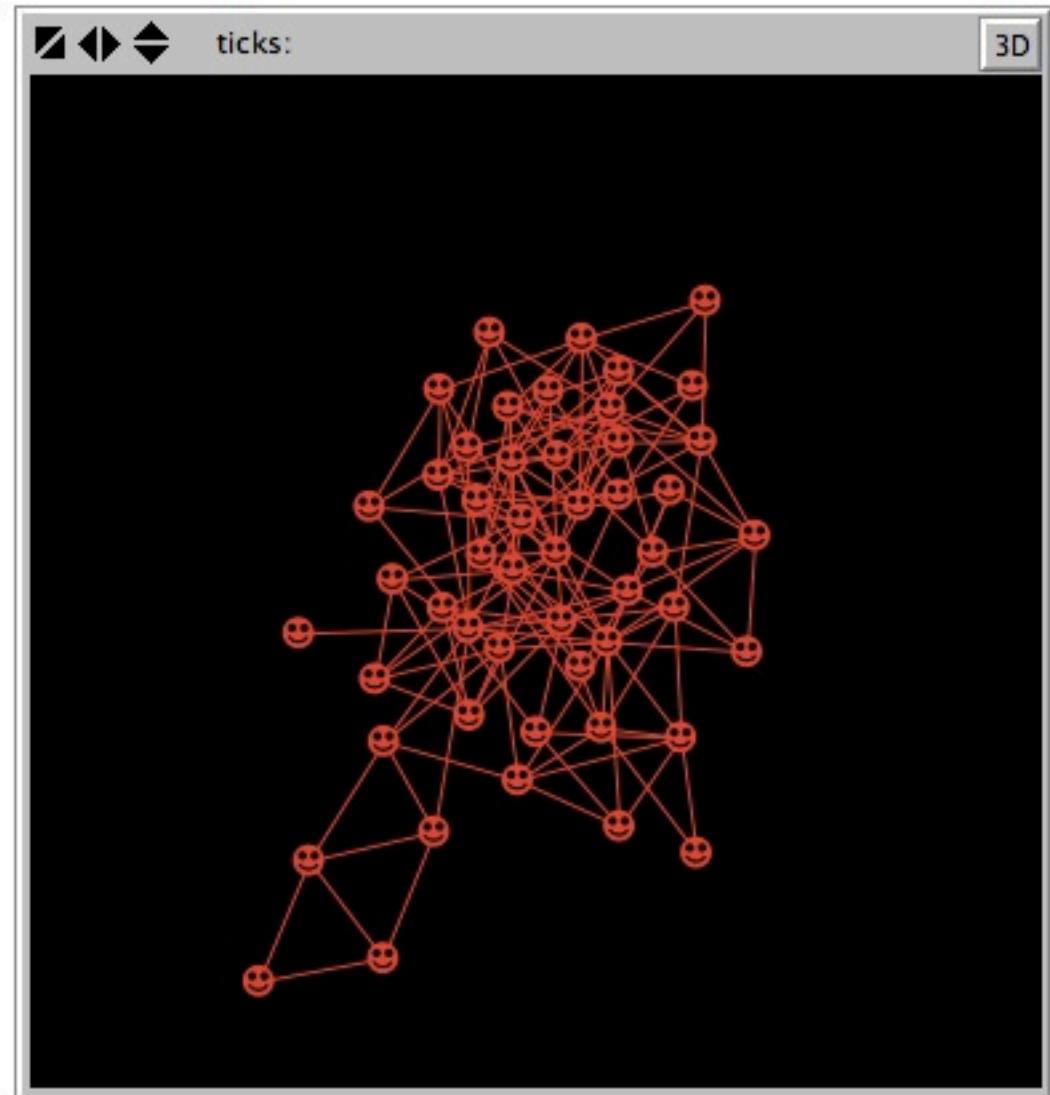
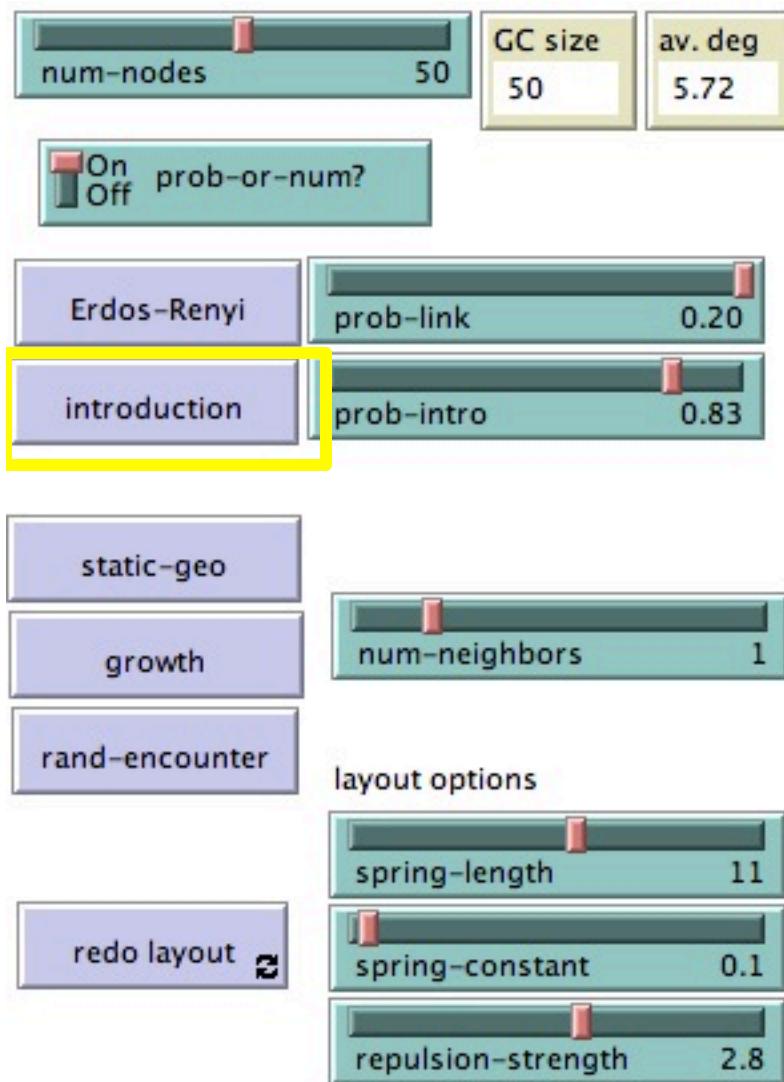
Realism

- ❑ Consider alternative mechanisms of constructing a network that are also fairly “random”.
- ❑ How do they stack up against Erdös-Renyi?
- ❑ [http://www.ladamic.com/netlearn/nw/
RandomGraphs.html](http://www.ladamic.com/netlearn/nw/RandomGraphs.html)

Introduction model

- ❑ Prob-link is the p (probability of any two nodes sharing an edge) that we are used to
- ❑ But, with probability prob-intro the other node is selected among one of our friends' friends and not completely at random

Introduction model



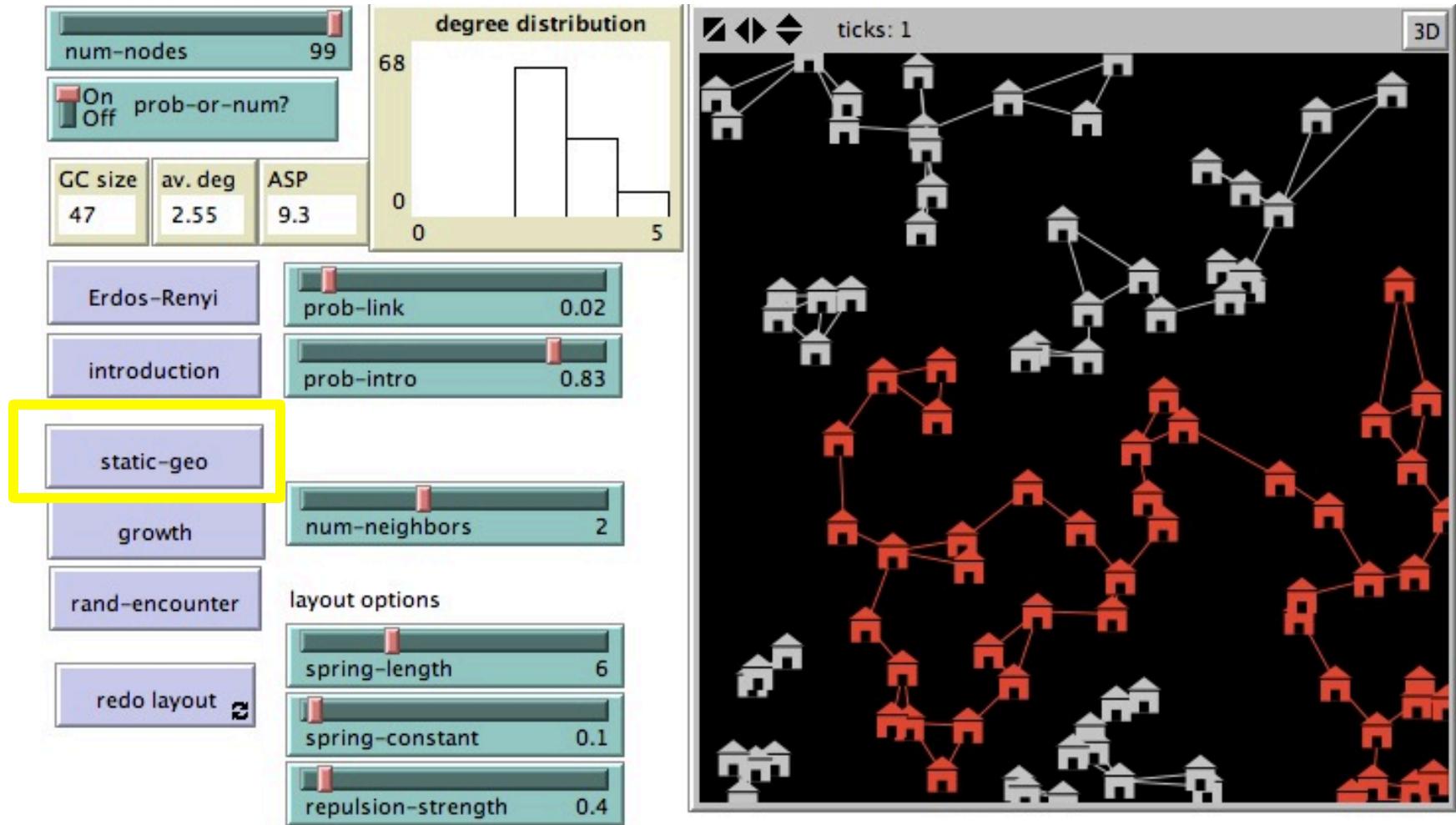
Quiz Q:

- ❑ Relative to ER, the introduction model has:
 - ❑ more edges
 - ❑ more closed triads
 - ❑ longer average shortest path
 - ❑ more uneven degree
 - ❑ smaller giant component at low p

Static Geographical model

- ❑ Each node connects to num-neighbors of its closest neighbors
- ❑ use the num-neighbors slider, and for comparison, switch PROB-OR-NUM to ‘off’ to have the ER model aim for num-neighbors as well
- ❑ turn off the layout algorithm while this is running, you can apply it at the end

static geo



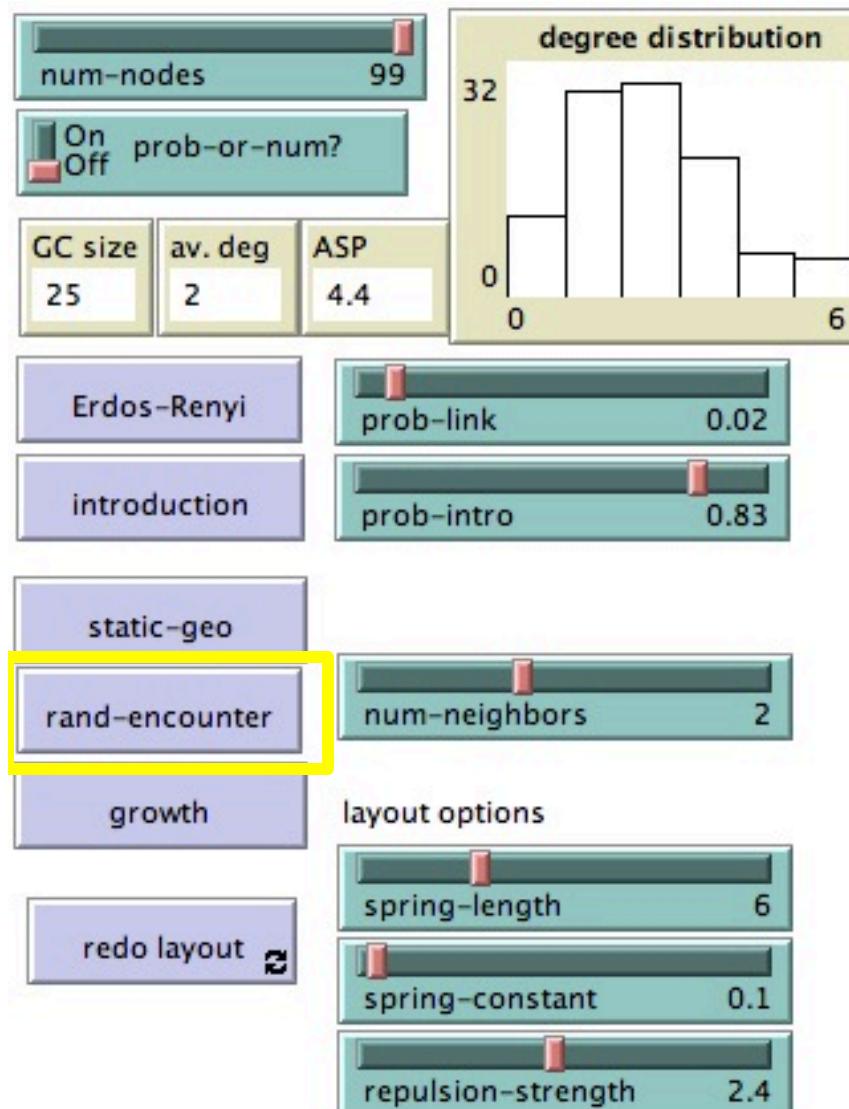
Quiz Q:

- ❑ Relative to ER, the static geographical model has :
 - ❑ longer average shortest path
 - ❑ shorter average shortest path
 - ❑ narrower degree distribution
 - ❑ broader degree distribution
 - ❑ smaller giant component at a low number of neighbors
 - ❑ larger giant component at a low number of neighbors

Random encounter

- ❑ People move around randomly and connect to people they bump into
- ❑ use the num-neighbors slider, and for comparison, switch PROB-OR-NUM to ‘off’ to have the ER model aim for num-neighbors as well
- ❑ turn off the layout algorithm while this is running (you can apply it at the end)

random encounters



Quiz Q:

- ❑ Relative to ER, the random encounters model has :
 - ❑ more closed triads
 - ❑ fewer closed triads
 - ❑ smaller giant component at a low number of neighbors
 - ❑ larger giant component at a low number of neighbors

Growth model

- ❑ Instead of starting out with a fixed number of nodes, nodes are added over time
- ❑ use the num-neighbors slider, and for comparison, switch PROB-OR-NUM to ‘off’ to have the ER model aim for num-neighbors as well

growth model

degree distribution

Degree	Frequency
0	32
1	28
2	18
3	12
4	8
5	5
6	3
7	2
8	1

ticks: 1

3D

num-nodes 99

On Off prob-or-num?

GC size 65 av. deg 1.88 ASP 3.9

Erdos-Renyi

introduction

static-geo

rand-encounter

growth

redo layout

prob-link 0.02

prob-intro 0.83

num-neighbors 2

layout options

spring-length 6

spring-constant 0.1

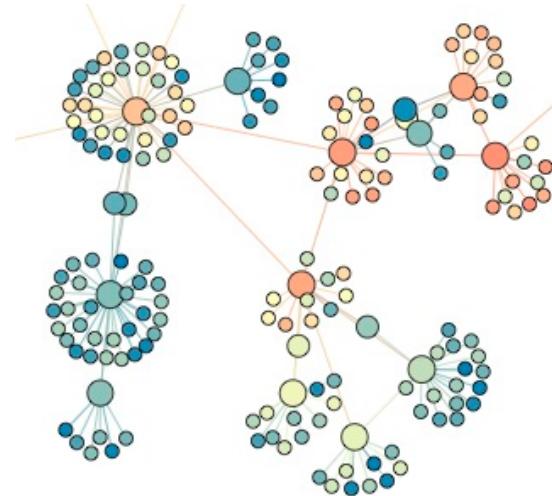
repulsion-strength 2.4

Quiz Q:

- ❑ Relative to ER, the growth model has :
 - ❑ more hubs
 - ❑ fewer hubs
 - ❑ smaller giant component at a low number of neighbors
 - ❑ larger giant component at a low number of neighbors

other models

- ❑ in some instances the ER model is plausible
- ❑ if dynamics are different, ER model may be a poor fit

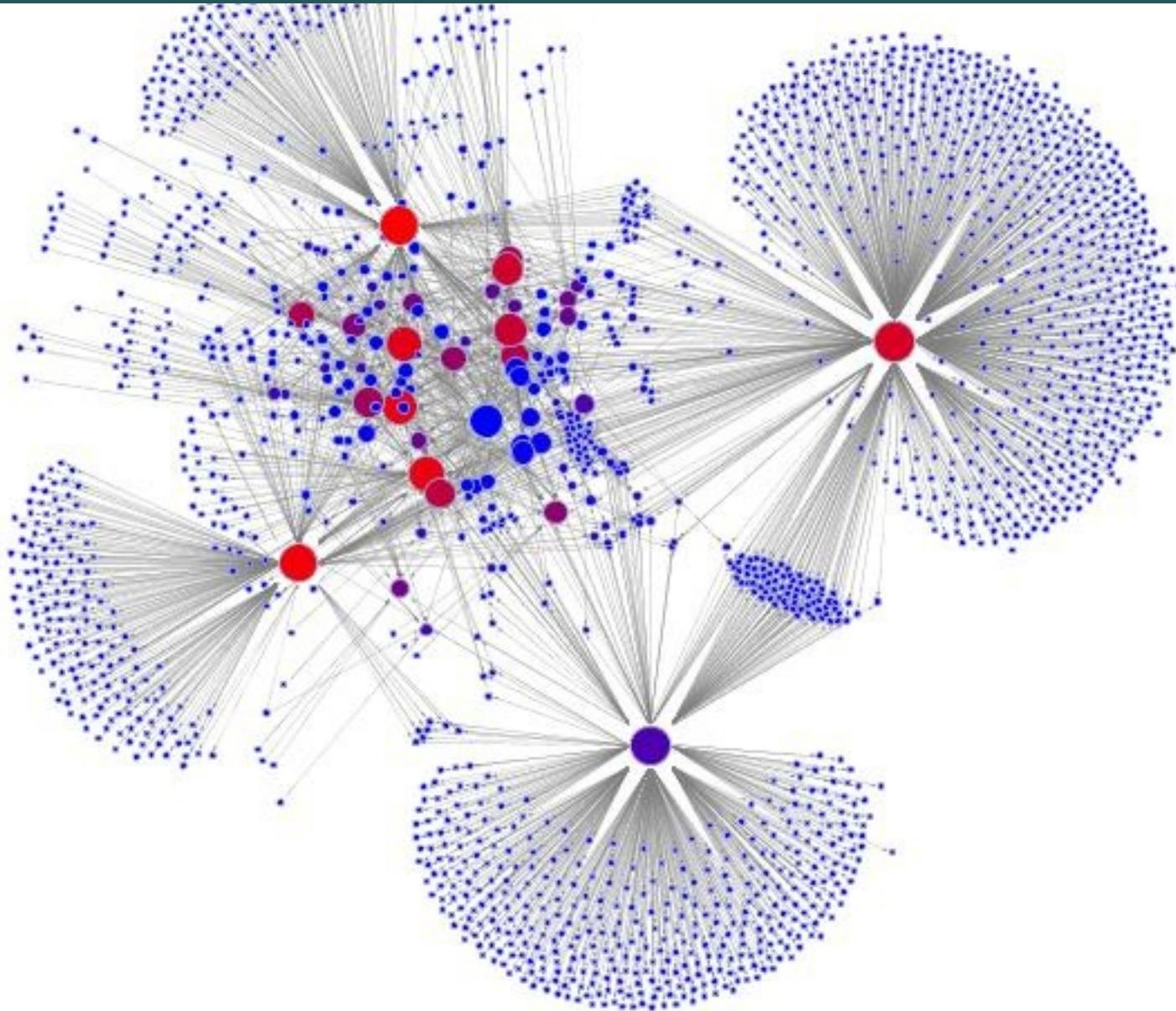


SNA 2C: Growth & Preferential Attachment Models

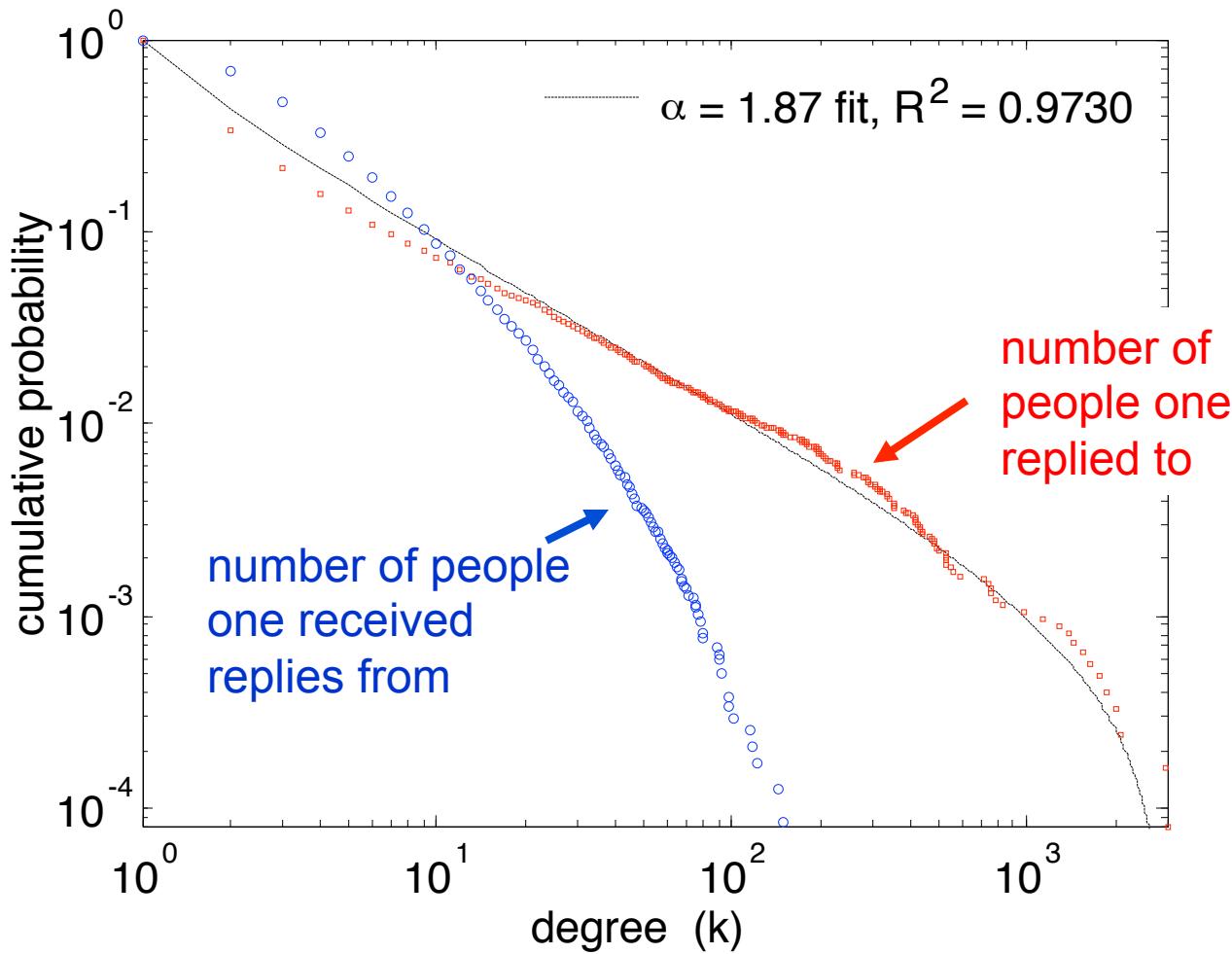
Lada Adamic



Online Question & Answer Forums



Uneven participation

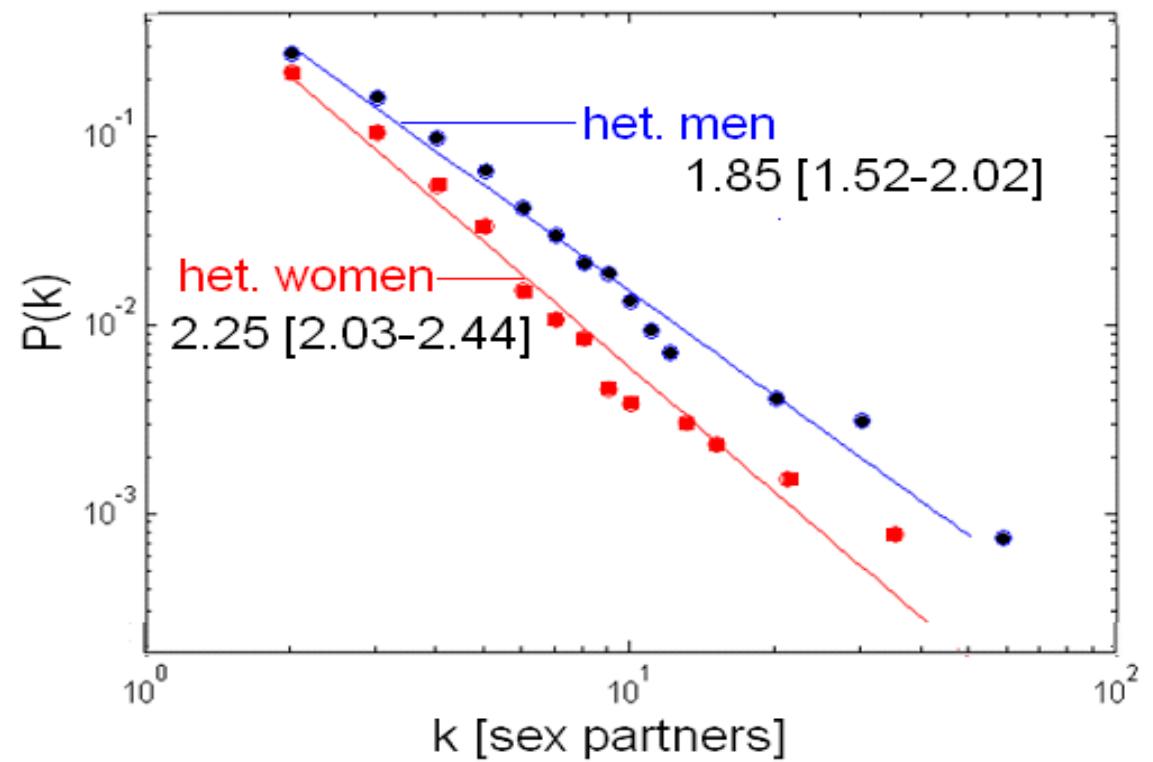


- ‘answer people’ may reply to thousands of others
- ‘question people’ are also uneven in the number of repliers to their posts, but to a lesser extent

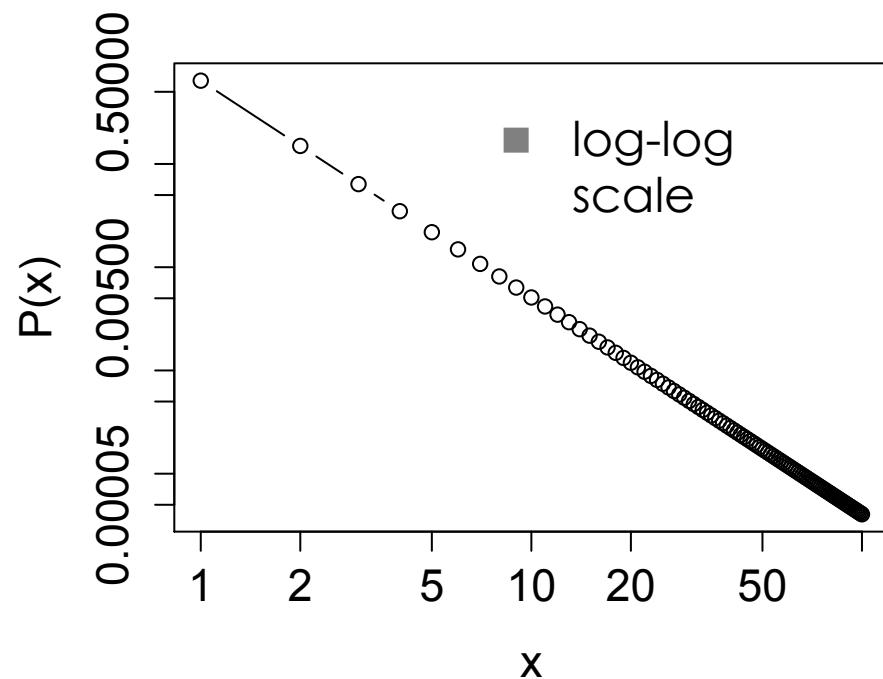
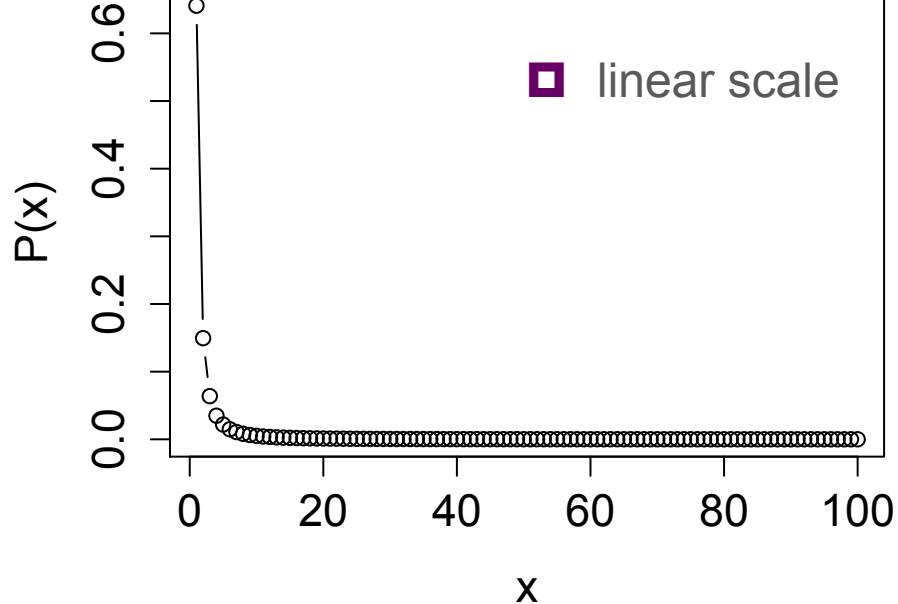
Real-world degree distributions

- Sexual networks

- Great variation in contact numbers

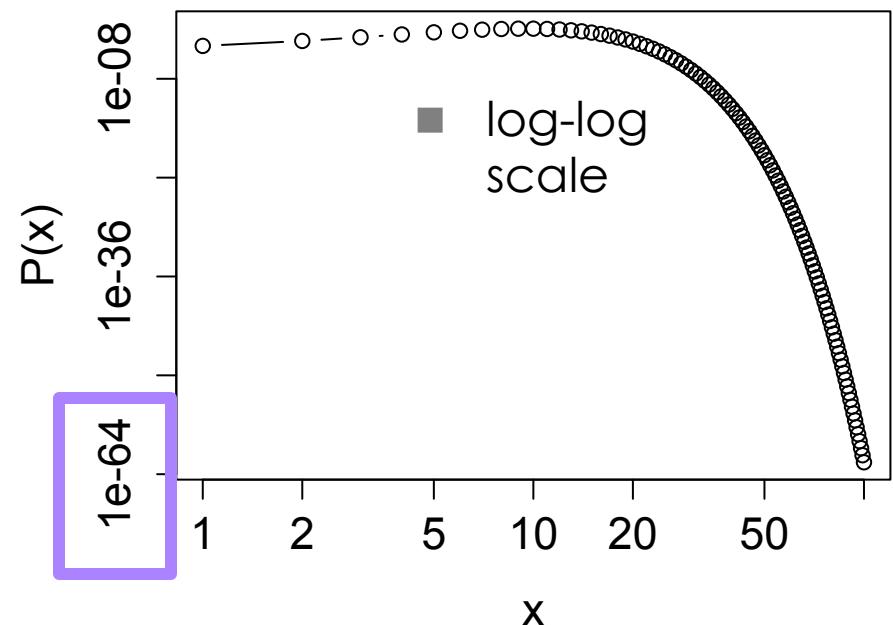
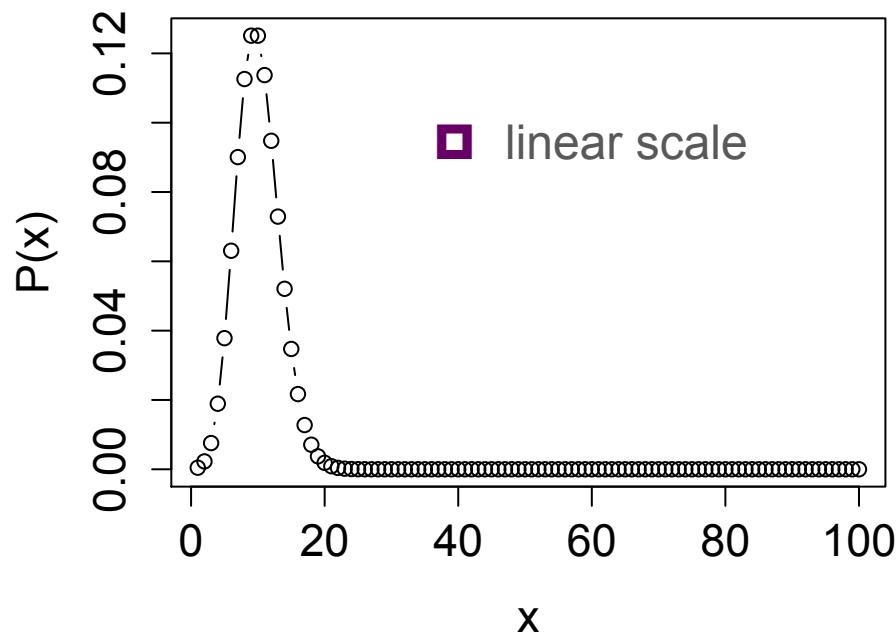


Power-law distribution



- high skew (asymmetry)
- straight line on a log-log plot

Poisson distribution



- little skew (asymmetry)
- curved on a log-log plot

Power law distribution

- Straight line on a log-log plot

$$\ln(p(k)) = c - \alpha \ln(k)$$

- Exponentiate both sides to get that $p(k)$, the probability of observing a node of degree ‘ k ’ is given by

$$p(k) = Ck^{-\alpha}$$

normalization constant (probabilities over all k must sum to 1)

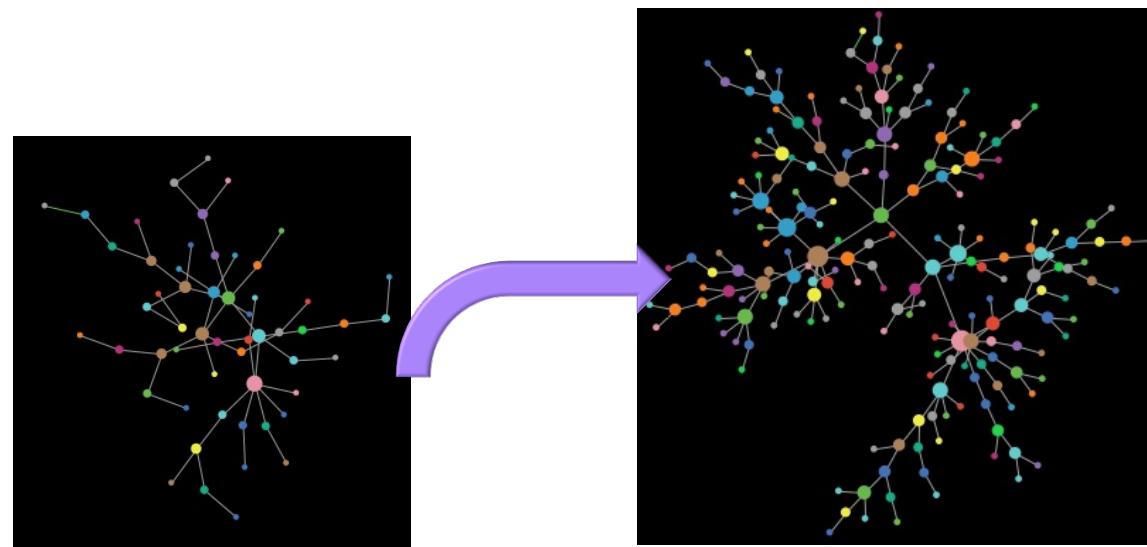
power law exponent α

Quiz Q:

- ❑ As the exponent α increases, the downward slope of the line on a log-log plot
 - ❑ stays the same
 - ❑ becomes milder
 - ❑ becomes steeper

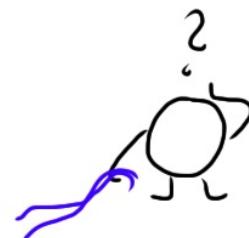
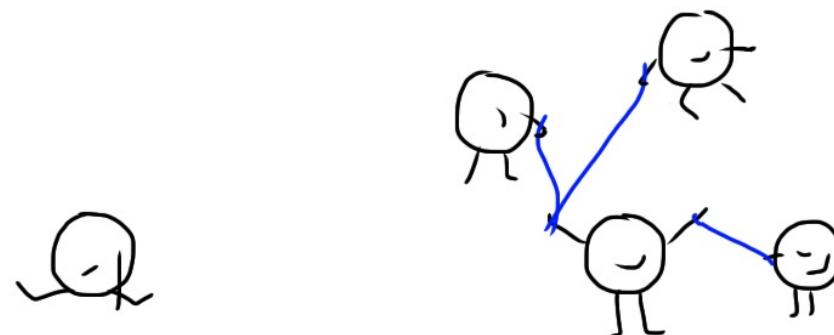
2 ingredients in generating power-law networks

- nodes appear over time (growth)



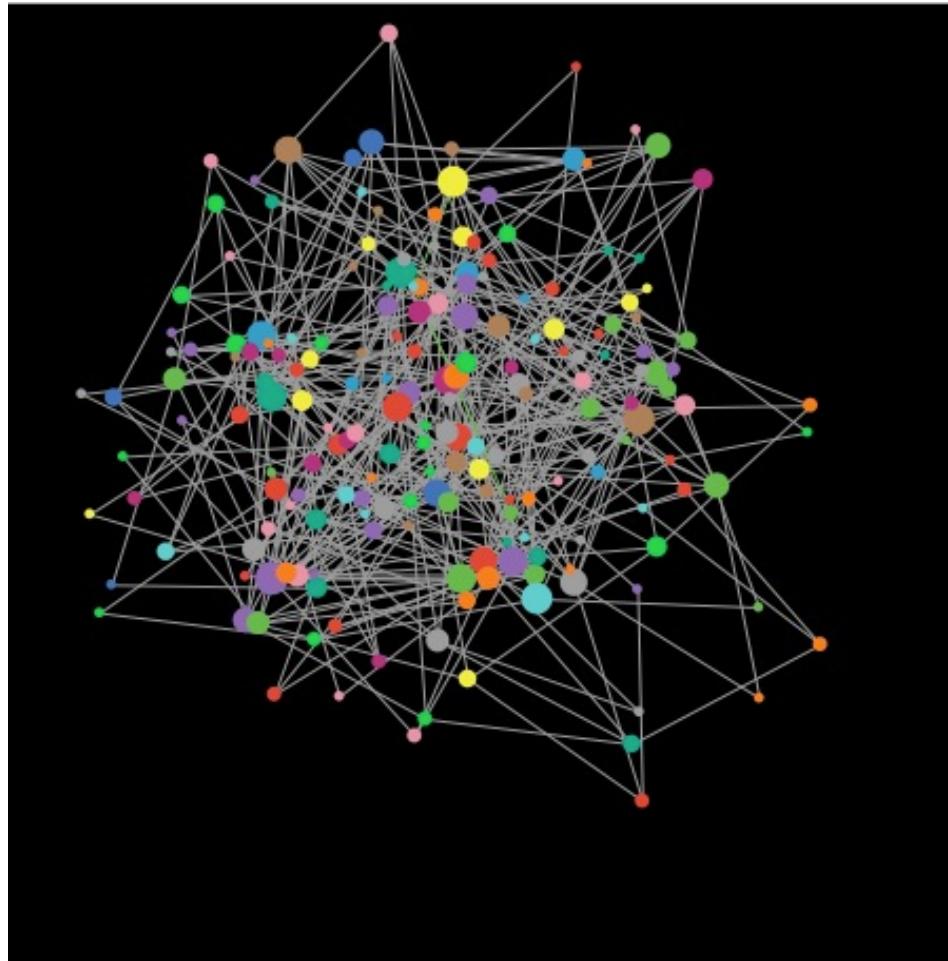
2 ingredients in generating power-law networks

- nodes prefer to attach to nodes with many connections (preferential attachment, cumulative advantage)



Ingredient # 1: growth over time

- nodes appear one by one, each selecting m other nodes at random to connect to



$m = 2$

random network growth

- ❑ one node is born at each time tick
- ❑ at time t there are t nodes
- ❑ change in degree k_i of node i (born at time i , with $0 < i < t$)

$$\frac{dk_i(t)}{dt} = \frac{m}{t}$$

there are m new edges being added per unit time (with 1 new node)

the m edges are being distributed among t nodes

a node in a randomly grown network

- ❑ how many new edges does a node accumulate since it's birth at time i until time t ?
- ❑ integrate from i to t

$$\frac{dk_i(t)}{dt} = \frac{m}{t}$$

to get

$$k_i(t) = m + m \log\left(\frac{t}{i}\right)$$

born with \mathbf{m} edges

age and degree

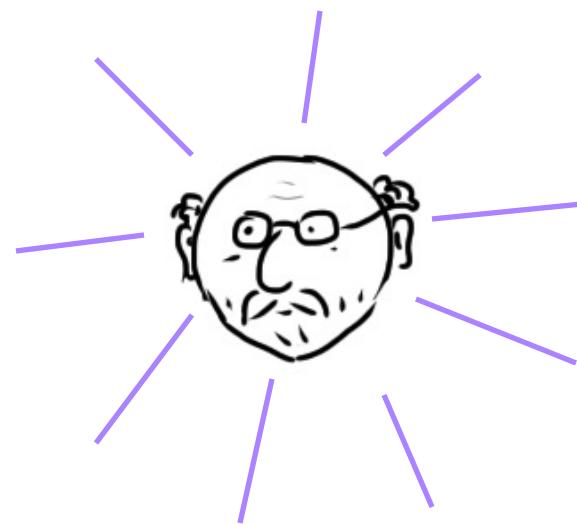
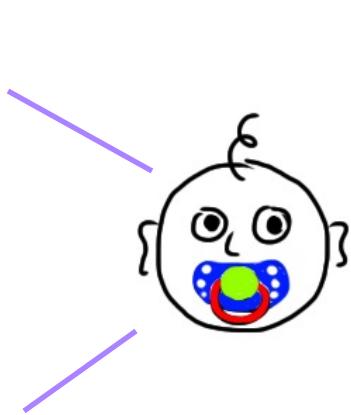
on average

$$k_i(t) > k_j(t)$$

if

$$i < j$$

i.e. older nodes on average have more edges



Quiz Q:

- ❑ How could one make the growth model more realistic for social networks?
 - ❑ old nodes die
 - ❑ some nodes are more sociable
 - ❑ friendships vane over time
 - ❑ all of the above

growing random networks

Let $\tau(100)$ be the time at which node with degree e.g. 100 is born. The the fraction of nodes that have degree ≤ 100 is $(t - \tau)/t$

$$k_\tau(t) = m + m \log\left(\frac{t}{\tau}\right)$$

random growth: degree distribution

□ continuing...

$$\log\left(\frac{t}{\tau}\right) = \frac{k - m}{m}$$

$$\frac{\tau}{t} = e^{-\frac{k-m}{m}}$$

exponential distribution in degree

The probability that a node has degree k or less is
 $1 - \tau/t$

$$P(k < k') = 1 - e^{-\frac{k' - m}{m}}$$

Quiz Q:

- ❑ The degree distribution for a growth model where new nodes attach to old nodes at random will be
 - ❑ a curved line on a log-log plot
 - ❑ a straight line on a log-log plot

2nd ingredient: preferential attachment

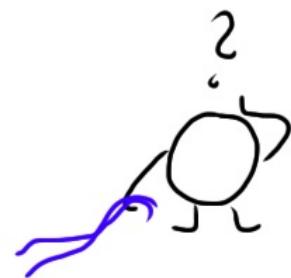
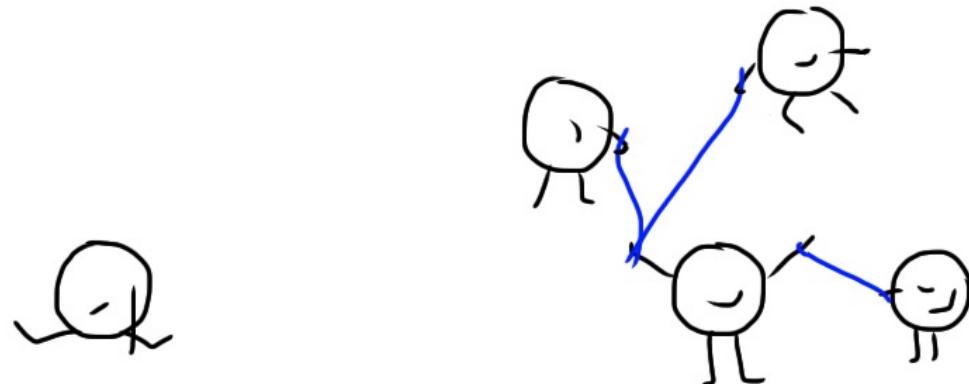
- Preferential attachment:
 - new nodes prefer to attach to well-connected nodes over less-well connected nodes

- Process also known as
 - cumulative advantage
 - rich-get-richer
 - Matthew effect

Price's preferential attachment model for citation networks

- [Price 65]
 - each new paper is generated with m citations (mean)
 - new papers cite previous papers with probability proportional to their indegree (citations)
 - what about papers without any citations?
 - each paper is considered to have a “default” citation
 - probability of citing a paper with degree k , proportional to $k+1$
- Power law with exponent $\alpha = 2+1/m$

Preferential attachment



Cumulative advantage: how?

File Size: 100K



Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference

Steve Rasmussen, Ph.D., FRCR

Received 11 December 1991; accepted 14 April 1992.

Key words: *Streptomyces* — *bioactive substances* — *antibiotic resistance* — *bioactive substances* — *bioactive substances* — *bioactive substances*

www.ijerph.com

Page 1

(8) C.R. Raman and A. Bhowmik, *Polymer*, **2000**, *41*, 103-108.

classical change occurs in only one direction (upward or downward) in the spectrum of energy levels. The energy levels of the system are discrete, as they are in any regular or periodic motion; each level will be filled from below upwards, from the ground state up to the highest occupied level. Once the paramagnetic resonance signal disappears, the magnetization may still hold its value for a short time, but it will then decrease, and finally reach zero.

A parametric insurance and possibly more realistic model of insurance availability has been suggested by Fox (2001). This available solution can distinguish that whether or not the agent can make his/her insurance available.

newer, more advanced strains have substantially reduced the amount of glucose consumption by lysed mycoplasmas, although remaining cell lysis rates are often still about 10% of those in the absence of glucose. In contrast to these in vitro findings, the glucose tolerance of a pure strain of *Mycoplasma genitalium*, which was described as noncapable of glucose metabolism, was found to grow on glucose at a rate similar to that of *M. pneumoniae*.

biochemical), which was developed to compare the results of cross-species and contrast studies for phylogenies of tree species, has now been developed to deal with lineage relationships. We shall report the results of such analyses elsewhere.

“I have had three major publications in refereed journals that examine the range of maximum trade demanded from different types of business, one purpose has been to examine very closely the data used.

and the χ^2 statistic. However, the proportion of cases with a parametric estimate of standard error in these systematic analyses is not that much smaller than the proportion of cases with a nonparametric estimate. Therefore, to get more idea of the magnitude of parametric estimates we selected that $N = 10$ (as per our previous studies) and we used the results of $N = 10$ as estimates of the mean of the distribution of estimates.

at certain points an increase in infectious reservoirs may bring R_0 above the epidemic rate in the plethoric calamineous programs, which is due to many workers in a case rate for complete eradication. Furthermore, the specific control areas, such as urbanized concentrations, individual

covert super-motility and toxins-mobilizing which a terminal signature in a series of toxins mentioned will lead or predominance of biological warfare virus. Both recent cases infections and toxins-mobilizing are likely to be common less involved in regional and international conflicts.

www.scholarone.com

- ❑ copying mechanism
 - ❑ visibility

Barabasi-Albert model

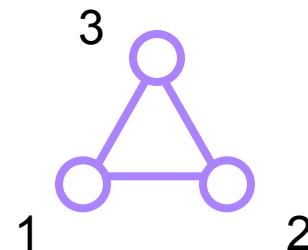
- First used to describe skewed degree distribution of the World Wide Web
- Each node connects to other nodes with probability proportional to their degree
 - the process starts with some initial subgraph
 - each new node comes in with m edges
 - probability of connecting to node i

$$\Pi(i) = m \frac{k_i}{\sum_j k_j}$$

- Results in power-law with exponent $\alpha = 3$

Basic BA-model

- Very simple algorithm to implement
 - start with an initial set of m_0 fully connected nodes
 - e.g. $m_0 = 3$



- now add new vertices one by one, each one with exactly m edges
- each new edge connects to an existing vertex in proportion to the number of edges that vertex already has → ***preferential attachment***
- easiest if you keep track of edge endpoints in one large array and select an element from this array at random
 - the probability of selecting any one vertex will be proportional to the number of times it appears in the array – which corresponds to its degree

1	1	2	2	2	3	3	4	5	6	6	7	8
---	---	---	---	---	---	---	---	---	---	---	---	---	------

generating BA graphs – cont'd

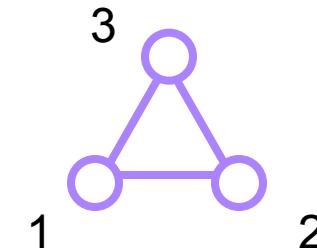
- To start, each vertex has an equal number of edges (2)
 - the probability of choosing any vertex is $1/3$

- We add a new vertex, and it will have m edges, here take $m=2$
 - draw 2 random elements from the array – suppose they are 2 and 3

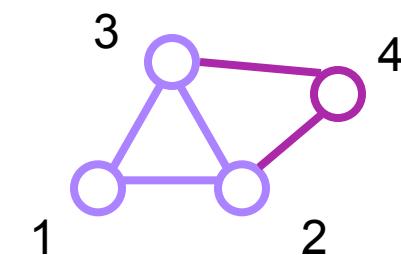
- Now the probabilities of selecting 1,2,3,or 4 are $1/5$, $3/10$, $3/10$, $1/5$

- Add a new vertex, draw a vertex for it to connect from the array
 - etc.

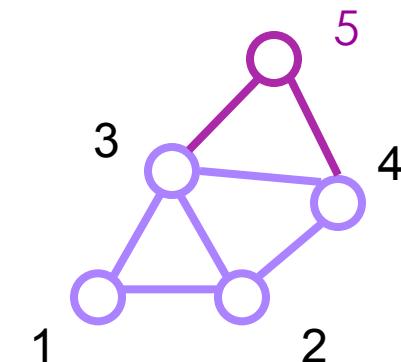
1 1 2 2 3 3



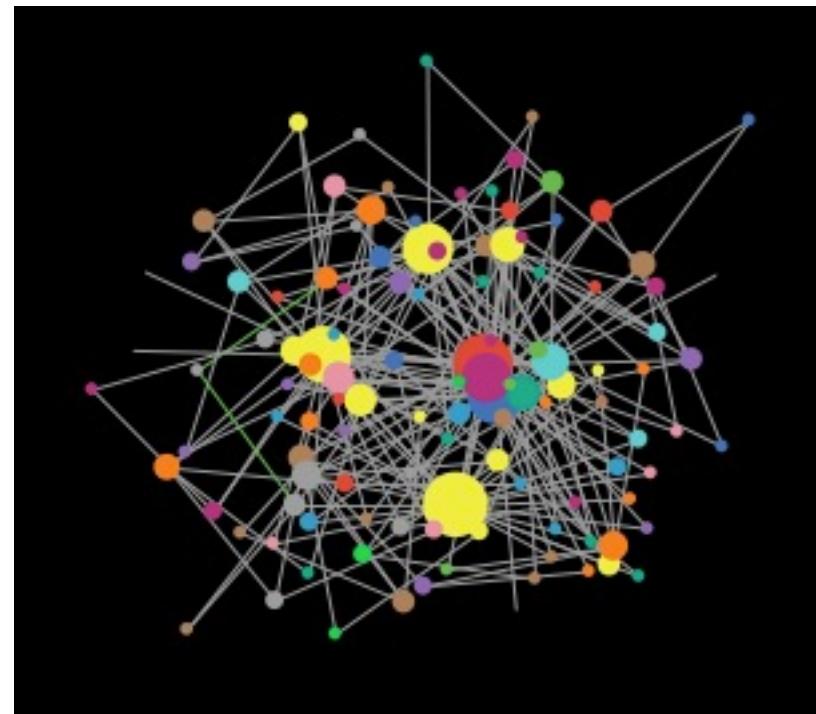
1 1 2 2 2 3 3 3 4 4



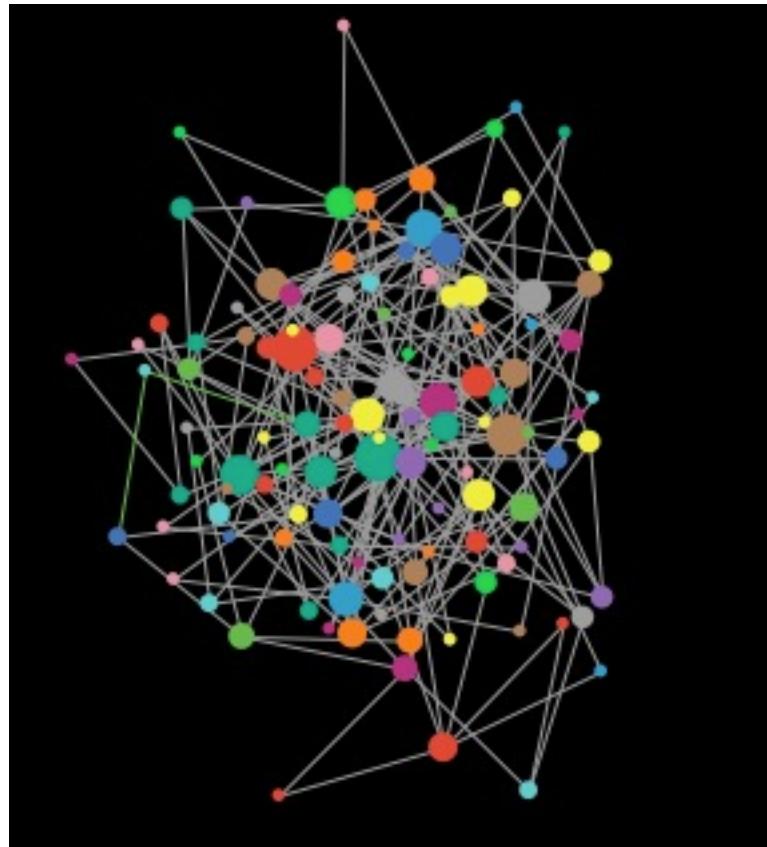
1 1 2 2 2 3 3 3 3 4 4 4 5 5



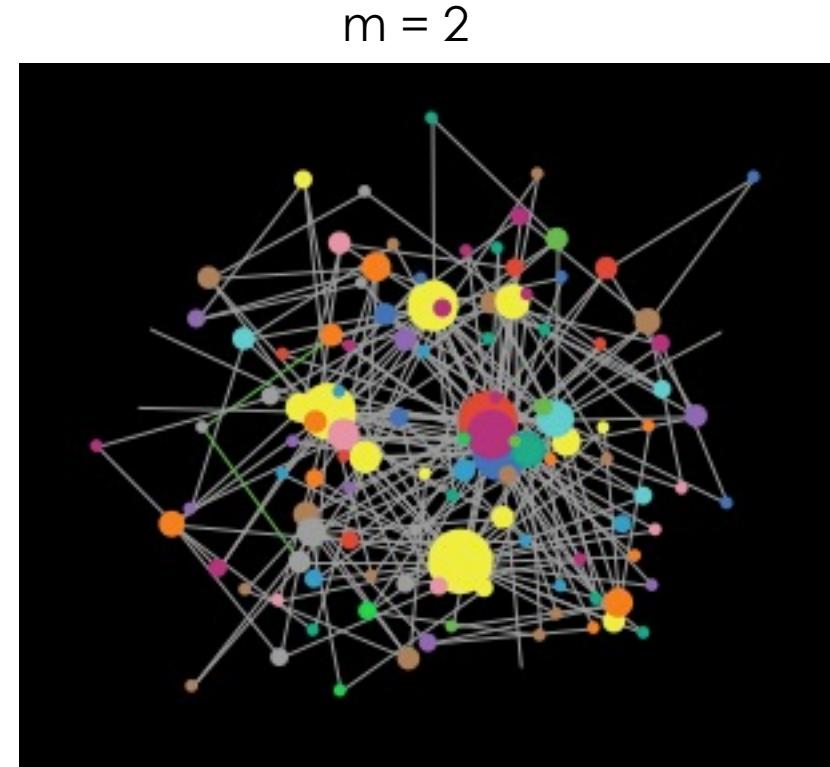
after a while...



contrasting with random (non-preferential) growth



random



preferential

mean field approximation

- ❑ probability that node i acquires a new link at time t

$$\frac{dk_i(t)}{dt} = m \frac{k_i}{2tm} = \frac{k_i}{2t} \quad \text{with} \quad k_i(i) = m$$

$$k_i(t) = m \left(\frac{t}{i}\right)^{1/2}$$

BA model degree distribution

□ time of birth of node of degree k' : τ

$$\frac{\tau}{t} = \left(\frac{m}{k'} \right)^2$$

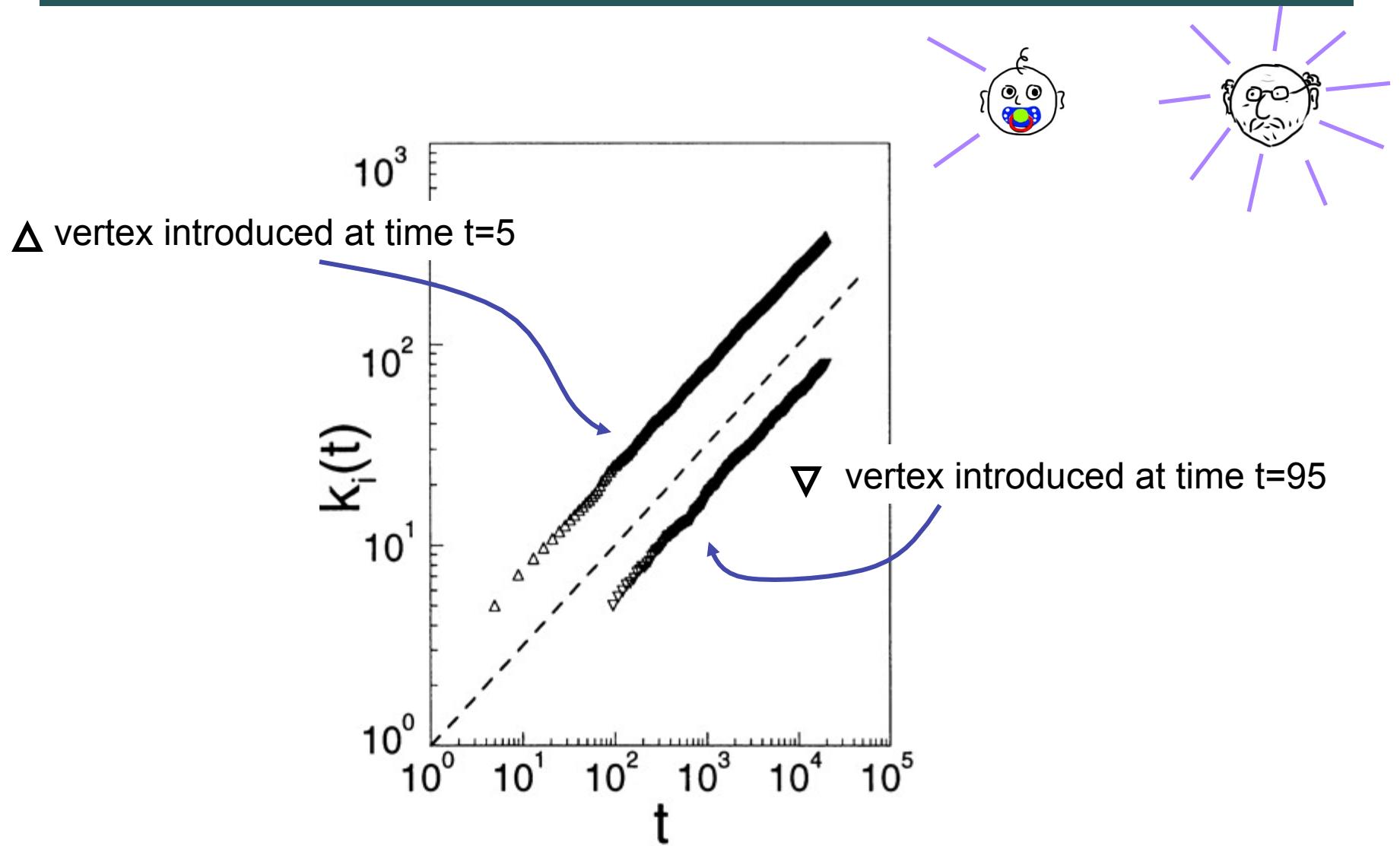
$$P(k < k') = 1 - \frac{m^2}{k'^2}$$

$$p(k) = \frac{2m^2}{k^3}$$

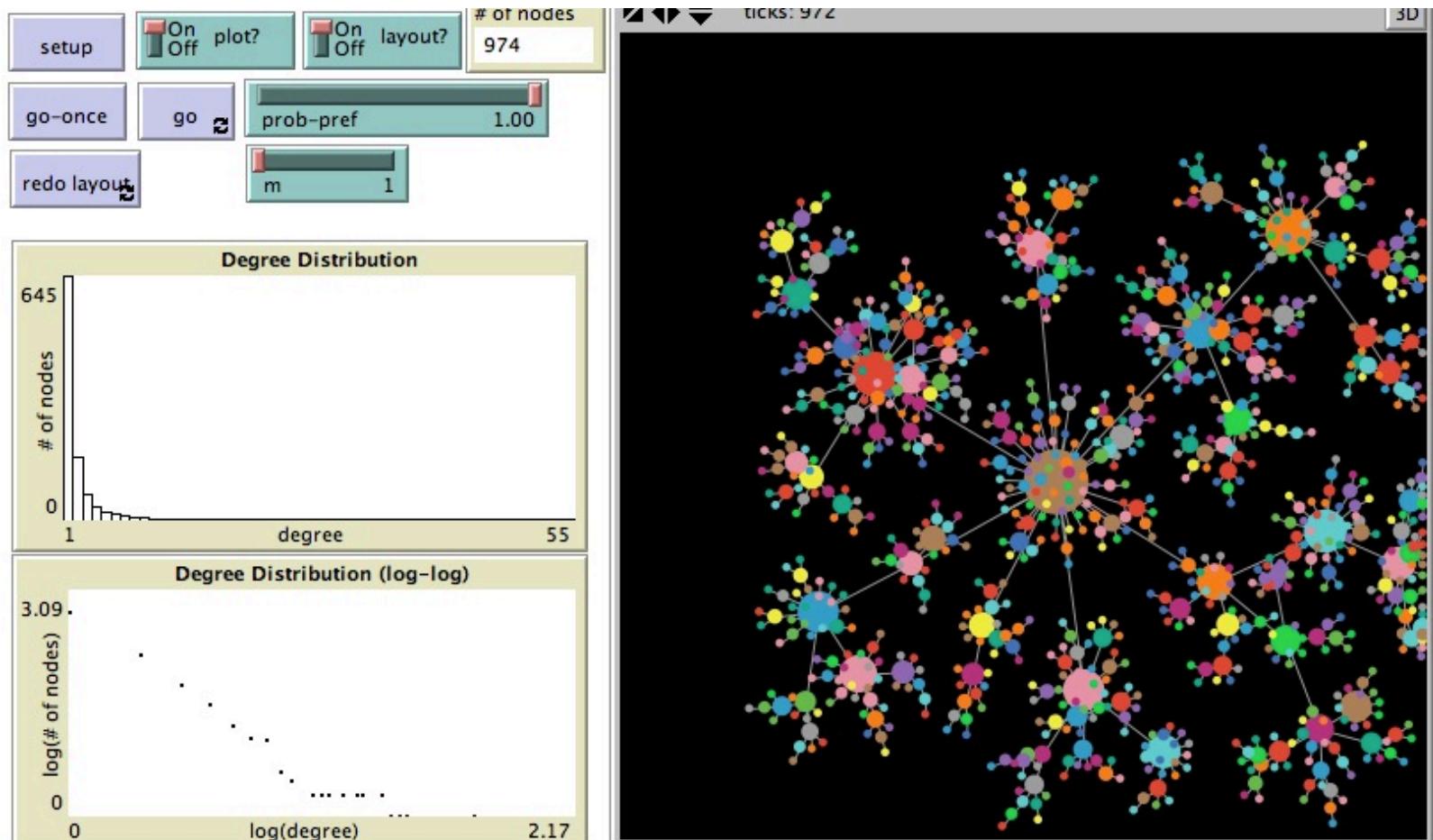
Properties of the BA graph

- The distribution is scale free with exponent $\alpha = 3$
 $P(k) = 2 m^2/k^3$
- The graph is connected
 - Every new vertex is born with a link or several links (depending on whether $m = 1$ or $m > 1$)
 - It then connects to an ‘older’ vertex, which itself connected to another vertex when it was introduced
 - And we started from a connected core
- The older are richer
 - Nodes accumulate links as time goes on, which gives older nodes an advantage since newer nodes are going to attach preferentially – and older nodes have a higher degree to tempt them with than some new kid on the block

Young vs. old in BA model



try it yourself



<http://www.ladamic.com/netlearn/NetLogo501/RAndPrefAttachment.html>

Quiz Q:

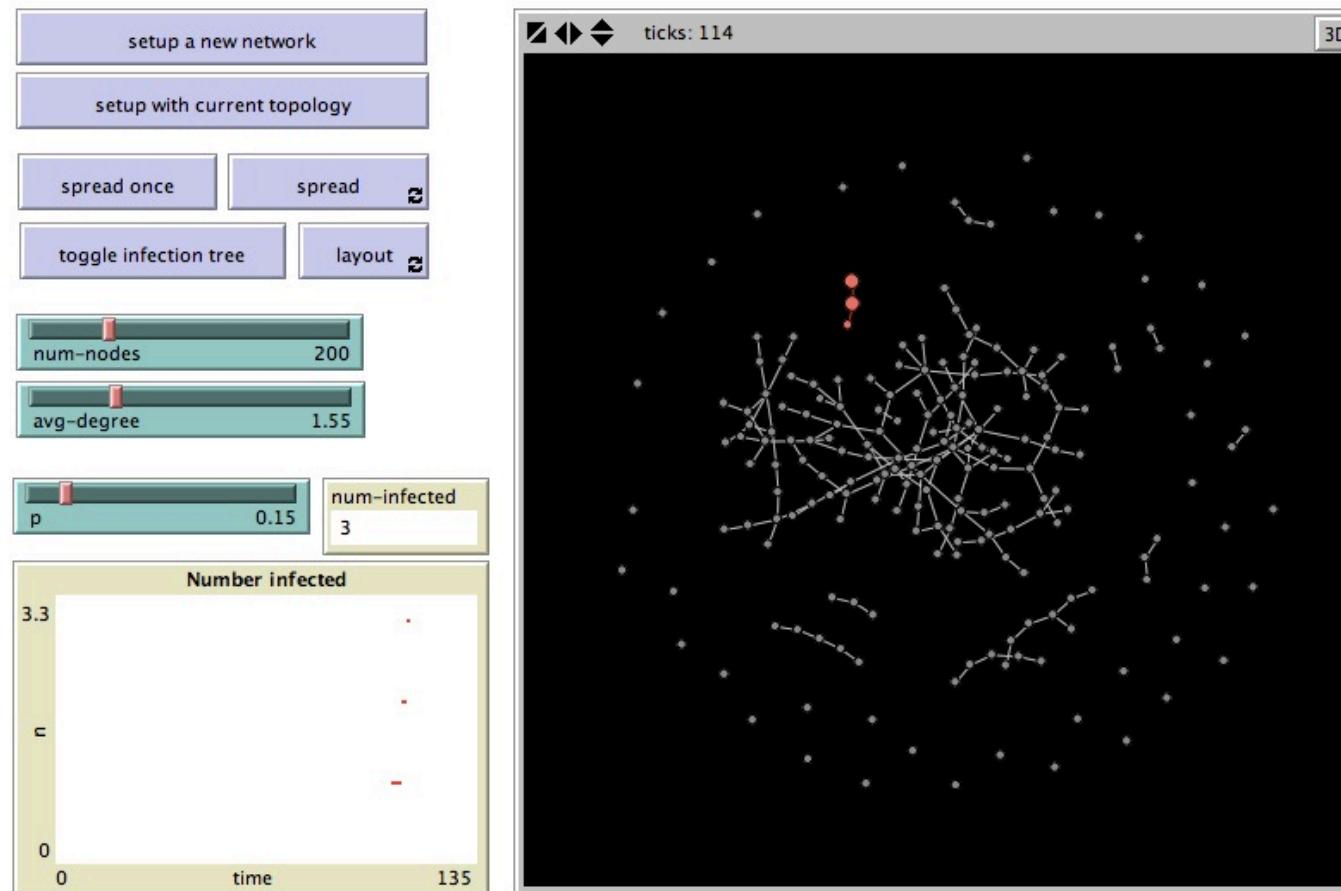
- ❑ Relative to the random growth model, the degree distribution in the preferential attachment model
 - ❑ resembles a power-law distribution less
 - ❑ resembles a power-law distribution more

Summary: growth models

- ❑ Most networks aren't 'born', they are made.
- ❑ Nodes being added over time means that older nodes can have more time to accumulate edges
- ❑ Preference for attaching to 'popular' nodes further skews the degree distribution toward a power-law

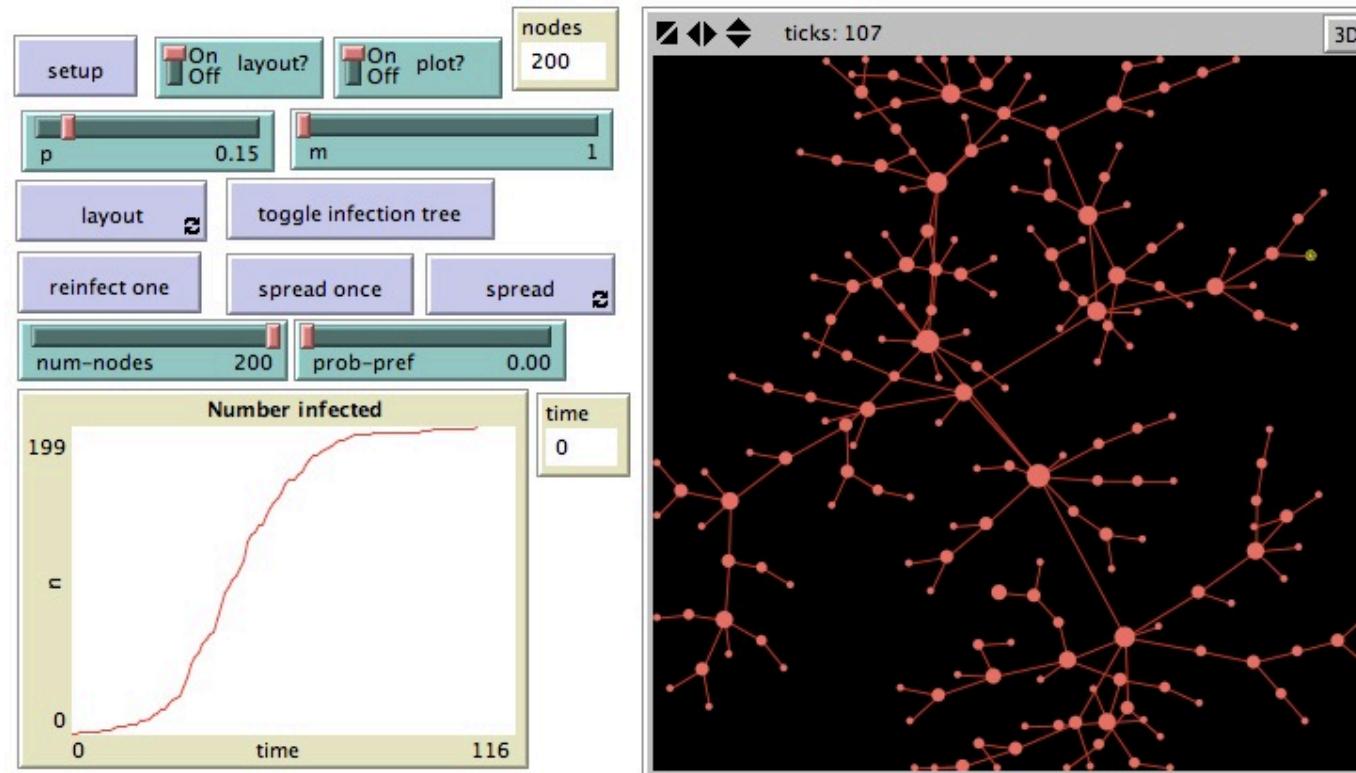
Assignment: implications for diffusion

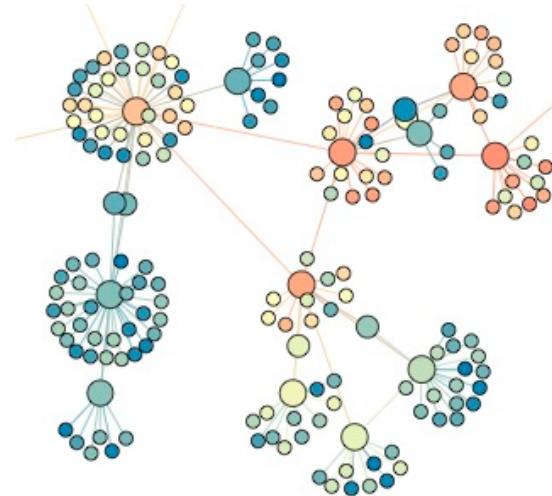
- How does the size of the giant component influence diffusion?



Assignment: implications for diffusion

- How do growth and preferential attachment influence diffusion?



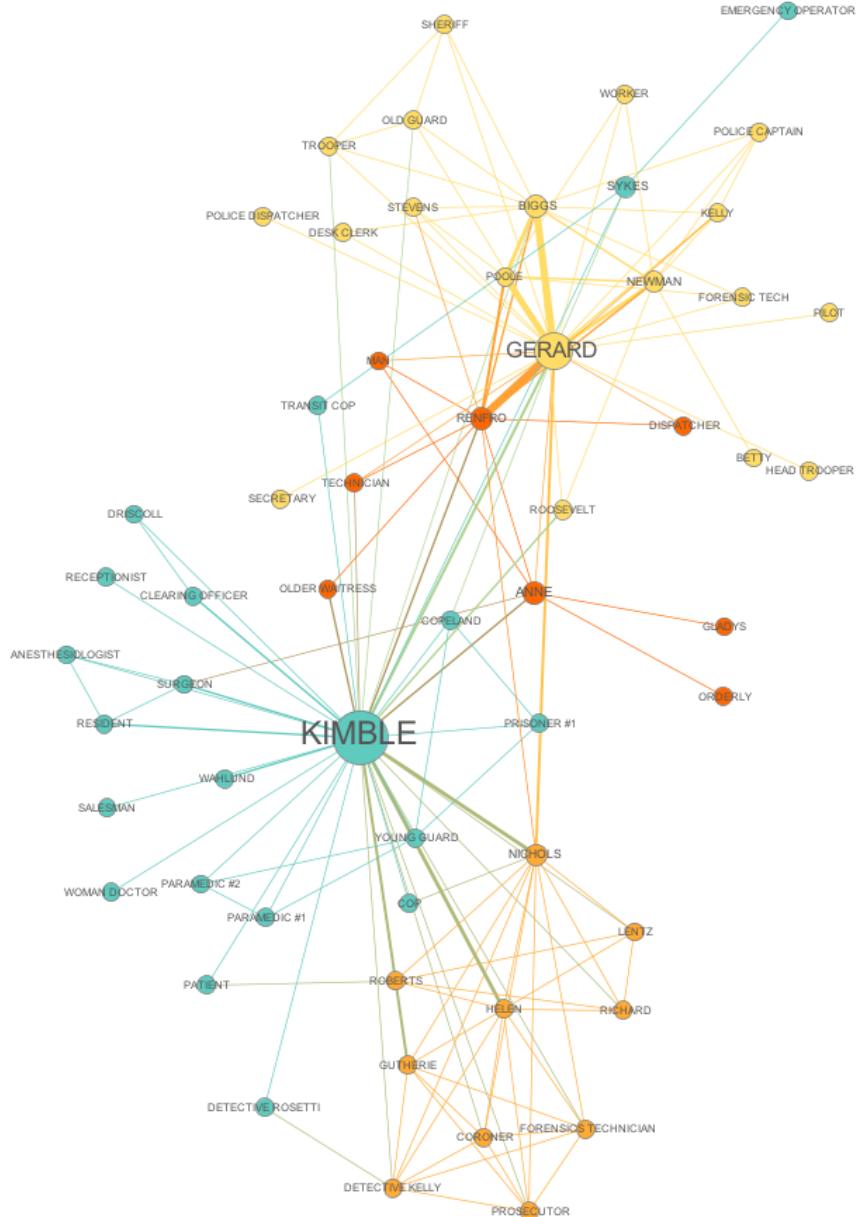


SNA 3A: Centrality

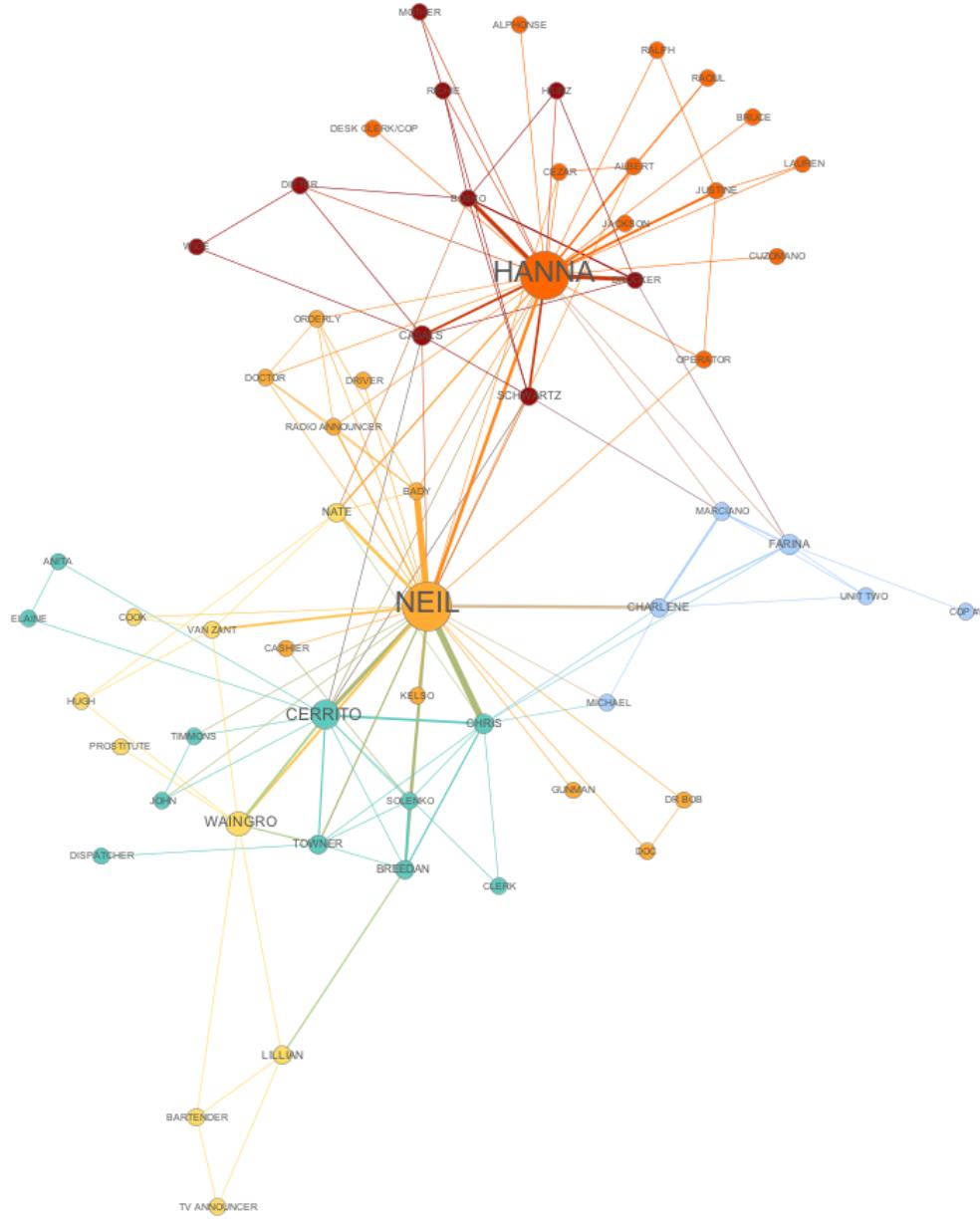
Lada Adamic

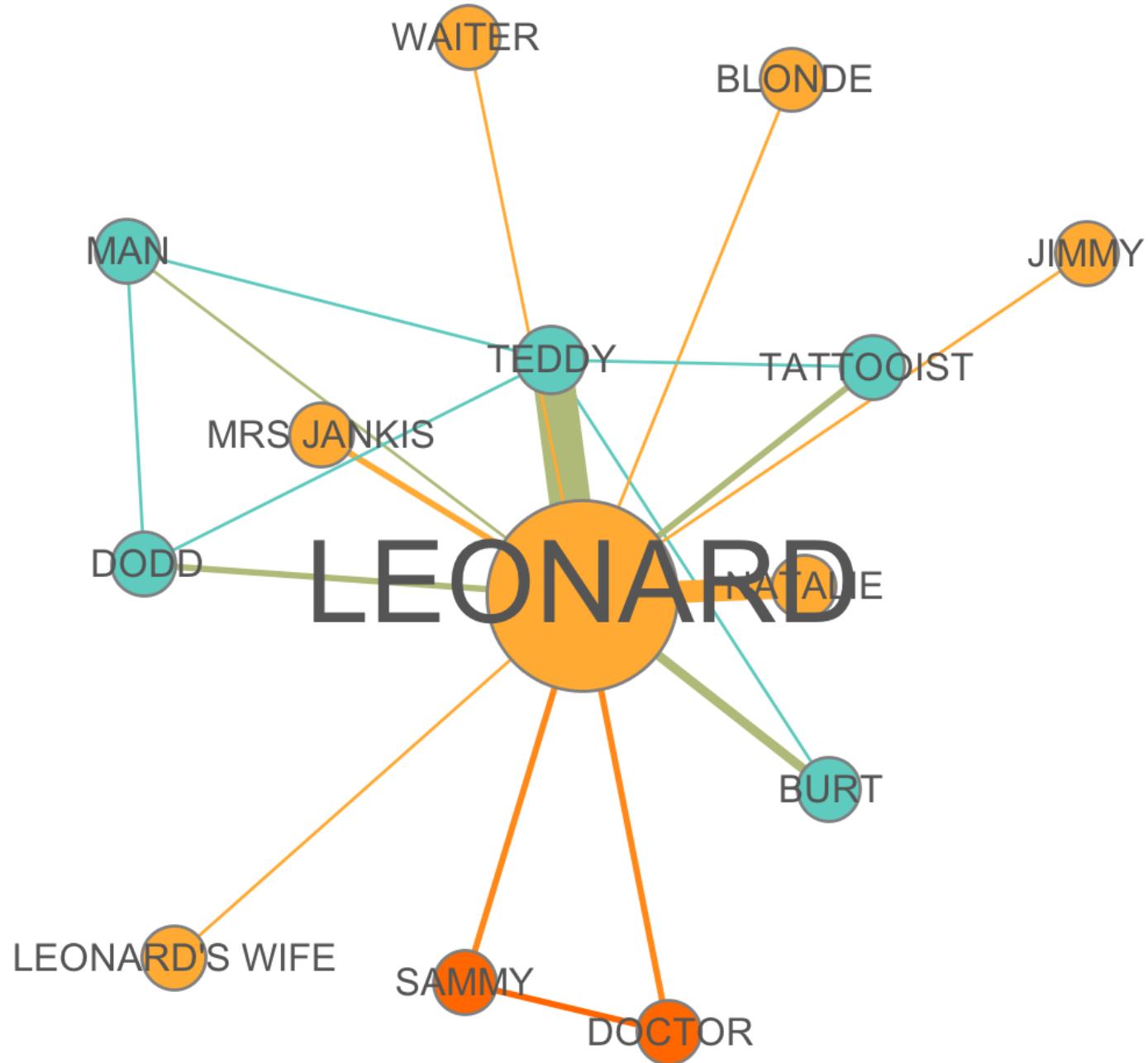


The Fugitive (1993)

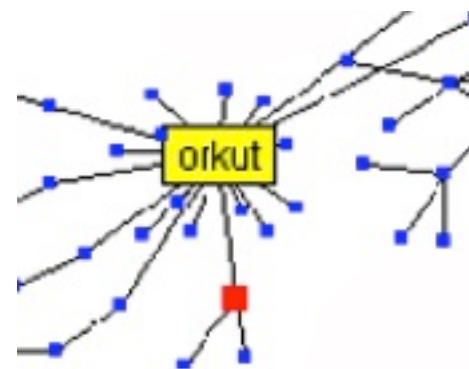


Heat (1995)

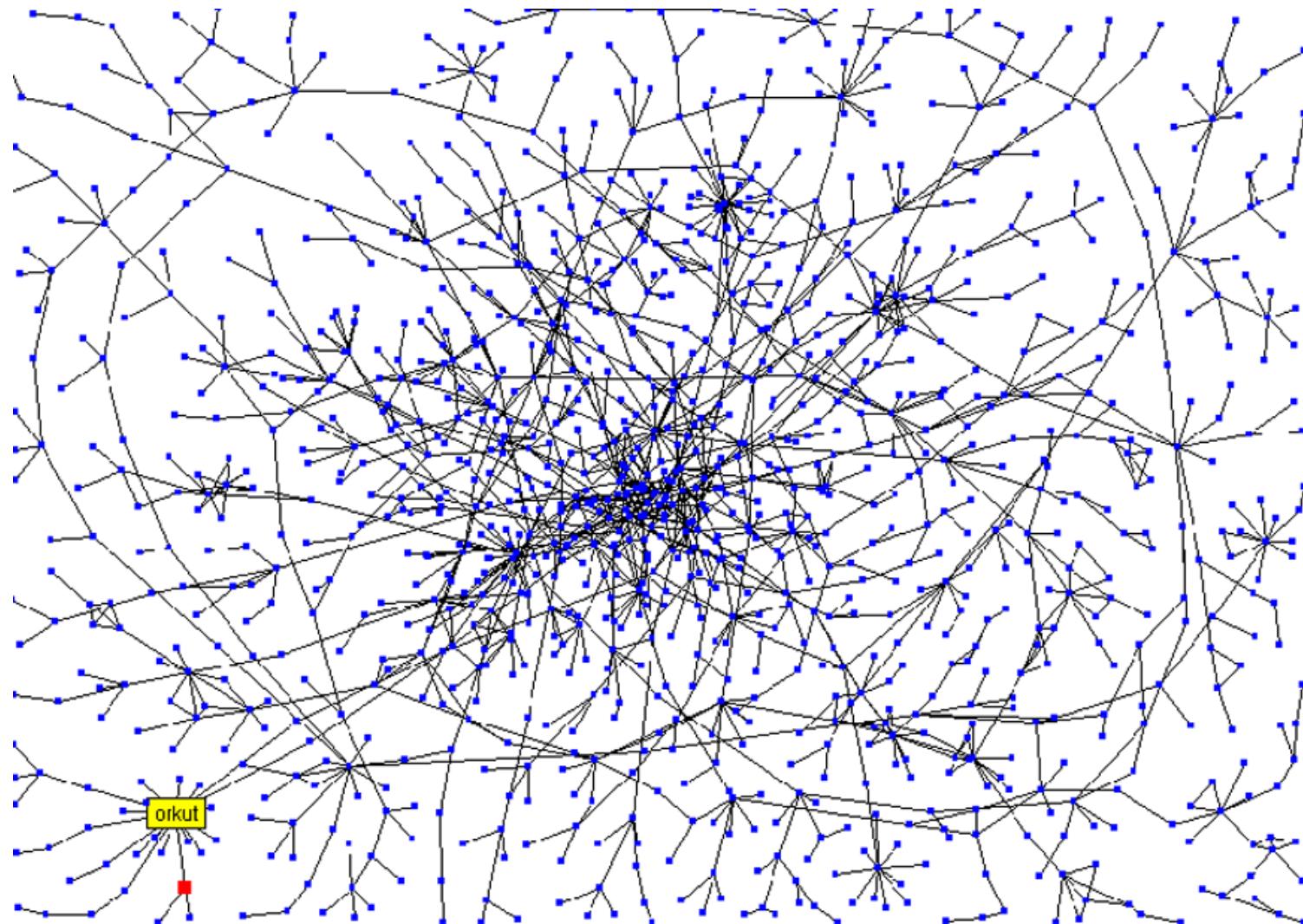




is counting the edges enough?



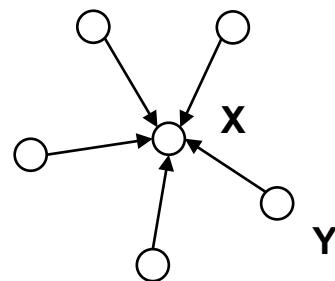
Stanford Social Web (ca. 1999)



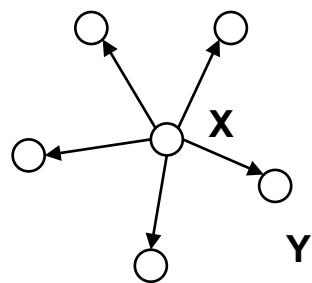
network of personal homepages at Stanford

different notions of centrality

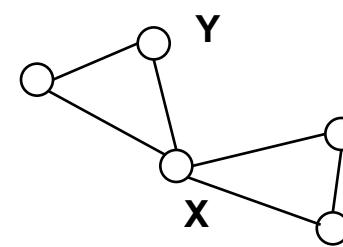
In each of the following networks, X has higher centrality than Y according to a particular measure



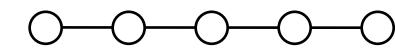
indegree



outdegree

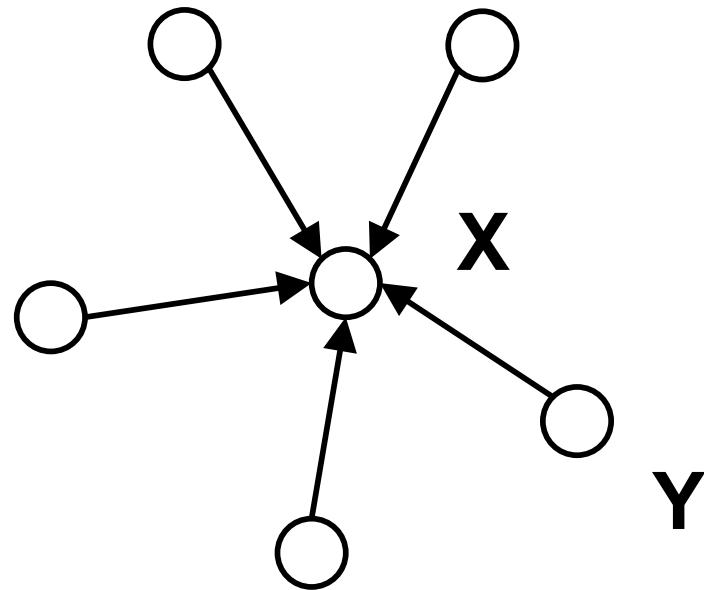


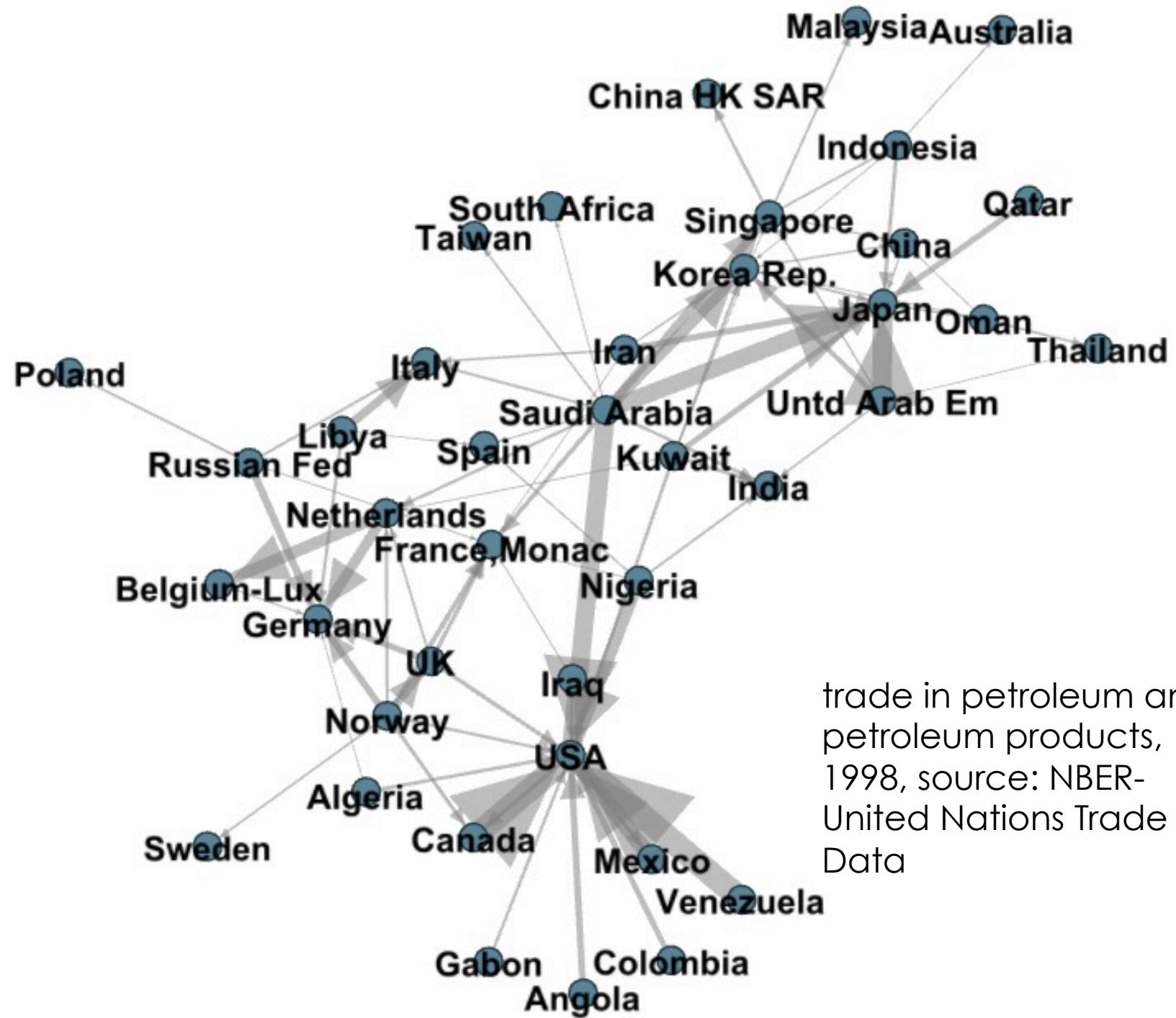
betweenness



closeness

review: indegree



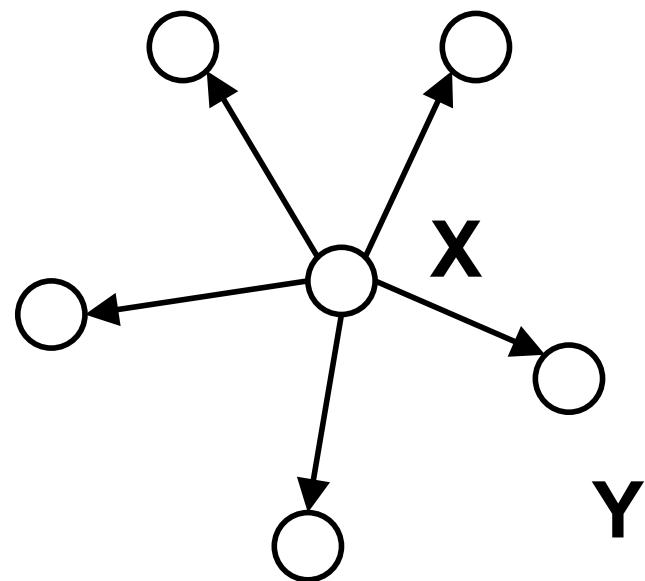


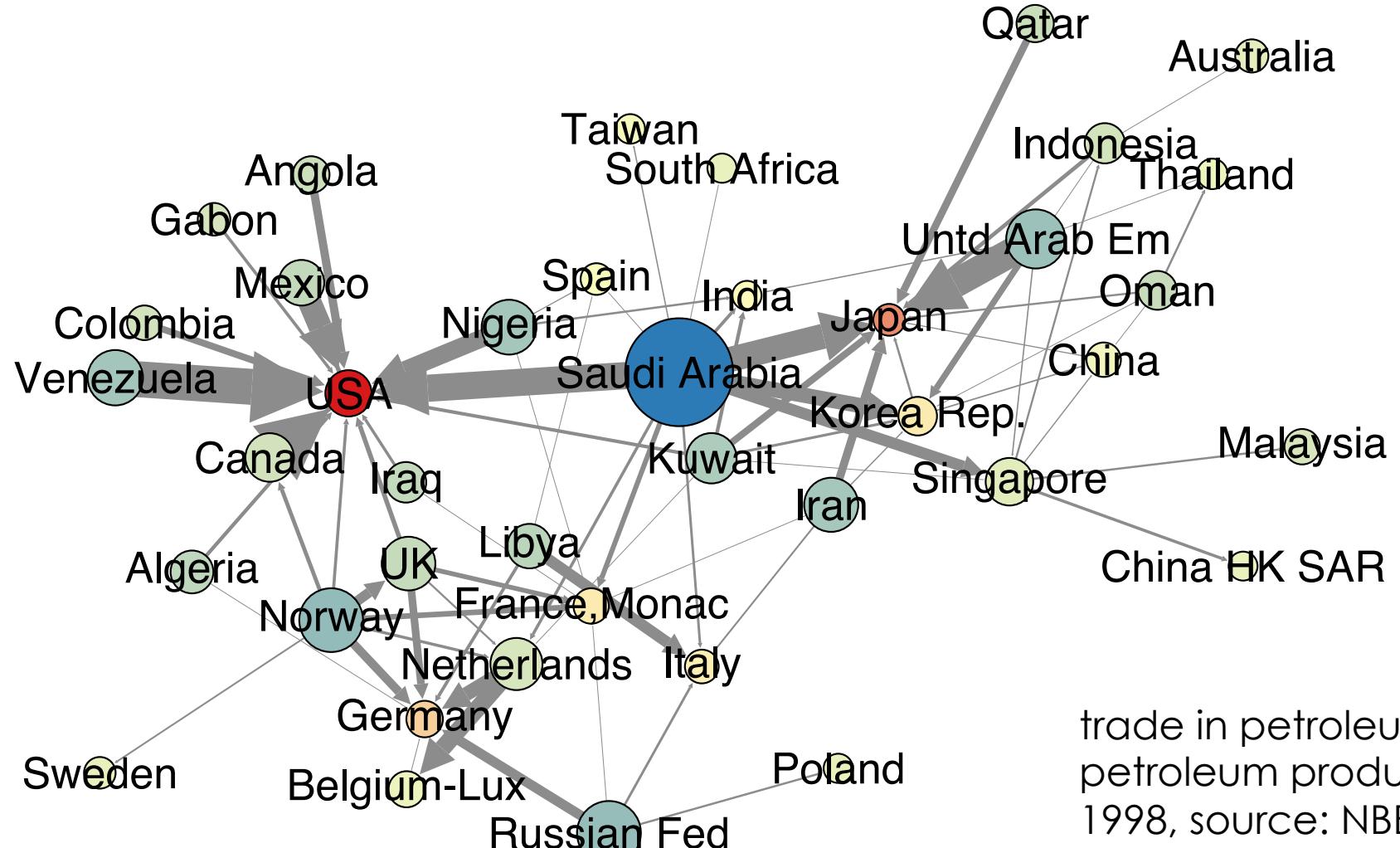
trade in petroleum and
petroleum products,
1998, source: NBER-
United Nations Trade
Data

Quiz Q:

- ❑ Which countries have high indegree
(import petroleum and petroleum products from many others)
 - ❑ Saudi Arabia
 - ❑ Japan
 - ❑ Iraq
 - ❑ USA
 - ❑ Venezuela

review: outdegree

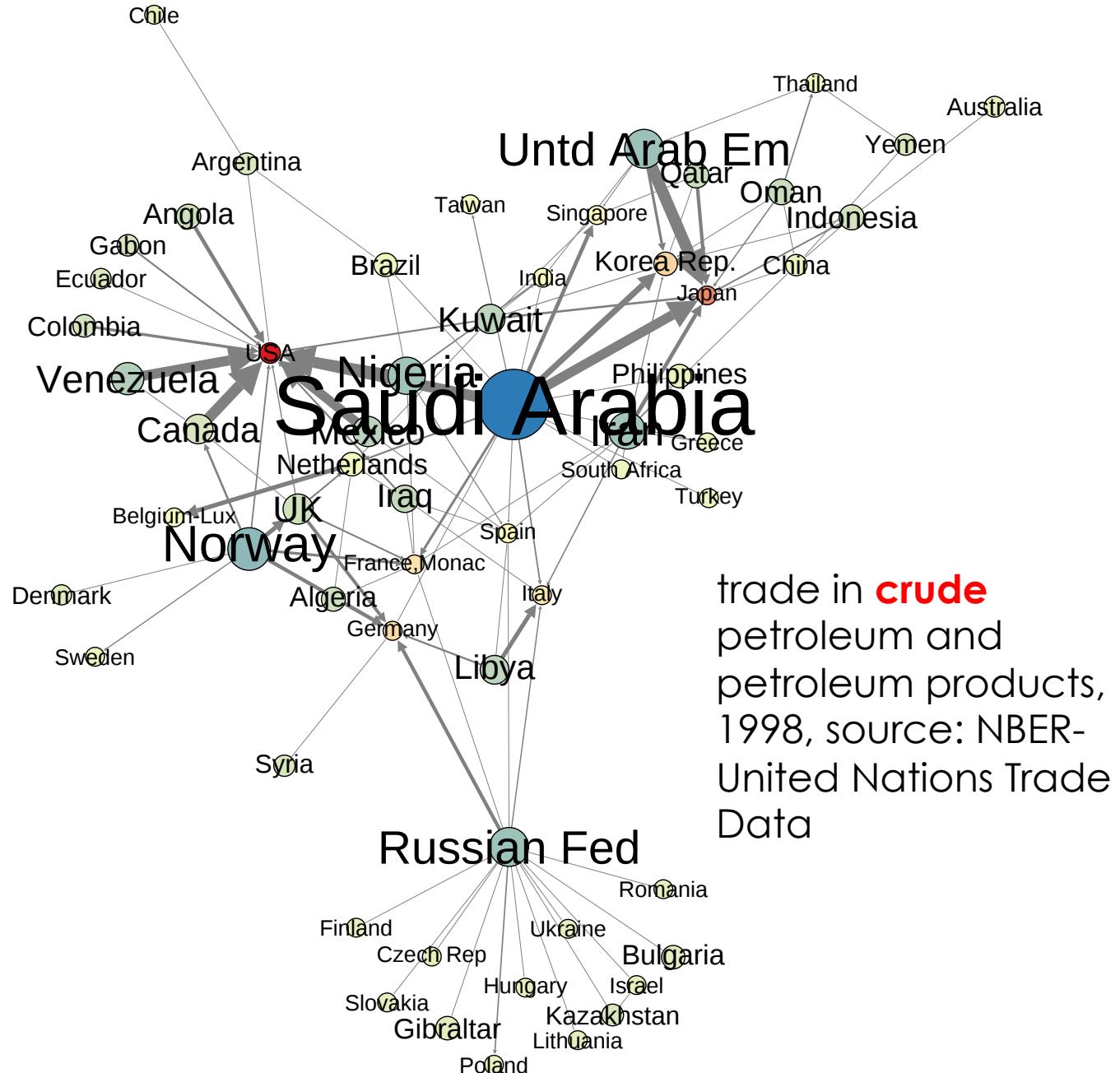




trade in petroleum and
petroleum products,
1998, source: NBER-
United Nations Trade
Data

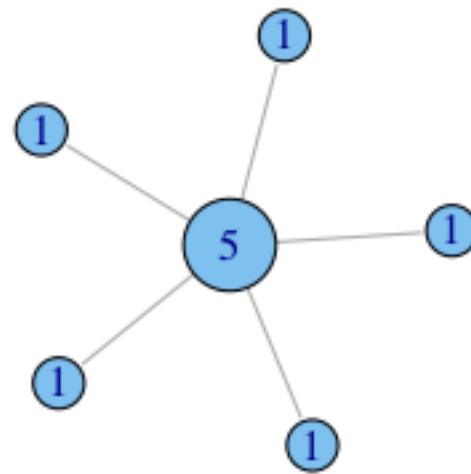
Quiz Q:

- ❑ Which country has low outdegree but exports a significant quantity (thickness of the edges represents \$\$ value of export) of petroleum products
 - ❑ Saudi Arabia
 - ❑ Japan
 - ❑ Iraq
 - ❑ USA
 - ❑ Venezuela



putting numbers to it

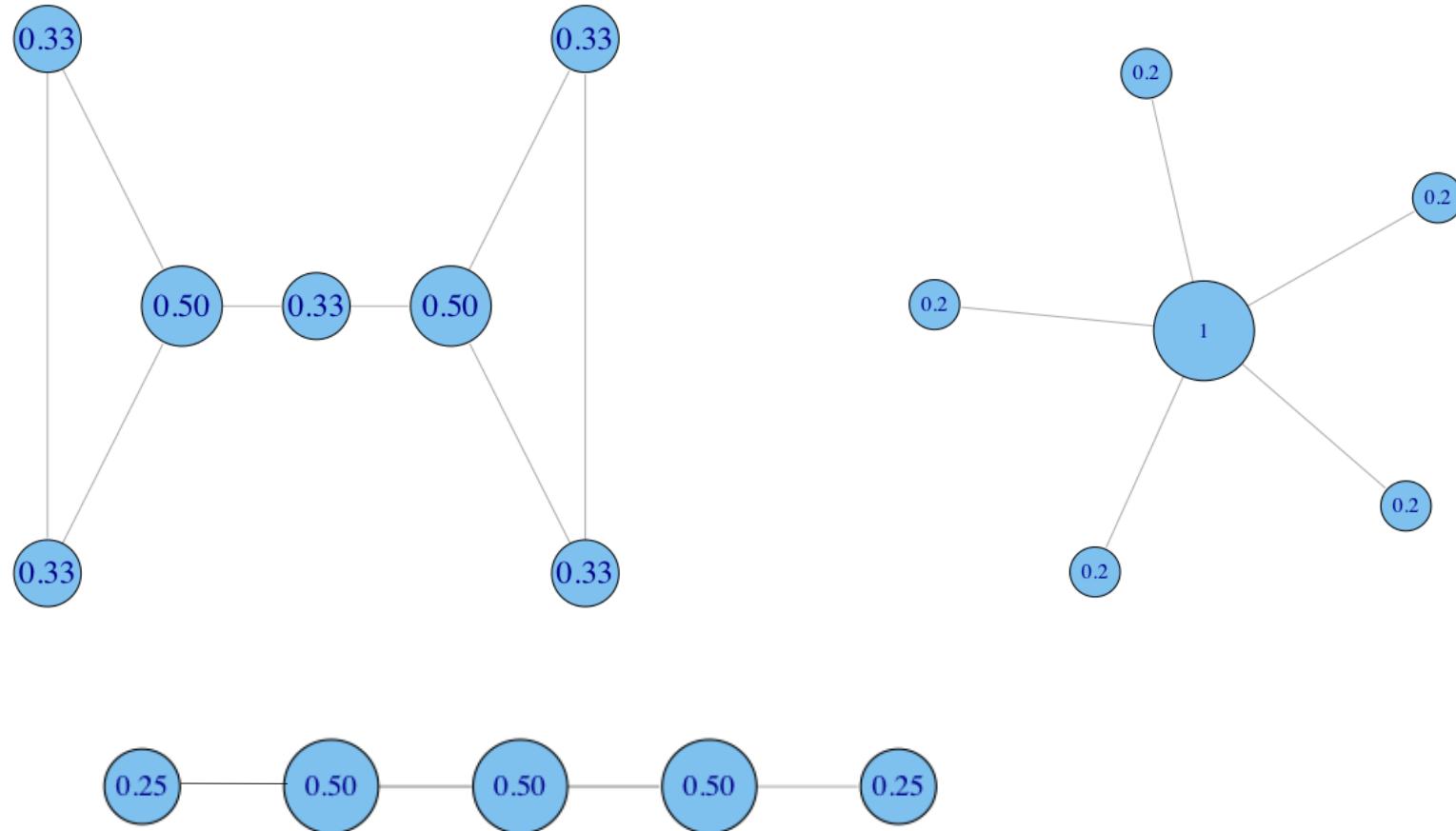
Undirected degree, e.g. nodes with more friends are more central.



Assumption: the connections that your friend has don't matter, it is what they can do directly that does (e.g. go have a beer with you, help you build a deck...)

normalization

divide degree by the max. possible, i.e. $(N-1)$



centralization: skew in distribution

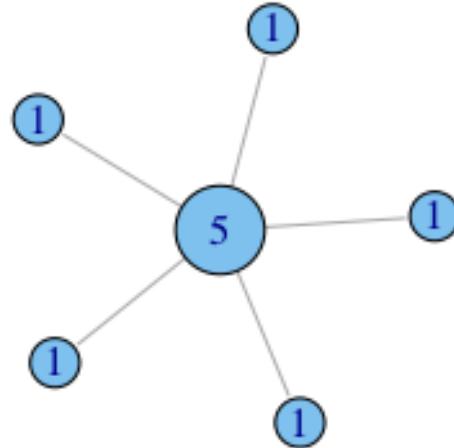
How much variation is there in the centrality scores among the nodes?

Freeman's general formula for centralization (can use other metrics, e.g. gini coefficient or standard deviation):

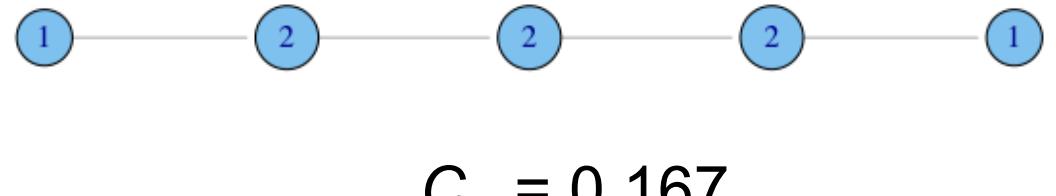
$$C_D = \frac{\sum_{i=1}^g [C_D(n^*) - C_D(i)]}{[(N-1)(N-2)]}$$

maximum value in the network

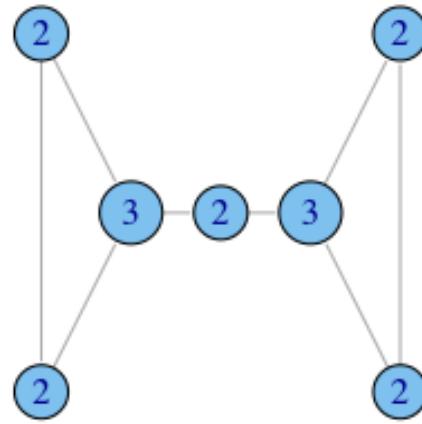
degree centralization examples



$$C_D = 1.0$$



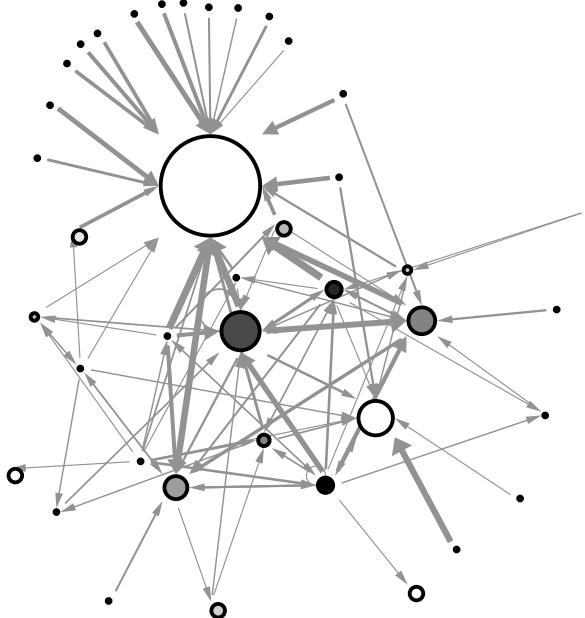
$$C_D = 0.167$$



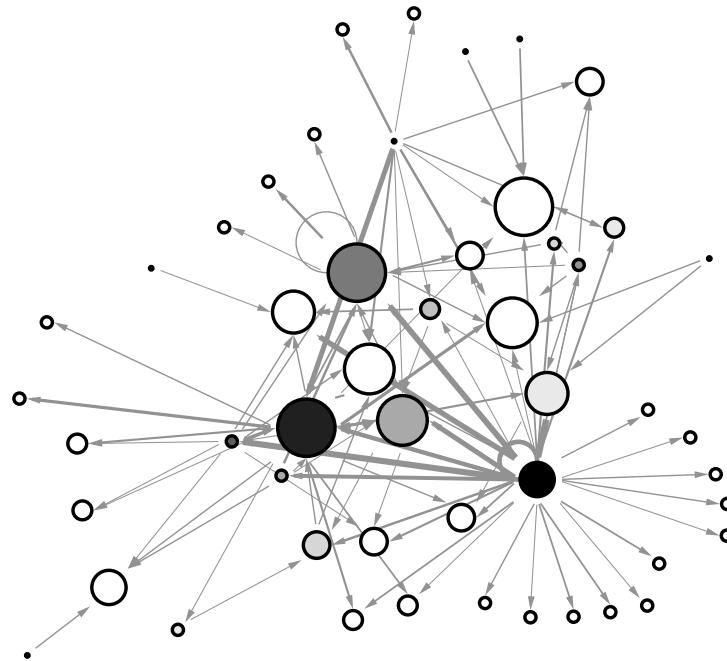
$$C_D = 0.167$$

real-world examples

example financial trading networks



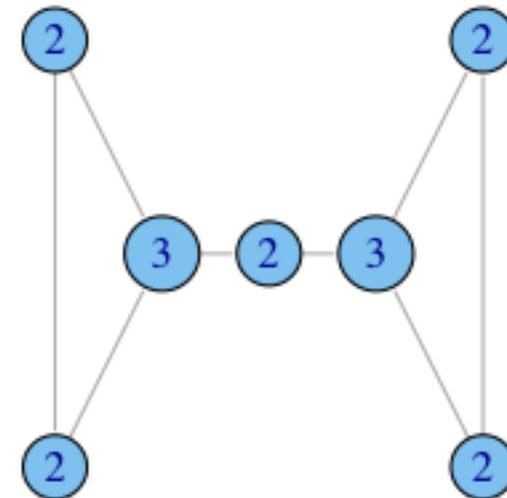
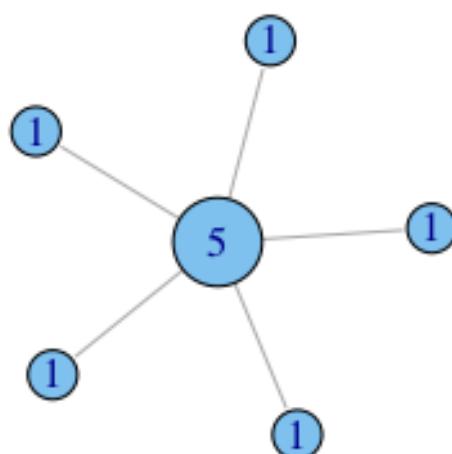
high in-centralization:
one node buying from
many others



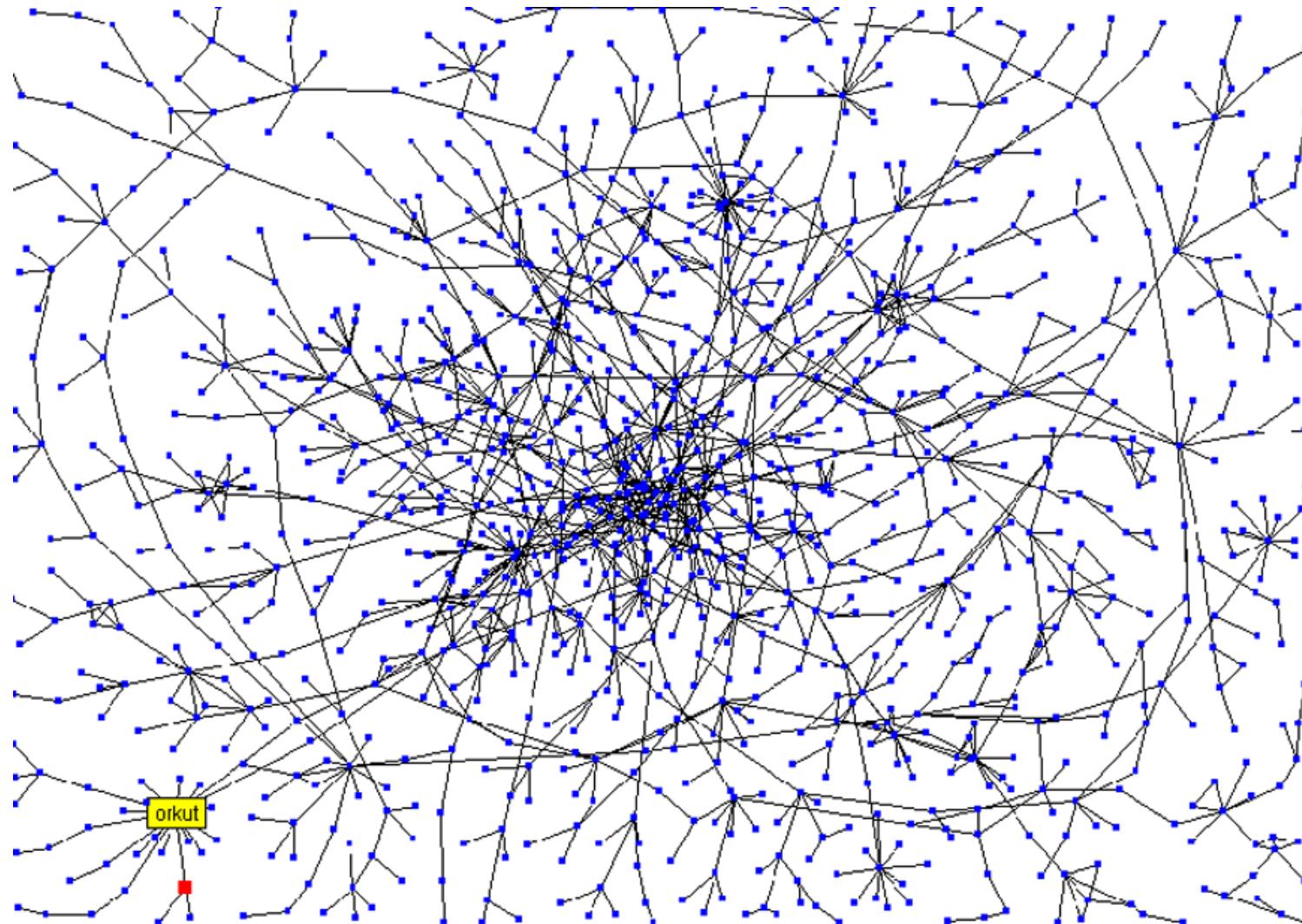
low in-centralization:
buying is more evenly
distributed

what does degree not capture?

In what ways does degree fail to capture centrality in the following graphs?

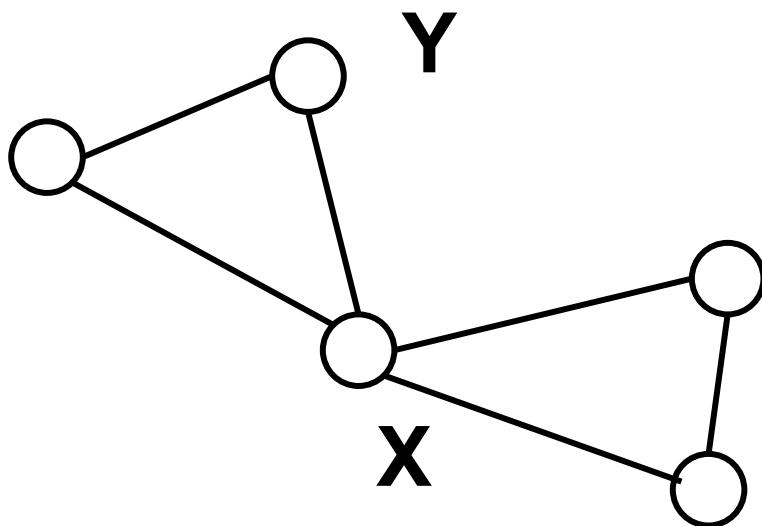


Stanford Social Web (ca. 1999)

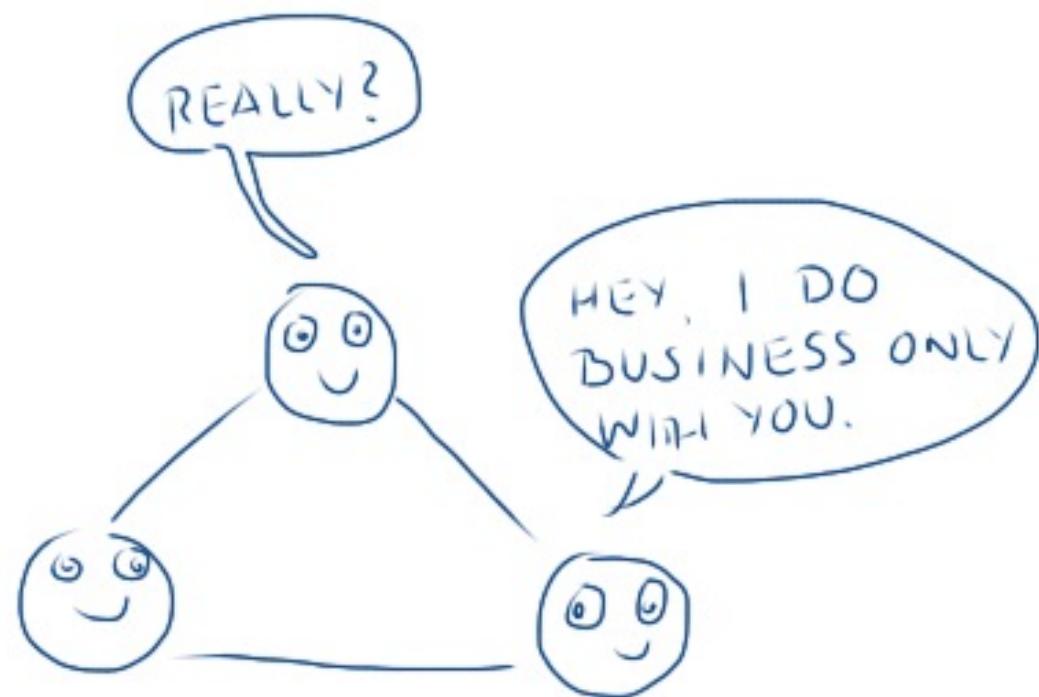


network of personal homepages at Stanford

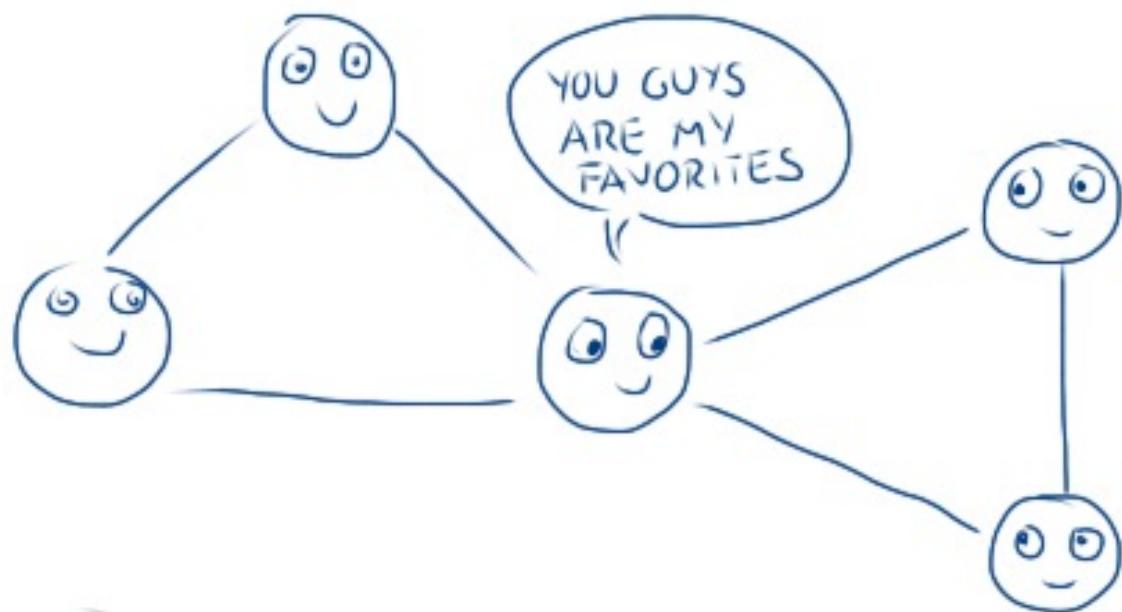
Brokerage not captured by degree



constraint

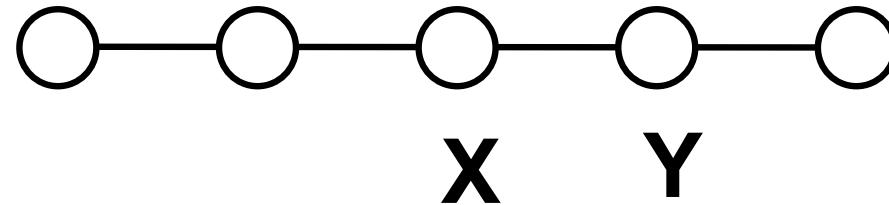


constraint



betweenness: capturing brokerage

- ❑ intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?



betweenness: definition

$$C_B(i) = \sum_{j < k} g_{jk}(i) / g_{jk}$$

Where g_{jk} = the number of shortest paths connecting jk
 $g_{jk}(i)$ = the number that actor i is on.

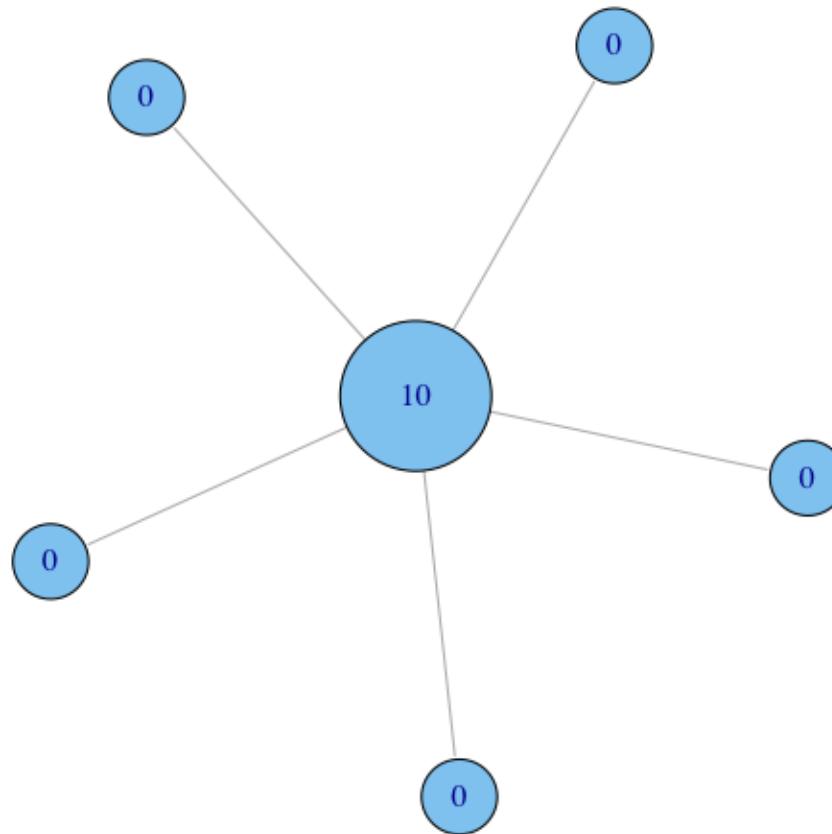
Usually normalized by:

$$C'_B(i) = C_B(i) / [(n - 1)(n - 2)/2]$$

number of pairs of vertices
excluding the vertex itself

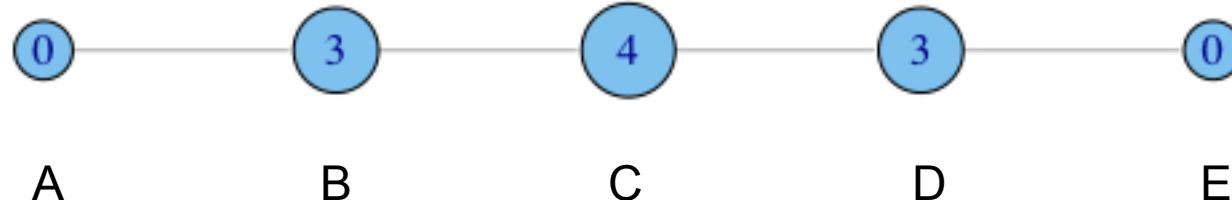
betweenness on toy networks

□ non-normalized version:



betweenness on toy networks

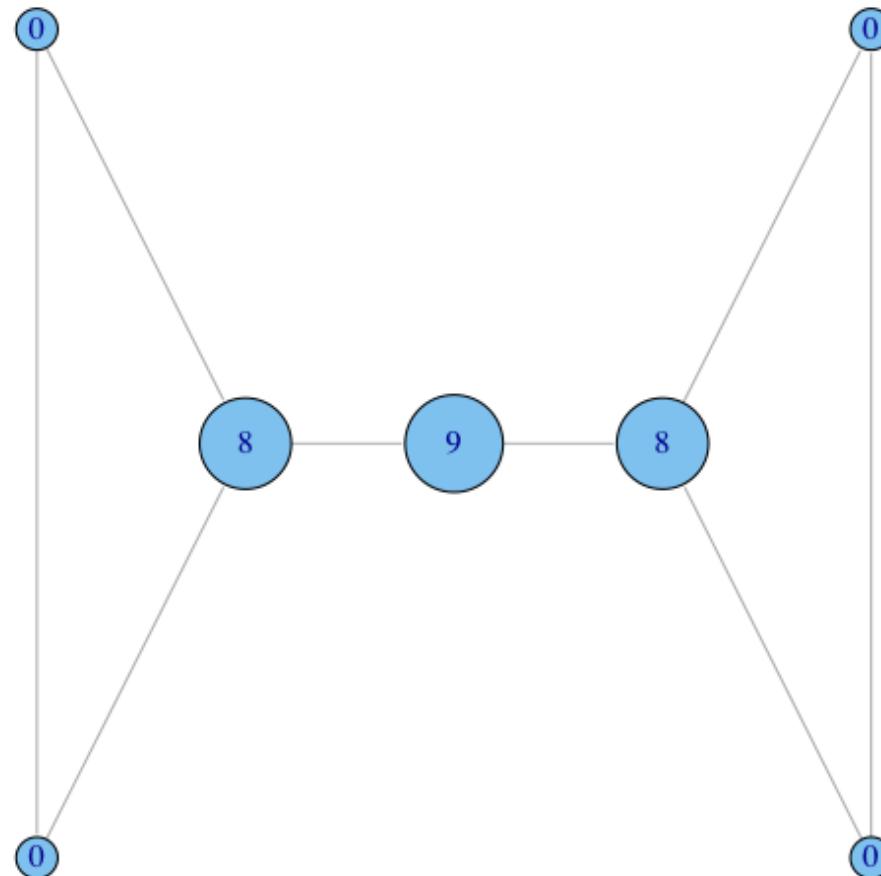
■ non-normalized version:



- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)
- note that there are no alternate paths for these pairs to take, so C gets full credit

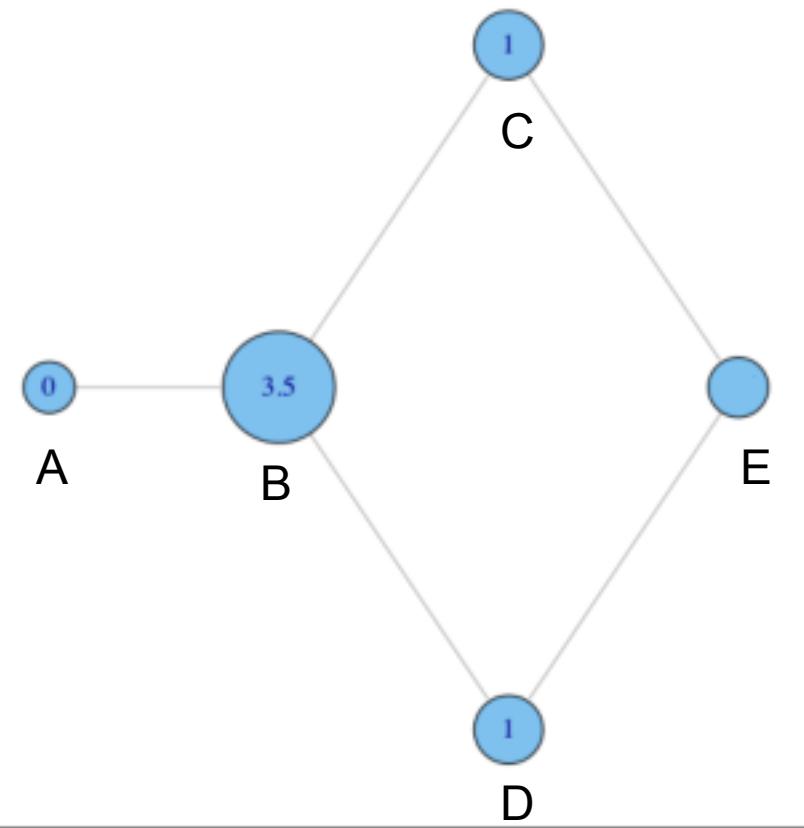
betweenness on toy networks

□ non-normalized version:



betweenness on toy networks

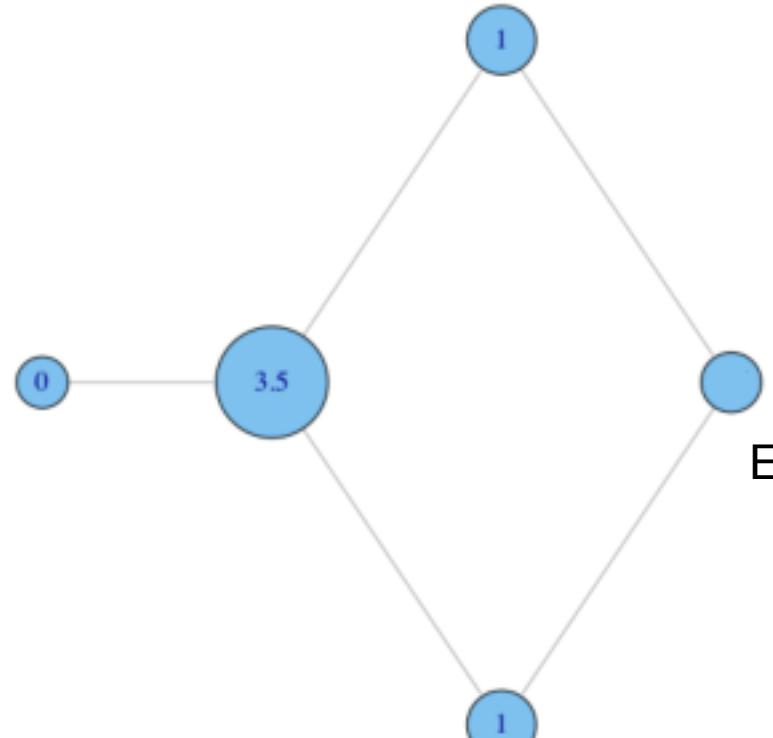
□ non-normalized version:



- why do C and D each have betweenness 1?
- They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
 - $\frac{1}{2} + \frac{1}{2} = 1$

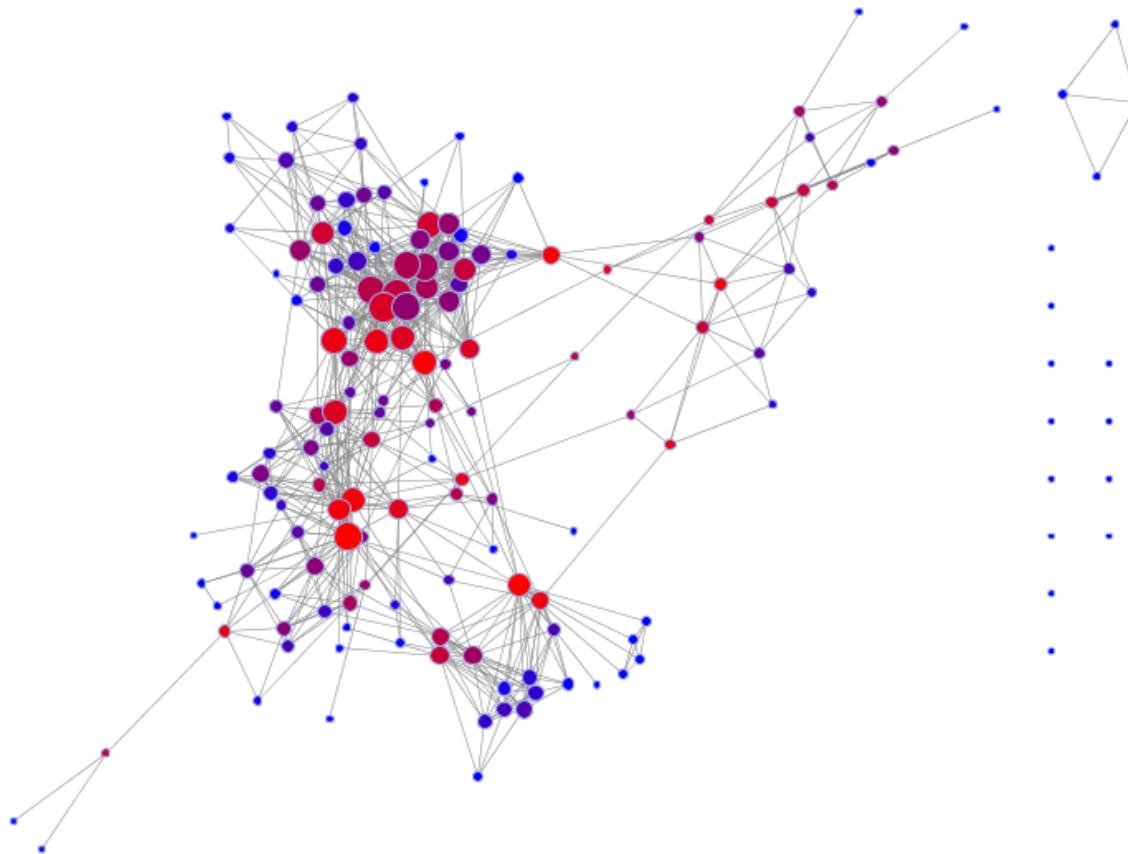
Quiz Question

❑ What is the betweenness of node E?



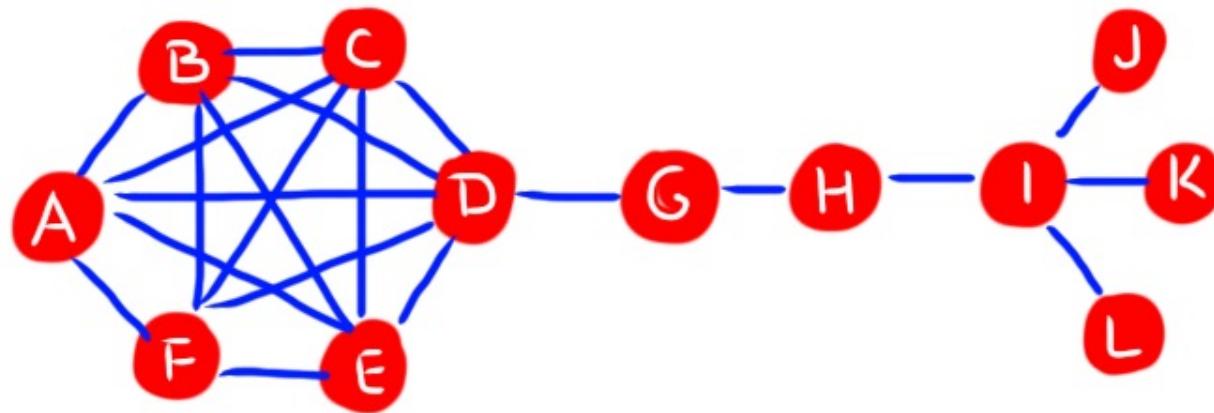
betweenness: example

Lada's old Facebook network: nodes are sized by degree, and colored by betweenness.



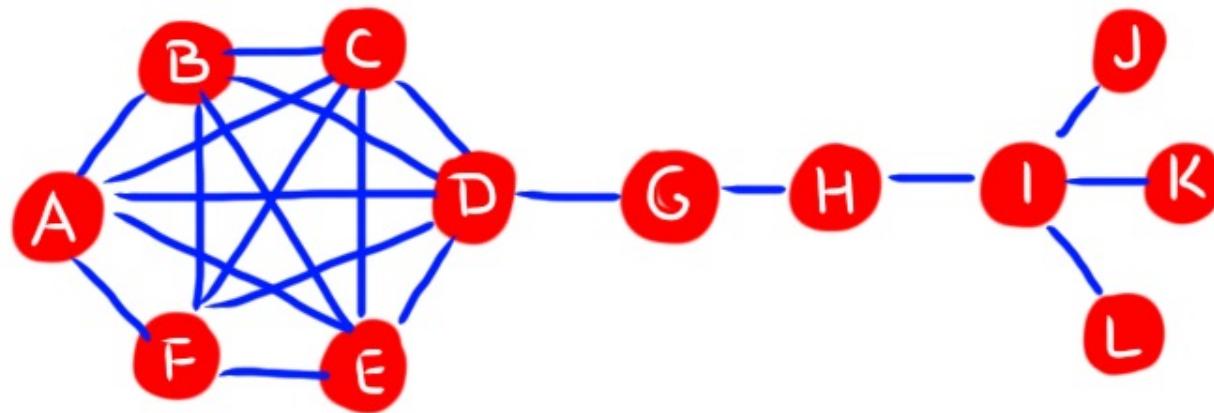
Quiz Q:

- Find a node that has high betweenness but low degree



Quiz Q:

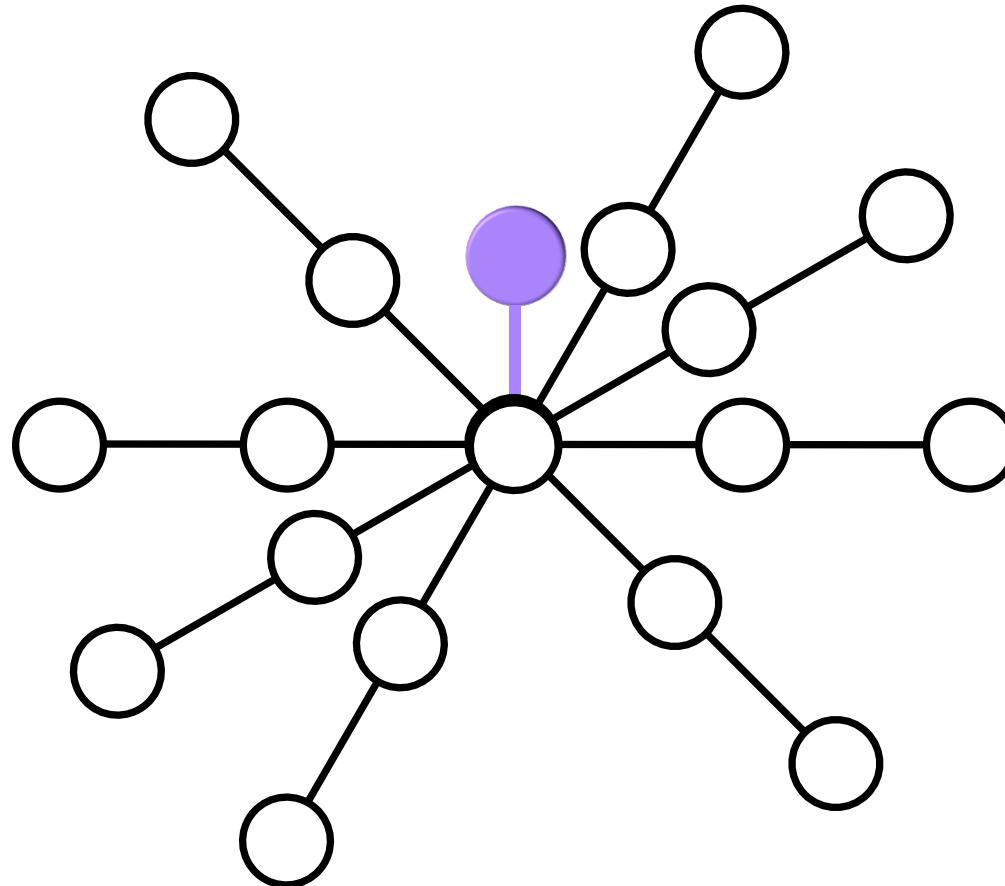
- Find a node that has low betweenness but high degree



closeness

- ❑ What if it's not so important to have many direct friends?
- ❑ Or be “between” others
- ❑ But one still wants to be in the “middle” of things, not too far from the center

need not be in a brokerage position



closeness: definition

Closeness is based on the length of the average shortest path between a node and all other nodes in the network

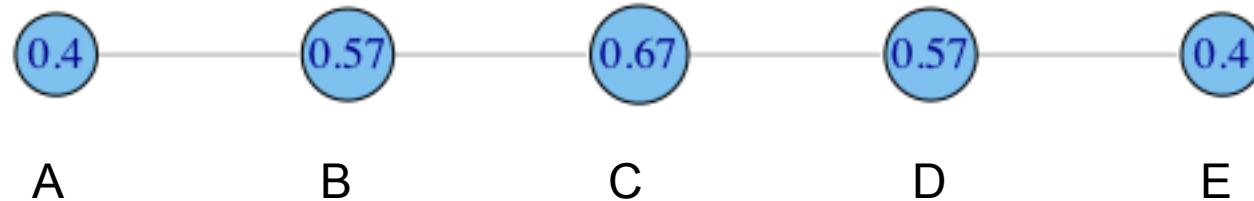
Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^N d(i,j) \right]^{-1}$$

Normalized Closeness Centrality

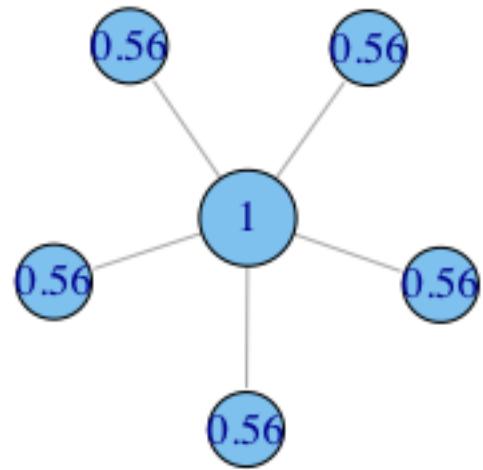
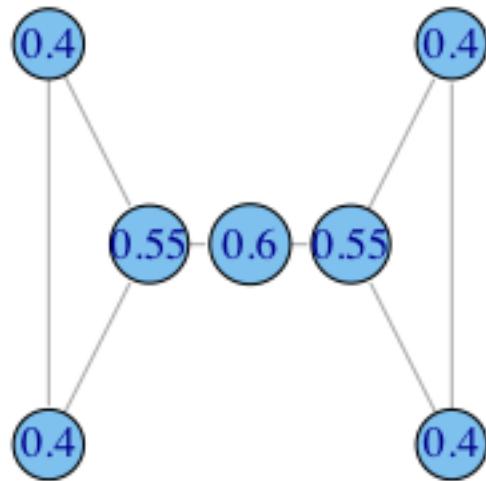
$$C'_c(i) = (C_c(i)) / (N - 1)$$

closeness: toy example



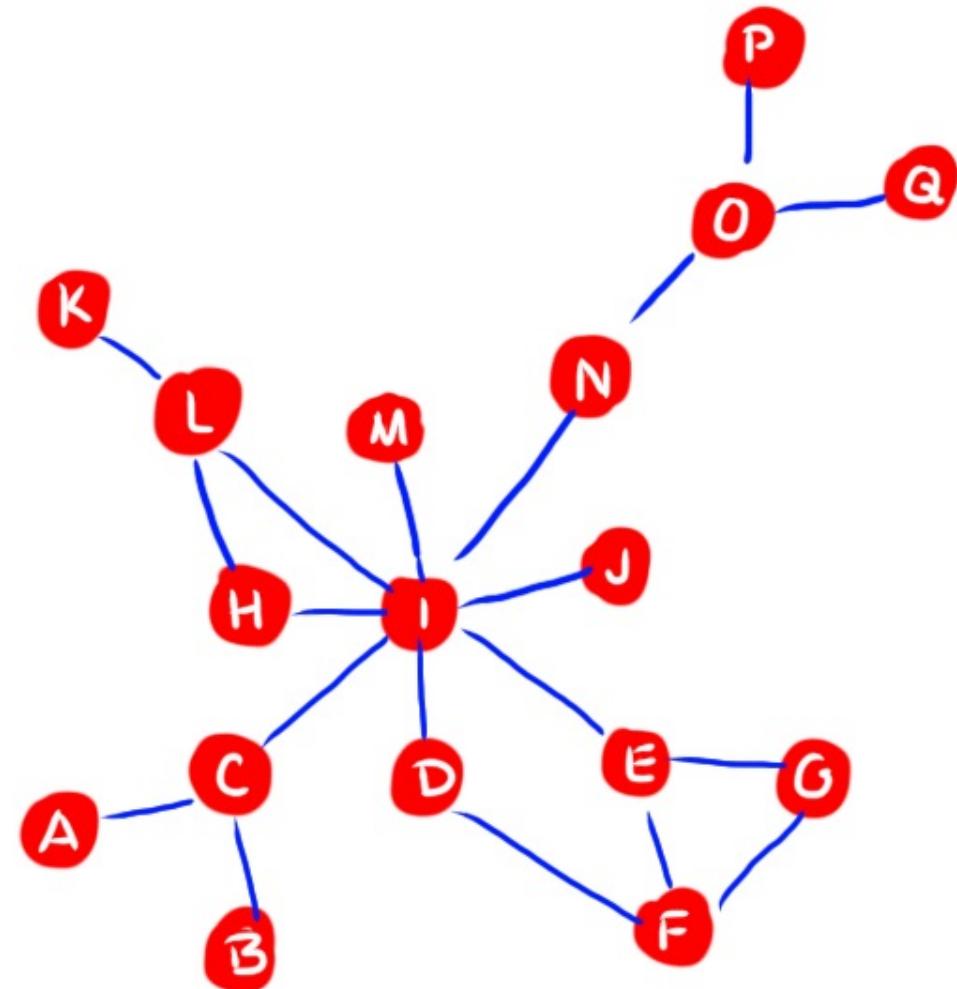
$$C_c(A) = \left[\frac{\sum_{j=1}^N d(A, j)}{N - 1} \right]^{-1} = \left[\frac{1 + 2 + 3 + 4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

closeness: more toy examples

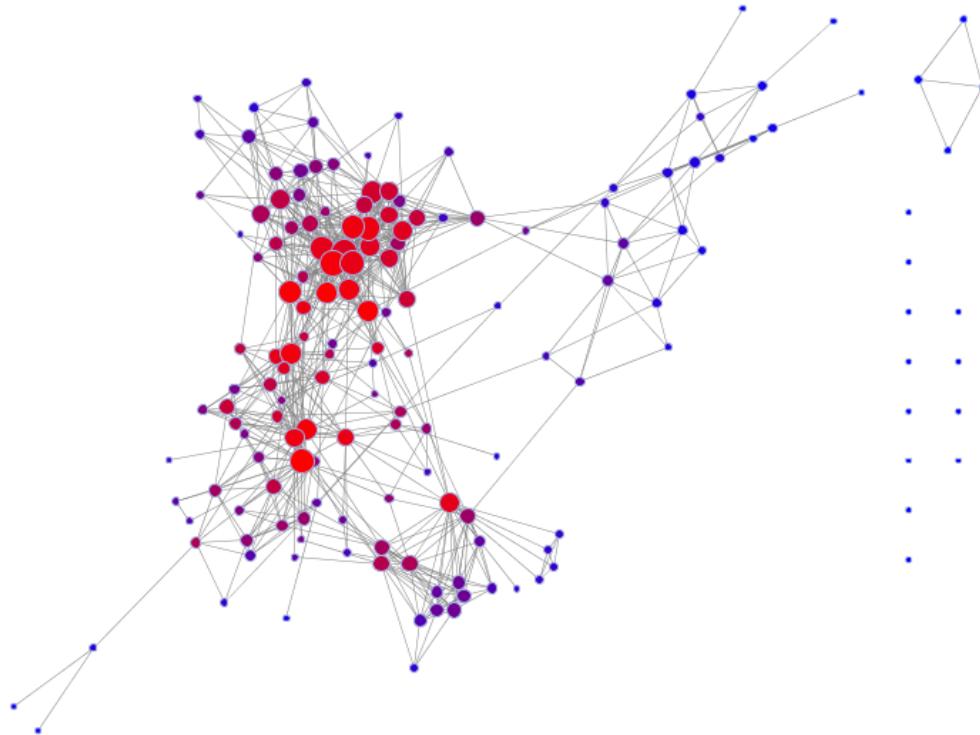


Quiz Q:

Which node has relatively high degree but low closeness?



Closeness and Lada's fb network

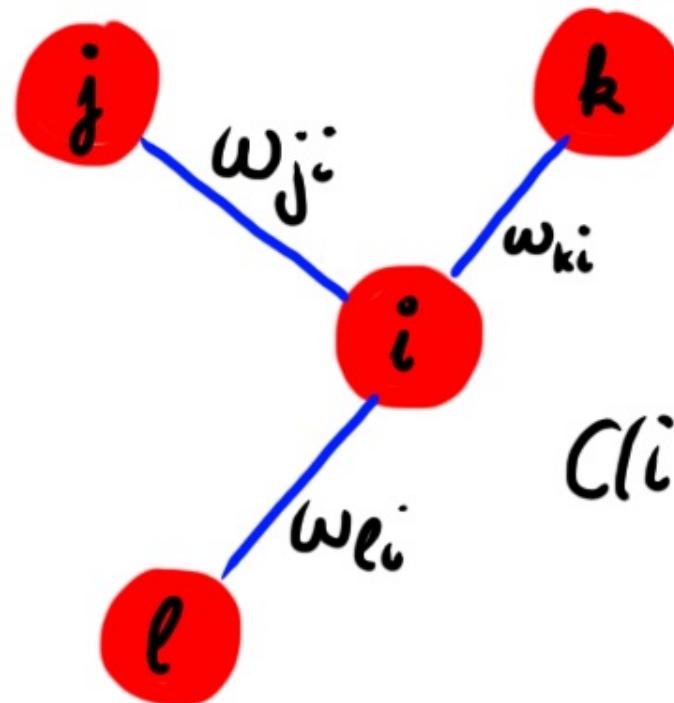


□ **degree** (number of connections) denoted by size

□ **closeness** (length of shortest path to all others) denoted by color

Eigenvector centrality

- How central you are depends on how central your neighbors are



$$C(i) = w_{ji} \cdot C(j) + w_{ki} \cdot C(k) + w_{li} \cdot C(l)$$

Bonacich eigenvector centrality

$$c_i(\beta) = \sum_j (\alpha + \beta c_j) A_{ji}$$

$$c(\beta) = \alpha(I - \beta A)^{-1} A \mathbf{1}$$

- α is a normalization constant
- β determines how important the centrality of your neighbors is
- A is the adjacency matrix (can be weighted)
- I is the identity matrix (1s down the diagonal, 0 off-diagonal)
- $\mathbf{1}$ is a matrix of all ones.

Bonacich Power Centrality: attenuation factor β

small $\beta \rightarrow$ high attenuation

only your immediate friends matter, and
their importance is factored in only a bit

high $\beta \rightarrow$ low attenuation

global network structure matters (your
friends, your friends' of friends etc.)

$= 0$ yields simple degree centrality

$$c_i(\beta) = \sum_j (\alpha \boxed{}) A_{ji}$$

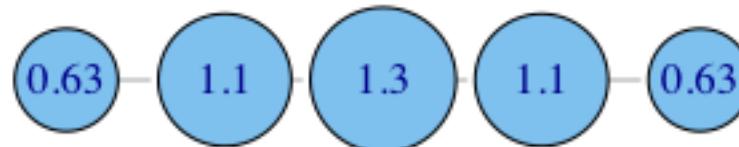
Bonacich Power Centrality: attenuation factor β

If $\beta > 0$, nodes have higher centrality when they have edges to other central nodes.

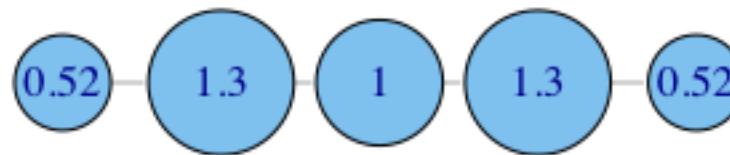
If $\beta < 0$, nodes have higher centrality when they have edges to less central nodes.

Bonacich Power Centrality: examples

$\beta = .25$



$\beta = -.25$



Why does the middle node have lower centrality than its neighbors when β is negative?

Centrality in directed networks

- ❑ WWW
- ❑ food webs
- ❑ population dynamics
- ❑ influence
- ❑ hereditary
- ❑ citation
- ❑ transcription regulation networks
- ❑ neural networks

Betweenness centrality in directed networks

- We now consider the fraction of all directed paths between any two vertices that pass through a node

betweenness of vertex i

paths between j and k that pass through i

all paths between j and k

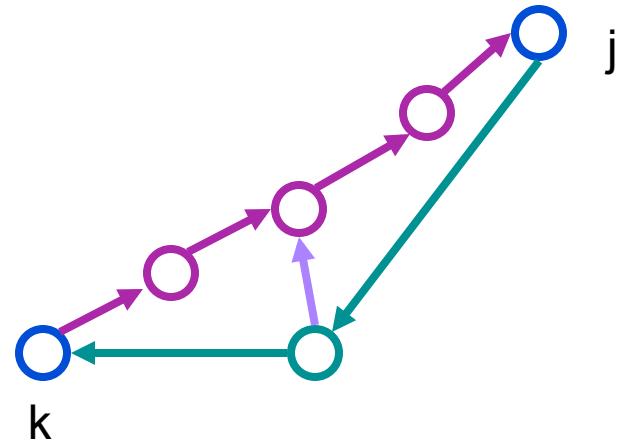
$$C_B(i) = \sum_{j,k} g_{jk}(i) / g_{jk}$$

- Only modification: when normalizing, we have $(N-1)*(N-2)$ instead of $(N-1)*(N-2)/2$, because we have twice as many ordered pairs as unordered pairs

$$C'_B(i) = C_B(i) / [(N-1)(N-2)]$$

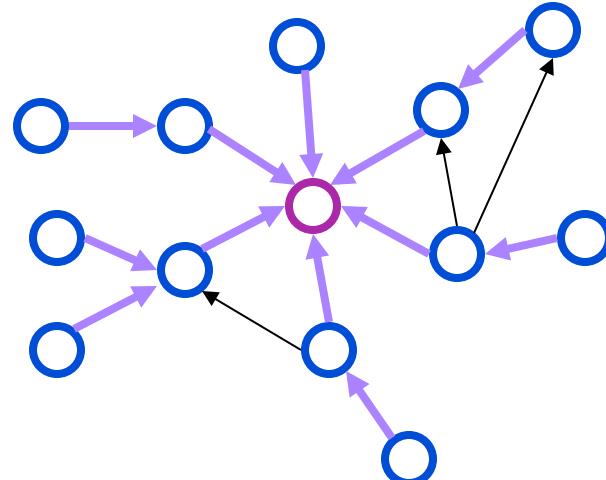
Directed geodesics

- A node does not necessarily lie on a geodesic from j to k if it lies on a geodesic from k to j



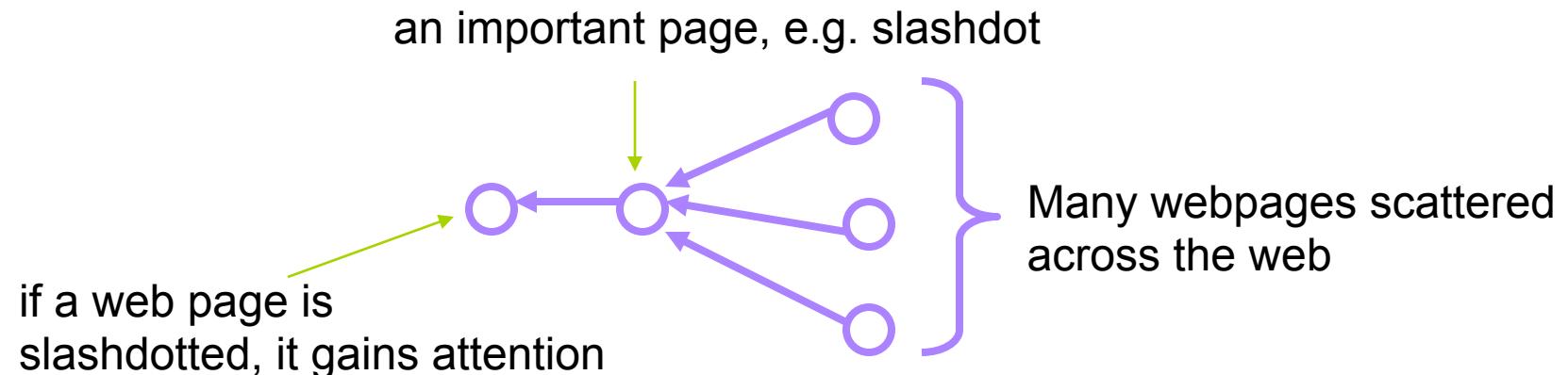
Directed closeness centrality

- ❑ choose a direction
 - ❑ in-closeness (e.g. prestige in citation networks)
 - ❑ out-closeness
- ❑ usually consider only vertices from which the node i in question can be reached



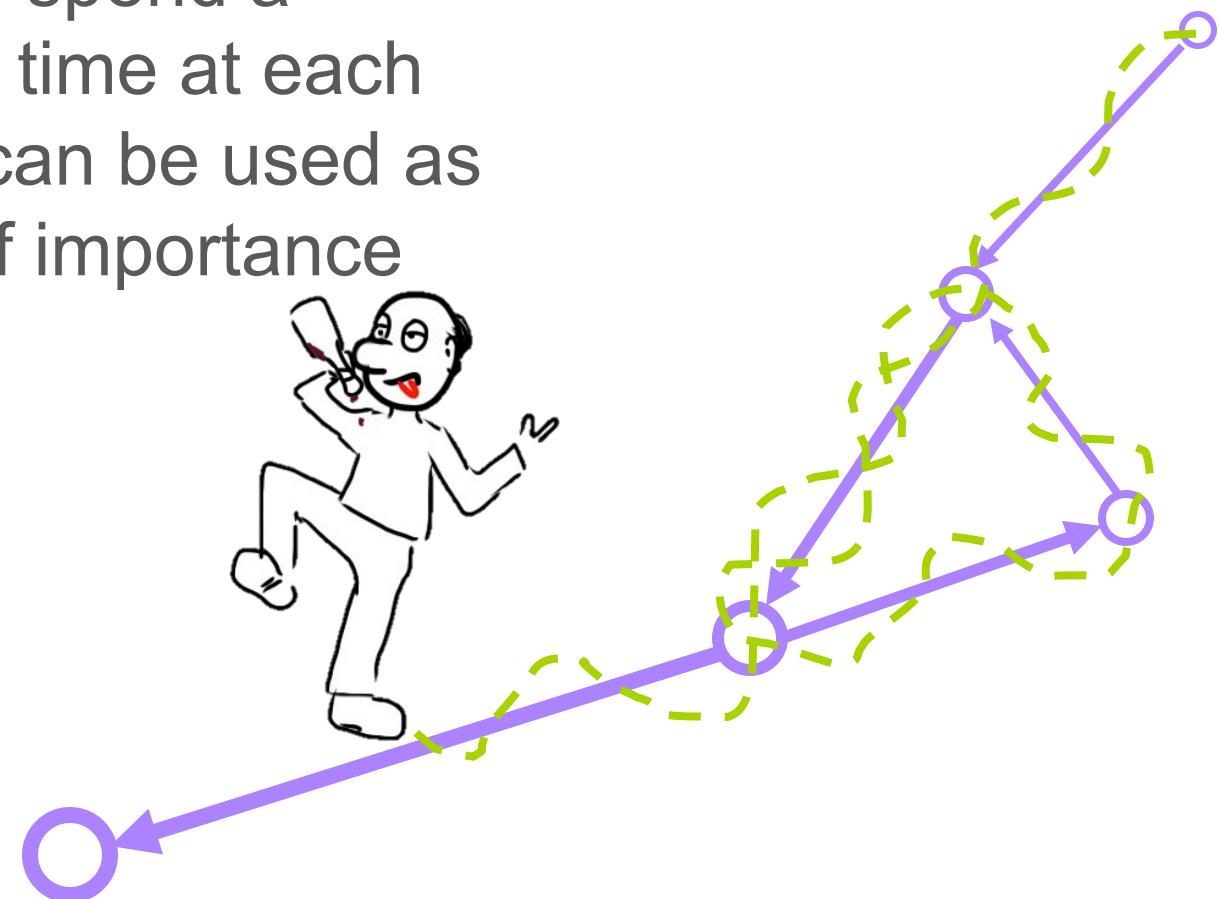
Eigenvector centrality in directed networks

- ❑ PageRank brings order to the Web:
 - ❑ it's not just the pages that point to you, but how many pages point to those pages, etc.
 - ❑ more difficult to artificially inflate centrality with a recursive definition



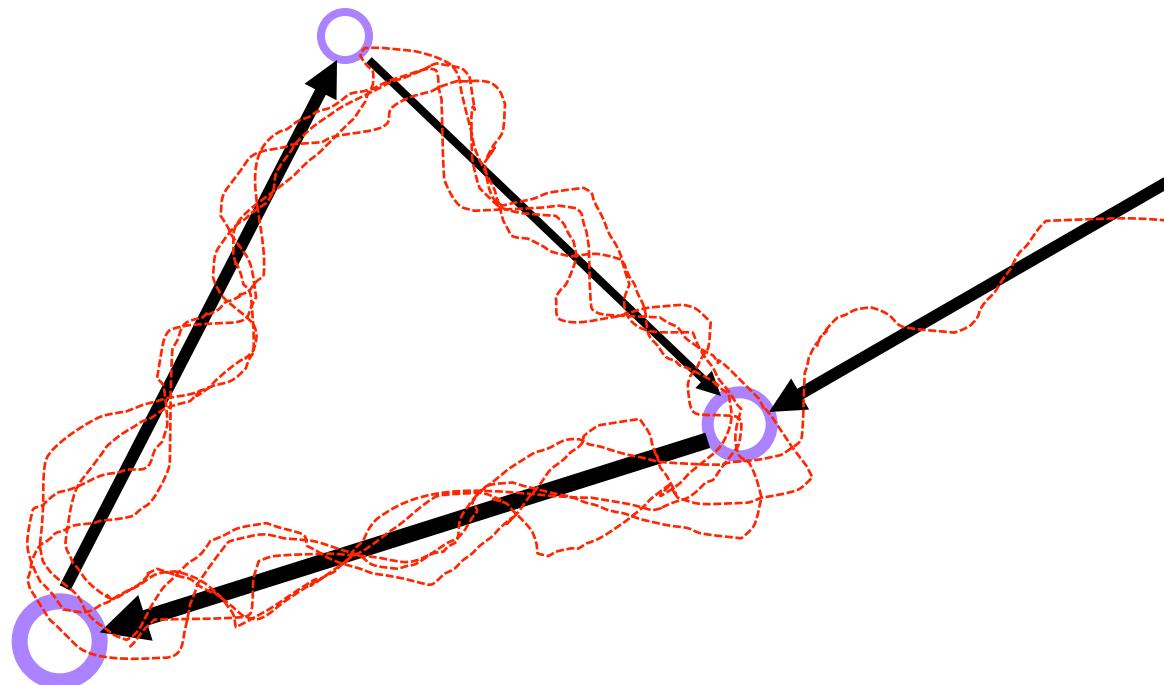
Ranking pages by tracking a drunk

- ❑ A random walker following edges in a network for a very long time will spend a proportion of time at each node which can be used as a measure of importance



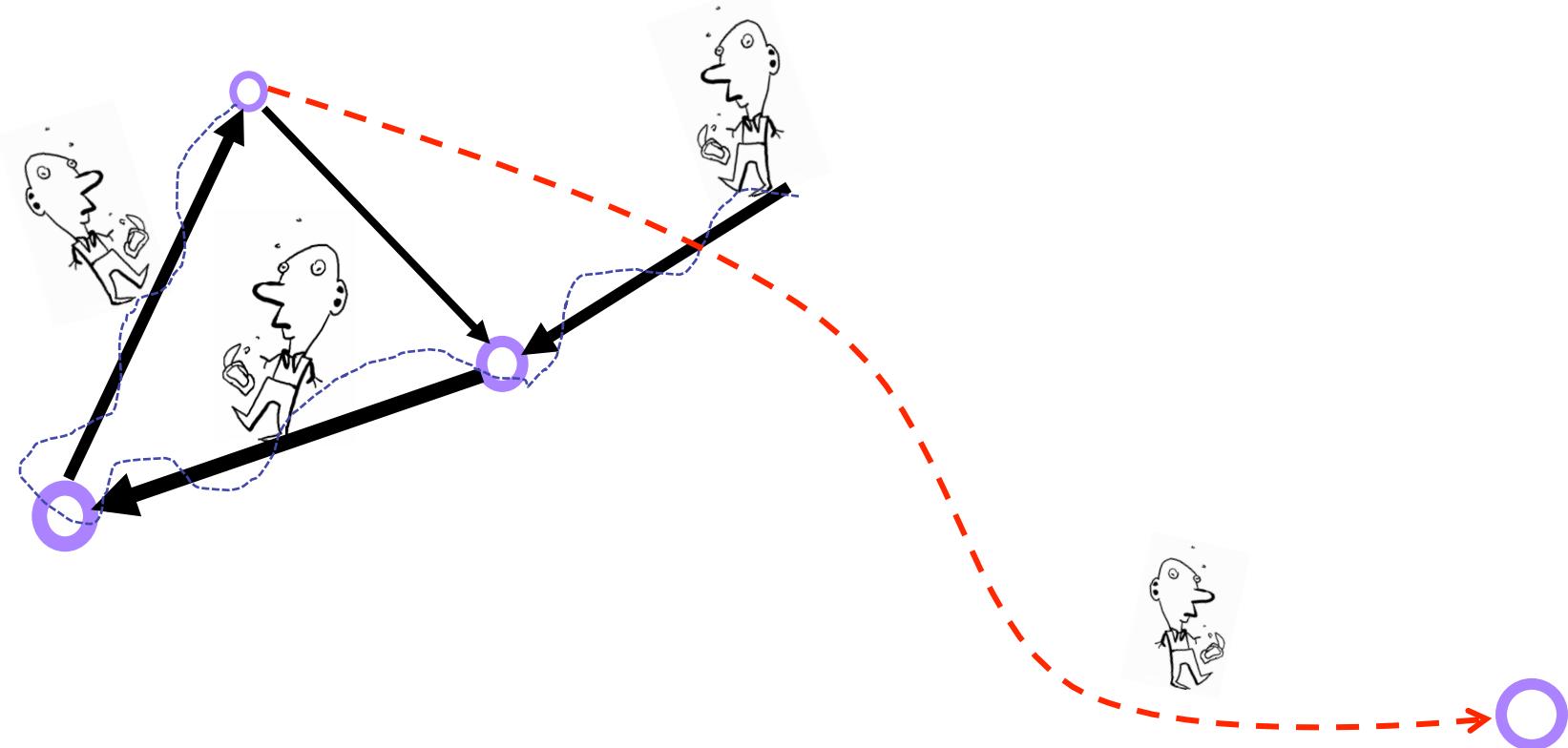
Trapping a drunk

- Problem with pure random walk metric:
 - Drunk can be “trapped” and end up going in circles

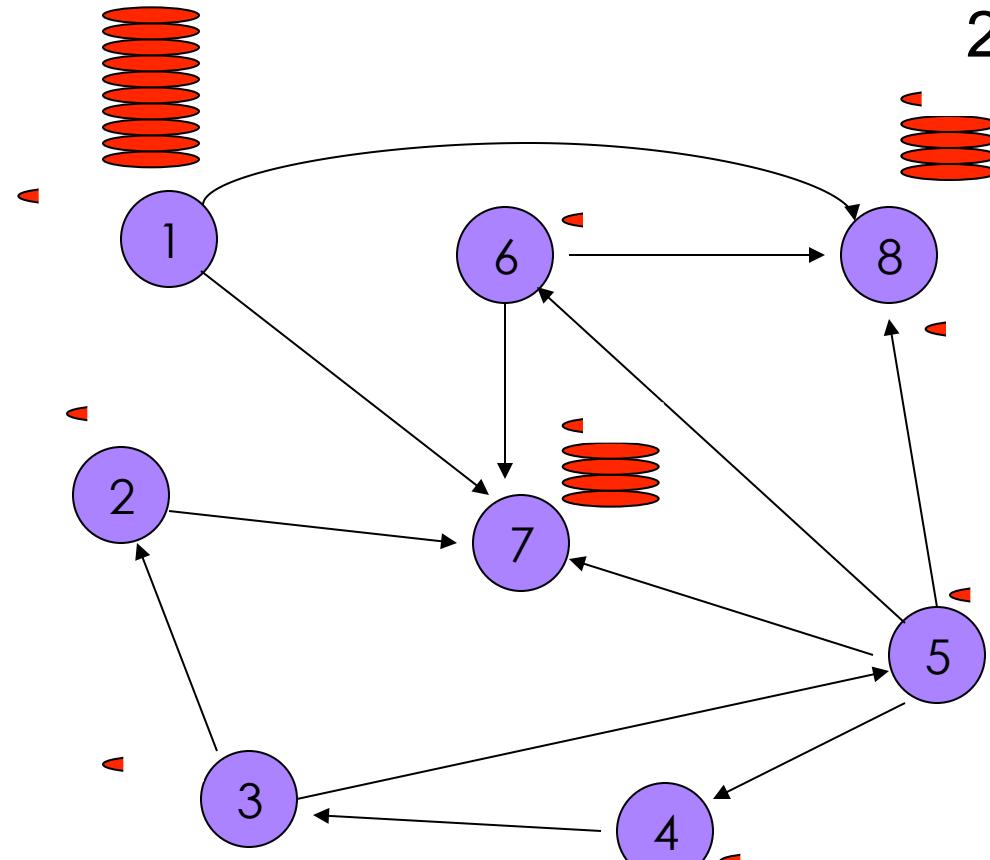


Ingenuity of the PageRank algorithm

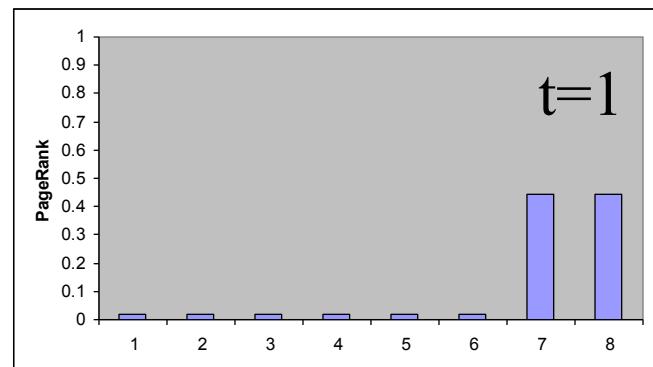
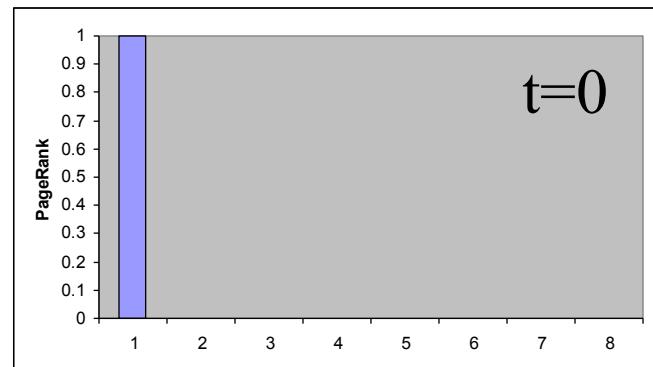
- ❑ Allow drunk to teleport with some probability
 - ❑ e.g. random websurfer follows links for a while, but with some probability teleports to a “random” page (bookmarked page or uses a search engine to start anew)



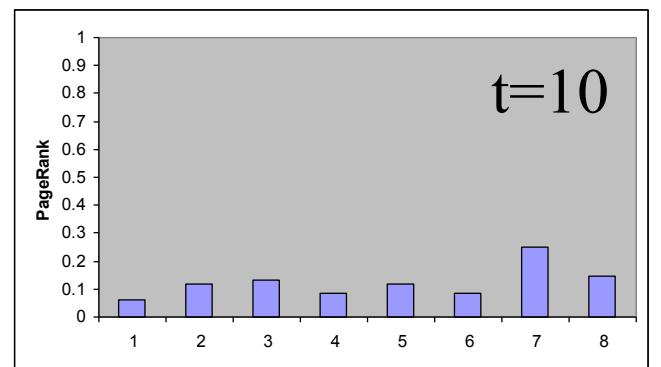
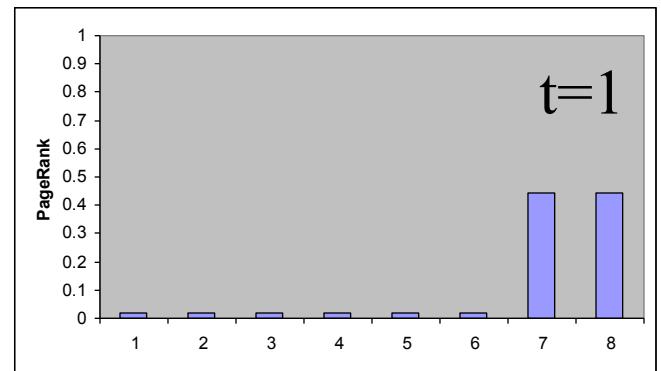
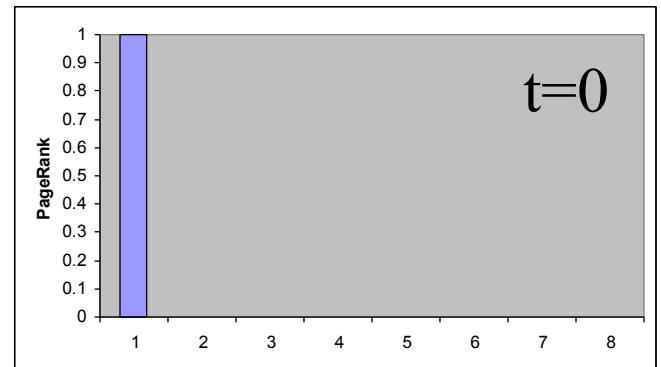
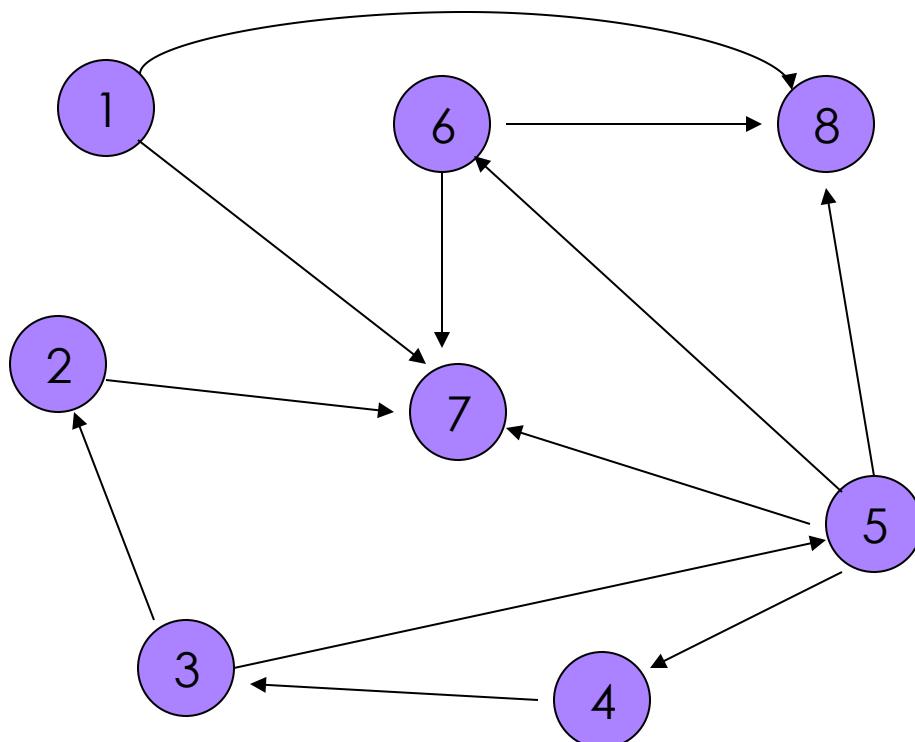
example: probable location of random walker after 1 step



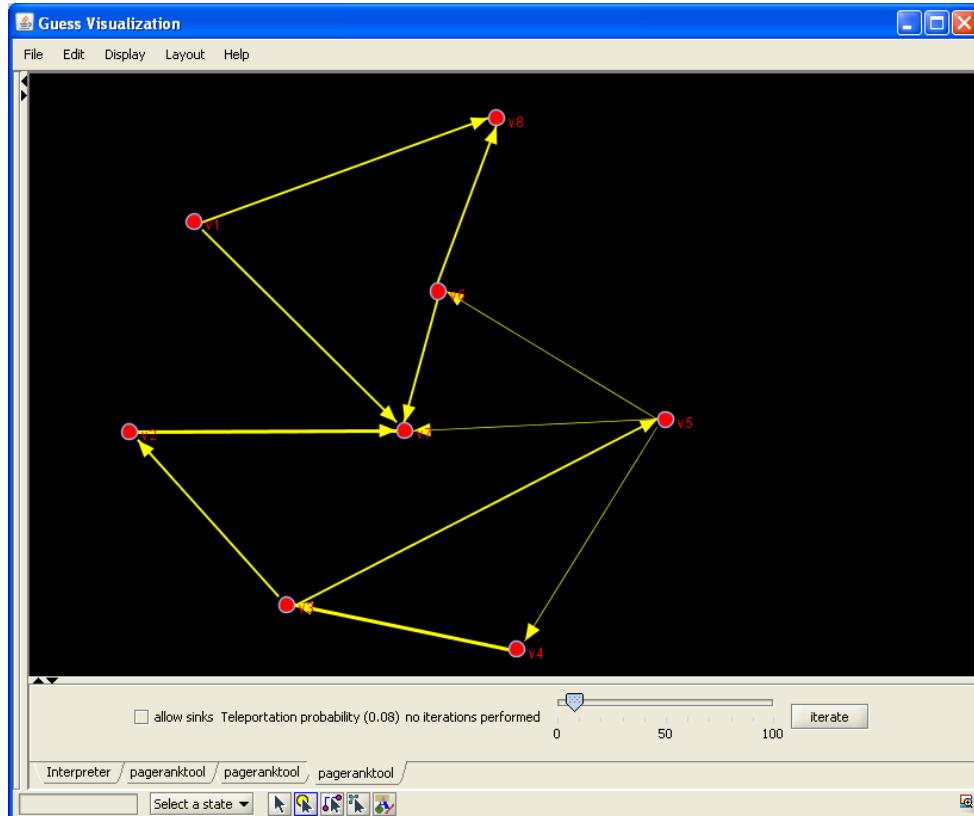
20% teleportation probability



example: probable location of random walker after 10 steps



Quiz Q:



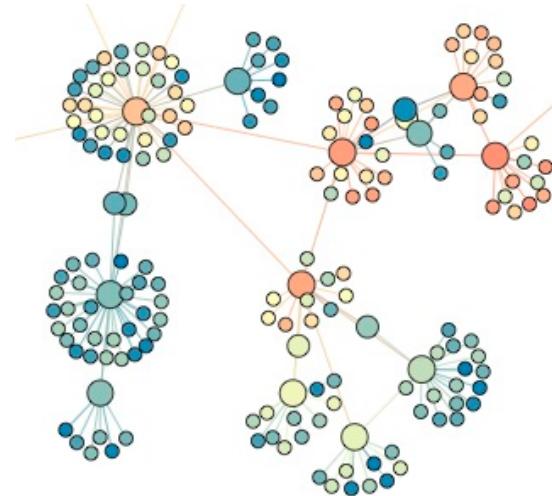
- ❑ What happens to the relative PageRank scores of the nodes as you increase the teleportation probability?
 - ❑ they equalize
 - ❑ they diverge
 - ❑ they are unchanged

<http://www.ladamic.com/netlearn/GUESS/pagerank.html>

wrap up

❑ Centrality

- ❑ many measures: degree, betweenness, closeness, eigenvector
- ❑ may be unevenly distributed
 - ❑ measure via distributions and centralization
- ❑ in directed networks
 - ❑ indegree, outdegree, PageRank
- ❑ consequences:
 - ❑ benefits & risks (Baker & Faulkner)
 - ❑ information flow & productivity (Aral & Van Alstyne)

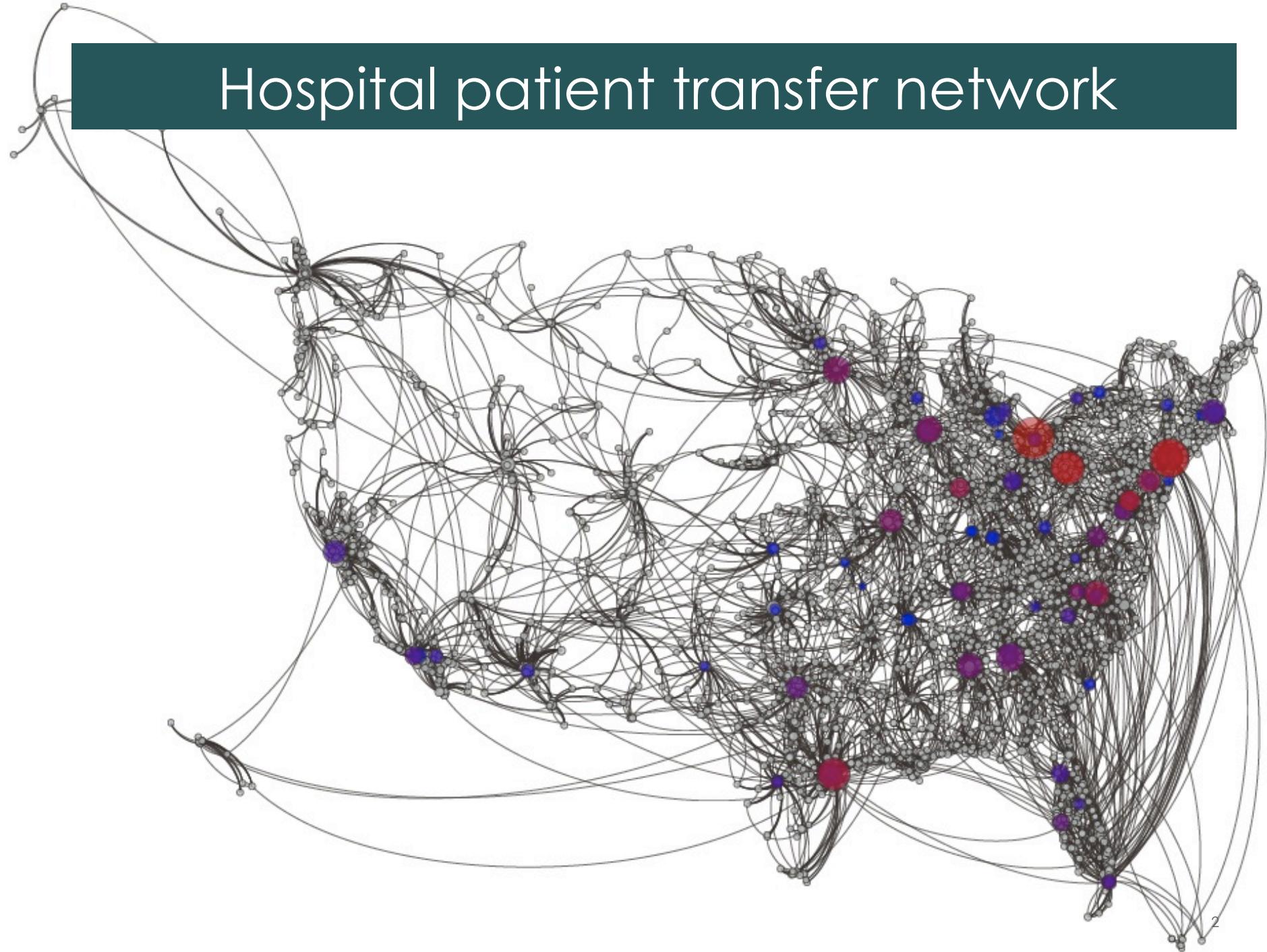


SNA 3C: Applications of network centrality

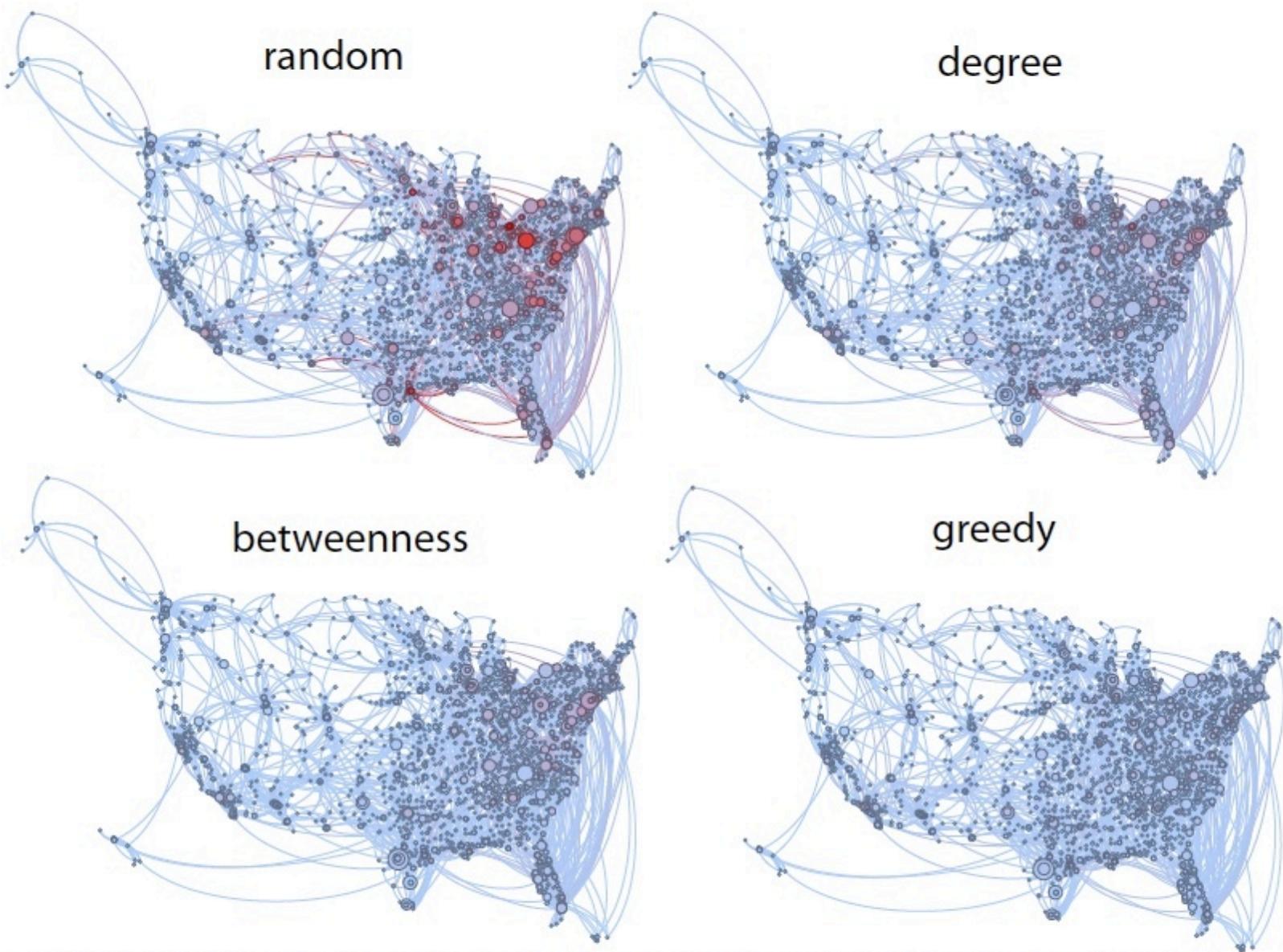
Lada Adamic



Hospital patient transfer network



Infection prevention strategies in a hospital patient transfer network



probability of becoming infected within a year from a random starting hospital

0

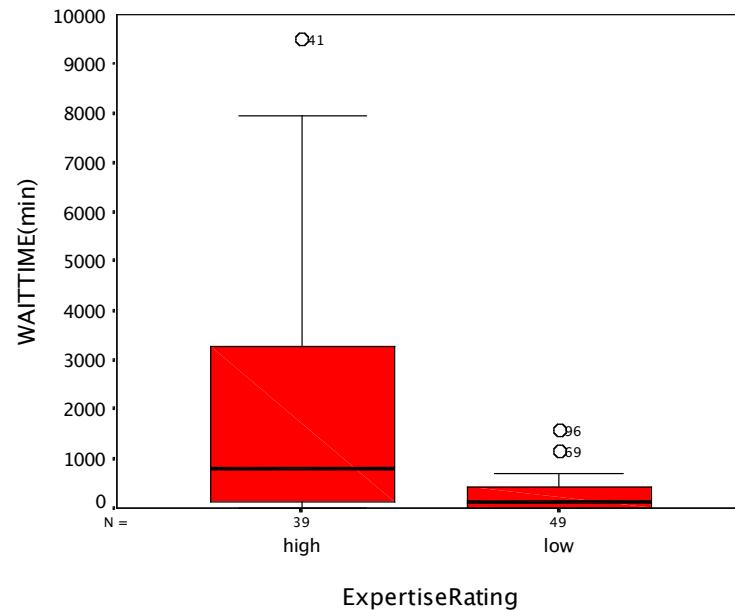
$\frac{1}{6}$

$\frac{1}{3}$

3

Identifying expertise

❑ The Response Time Gap



- The Expertise Gap
- Difficult to infer reliability of answers

Automatically ranking expertise may be helpful.

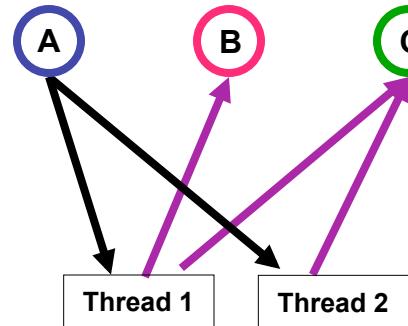
Java Forum

- 87 sub-forums
- 1,438,053 messages
- community expertise network constructed:
 - 196,191 users
 - 796,270 edges

The screenshot shows the Sun Developer Network Java Technology Forums. The top navigation bar includes links for Java, Solaris, Communities, Sun Store, Join SDN, My Profile, and Why Join? A search bar is also present. The main content area features a sidebar with links for Welcome, Login, Watch List, Duke Stars Program, My Forums, Feedback, and user statistics (Users Online: 75, Guests: 1,193). Below the sidebar is a search bar for forums. The main content area is titled "Java Technology Forums" and displays a list of forums categorized by Java version (e.g., Java 5, Java 6, Java SE). Each forum entry includes the name, a brief description, the number of topics and messages, and the last post date. A "Duke Stars" logo is visible on the right side of the page.

Forum	Topics / Messages	Last Post
Java Essentials General discussions of Java for beginners (New to Java) and advanced users (Java Programming)	27,420 / 189,572	May 22, 2007 3:13 AM
New To Java	76,584 / 482,710	May 22, 2007 4:25 AM
Java Programming	200 / 1,400	May 22, 2007 4:23 AM
Training / Learning / Certification		
Core		
Discussions of the core APIs in the Java SE, including tools for monitoring performance		
Core APIs		
Concurrency	548 / 2,704	May 21, 2007 9:52 AM
Concurrent & Interrupt I/O	435 / 1,813	May 22, 2007 4:25 AM
Networking	1,582 / 6,548	May 22, 2007 4:19 AM
Optimization	1,227 / 6,060	May 22, 2007 4:27 AM

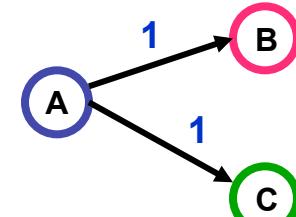
Constructing an expertise network



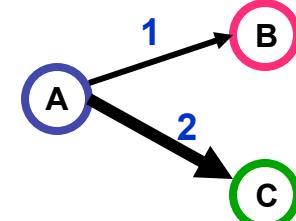
Thread 1: [Large Data, binary search or hashtable? user A](#)
 [Re: Large... user B](#)
 [Re: Large... user C](#)

Thread 2: [Binary file with ASCII data user A](#)
 [Re: File with... user C](#)

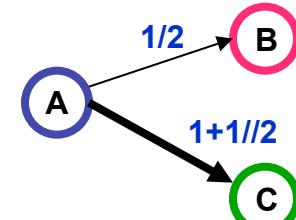
unweighted



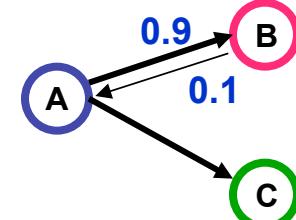
weighted by # threads



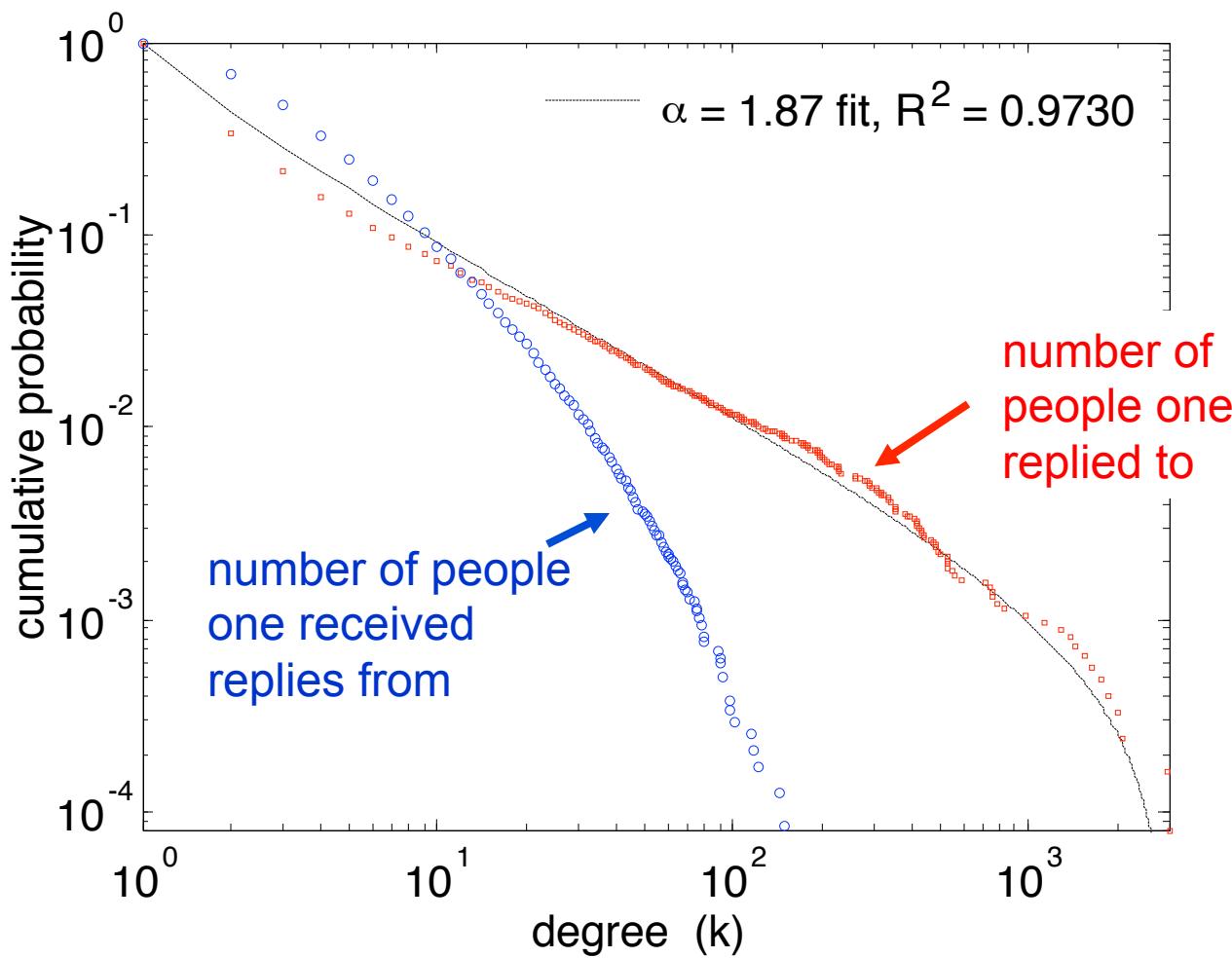
weighted by shared credit



weighted with backflow

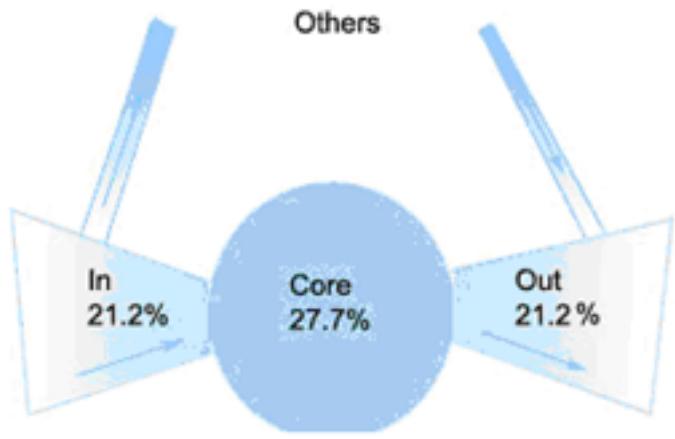


Uneven participation

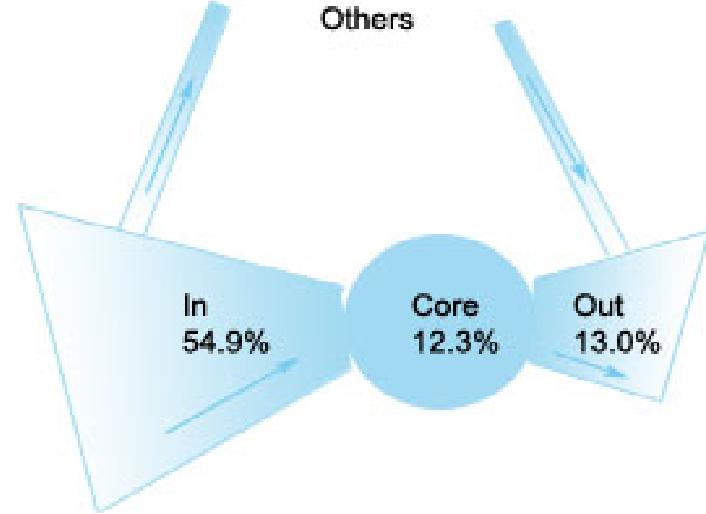


- ❑ ‘answer people’ may reply to thousands of others
- ❑ ‘question people’ are also uneven in the number of repliers to their posts, but to a lesser extent

Not Everyone Asks/Replies



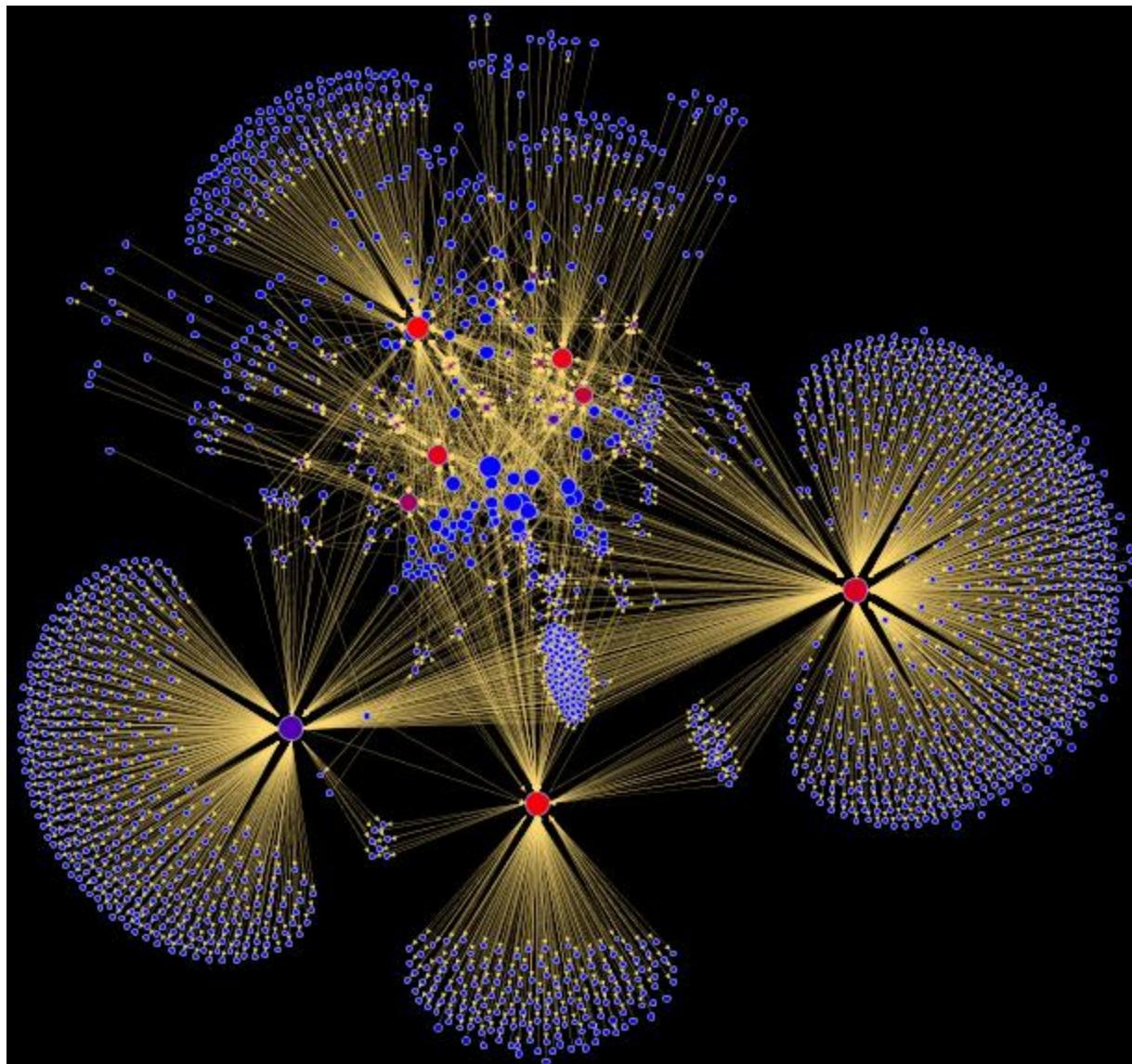
The Web is a bow tie



The Java Forum network is
an uneven bow tie

- Core: A strongly connected component, in which everyone asks and answers
- IN: Mostly askers.
- OUT: Mostly Helpers

fragment of the Java Forum



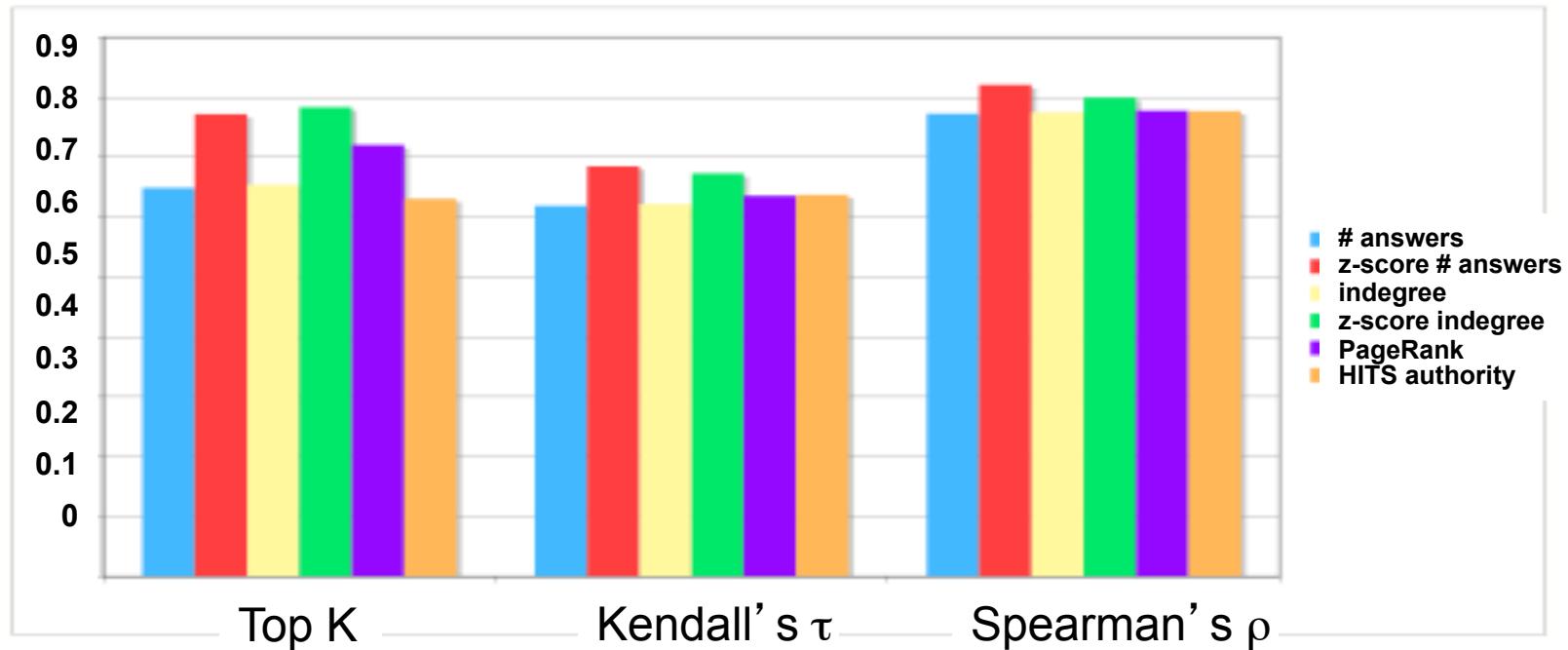
Relating network structure to expertise

□ Human-rated expertise levels

- 2 raters
- 135 JavaForum users with ≥ 10 posts
- inter-rater agreement ($\tau = 0.74$, $\rho = 0.83$)
- for evaluation of algorithms, omit users where raters disagreed by more than 1 level ($\tau = 0.80$, $\rho = 0.83$)

L	Category	Description
5	Top Java expert	Knows the core Java theory and related advanced topics deeply.
4	Java professional	Can answer all or most of Java concept questions. Also knows one or some sub topics very well,
3	Java user	Knows advanced Java concepts. Can program relatively well.
2	Java learner	Knows basic concepts and can program, but is not good at advanced topics of Java.
1	Newbie	Just starting to learn java.

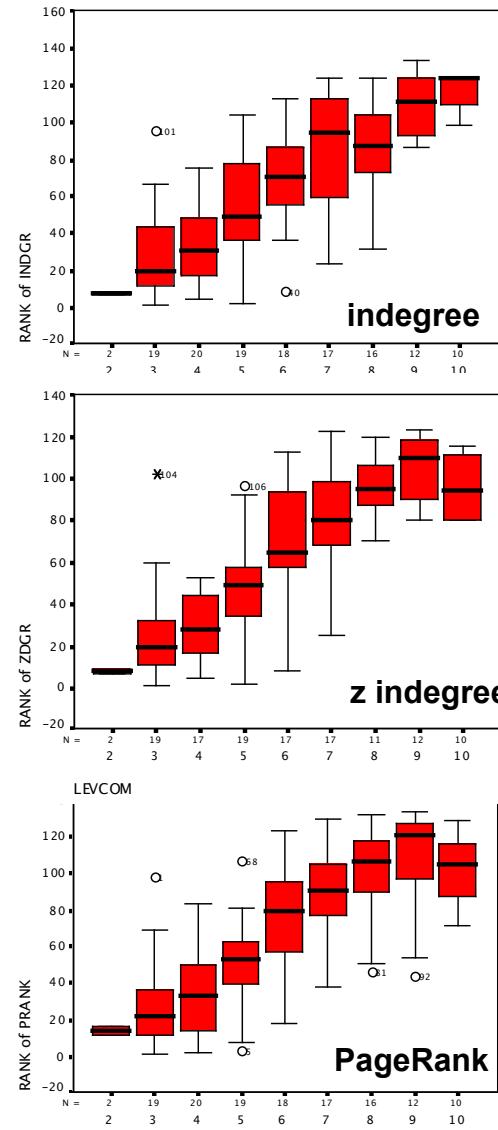
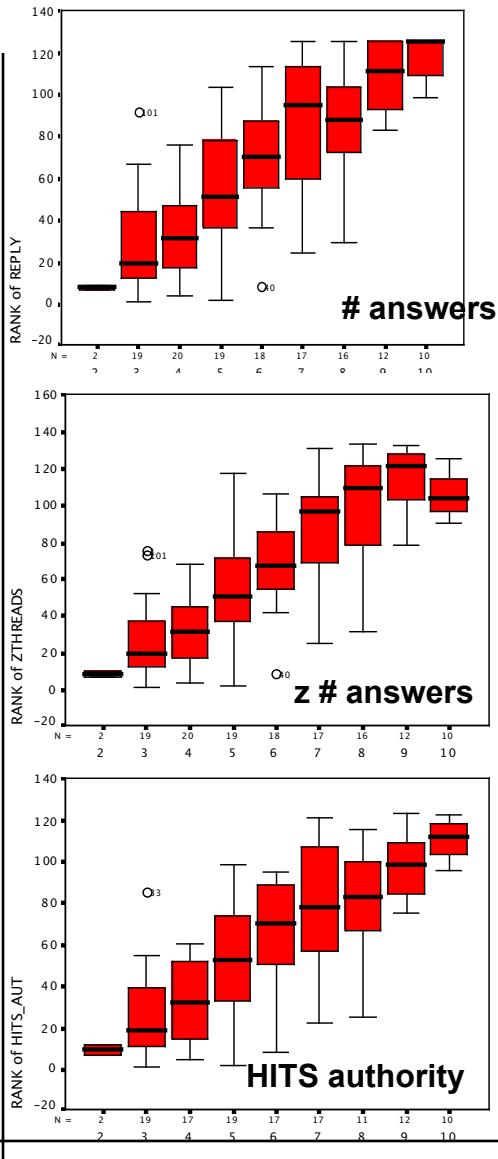
Algorithm Rankings vs. Human Ratings



simple local measures do as well (and better) than measures incorporating the wider network topology

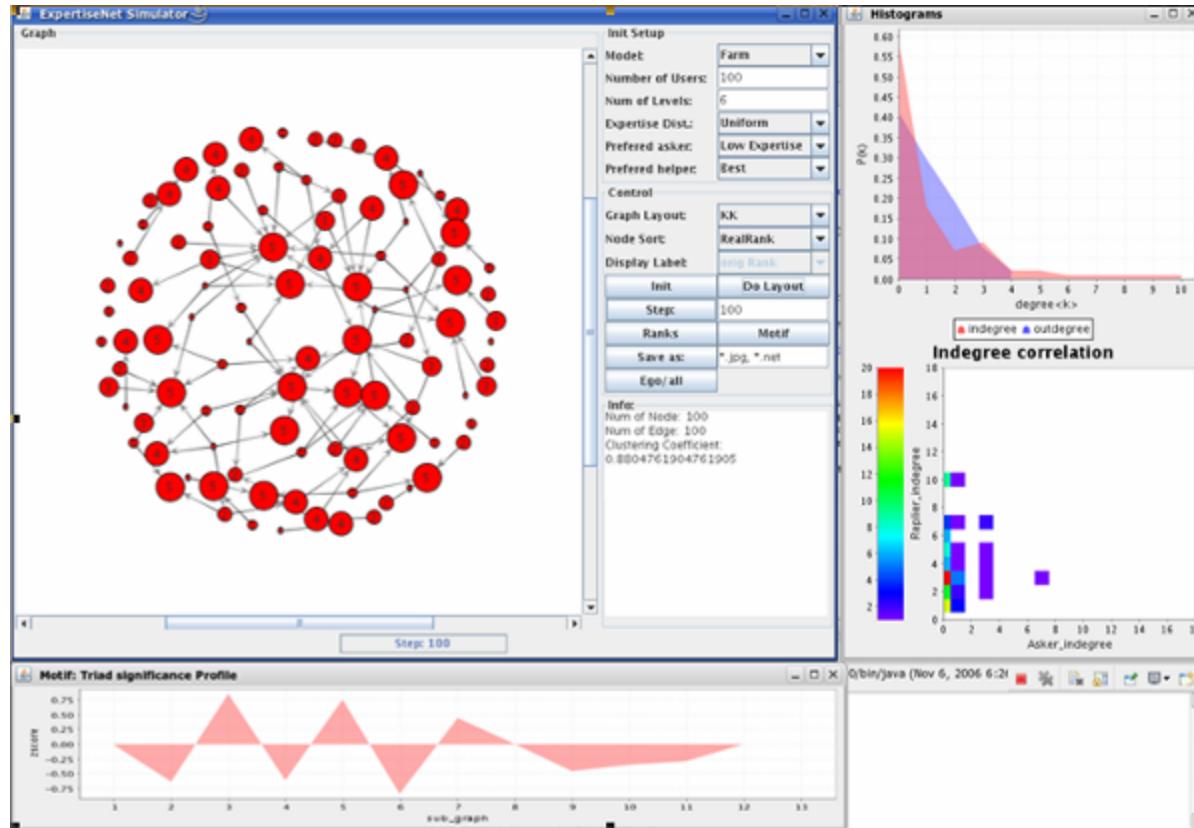
automated vs. human ratings

automated ranking



human rating

Modeling expertise network formation



ExpertiseNet Simulator

Control Parameters:

- Distribution of expertise
- Who asks questions most often?
- Who answers questions most often?
 - best expert most likely
 - someone a bit more expert

Simulating probability of expertise pairing

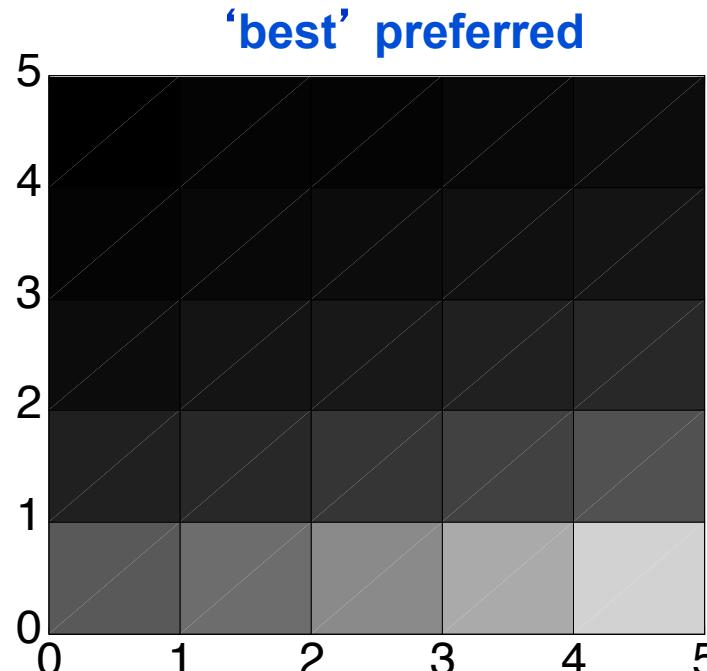
suppose:

expertise is uniformly distributed

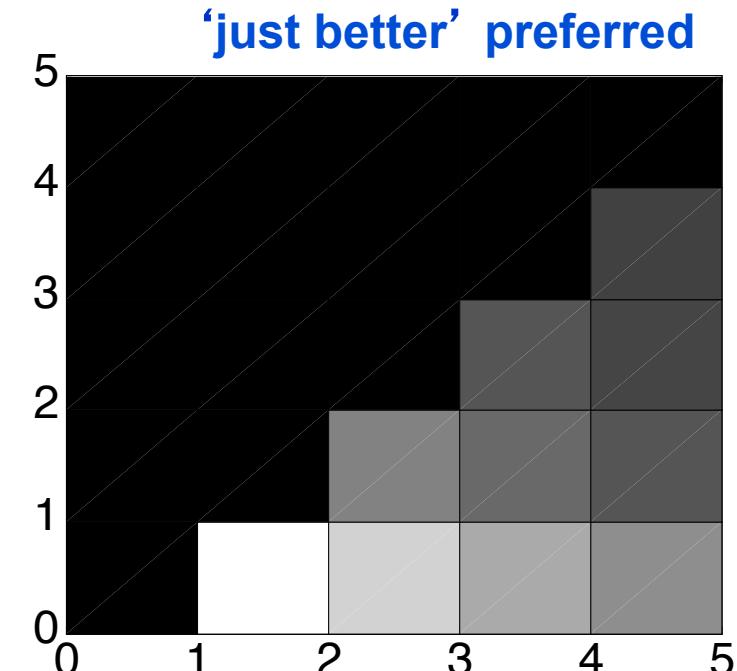
probability of posing a question is inversely proportional to expertise

p_{ij} = probability a user with expertise j replies to a user with expertise i

2 models:

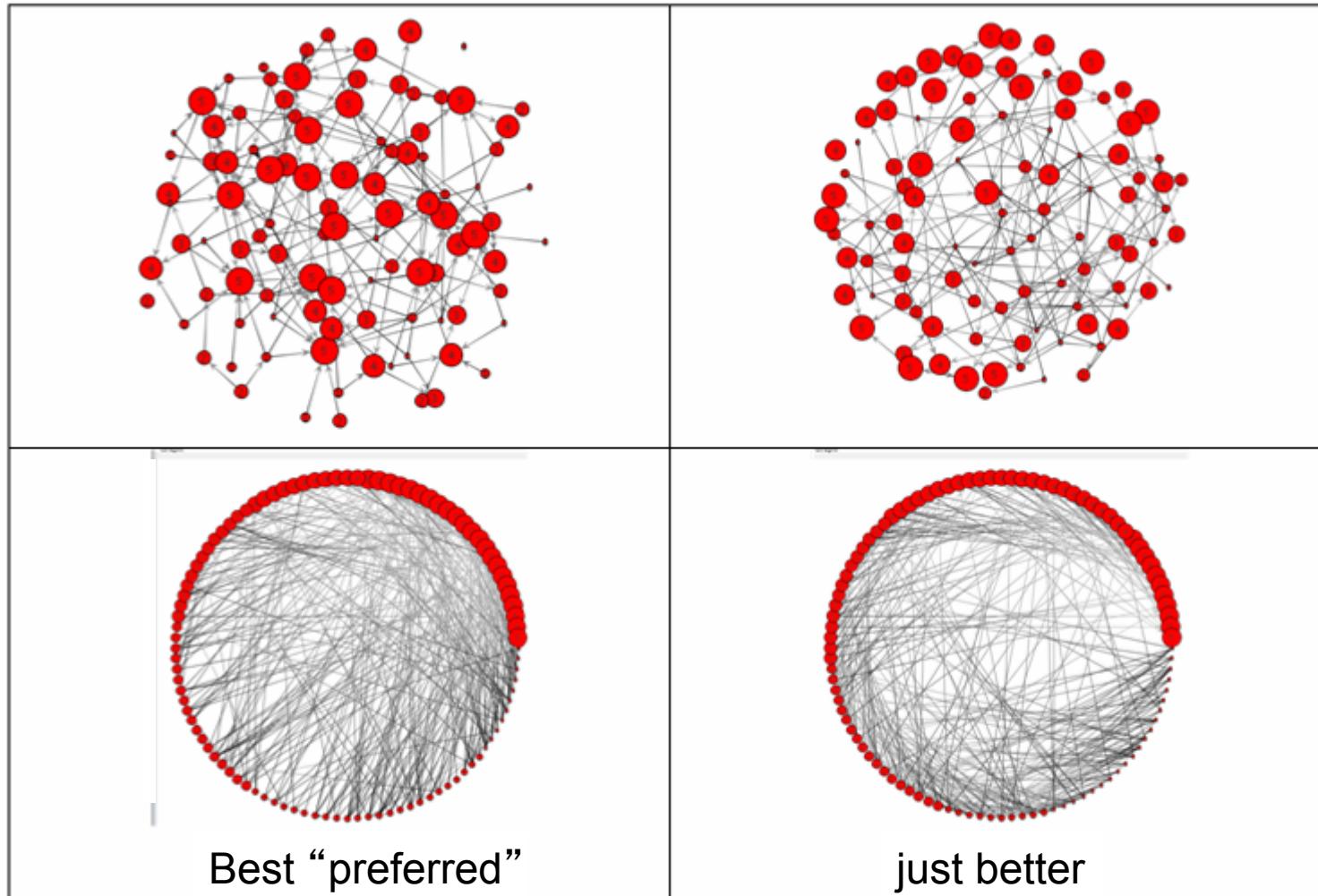


$$p_{ij} \sim e^{\beta(j-i)} / i$$

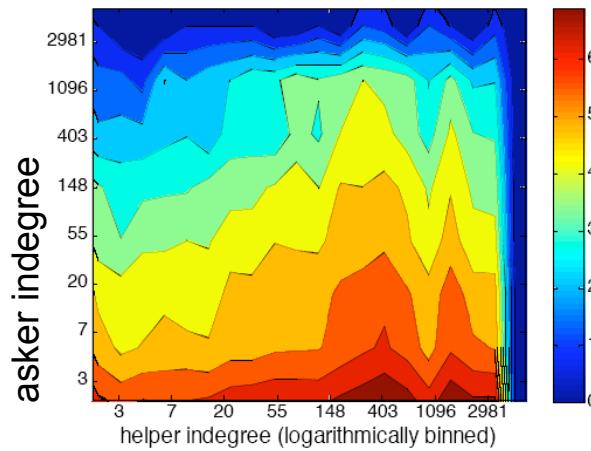


$$p_{ij} \sim e^{\gamma(i-j)} / i \quad j > i$$

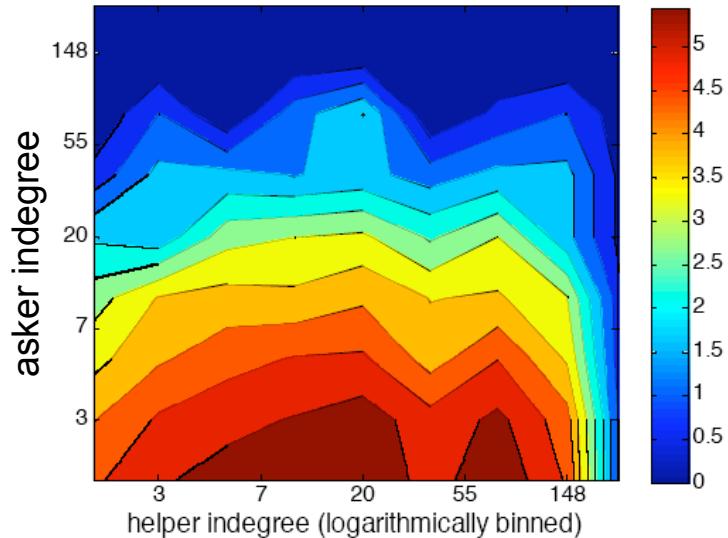
Visualization



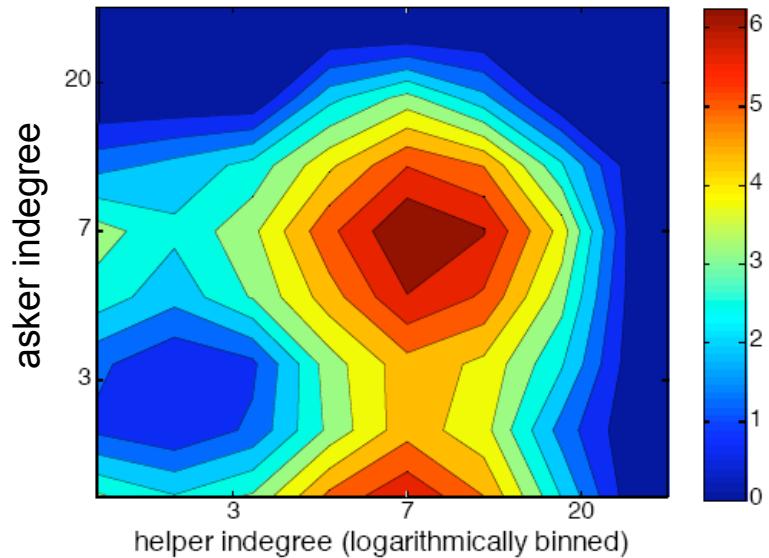
Degree correlation profiles



Java Forum Network

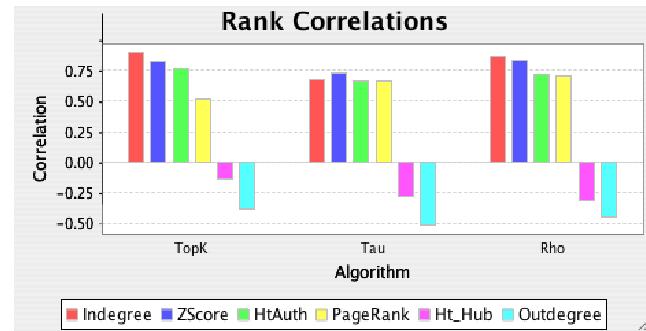
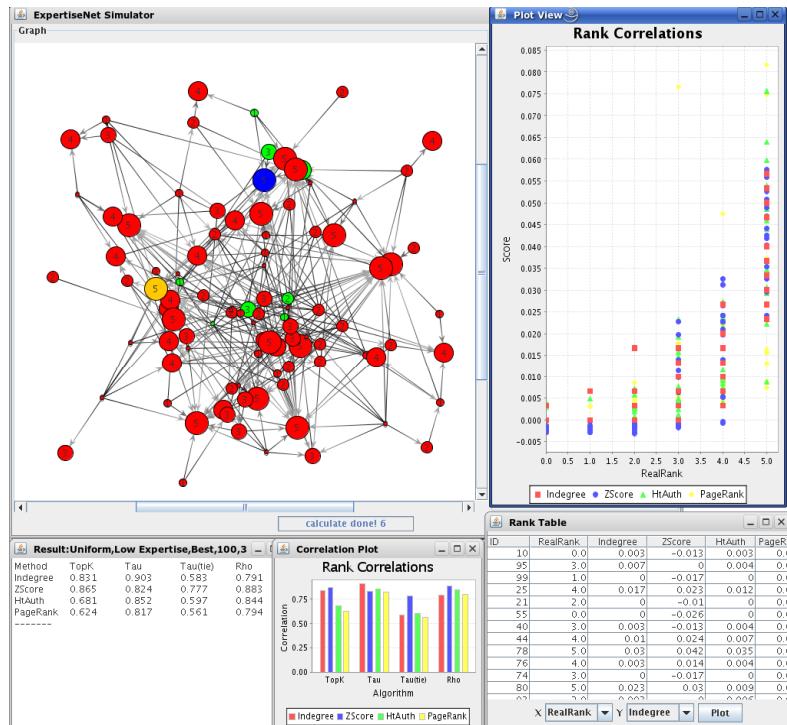


best preferred (simulation)

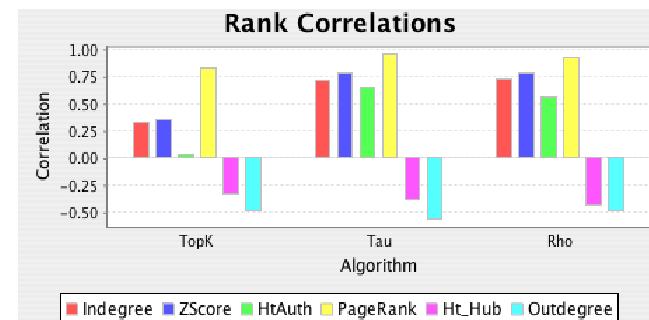


just better (simulation)

Algorithm selection

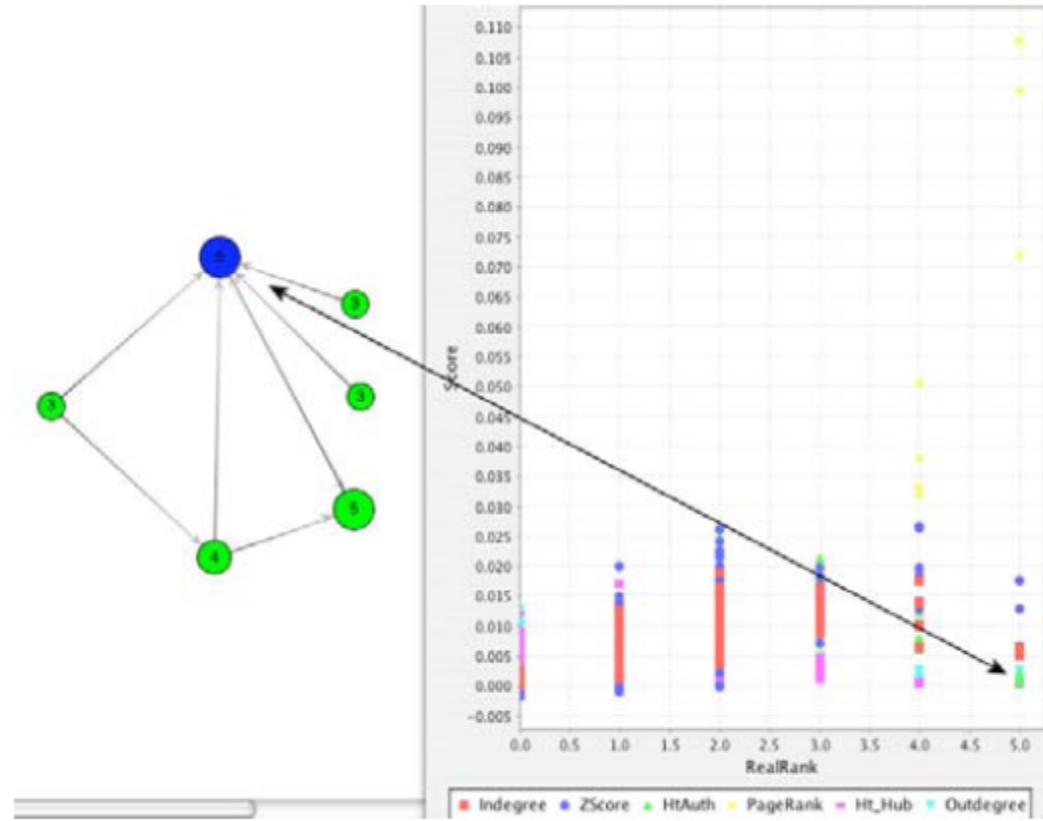


Preferred Helper: 'best available'

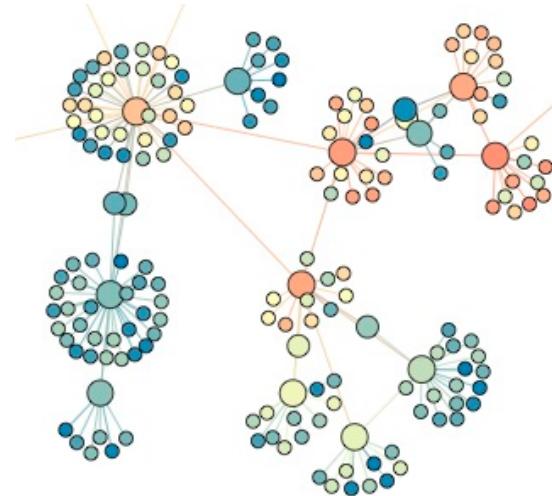


Preferred Helper: 'just better'

Algorithm evaluation



In the ‘just better’ model, a node is correctly ranked by PageRank but not by HITS



SNA 3D: Power laws

Lada Adamic

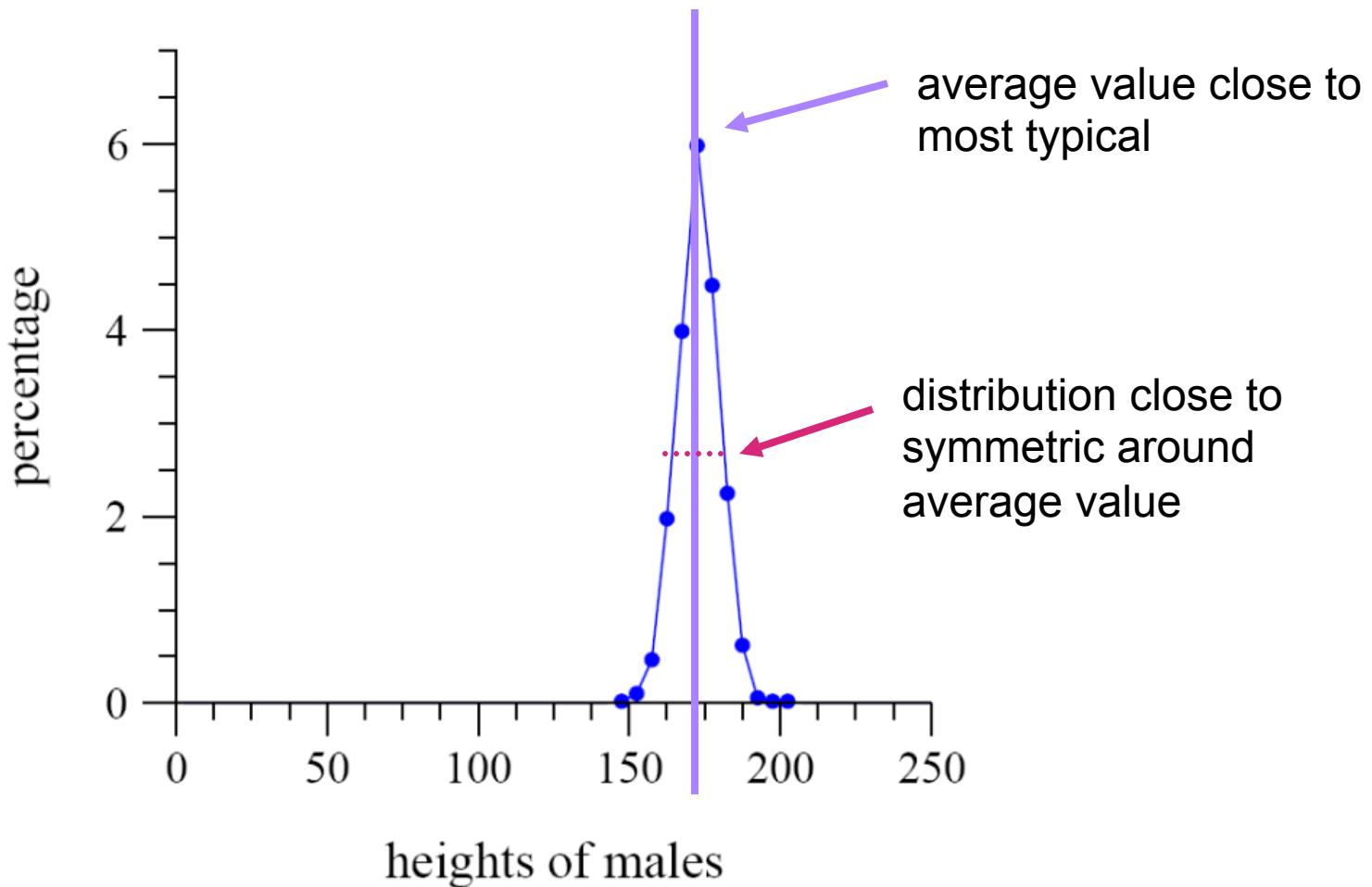


Heavy tails: right skew

❑ Right skew

- ❑ normal distribution (not heavy tailed)
 - ❑ e.g. heights of human males: centered around 180cm (5' 11'')
- ❑ Zipf's or power-law distribution (heavy tailed)
 - ❑ e.g. city population sizes: NYC 8 million, but many, many small towns

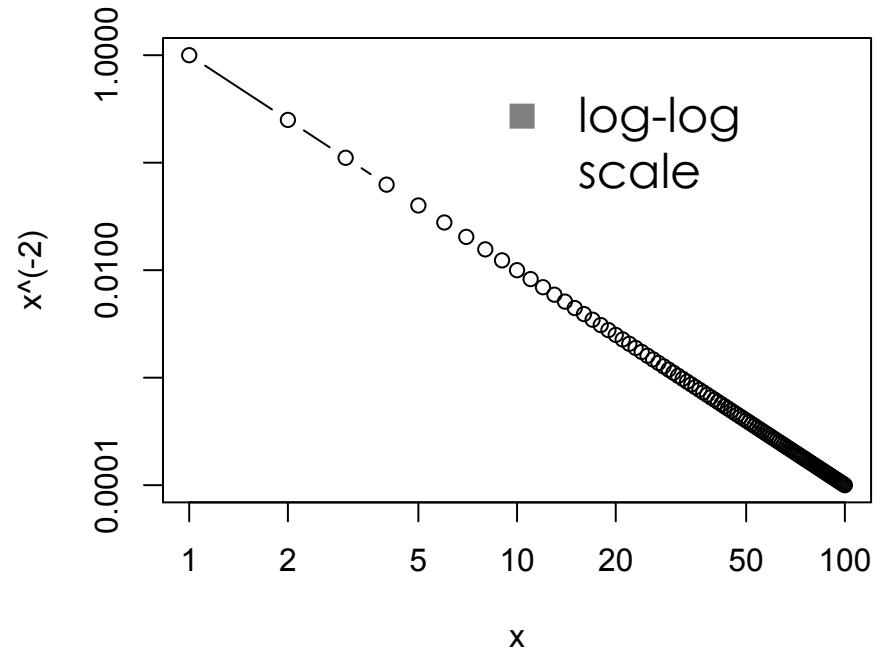
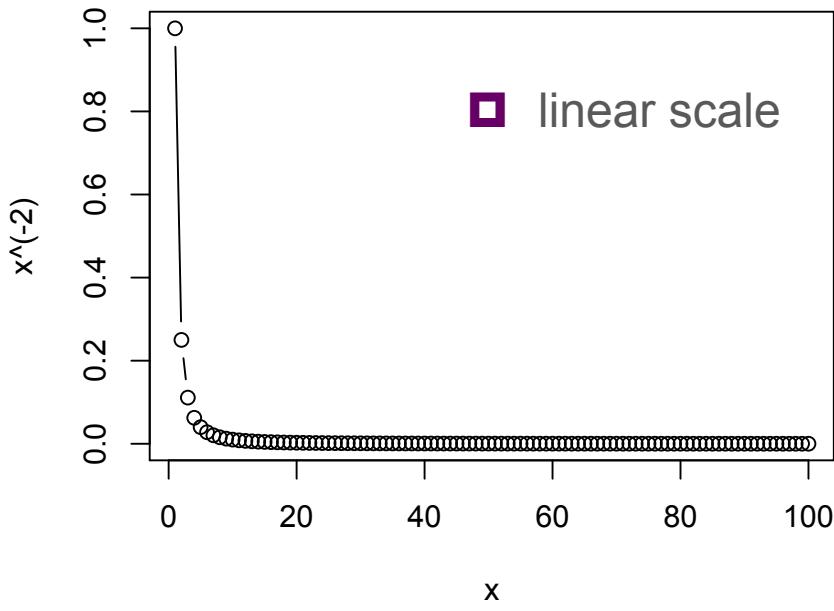
Normal distribution (human heights)



Heavy tails: max to min ratio

- High ratio of max to min
 - human heights
 - tallest man: 272cm (8' 11"), shortest man: (1' 10")
ratio: 4.8
from the Guinness Book of world records
 - city sizes
 - NYC: pop. 8 million, Duffield, Virginia pop. 52, *ratio: 150,000*

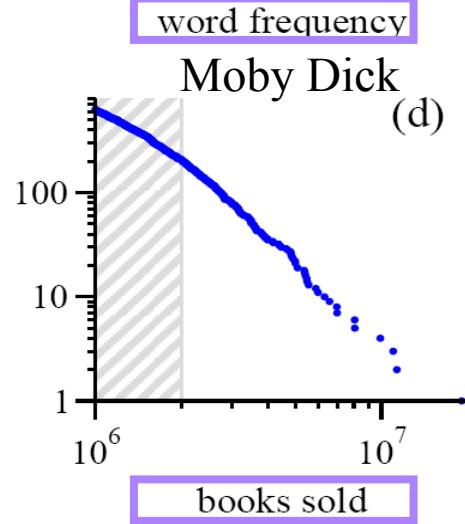
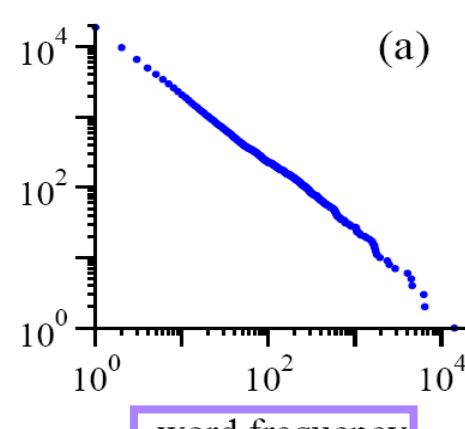
Power-law distribution



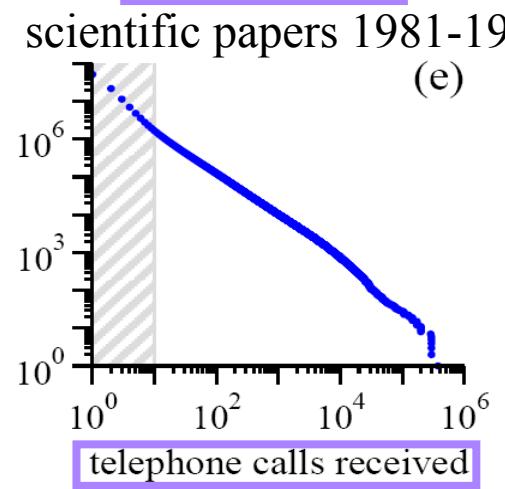
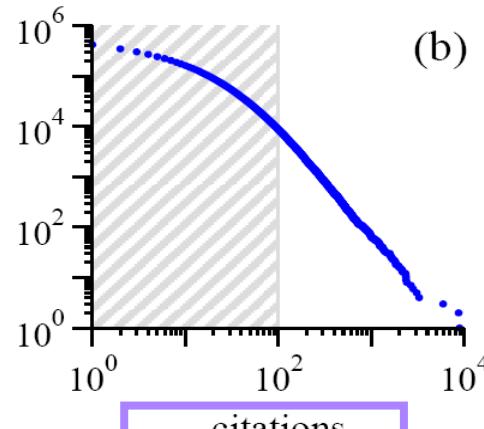
- high skew (asymmetry)
- straight line on a log-log plot

Power laws are seemingly everywhere

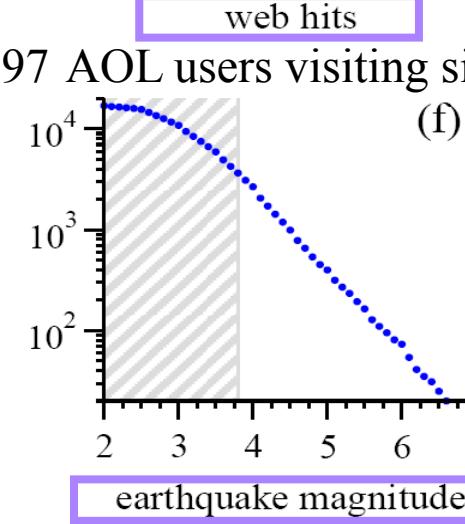
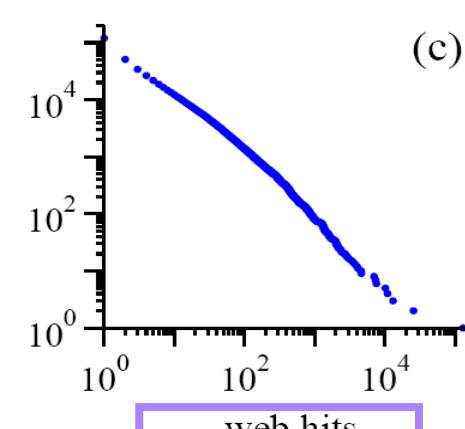
note: these are cumulative distributions, more about this in a bit...



bestsellers 1895-1965



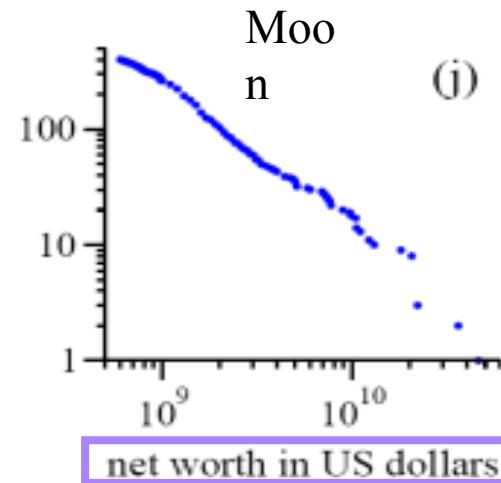
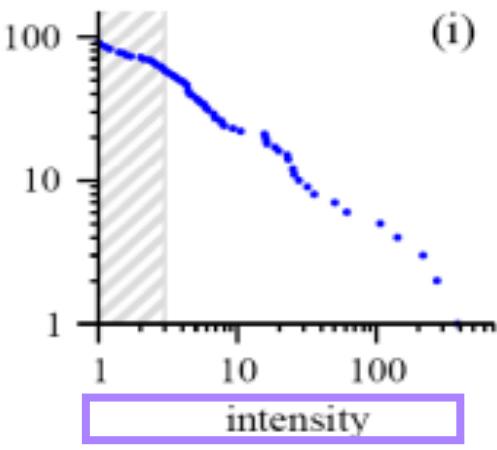
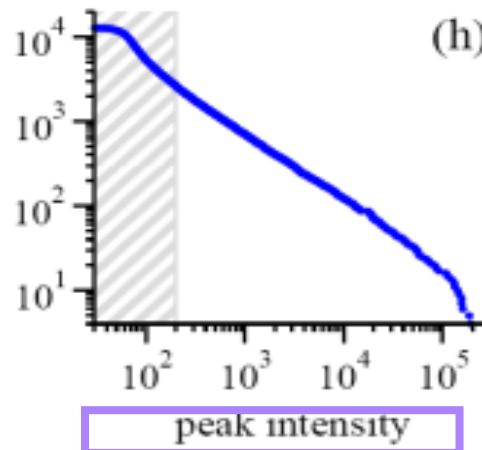
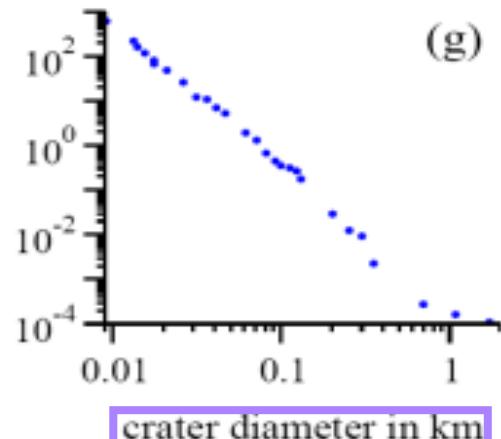
AT&T customers on 1 day



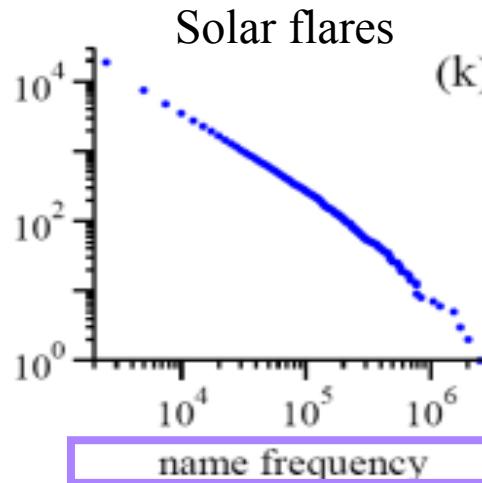
California 1910-1992

Source: M.E.J. Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

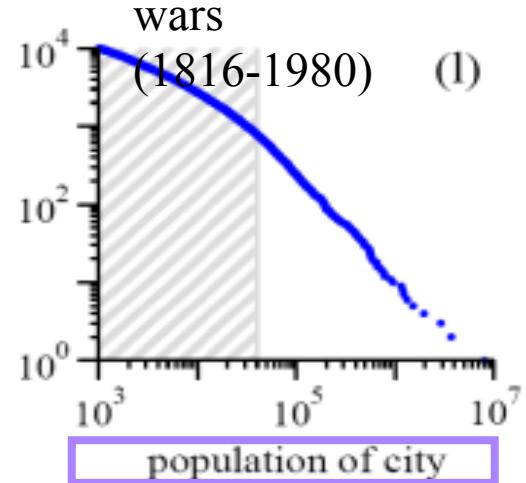
Yet more power laws



richest individuals
2003



US family names
1990



US cities 2003

Source: M.E.J. Newman, 'Power laws, Pareto distributions and Zipf's law', Contemporary Physics **46**, 323–351 (2005)

Power law distribution

- Straight line on a log-log plot

$$\ln(p(x)) = c - \alpha \ln(x)$$

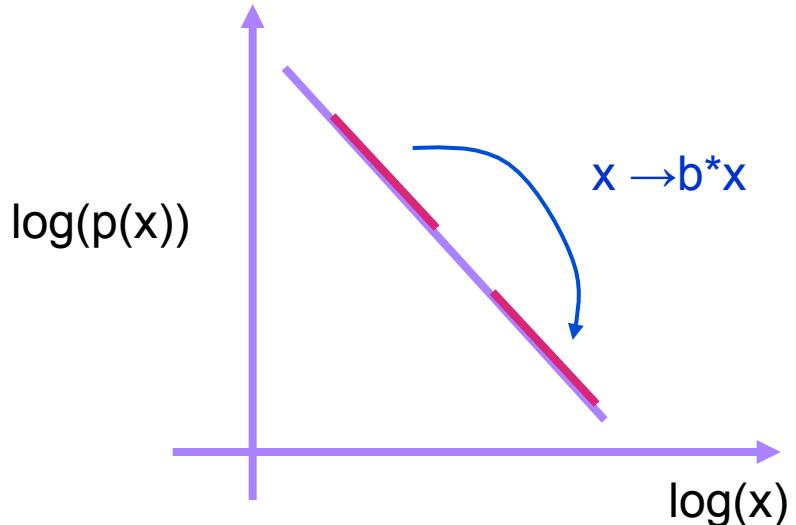
- Exponentiate both sides to get that $p(x)$, the probability of observing an item of size ‘x’ is given by

$$p(x) = Cx^{-\alpha}$$

normalization constant (probabilities over all x must sum to 1) power law exponent α

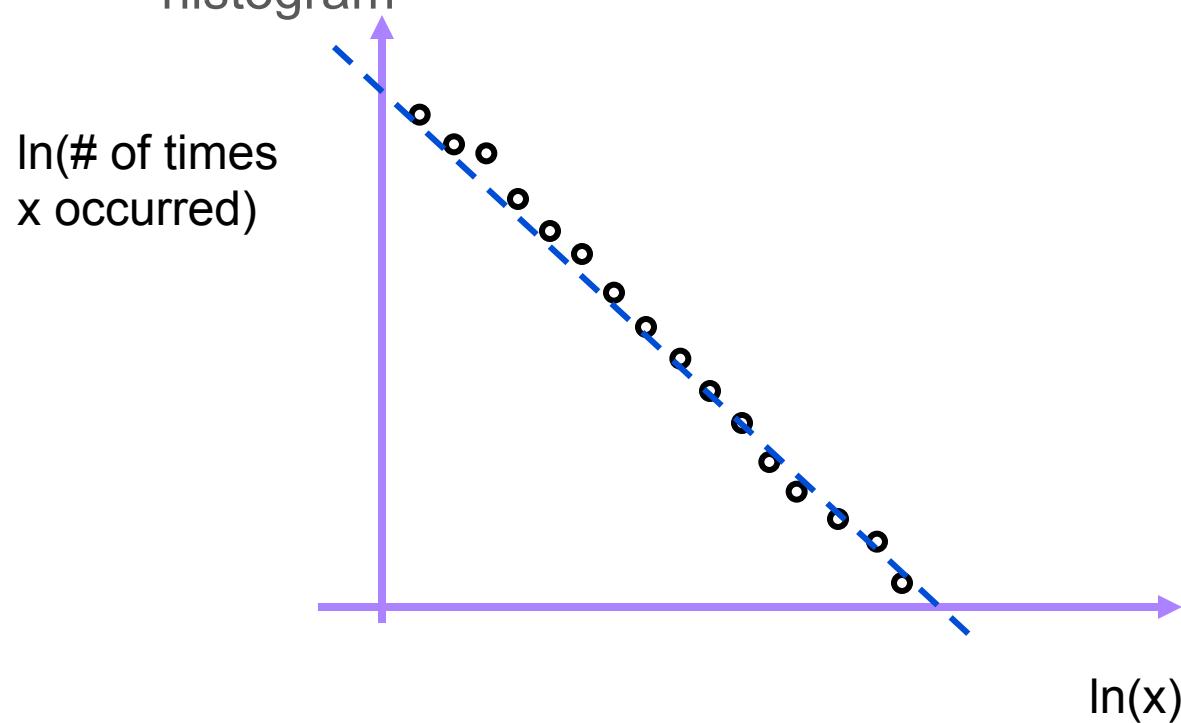
What does it mean to be scale free?

- A power law looks the same no mater what scale we look at it on (2 to 50 or 200 to 5000)
- Only true of a power-law distribution!
- $p(bx) = g(b) p(x)$ – shape of the distribution is unchanged except for a multiplicative constant
- $p(bx) = (bx)^{-\alpha} = b^{-\alpha} x^{-\alpha}$



Fitting power-law distributions

- ❑ Most common and not very accurate method:
 - ❑ Bin the different values of x and create a frequency histogram



$\ln(x)$ is the natural logarithm of x , but any other base of the logarithm will give the same exponent of α because $\log_{10}(x) = \ln(x)/\ln(10)$

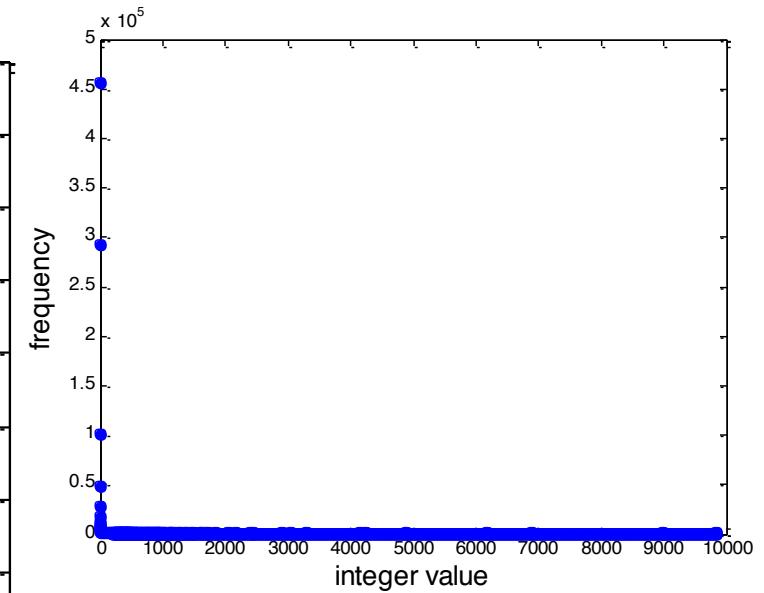
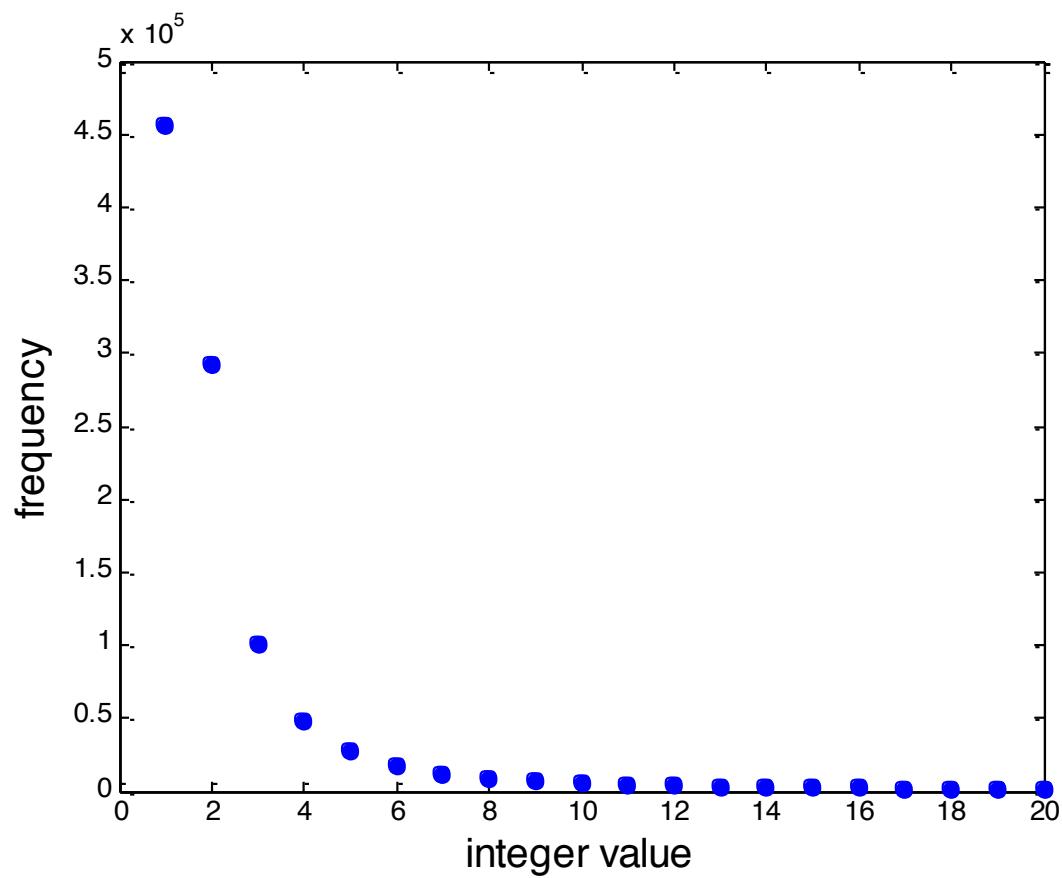
x can represent various quantities, the indegree of a node, the magnitude of an earthquake, the frequency of a word in text

Example on an artificially generated data set

- Take 1 million random numbers from a distribution with $\alpha = 2.5$
- Can be generated using the so-called ‘transformation method’
- Generate random numbers r on the unit interval $0 \leq r < 1$
- then $x = (1-r)^{-1/(\alpha-1)}$ is a random power law distributed real number in the range $1 \leq x < \infty$

Linear scale plot of straight bin of the data

- Number of times 1 or 3843 or 99723 occurred
- Power-law relationship not as apparent
- Only makes sense to look at smallest bins

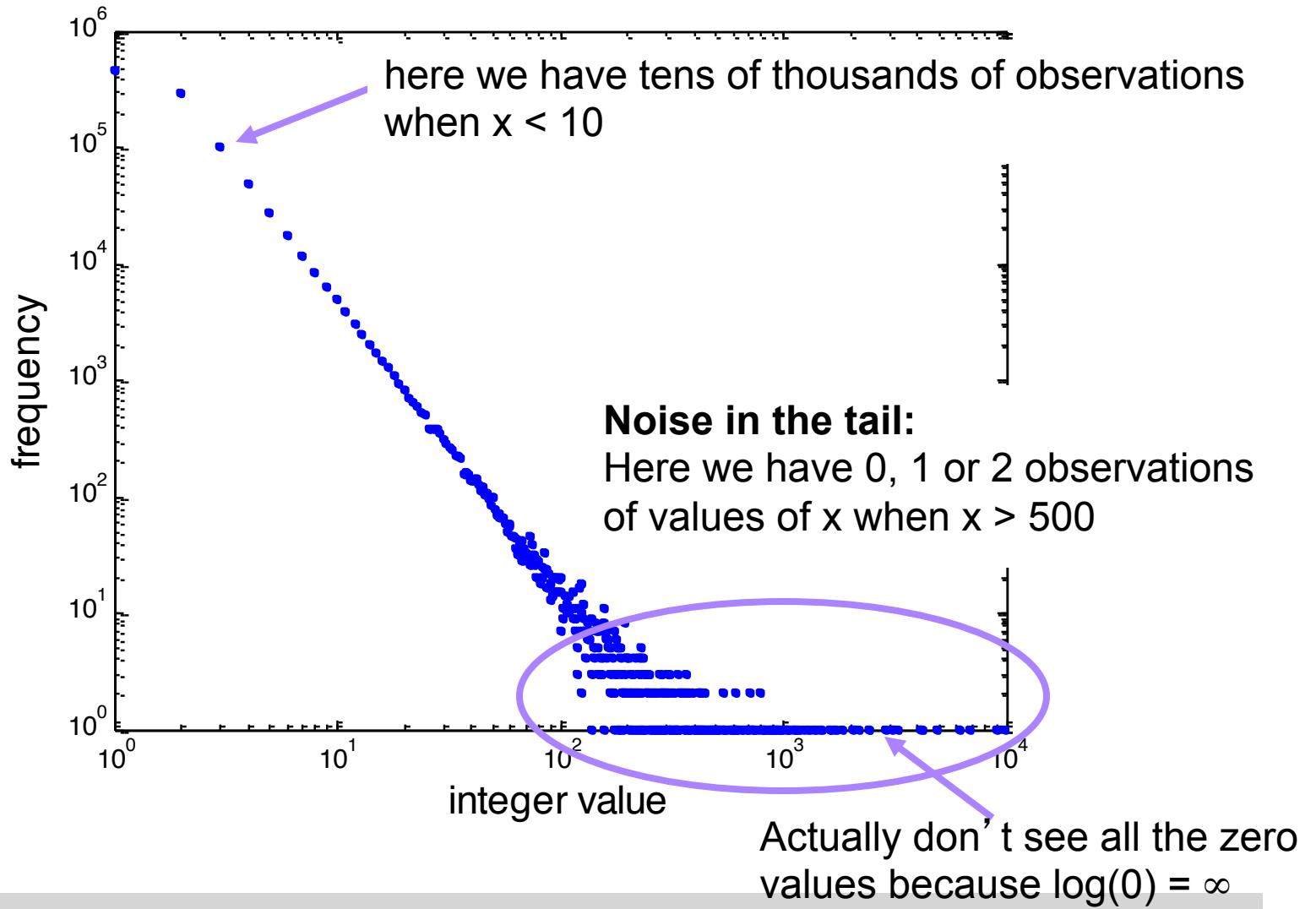


whole range

first few bins

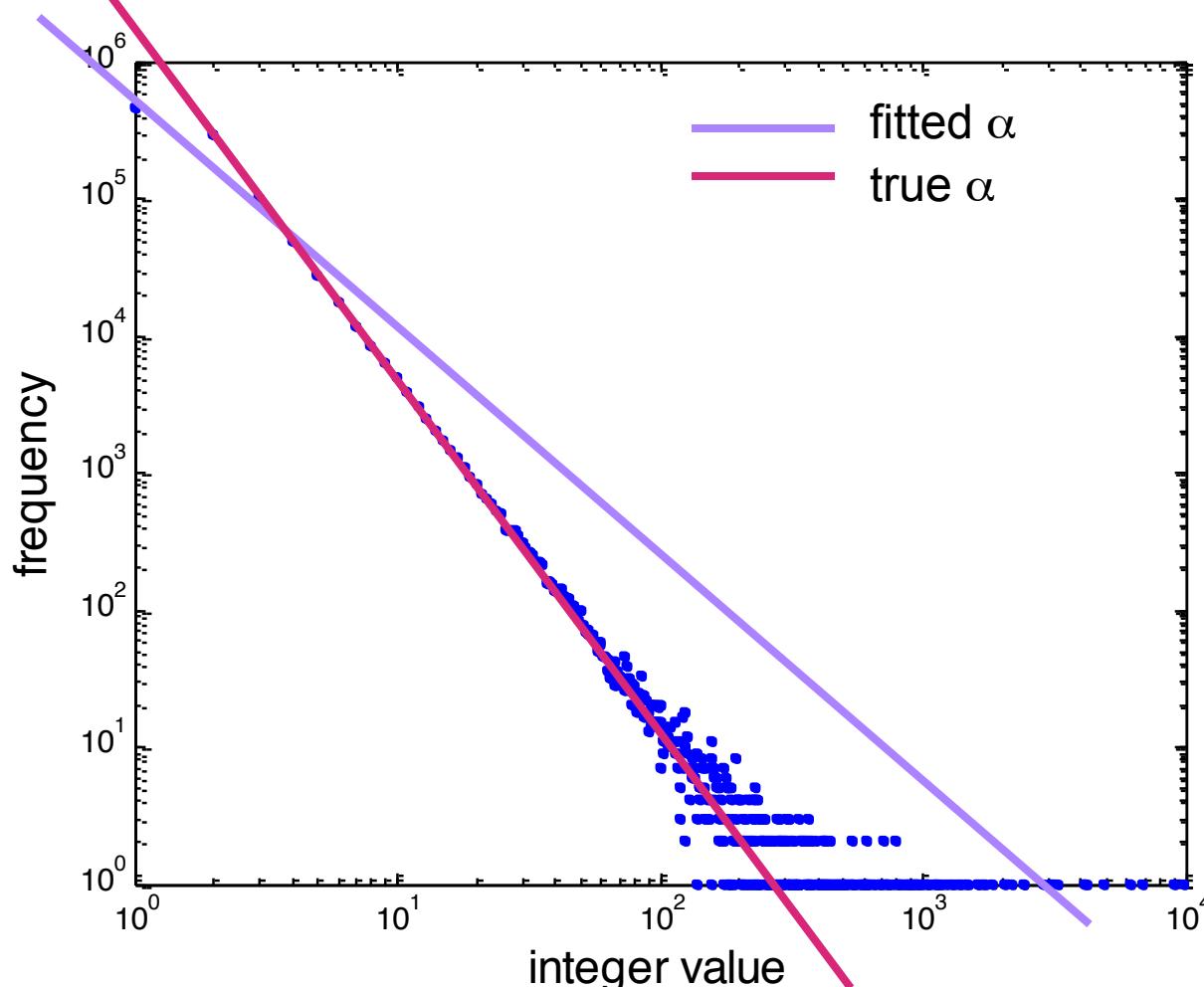
Log-log scale plot of simple binning of the data

- Same bins, but plotted on a log-log scale



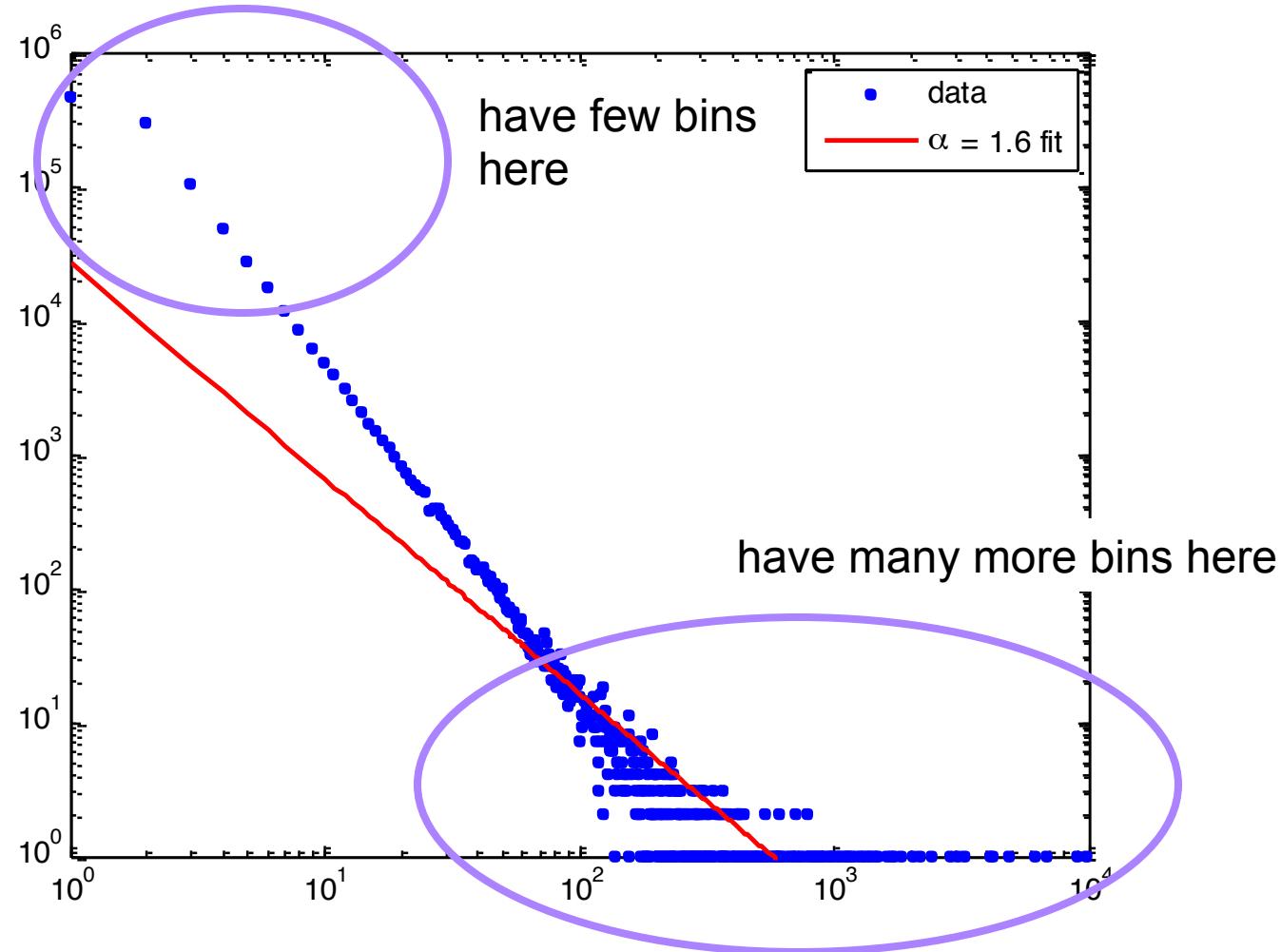
Log-log scale plot of straight binning of the data

- Fitting a straight line to it via least squares regression will give values of the exponent α that are too low



What goes wrong with straightforward binning

- ❑ Noise in the tail skews the regression result

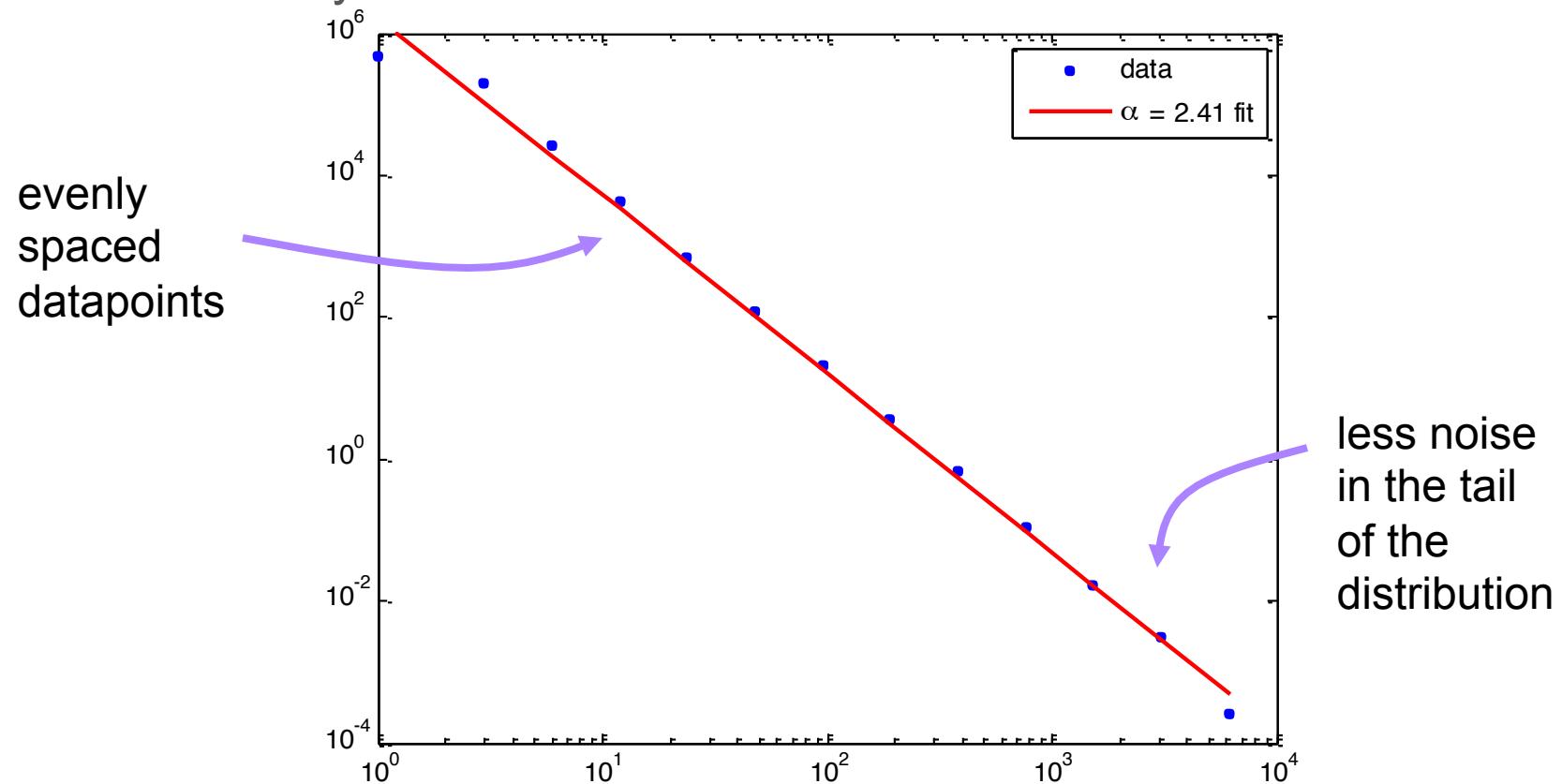


First solution: logarithmic binning

- bin data into exponentially wider bins:

 - 1, 2, 4, 8, 16, 32, ...

- normalize by the width of the bin



- disadvantage: binning smoothes out data but also loses information

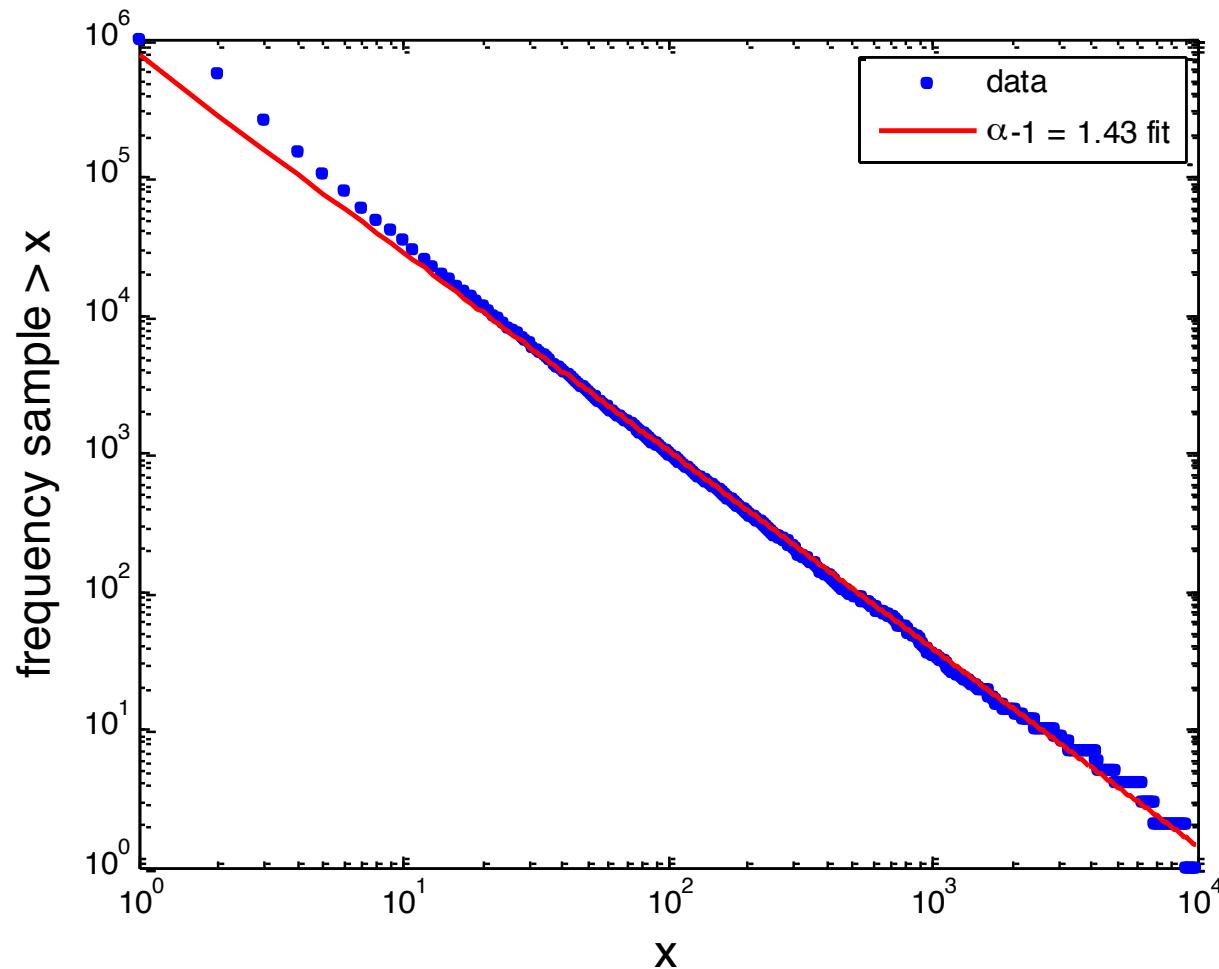
Second solution: cumulative binning

- No loss of information
 - No need to bin, has value at each observed value of x
- But now have cumulative distribution
 - i.e. how many of the values of x are at least X
- The cumulative probability of a power law probability distribution is also power law but with an exponent $\alpha - 1$

$$\int cx^{-\alpha} = \frac{c}{1-\alpha} x^{-(\alpha-1)}$$

Fitting via regression to the cumulative distribution

▣ fitted exponent (2.43) much closer to actual (2.5)

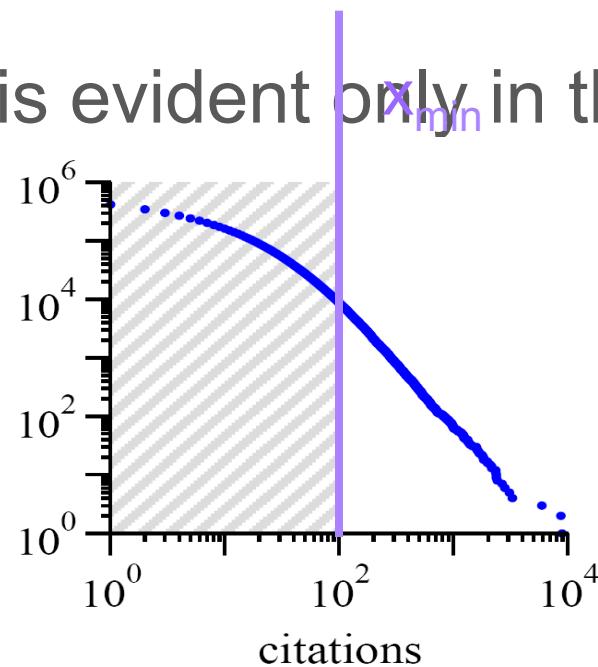


Where to start fitting?

- ❑ some data exhibit a power law only in the tail
- ❑ after binning or taking the cumulative distribution you can fit to the tail
- ❑ so need to select an x_{\min} the value of x where you think the power-law starts
- ❑ certainly x_{\min} needs to be greater than 0, because $x^{-\alpha}$ is infinite at $x = 0$

Example:

- Distribution of citations to papers
- power law is evident only in the tail ($x_{\min} > 100$ citation)



Source: M.E.J. Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

Maximum likelihood fitting – best

- You have to be sure you have a power-law distribution (this will just give you an exponent but not a goodness of fit)

$$\alpha = 1 + n \left[\sum_{i=1}^n \ln \frac{x_i}{x_{\min}} \right]^{-1}$$

- x_i are all your datapoints, and you have n of them
- for our data set we get $\alpha = 2.503$ – pretty close!

Some exponents for real world data

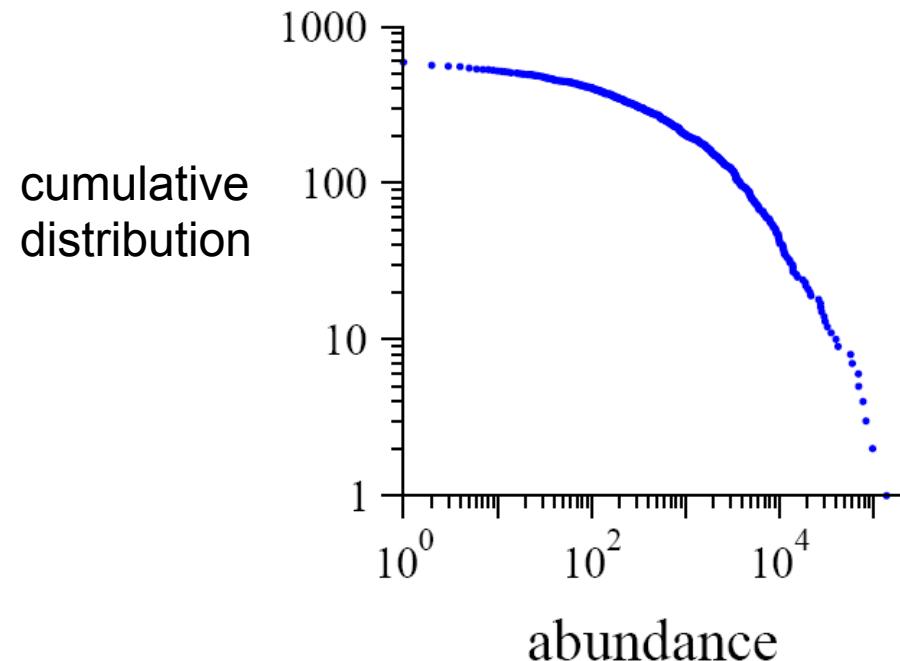
	x_{\min}	exponent α
frequency of use of words	1	2.20
number of citations to papers	100	3.04
number of hits on web sites	1	2.40
copies of books sold in the US	2 000 000	3.51
telephone calls received	10	2.22
magnitude of earthquakes	3.8	3.04
diameter of moon craters	0.01	3.14
intensity of solar flares	200	1.83
intensity of wars	3	1.80
net worth of Americans	\$600m	2.09
frequency of family names	10 000	1.94
population of US cities	40 000	2.30

Many real world networks are power law

	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
internet	2.5
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

Hey, not everything is a power law

- number of sightings of 591 bird species in the North American Bird survey in 2003.



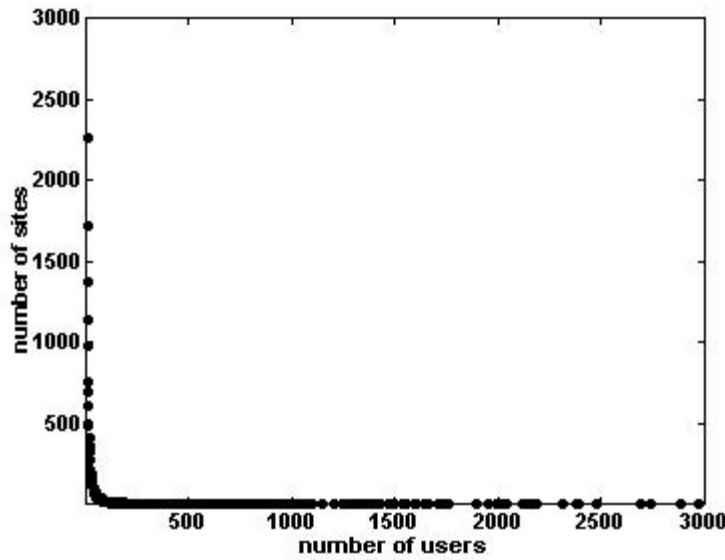
- another example:
 - size of wildfires (in acres)

Source: M.E.J. Newman, 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351 (2005)

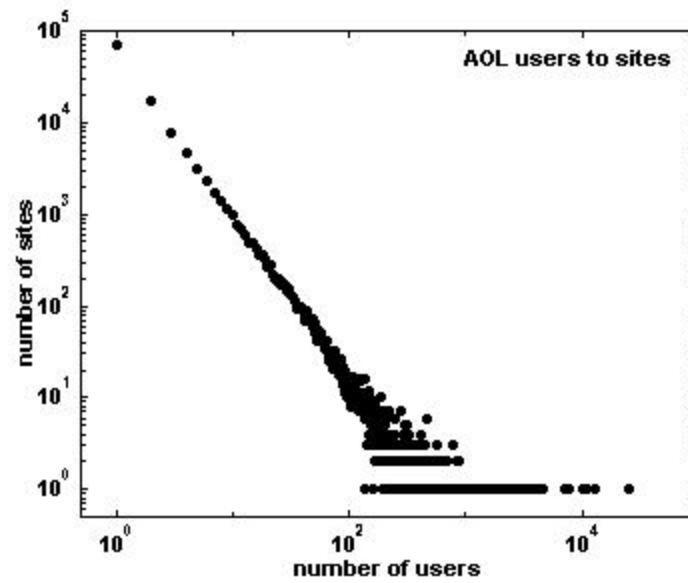
Not every network is power law distributed

- ❑ reciprocal, frequent email communication
- ❑ power grid
- ❑ Roget's thesaurus
- ❑ company directors...

Example on a real data set: number of AOL visitors to different websites back in 1997



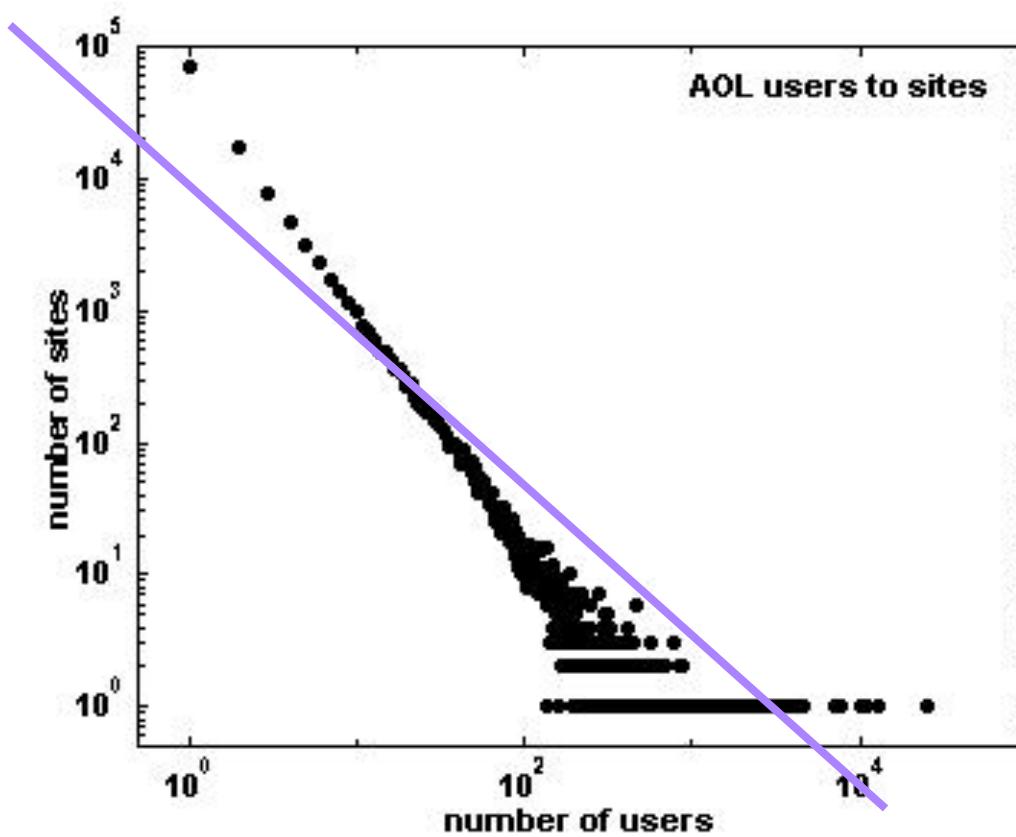
simple binning on a linear scale



simple binning on a log-log scale

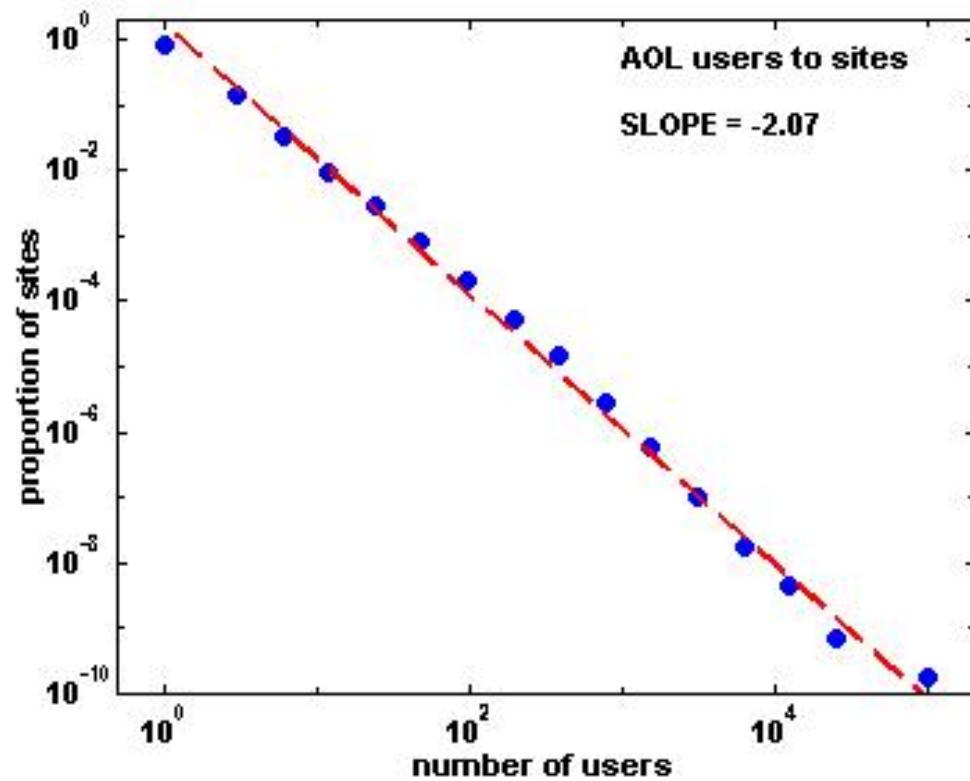
trying to fit directly...

- direct fit is too shallow: $\alpha = 1.17\dots$



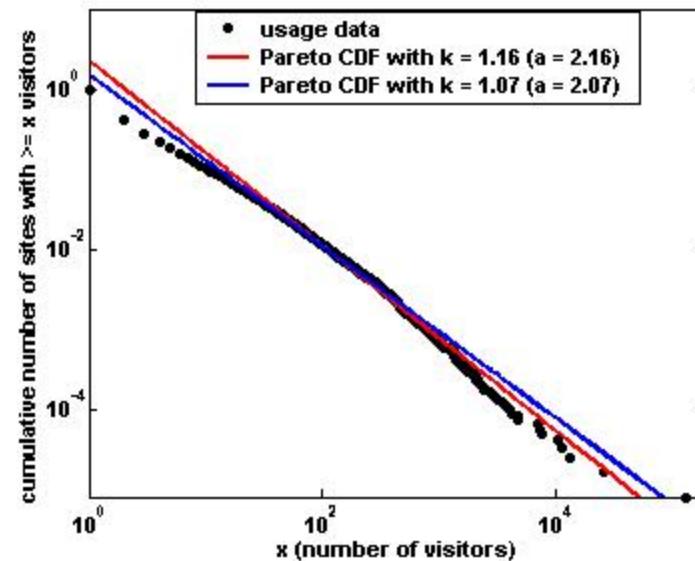
Binning the data logarithmically helps

- ❑ select exponentially wider bins
 - ❑ 1, 2, 4, 8, 16, 32,



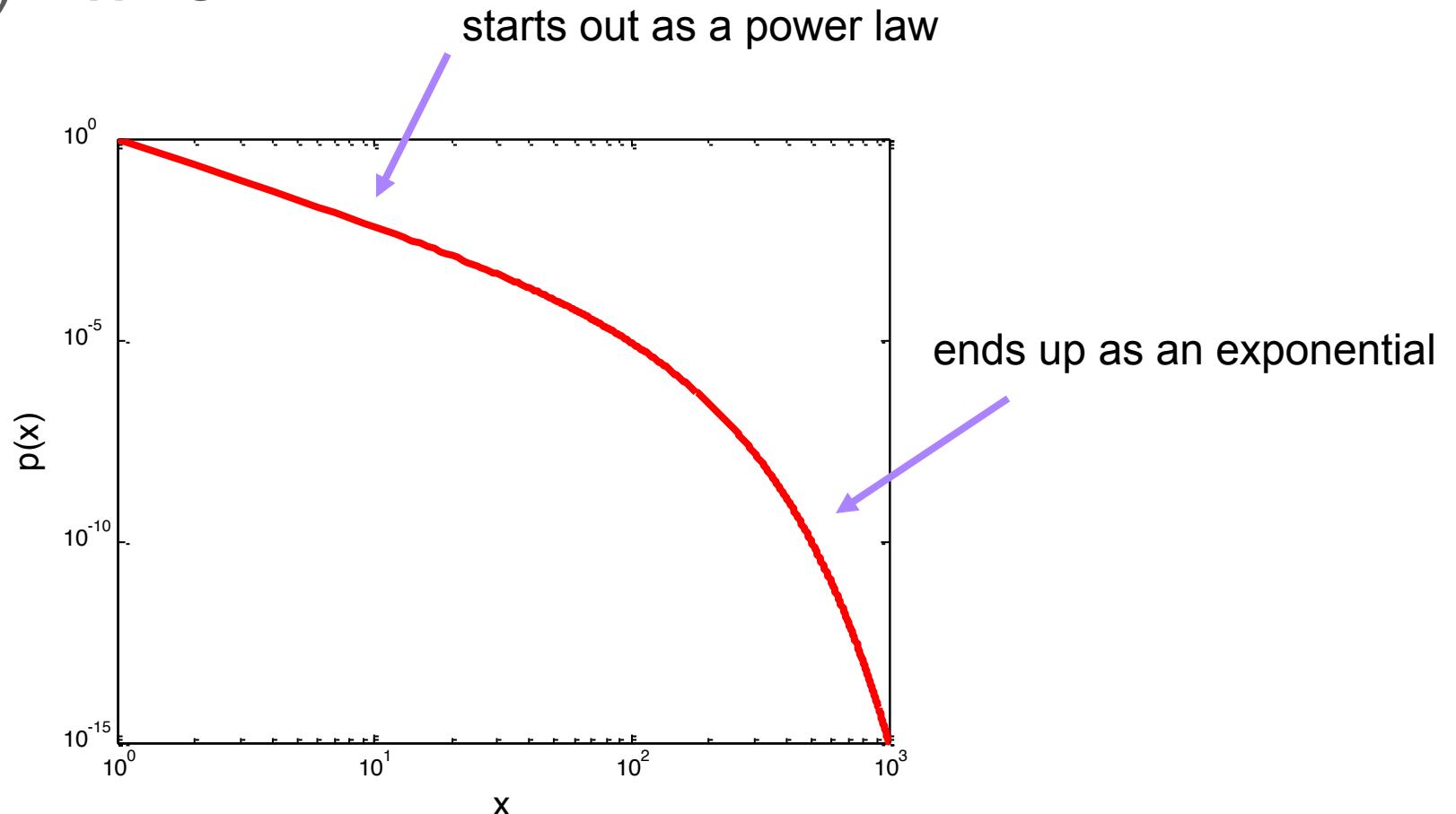
Or we can try fitting the cumulative distribution

- Shows perhaps 2 separate power-law regimes that were obscured by the exponential binning
- Power-law tail may be closer to 2.4



Another common distribution: power-law with an exponential cutoff

□ $p(x) \sim x^{-\alpha} e^{-x/\kappa}$



but could also be a lognormal or double exponential...

Zipf & Pareto: what they have to do with power-laws

■ Zipf

- George Kingsley Zipf, a Harvard linguistics professor, sought to determine the 'size' of the 3rd or 8th or 100th most common word.
- Size here denotes the frequency of use of the word in English text, and not the length of the word itself.
- Zipf's law states that the size of the r 'th largest occurrence of the event is inversely proportional to its rank:

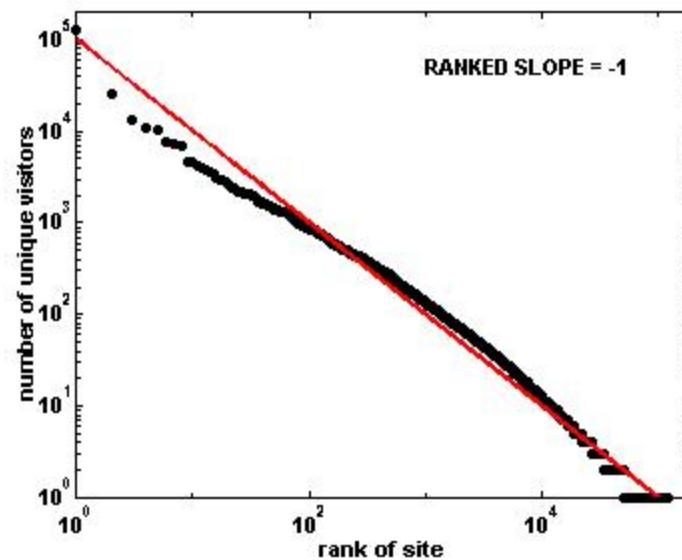
$$y \sim r^{-\beta}, \text{ with } \beta \text{ close to unity.}$$

So how do we go from Zipf to Pareto?

- The phrase "The r th largest city has n inhabitants" is equivalent to saying " r cities have n or more inhabitants".
- This is exactly the definition of the Pareto distribution, except the x and y axes are flipped. Whereas for Zipf, r is on the x-axis and n is on the y-axis, for Pareto, r is on the y-axis and n is on the x-axis.
- Simply inverting the axes, we get that if the rank exponent is β , i.e.
 $n \sim r^{-\beta}$ for Zipf, (n = income, r = rank of person with income n)
then the Pareto exponent is $1/\beta$ so that
 $r \sim n^{-1/\beta}$ (n = income, r = number of people whose income is n or higher)

Zipf's law & AOL site visits

- Deviation from Zipf's law
- slightly too few websites with large numbers of visitors



Zipf's Law and city sizes (~1930) [2]

not any more

Rank(k)	City	Population (1990)	Zips's Law $10,000,000/k$	Modified Zipf's law: (Mandelbrot) $5,000,000/(k - 2/5)^{3/4}$
1	New York	7,322,564	10,000,000	7,334,265
7	Detroit	1,027,974	1,428,571	1,214,261
13	Baltimore	736,014	769,231	747,693
19	Washington DC	606,900	526,316	558,258
25	New Orleans	496,938	400,000	452,656
31	Kansas City	434,829	322,581	384,308
37	Virginia Beach	393,089	270,270	336,015
49	Toledo	332,943	204,082	271,639
61	Arlington	261,721	163,932	230,205
73	Baton Rouge	219,531	136,986	201,033
85	Hialeah	188,008	117,647	179,243
97	Bakersfield	174,820	103,270	162,270

80/20 rule

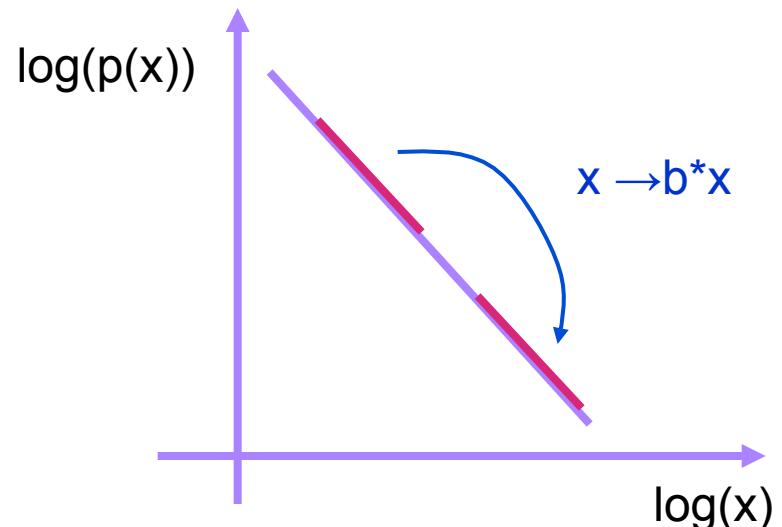
- The fraction W of the wealth in the hands of the richest P of the population is given by

$$W = P^{(\alpha-2)/(\alpha-1)}$$

- Example: US wealth: $\alpha = 2.1$
 - richest 20% of the population holds 86% of the wealth

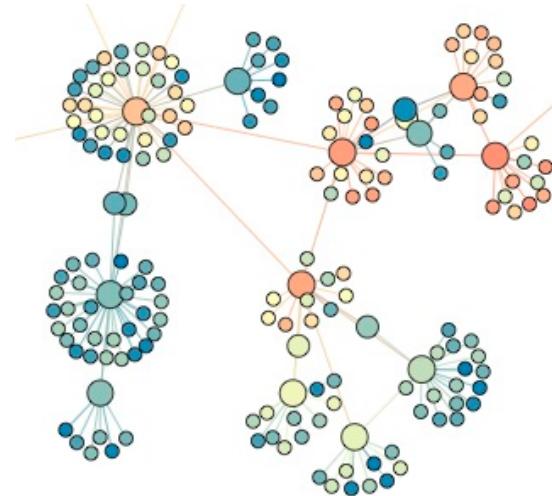
What does it mean to be scale free?

- A power law looks the same no mater what scale we look at it on (2 to 50 or 200 to 5000)
- Only true of a power-law distribution!
- $p(bx) = g(b) p(x)$ – shape of the distribution is unchanged except for a multiplicative constant
- $p(bx) = (bx)^{-\alpha} = b^{-\alpha} x^{-\alpha}$



Wrap up on power-laws

- ❑ Power-laws are cool and intriguing
- ❑ But make sure your data is actually power-law before boasting



SNA 4: community structure

Lada Adamic



Outline

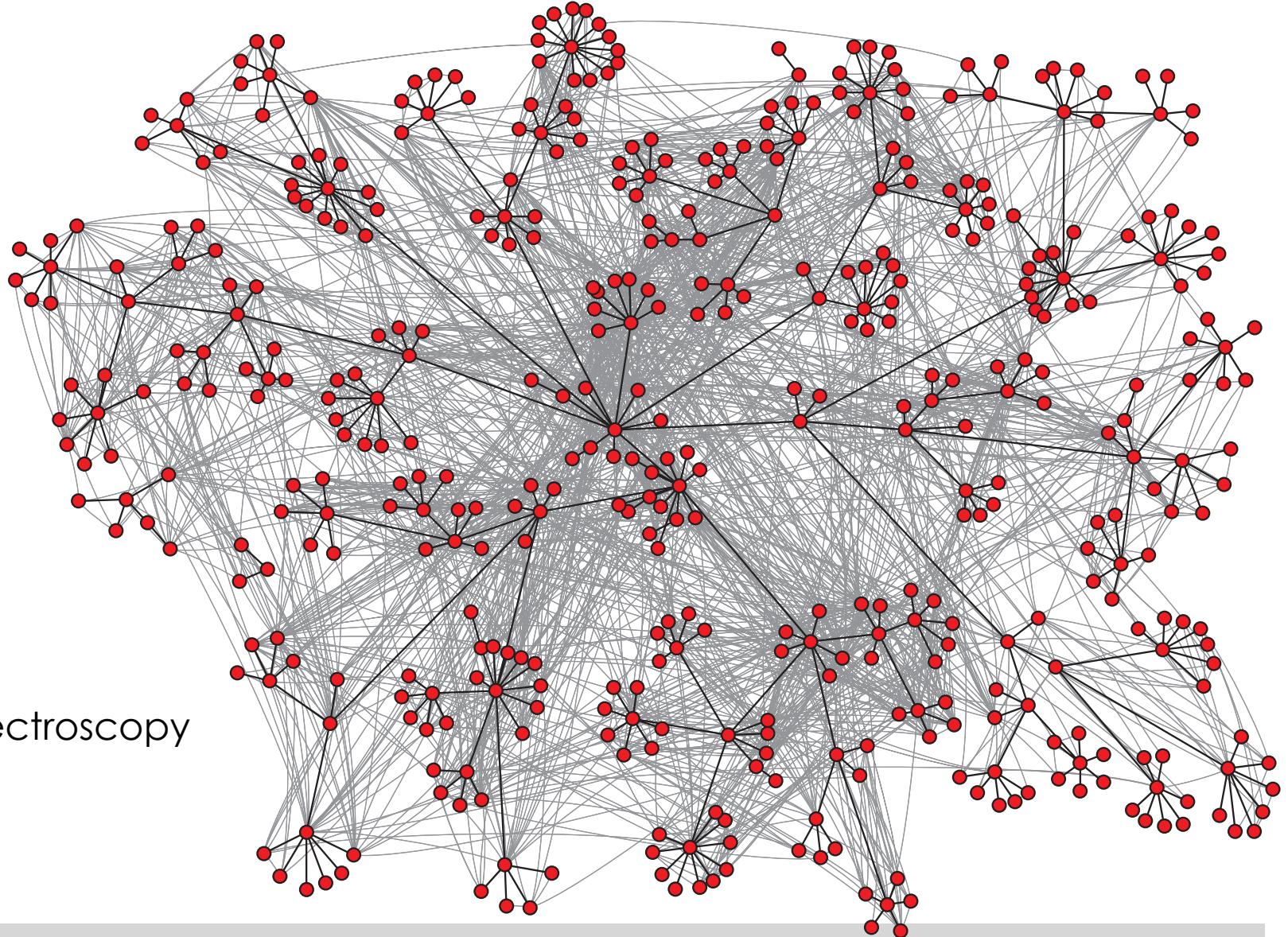
- ❑ why do we look for community structure?
- ❑ we need to define it in order to find it
- ❑ approaches to finding it

Why do it?

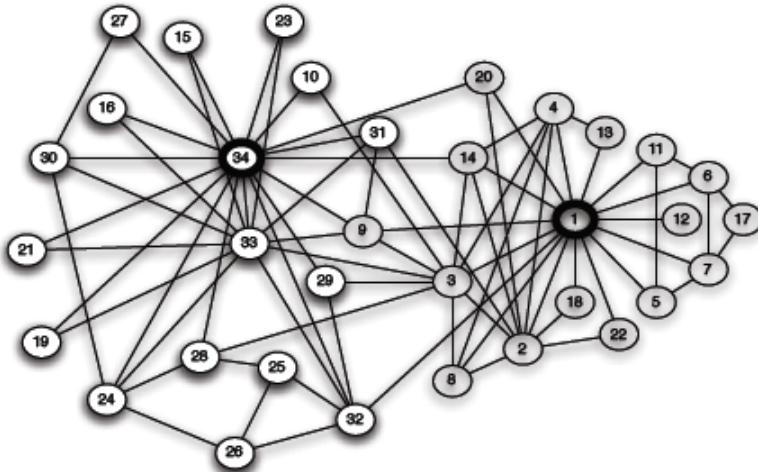
- Discover communities of practice
- Measure isolation of groups
- Understand opinion dynamics / adoption

Why look for community structure?

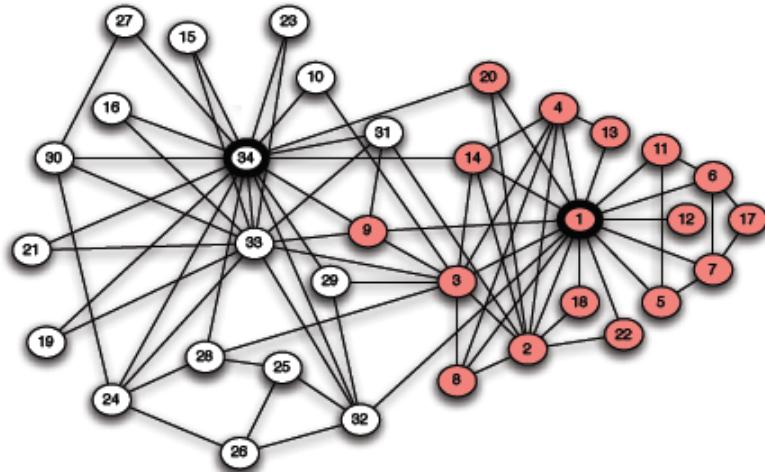
example:
email spectroscopy



Zachary Karate Club



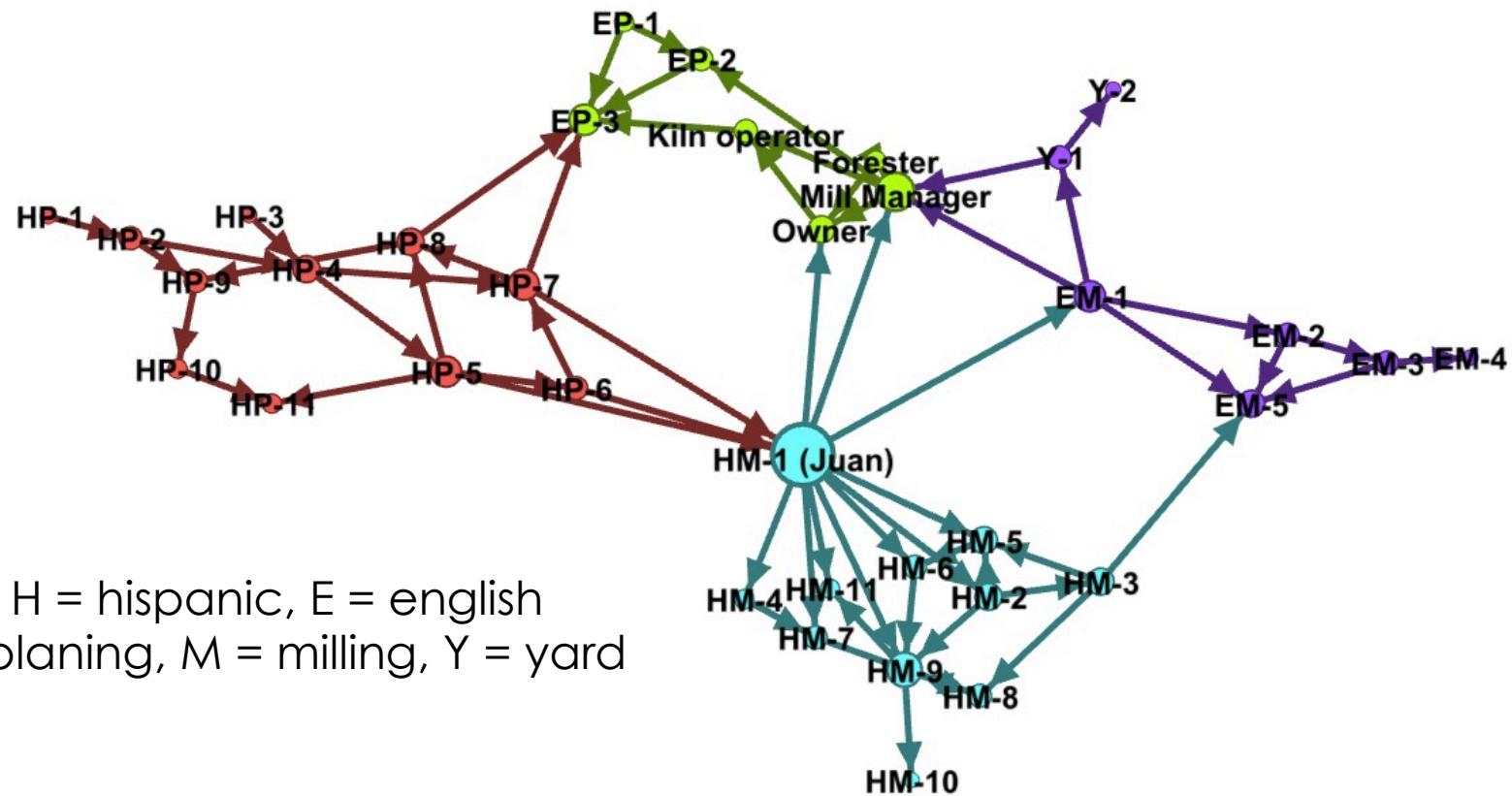
(a) *Karate club network*



(b) *After a split into two clubs*

source:Easley/Kleinberg

Why look for community structure?



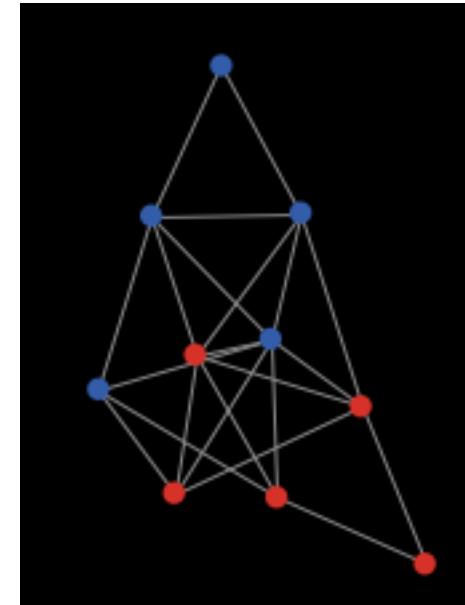
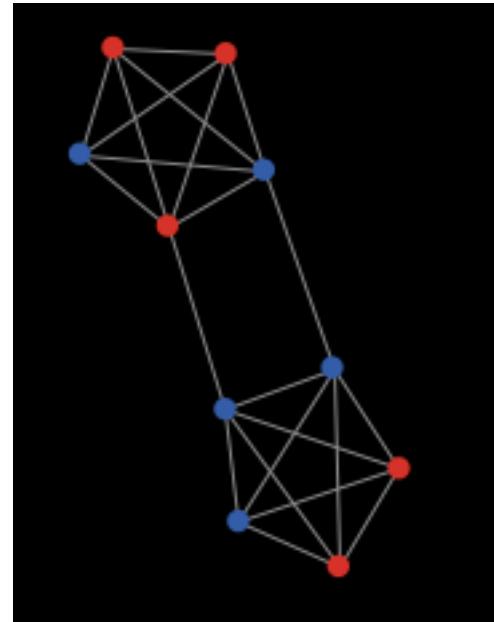
Sawmill network: source Exploratory Social Network Analysis with Pajek

Quiz Q:

- The management at the sawmill was having difficulty persuading the workers to adopt a new plan, even though everyone would benefit. In particular the Hispanic workers (H) were reluctant to agree. The management called in a sociologist who mapped out who talked to whom regularly. Then they suggested that the management talk to Juan and have him talk to the Hispanic workers. It was a success, promptly everyone was on board with the new plan. Why?

opinion formation and community structure

- [http://www.ladamic.com/netlearn/NetLogo502/
OpinionFormationModelToy.html](http://www.ladamic.com/netlearn/NetLogo502/OpinionFormationModelToy.html)
- each node adopts the majority opinion of its neighbors (flips a coin if it's a tie)

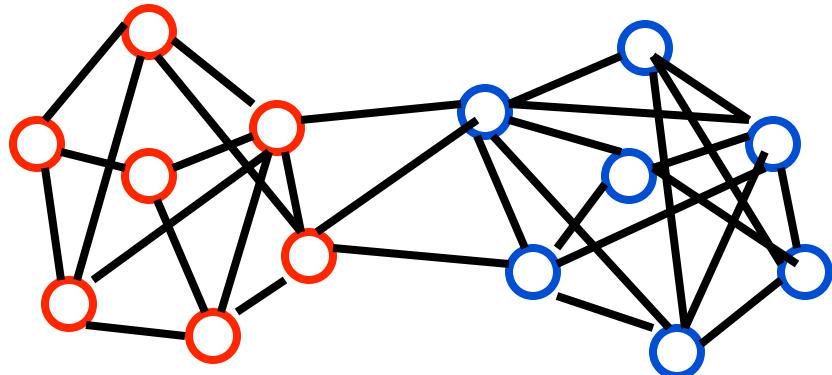


Quiz Q:

- ❑ Alternate between the 2 community and the Erdos-Renyi configuration. Which can maintain divergent opinions when you iterate opinion updates:
 - ❑ just Erdos-Renyi
 - ❑ just 2-community
 - ❑ both

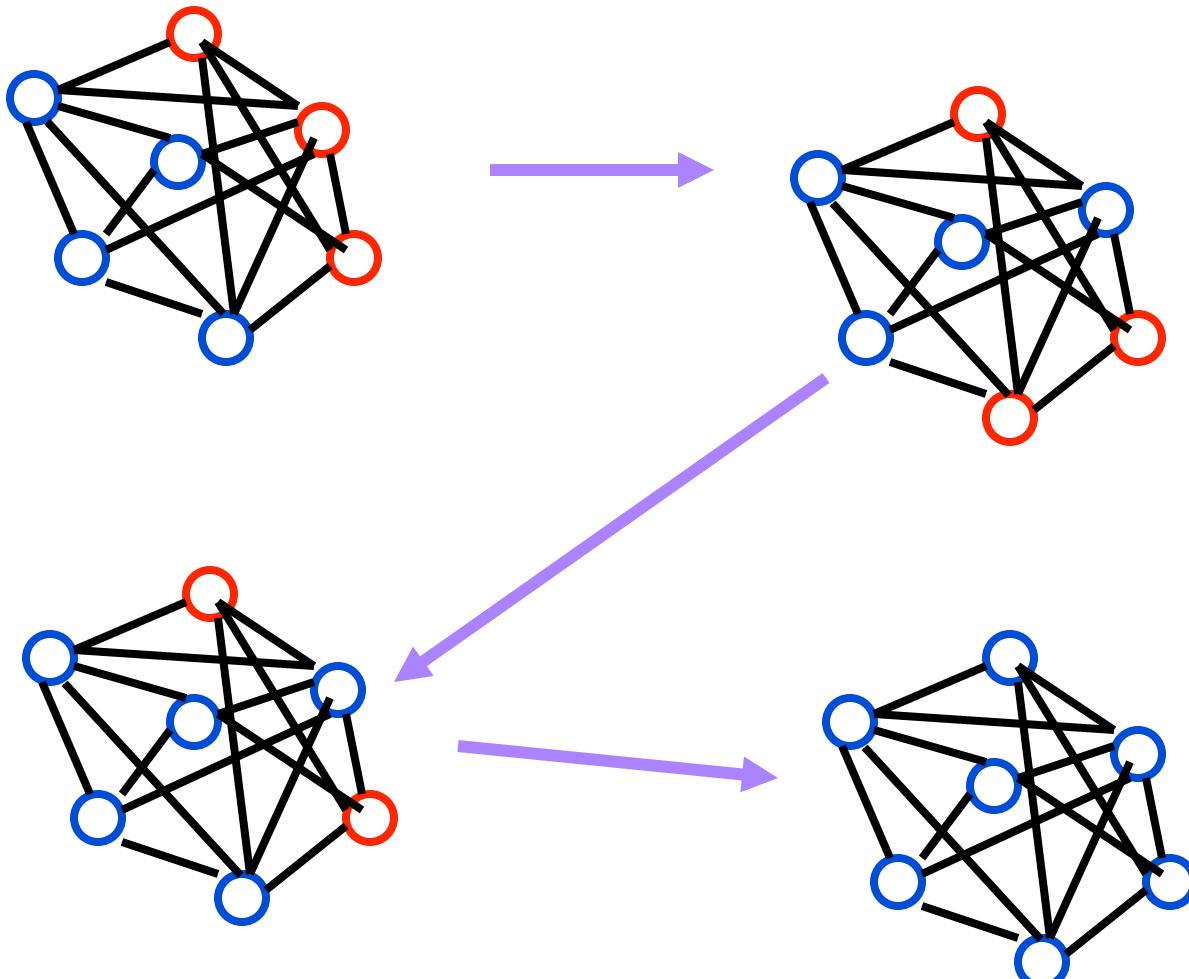
Why care about group cohesion?

- ❑ opinion formation and uniformity



- if each node adopts the opinion of the majority of its neighbors, it is possible to have different opinions in different cohesive subgroups

within a cohesive subgroup – greater uniformity



high-res maps of science

<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803>



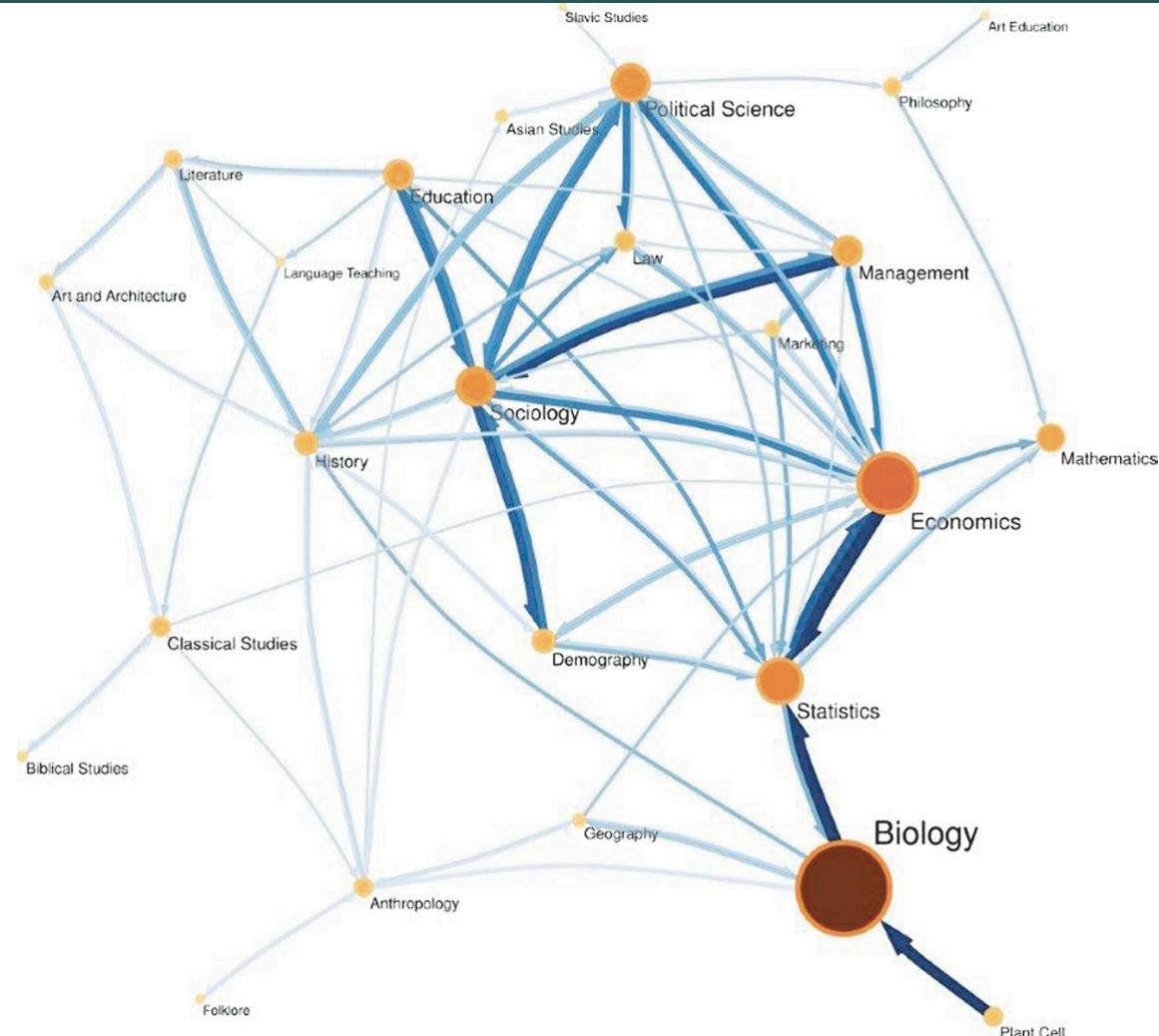
high-res maps of science

<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803>



high-res maps of science

<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803>

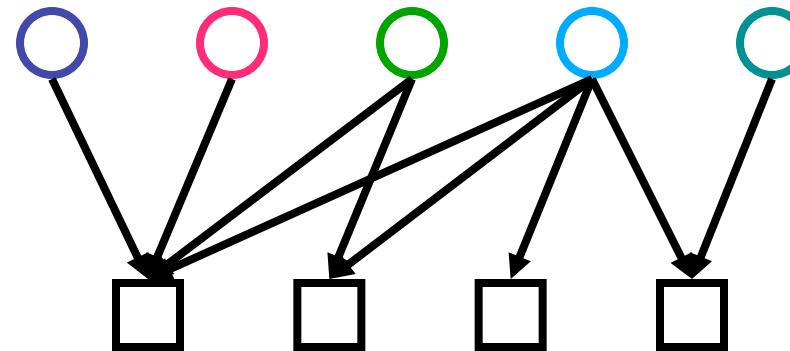
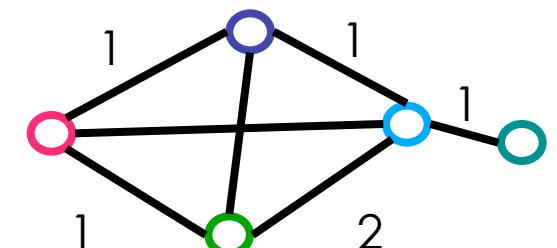


What makes a community?

- mutuality of ties
 - everybody in the group knows everybody else
- frequency of ties among members
 - everybody in the group has links to at least k others in the group
- closeness or reachability of subgroup members
 - individuals are separated by at most n hops
- relative frequency of ties among subgroup members compared to nonmembers

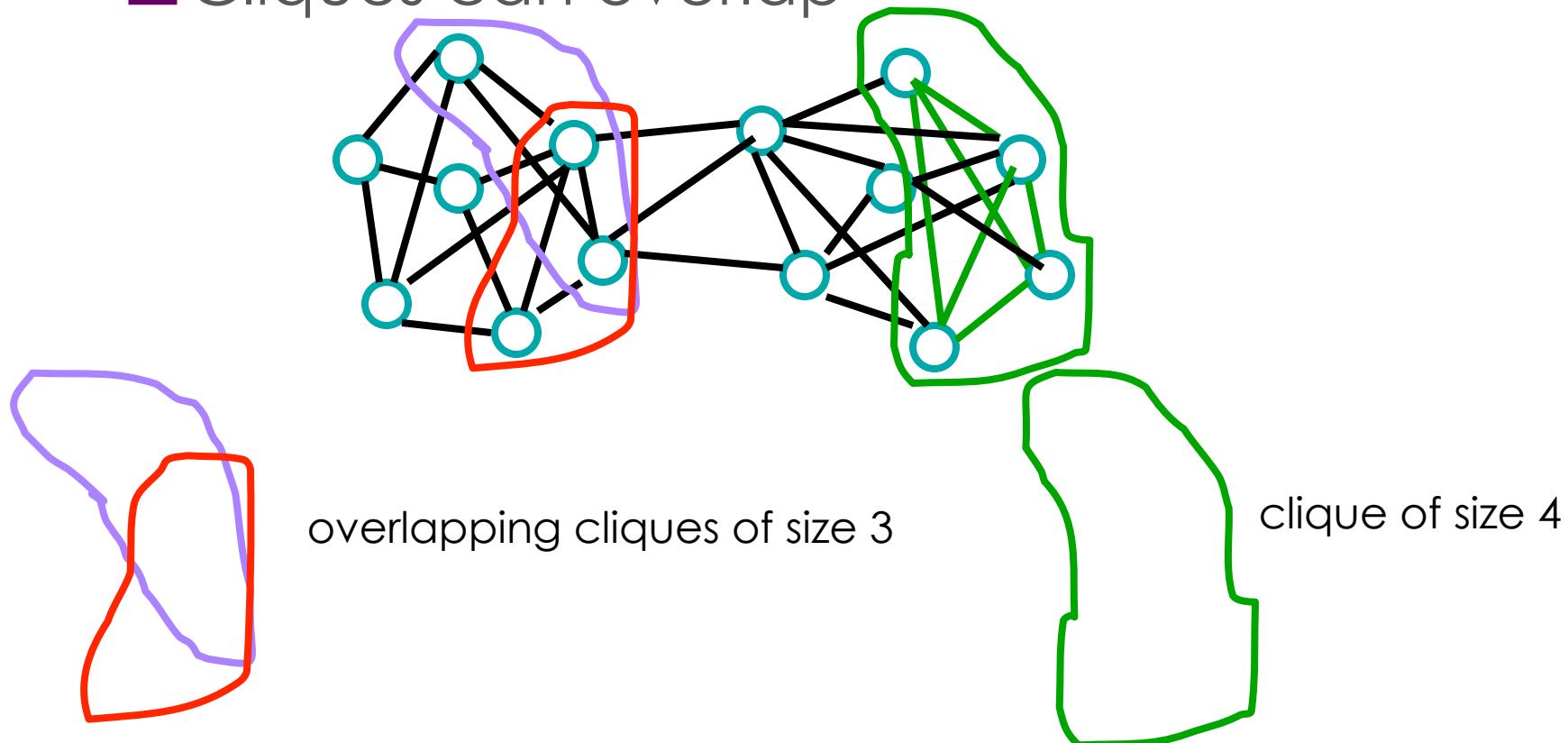
Affiliation networks

- ❑ otherwise known as
 - ❑ membership network
 - ❑ e.g. board of directors
 - ❑ hypernetwork or hypergraph
 - ❑ bipartite graphs
 - ❑ interlocks



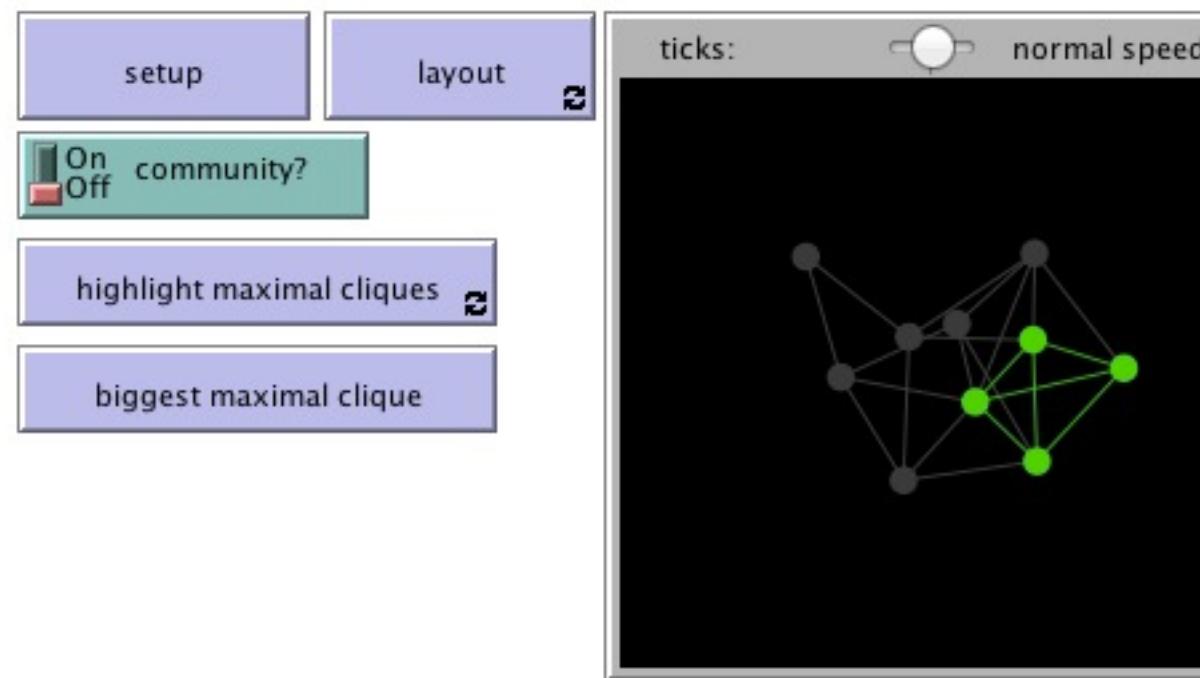
Cliques

- Every member of the group has links to every other member
- Cliques can overlap



Cliques betray community structure

- ❑ Go to
<http://www.ladamic.com/netlearn/nw/Cliques.html>
- ❑ Try the ER vs. community structure setup (they are the same as for the opinion formation model)



Quiz question

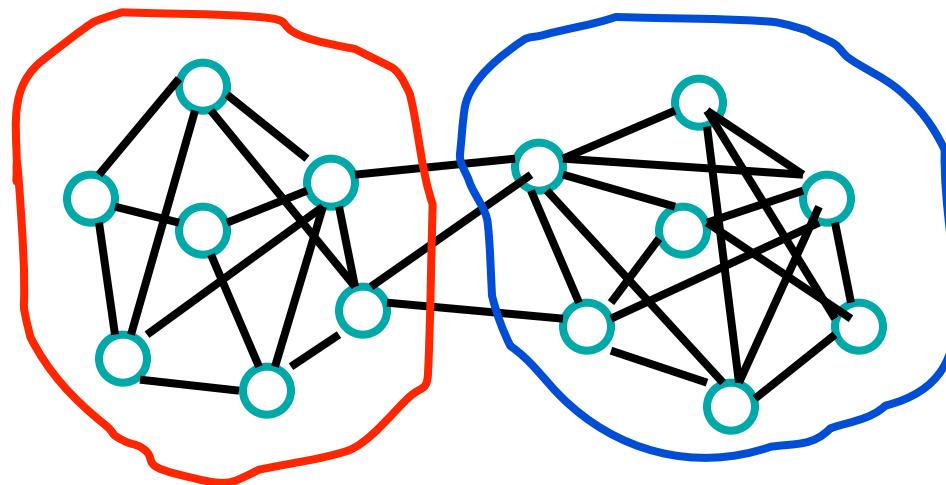
- ❑ Which has a larger maximal clique?
 - ❑ network with community structure
 - ❑ the equivalent ER random graph

Meaningfulness of cliques

- ❑ Not robust
 - ❑ one missing link can disqualify a clique
- ❑ Not interesting
 - ❑ everybody is connected to everybody else
 - ❑ no core-periphery structure
 - ❑ no centrality measures apply
- ❑ How cliques overlap can be more interesting than that they exist

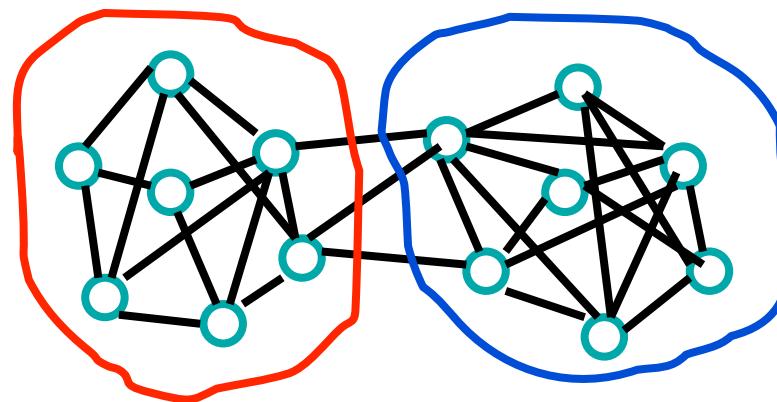
k-cores: similar idea, less stringent

- Each node within a group is connected to k other nodes in the group



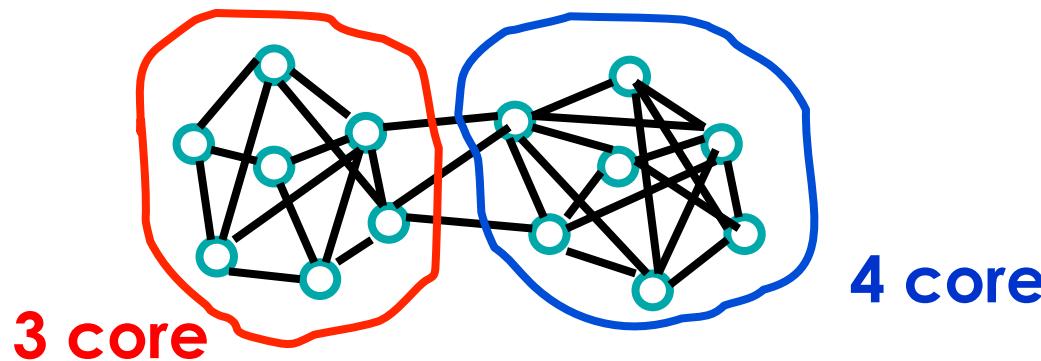
Quiz Question

- ❑ What is the “k” for the core circled in red?
- ❑ What is the “k” for the core circled in blue?

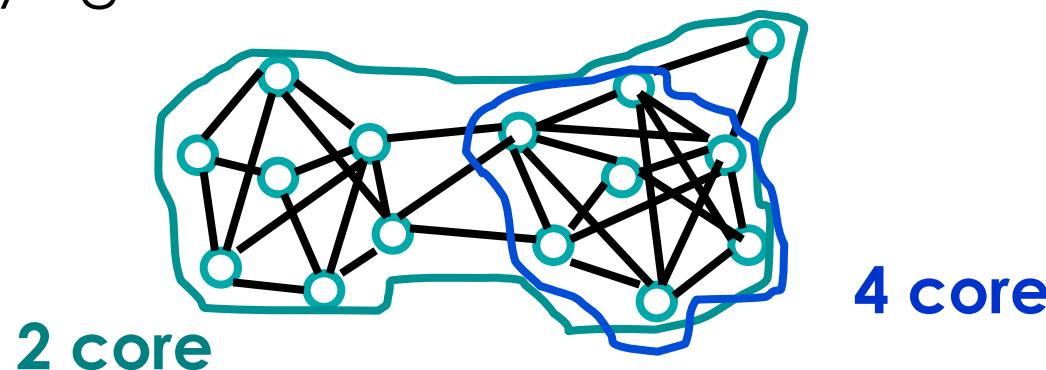


k-cores

- Each node within a group is connected to k other nodes in the group



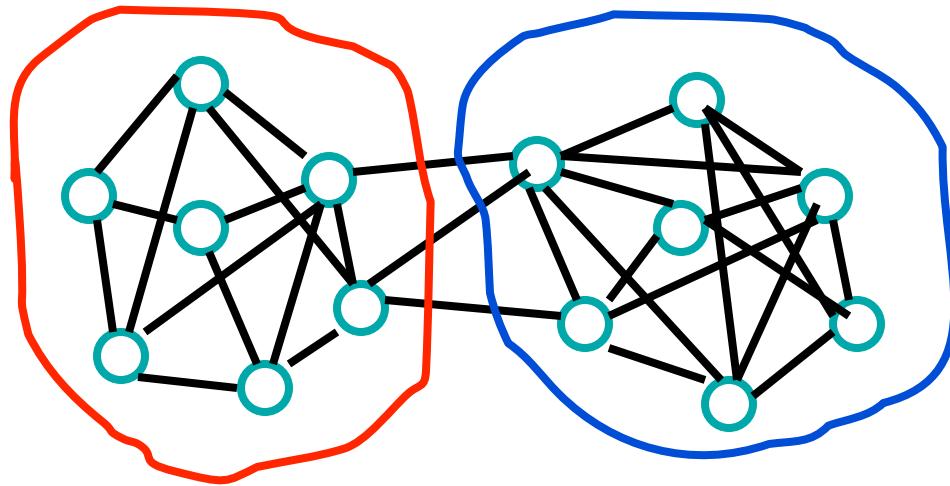
- but even this is too stringent of a requirement for identifying natural communities



subgroups based on reachability and diameter

- n – cliques

- maximal distance between any two nodes in subgroup is n



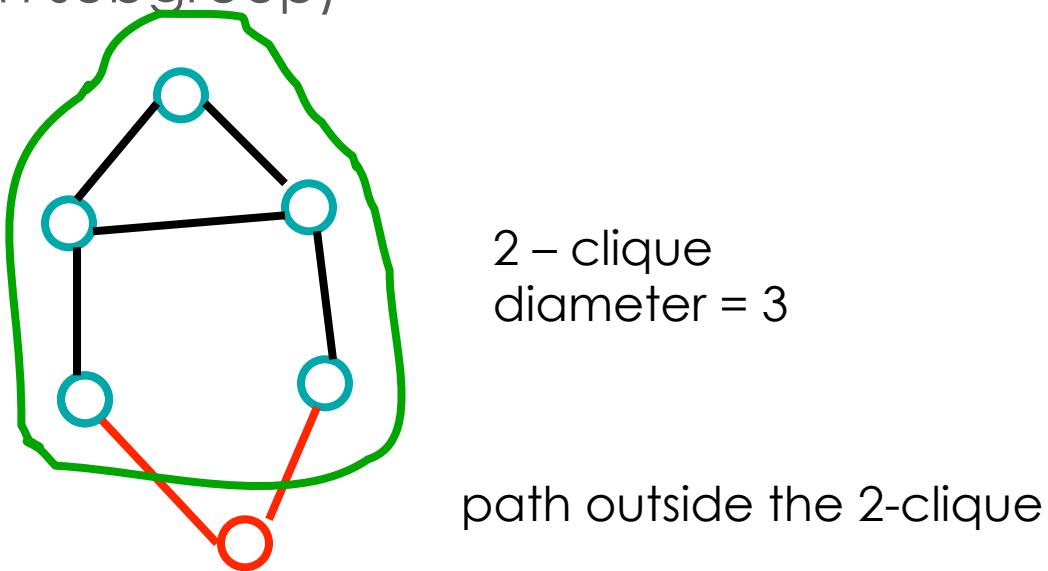
2-cliques

- theoretical justification
 - information flow through intermediaries

considerations with n-cliques

■ problem

- diameter may be greater than n
- n-clique may be disconnected (paths go through nodes not in subgroup)

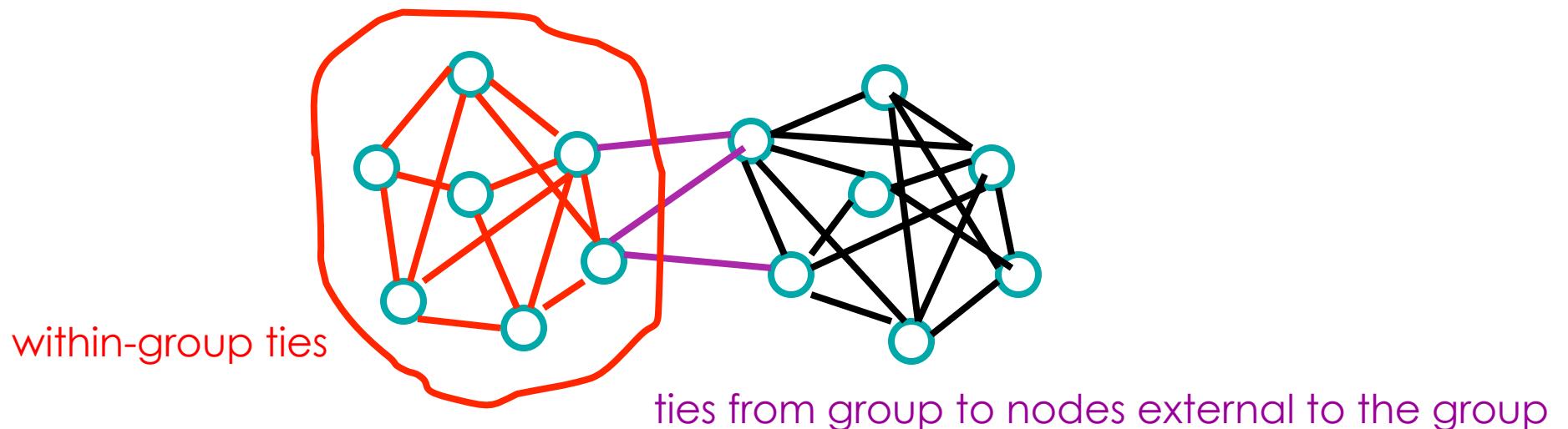


■ fix

- n-club: maximal subgraph of diameter 2

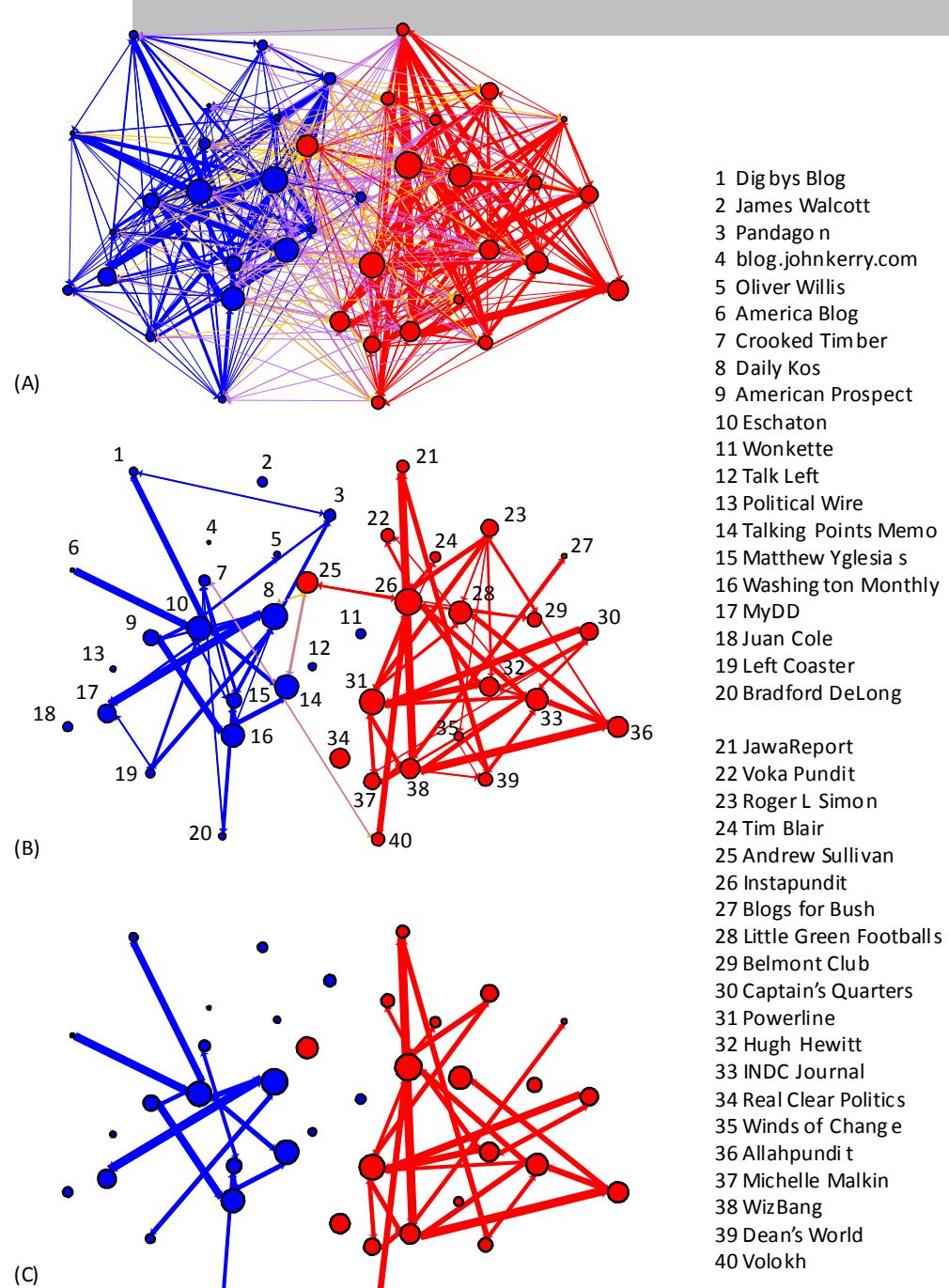
p-cliques: frequency of in group ties

- partition the network into clusters where vertices have at least a proportion p (number between 0 and 1) of neighbors inside the cluster.



cohesion in directed & weighted networks

- ❑ something we've already learned how to do:
 - ❑ find strongly connected components
- ❑ keep only a subset of ties before finding connected components
 - ❑ reciprocal ties
 - ❑ edge weight above a threshold



Example: political blogs

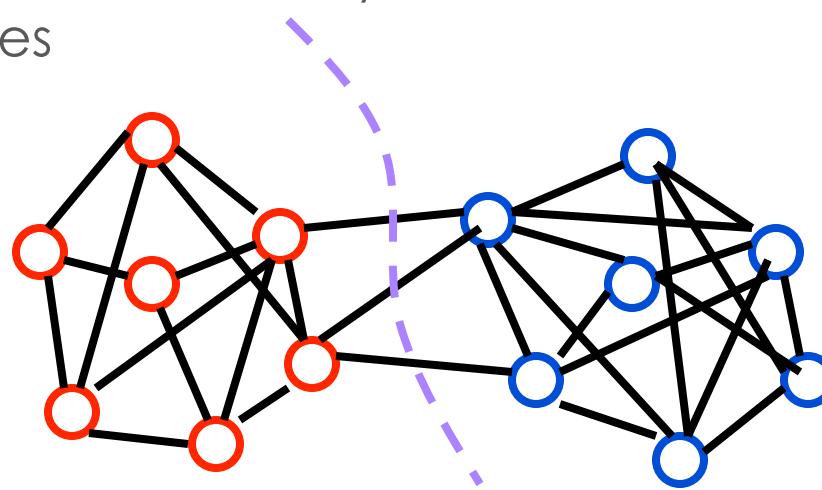
(Aug 29th – Nov 15th, 2004)

- A) all citations between A-list blogs in 2 months preceding the 2004 election
- B) citations between A-list blogs with at least 5 citations in both directions
- C) edges further limited to those exceeding 25 combined citations

only 15% of the citations bridge communities

Community finding vs. other approaches

- Social and other networks have a natural community structure
- We want to discover this structure rather than impose a certain size of community or fix the number of communities

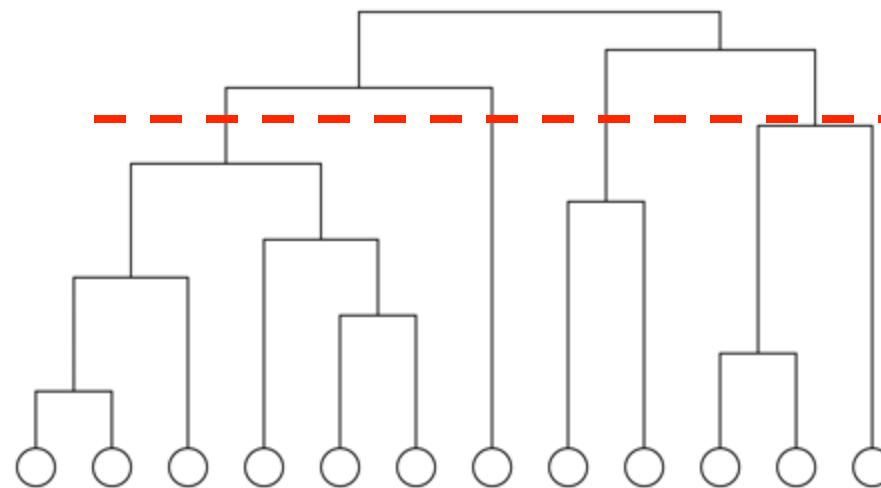


- Without “looking”, can we discover community structure in an automated way?

Hierarchical clustering

Process:

- ❑ after calculating the “distances” for all pairs of vertices
- ❑ start with all n vertices disconnected
- ❑ add edges between pairs one by one in order of decreasing weight
- ❑ result: nested components, where one can take a ‘slice’ at any level of the tree

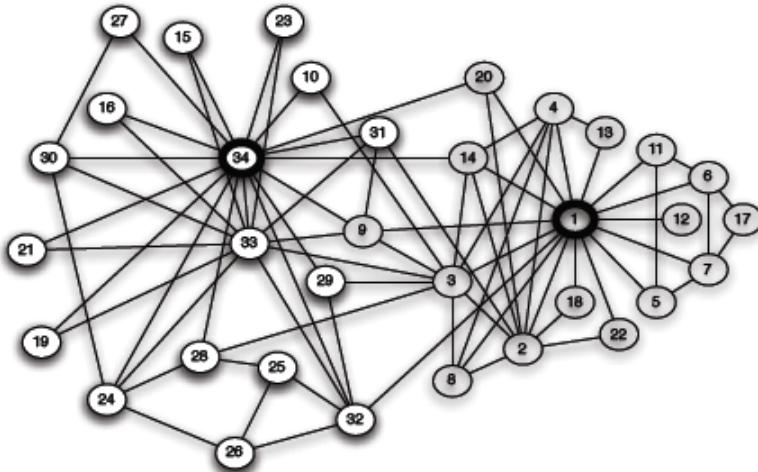


Hierarchical clustering

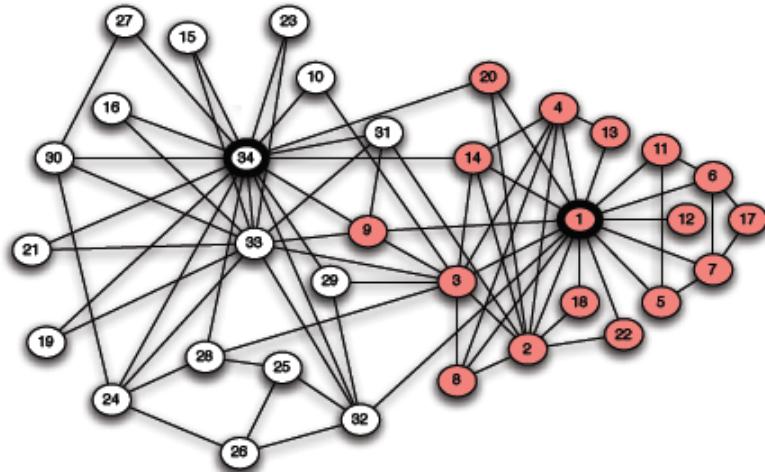
■ Process:

- after calculating the weights W for all pairs of vertices
- start with all n vertices disconnected
- add edges between pairs one by one in order of decreasing weight
- Efficient and successful implementation in Pajek...

Zachary Karate Club



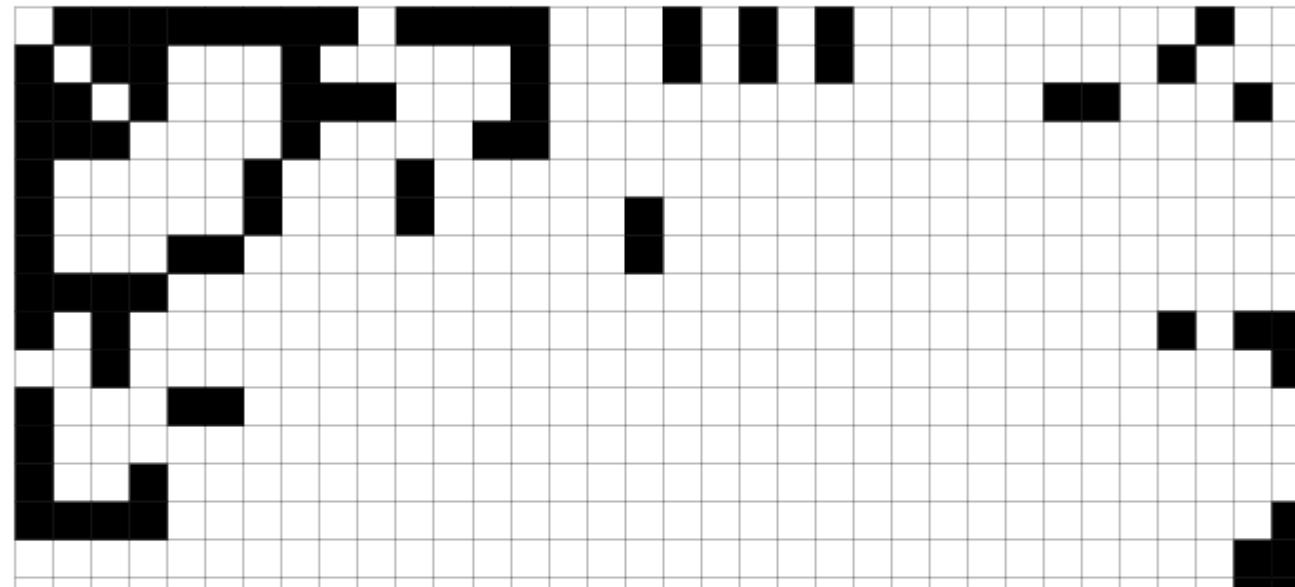
(a) *Karate club network*



(b) *After a split into two clubs*

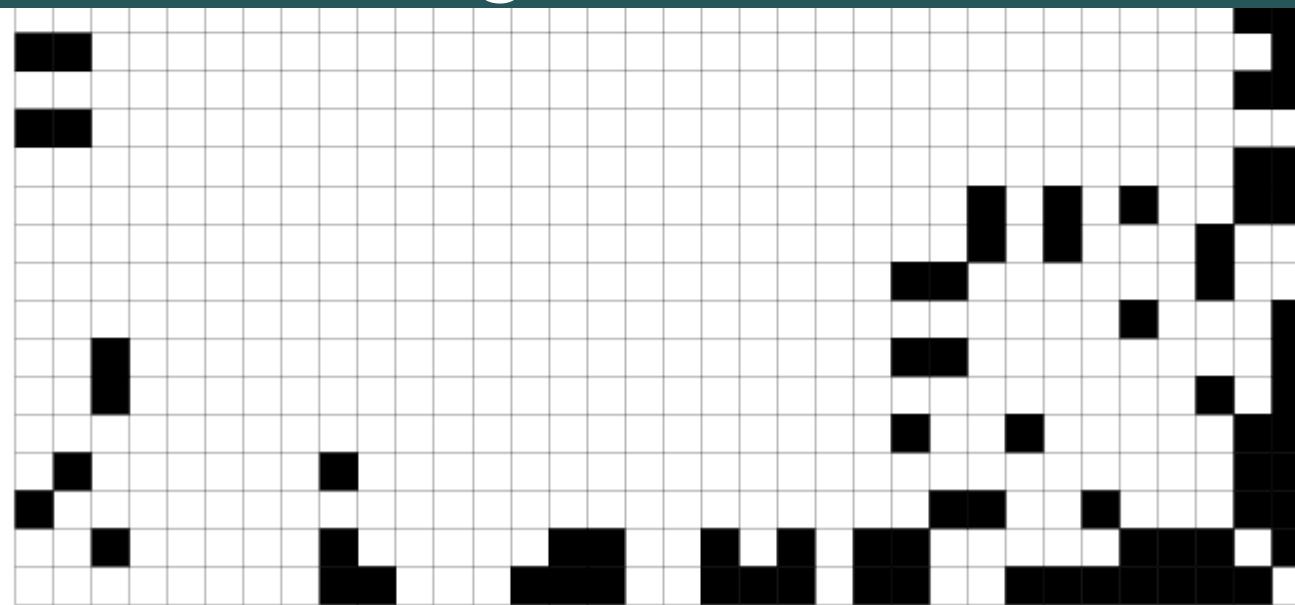
source:Easley/Kleinberg

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15



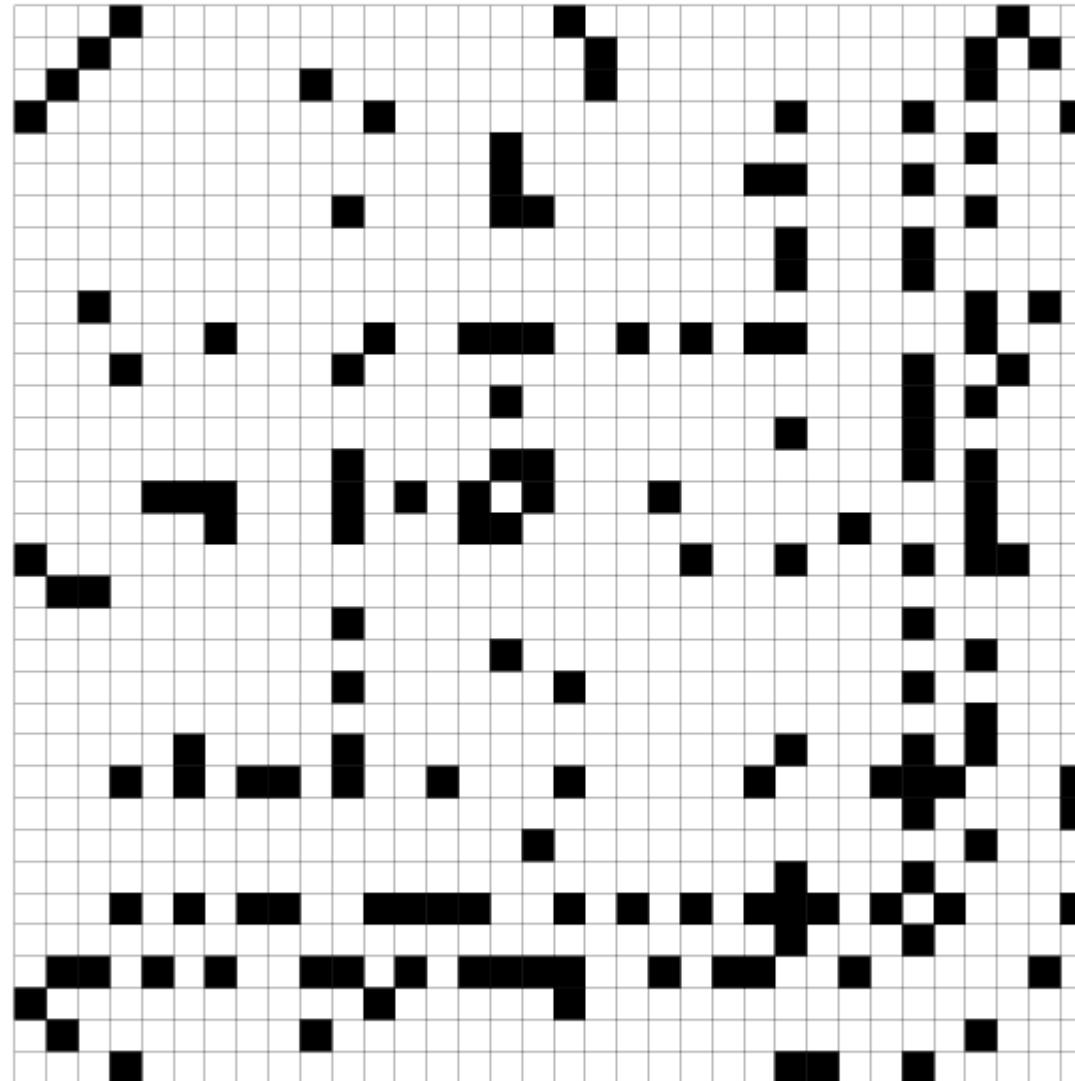
original matrix

19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34



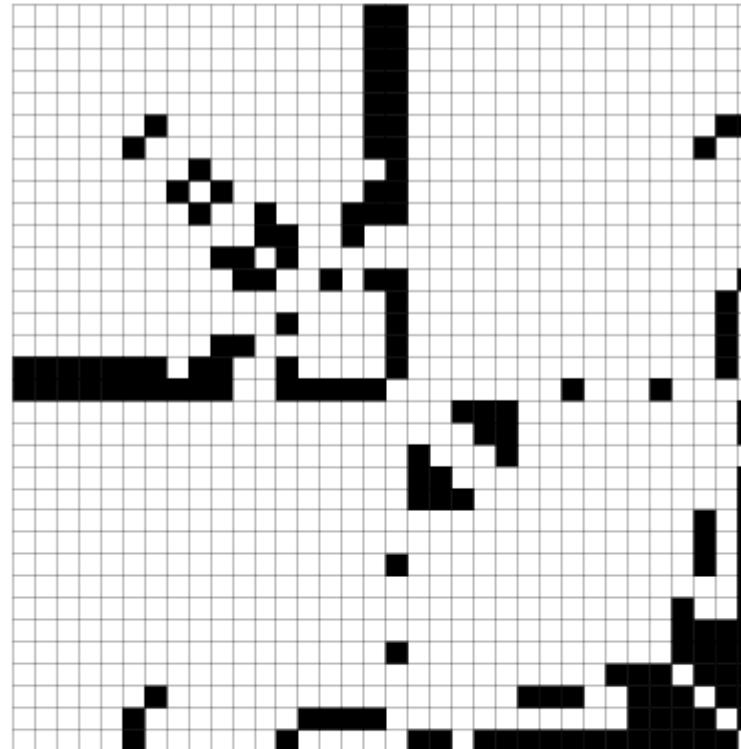
randomized karate club matrix

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34



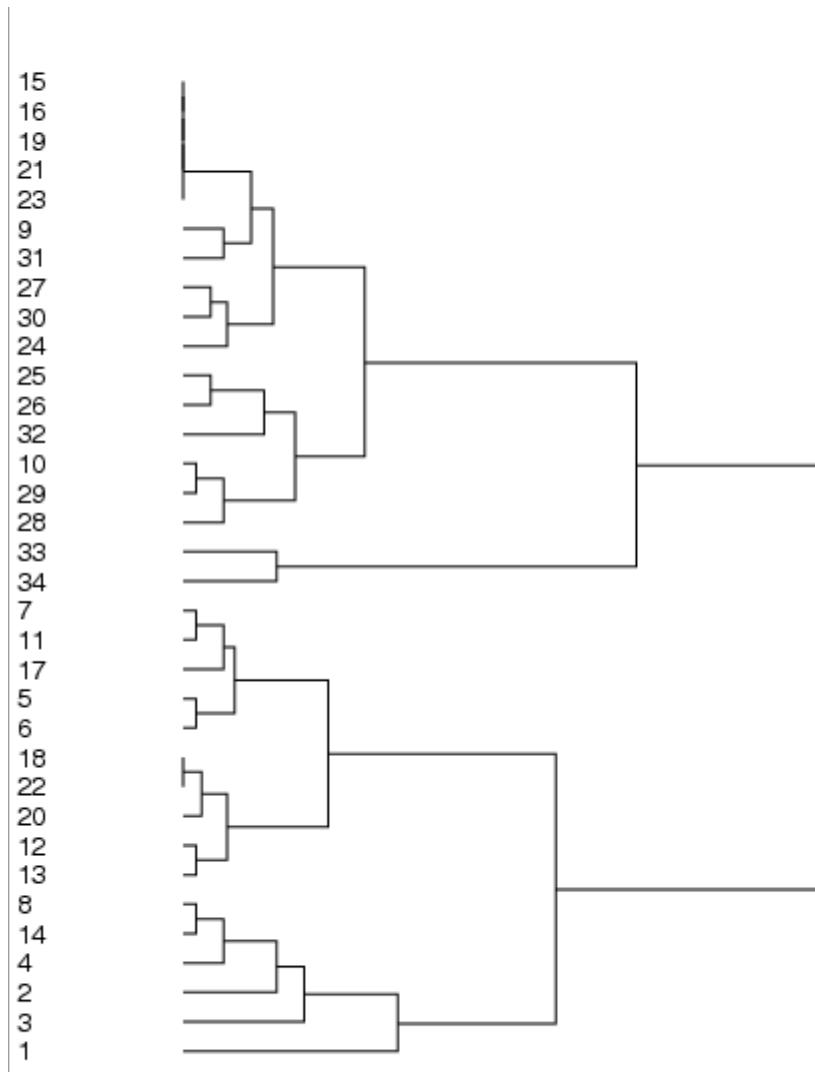
permuted matrix

15
16
19
21
23
9
31
27
30
24
25
26
32
10
29
28
33
34
7
11
17
5
6
18
22
20
12
13
8
14
4
2
3
1



15 16 19 21 23 9 31 27 30 24 25 26 32 10 29 28 33 34 7 11 17 5 6 18 22 20 12 13 8 14 4 2 3 1

dendrogram

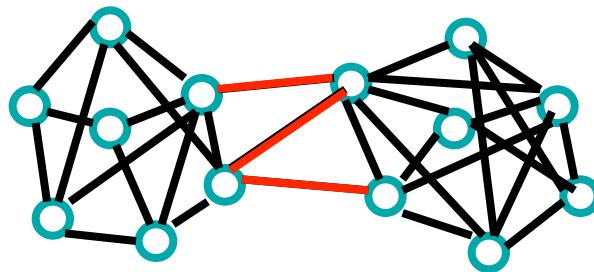


betweenness clustering

- Algorithm
 - compute the betweenness of all edges
 - while (betweenness of any edge > threshold):
 - remove edge with highest betweenness
 - recalculate betweenness

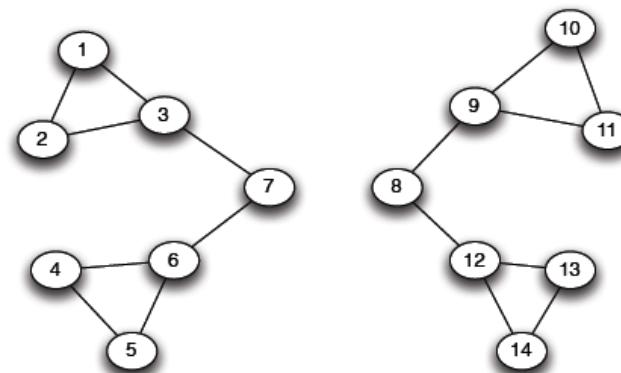
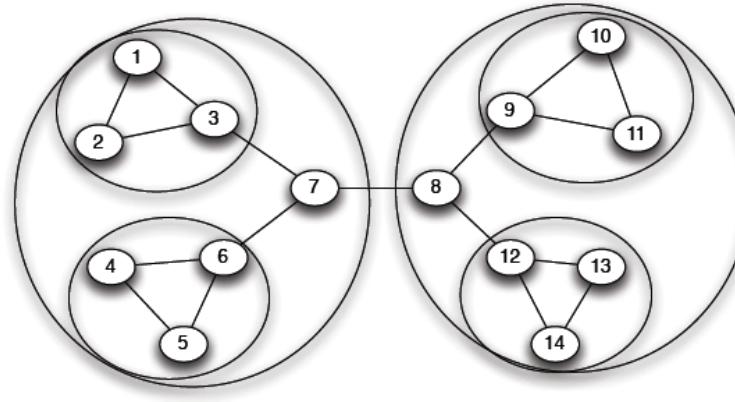
- Betweenness needs to be recalculated at each step
 - removal of an edge can impact the betweenness of another edge
 - very expensive: all pairs shortest path – $O(N^3)$
 - may need to repeat up to N times
 - does not scale to more than a few hundred nodes, even with the fastest algorithms

betweenness clustering algorithm

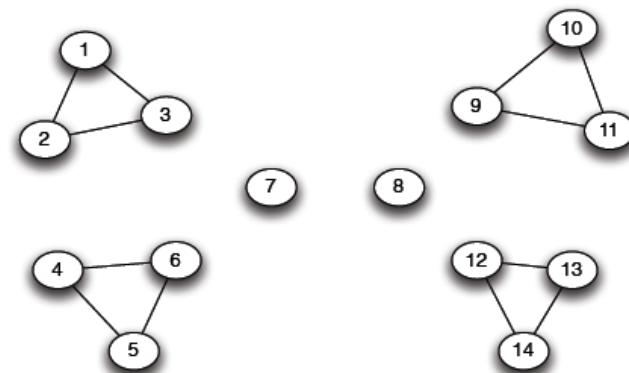


betweenness clustering:

- successively remove edges of highest betweenness (the bridges, or local bridges), breaking up the network into separate components

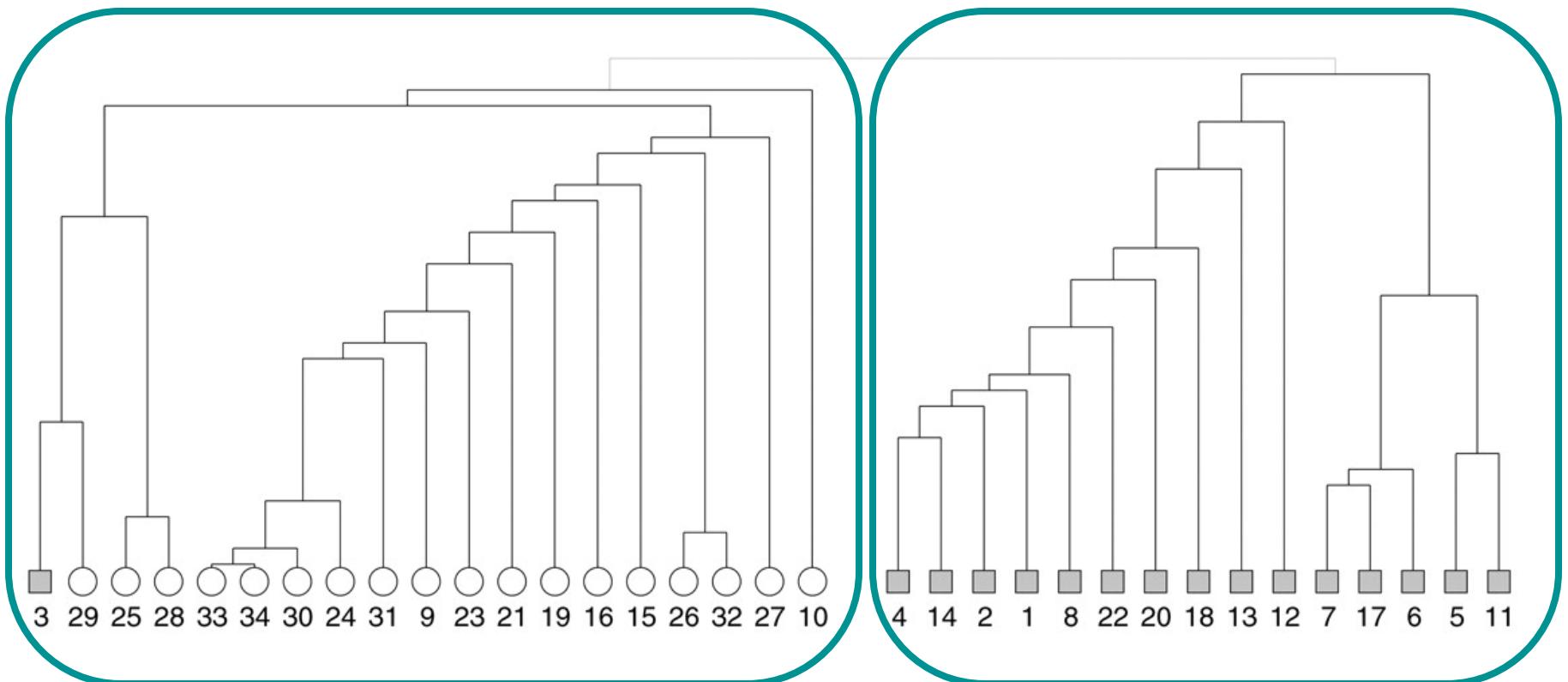


(a) *Step 1*



(b) *Step 2*

betweenness clustering algorithm & the karate club data set



source: Girvan and Newman, PNAS June 11, 2002 99(12):7821-7826

Modularity

- Consider edges that fall within a community or between a community and the rest of the network
- Define modularity:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \delta(c_v, c_w) \right]$$

adjacency matrix

if vertices are in the same community

probability of an edge between two vertices is proportional to their degrees

- For a random network, $Q = 0$
 - the number of edges within a community is no different from what you would expect

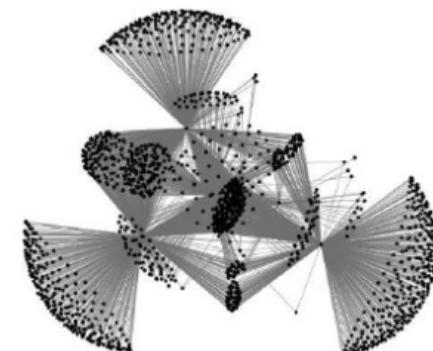
Finding community structure in very large networks

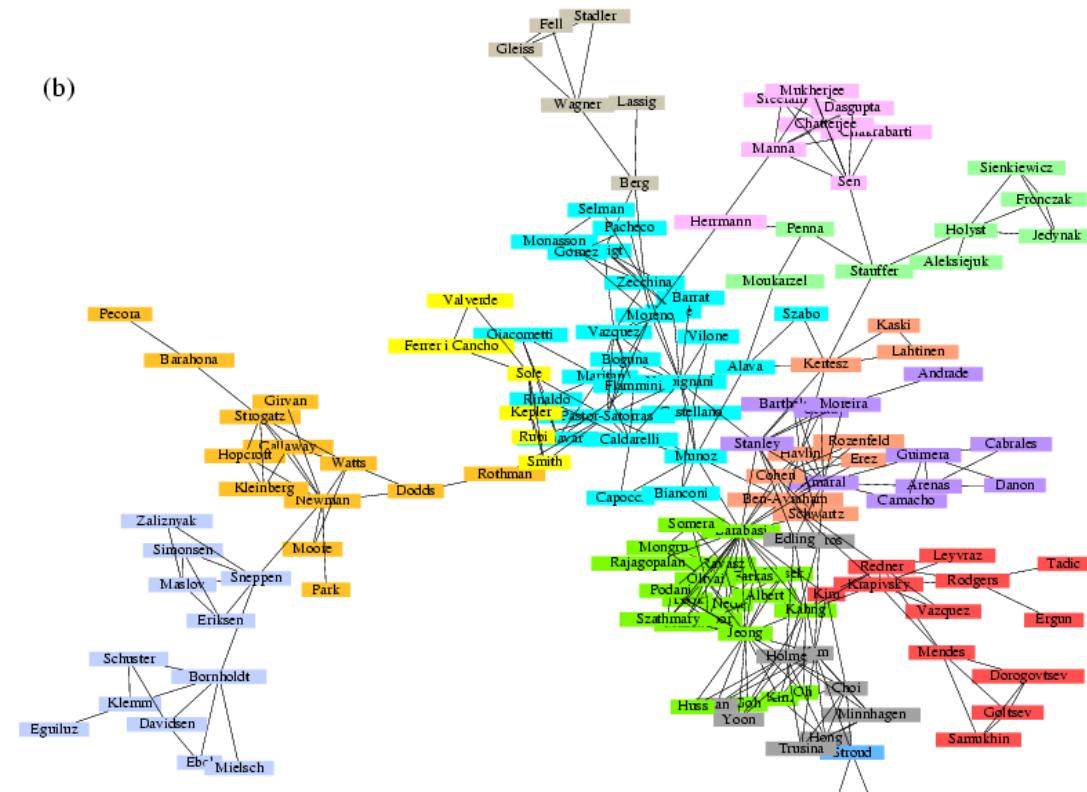
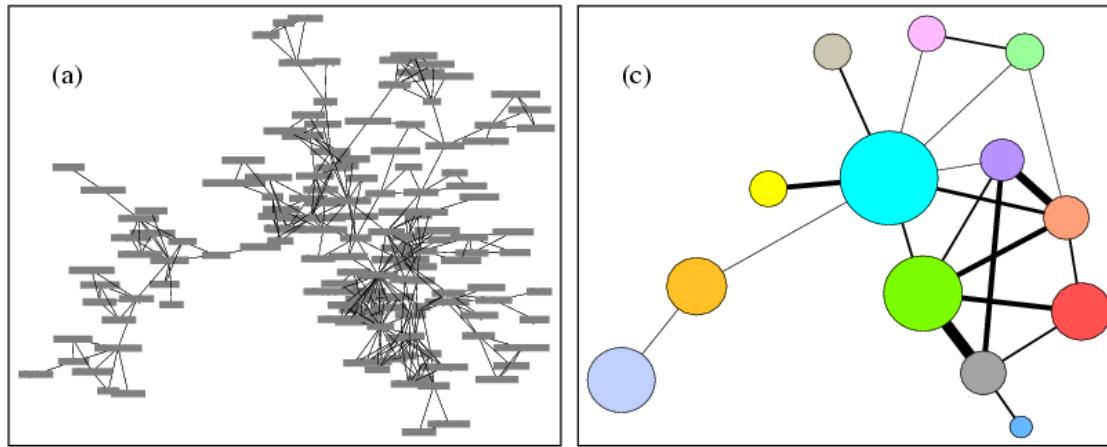
Authors: [Aaron Clauset](#), [M. E. J. Newman](#), [Cristopher Moore](#) 2004

Modularity

❑ Algorithm

- ❑ start with all vertices as isolates
- ❑ follow a greedy strategy:
 - ❑ successively join clusters with the greatest increase ΔQ in modularity
 - ❑ stop when the maximum possible $\Delta Q \leq 0$ from joining any two
- ❑ successfully used to find community structure in a graph with $> 400,000$ nodes with > 2 million edges
 - ❑ Amazon's people who bought this also bought that...
- ❑ alternatives to achieving optimum ΔQ :
 - ❑ simulated annealing rather than gre





**Reminder of
how
modularity
can help us
visualize large
networks**

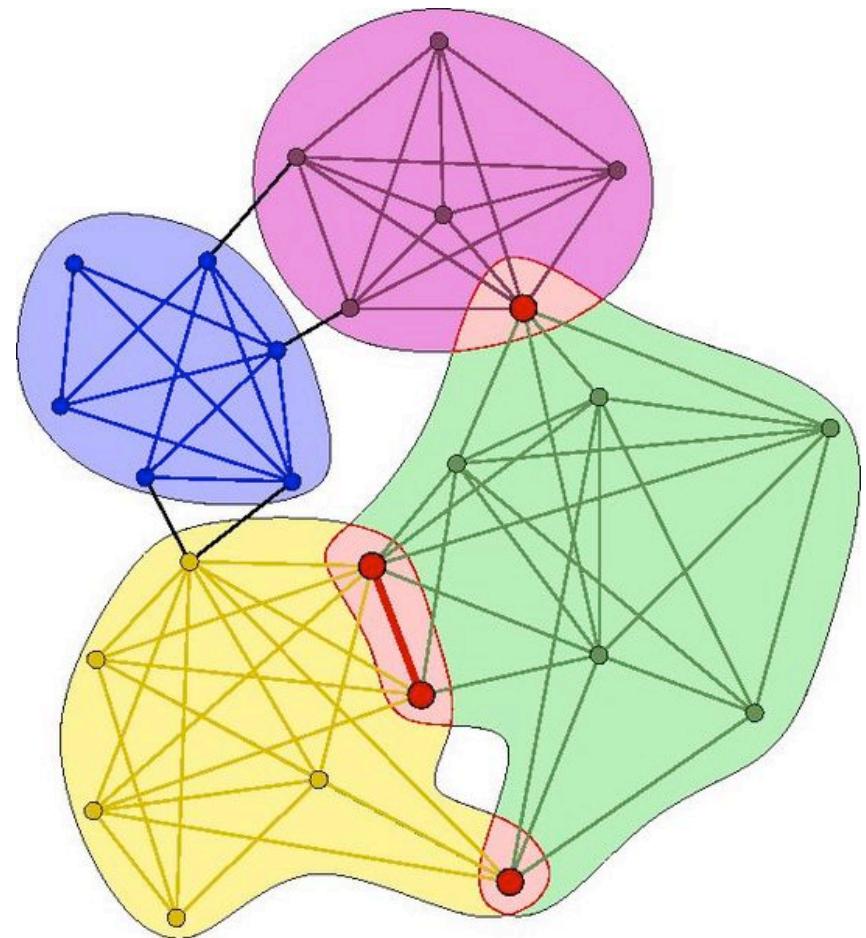
What if communities overlap?

- Recent research has found that for communities such as Orkut and Flickr, community finding algorithms cannot identify communities of more than ~100 nodes
- [Statistical Properties of Community Structure in Large Social and Information Networks](#) by J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. *International World Wide Web Conference (WWW)*, 2008. [[Video](#)]

Clique finder

■ <http://cfinder.org>

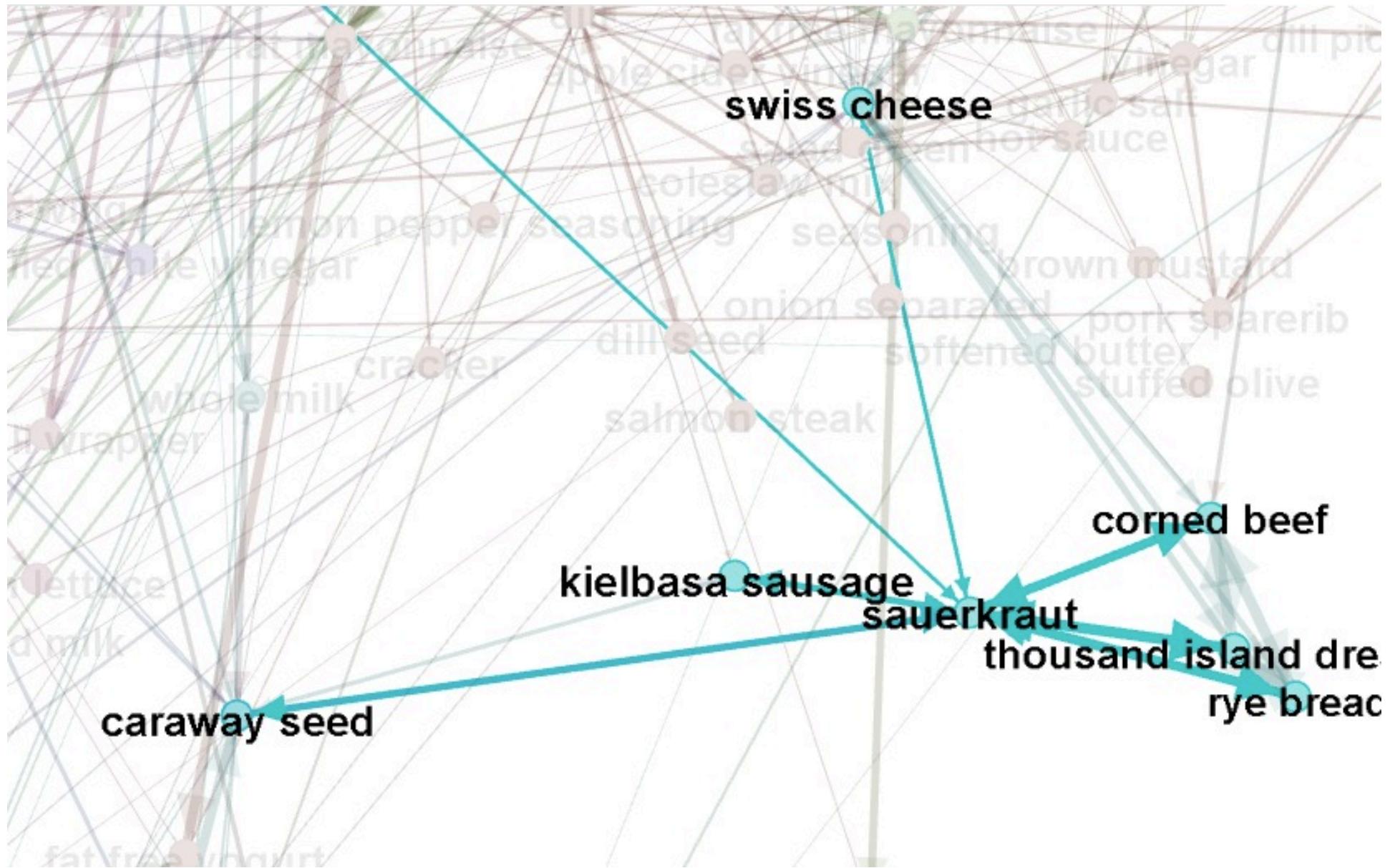
Uncovering the overlapping community structure of complex networks in nature and society G. Palla, I. Derényi, I. Farkas, and T. Vicsek: Nature 435, 814–818 (2005)



wrap up

- ❑ community structure is a way of ‘x-rayng’ the network, finding out what it’s made of
- ❑ you can look for specific structures
 - ❑ k-cliques, k-cores, etc.
- ❑ but most popular is to discover the “natural” community boundaries

For your assignment: community finding & ingredients ☺

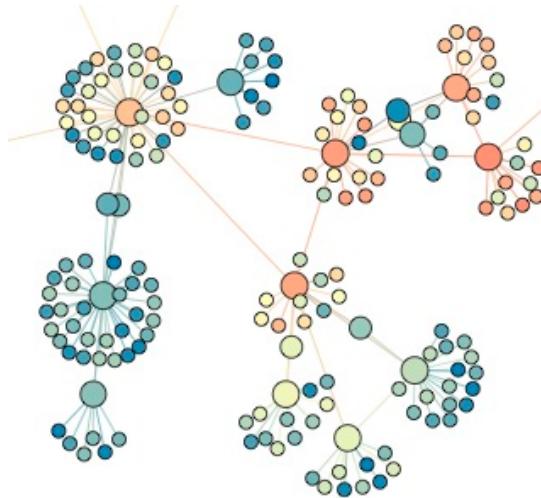


An information theoretic approach

- How to most concisely describe a random walk on the network using huffman codes? Prefixes become communities...
-

[http://www.pnas.org/content/
105/4/1118](http://www.pnas.org/content/105/4/1118)

[http://www.mapequation.org/
mapgenerator/index.html](http://www.mapequation.org/mapgenerator/index.html)



SNA 5: small world networks

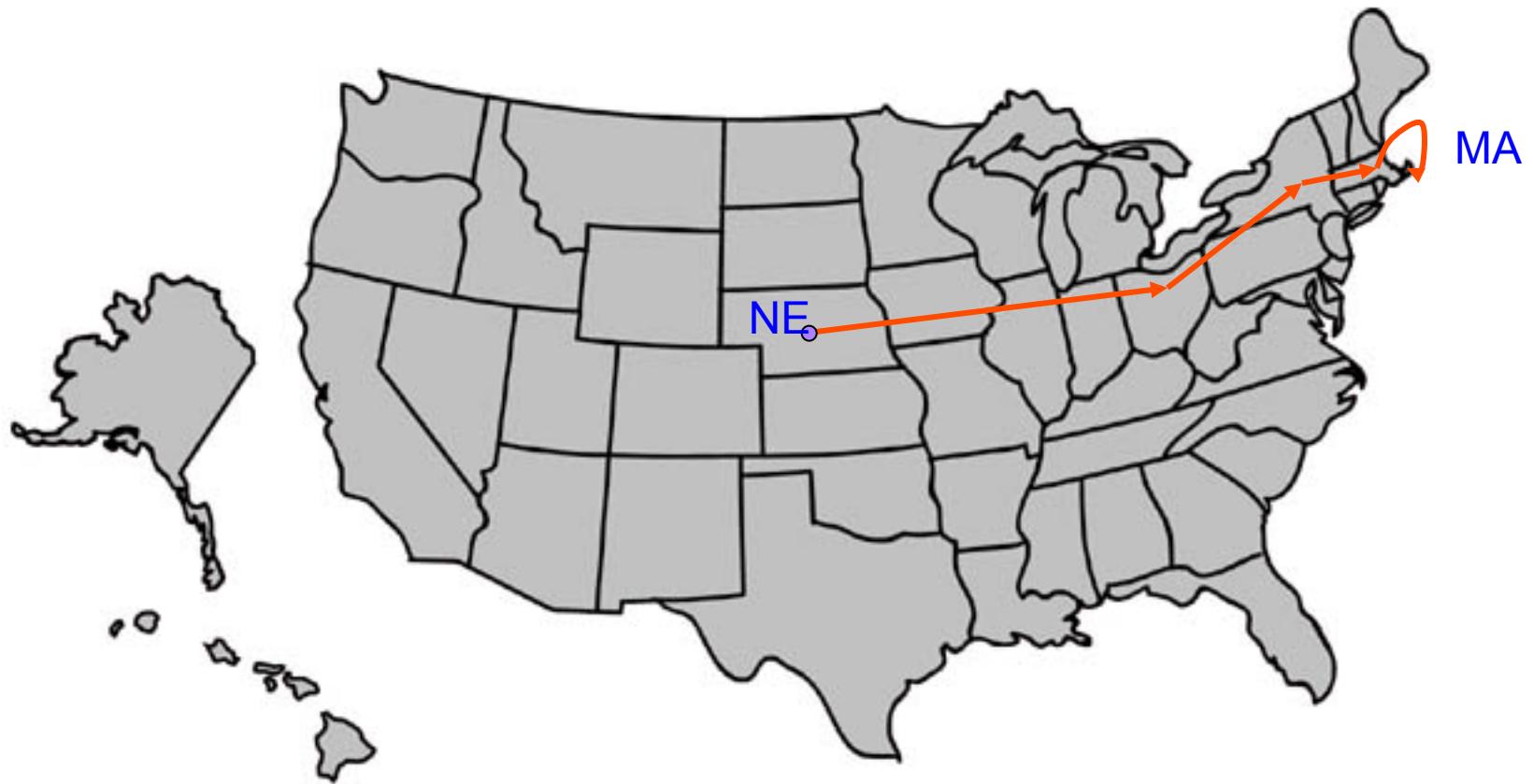
Lada Adamic



Outline

- Small world phenomenon
 - Milgram's small world experiment
- Local structure
 - clustering coefficient
 - motifs
- Small world network models:
 - Watts & Strogatz (clustering & short paths)
 - Kleinberg (geographical)
 - Kleinberg, Watts/Dodds/Newman (hierarchical)
- Small world networks: why do they arise?
- Next week: what are the consequences for diffusion, coordination and learning.

Small world phenomenon: Milgram's experiment



Milgram's experiment

Instructions:

Given a target individual (stockbroker in Boston), pass the message to a person you correspond with who is “closest” to the target.

Outcome:

**20% of initiated chains reached target
average chain length = 6.5**

- ▣ “Six degrees of separation”

Milgram's experiment repeated

email experiment
Dodds, Muhamad, Watts,
Science 301, (2003)
(optional reading)

- 18 targets
- 13 different countries
- 60,000+ participants
- 24,163 message chains
- 384 reached their targets
- average path length 4.0



Interpreting Milgram's experiment

- Is 6 is a ***surprising*** number?
 - In the 1960s? Today? Why?
- Pool and Kochen in (1978 established that the average person has between 500 and 1500 acquaintances)

Quiz Q:

- ▢ Ignore for the time being the fact that many of your friends' friends are your friends as well. If everyone has 500 friends, the average person would have how many friends of friends?

Quiz Q:

- With an average degree of 500, a node in a random network would have this many friends-of-friends-of-friends (3rd degree neighbors):

Interpreting Milgram's experiment

- Is 6 is a **surprising** number?
 - In the 1960s? Today? Why?
- If social networks were random... ?
 - Pool and Kochen (1978) - ~500-1500 acquaintances/person
 - ~ 500 choices 1st link
 - ~ $500^2 = 250,000$ potential 2nd degree neighbors
 - ~ $500^3 = 125,000,000$ potential 3rd degree neighbors
- If networks are completely cliquish?
 - all my friends' friends are my friends
 - what would happen?

Quiz Q:

- ❑ If the network were completely cliquish, that is all of your friends of friends were also directly your friends, what would be true:

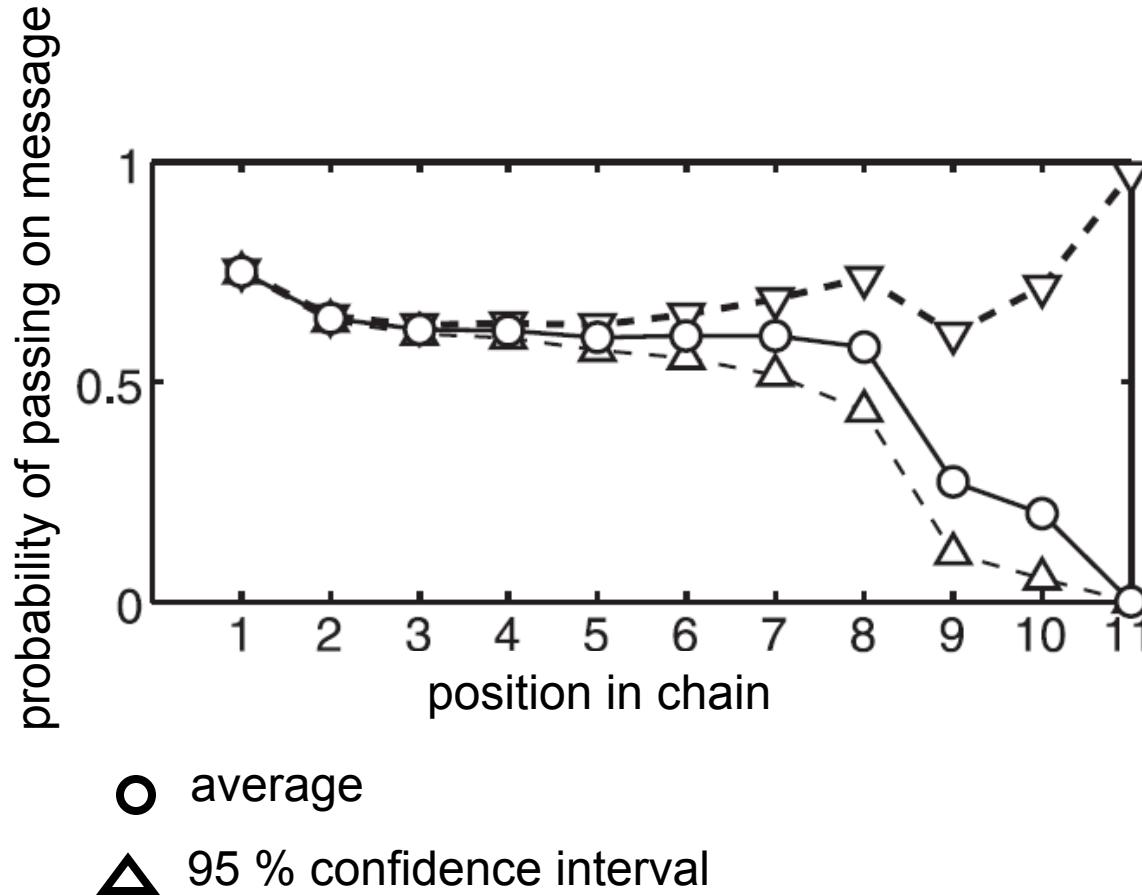
complete cliquishness

- If all your friends of friends were also your friends, you would be part of an isolated clique.

Uncompleted chains and distance

- Is 6 an ***accurate*** number?
- What bias is introduced by uncompleted chains?
 - are longer or shorter chains more likely to be completed?

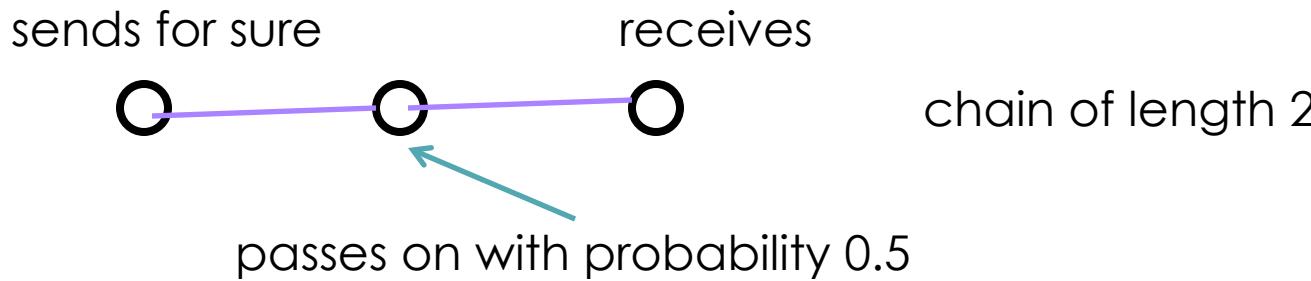
Attrition



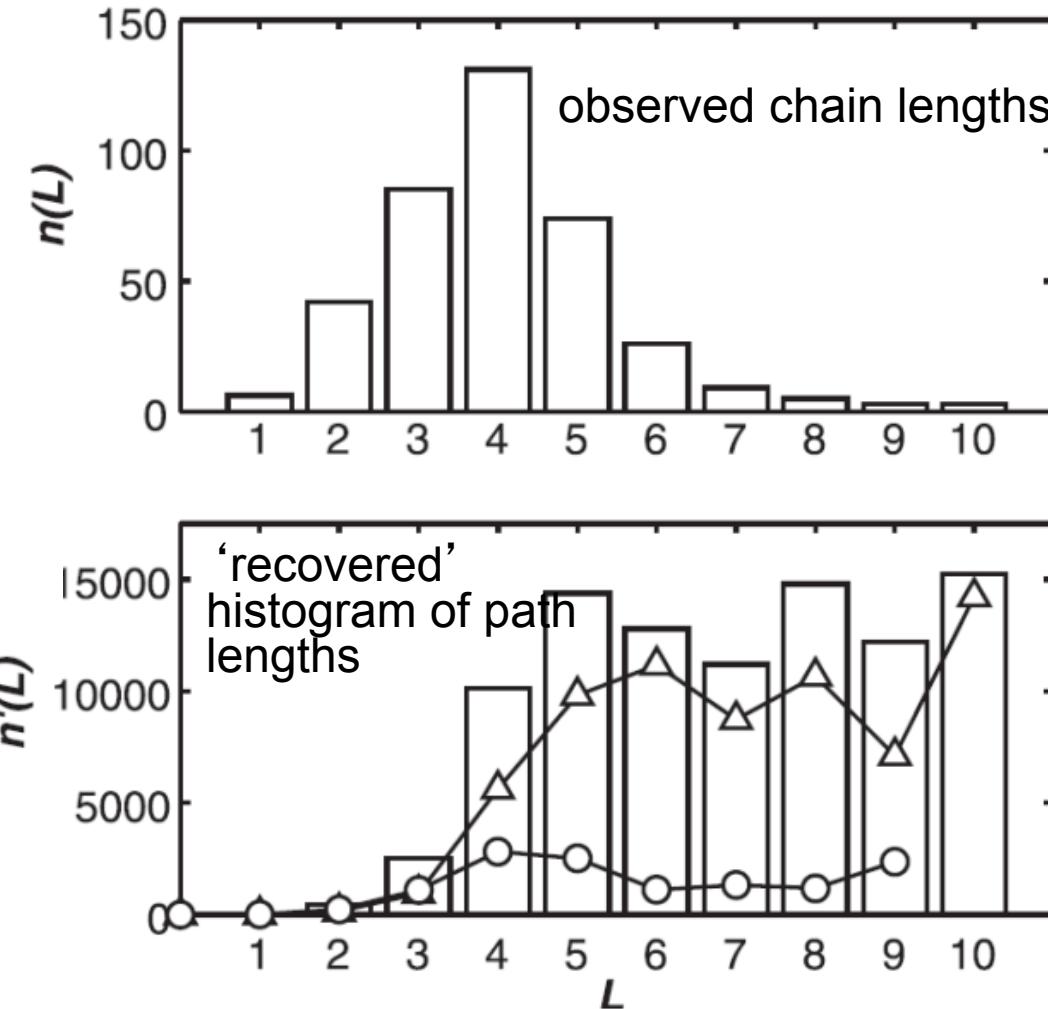
Source: An Experimental Study of Search in Global Social Networks: Peter Sheridan Dodds, Roby Muhammad, and Duncan J. Watts (8 August 2003); Science 301 (5634), 827.

Quiz Q:

- if each intermediate person in the chain has 0.5 probability of passing the letter on, what is the likelihood of a chain being completed
 - of length 2?
 - of length 5?



Estimating the true distance



Source: An Experimental Study of Search in Global Social Networks: Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts (8 August 2003); Science 301 (5634), 827.

Navigation and accuracy

- ❑ Is 6 an *accurate* number?

- ❑ Do people find the *shortest* paths?
 - ❑ Killworth, McCarty ,Bernard, & House (2005, optional):
 - ❑ less than optimal choice for next link in chain is made $\frac{1}{2}$ of the time

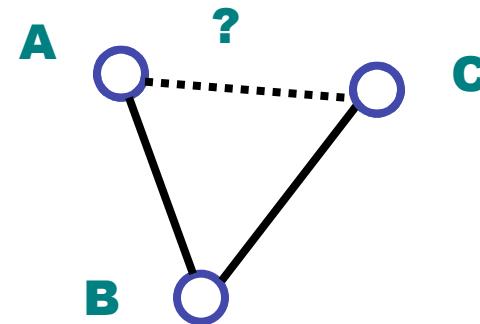
Small worlds & networking

What does it mean to be 1, 2, 3 hops apart on Facebook, Twitter, LinkedIn, Google Plus?

Transitivity, triadic closure, clustering

❑ Transitivity:

- ❑ if A is connected to B and B is connected to C
what is the probability that A is connected to C?
- ❑ my friends' friends are likely to be my friends



Clustering

■ Global clustering coefficient

$\frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples of vertices}}$

$$C = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples}}$$

Local clustering coefficient (Watts&Strogatz 1998)

■ For a vertex i

- The fraction pairs of neighbors of the node that are themselves connected
- Let n_i be the number of neighbors of vertex i

$$C_i = \frac{\text{# of connections between } i\text{'s neighbors}}{\max \text{ # of possible connections between } i\text{'s neighbors}}$$

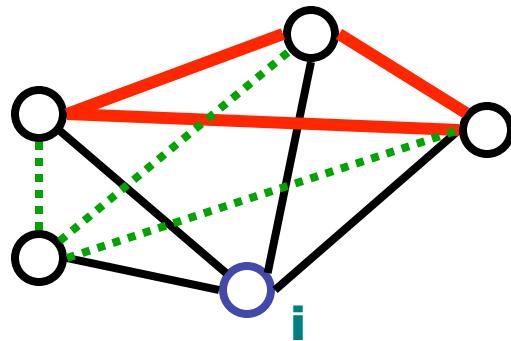
$$C_{i \text{ directed}} = \frac{\text{# directed connections between } i\text{'s neighbors}}{n_i * (n_i - 1)}$$

$$C_{i \text{ undirected}} = \frac{\text{# undirected connections between } i\text{'s neighbors}}{n_i * (n_i - 1) / 2}$$

Local clustering coefficient (Watts&Strogatz 1998)

- ❑ Average over all n vertices

$$C = \frac{1}{n} \sum_i C_i$$



— link present
- - - link absent

$$n_i = 4$$

max number of connections:

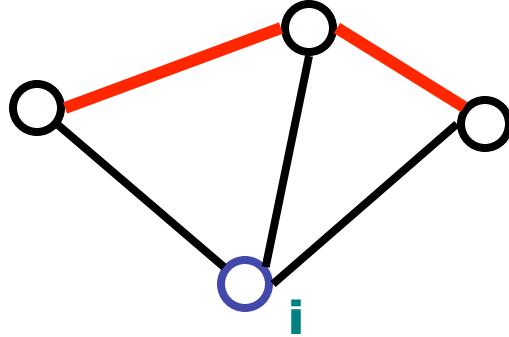
$$4 * 3 / 2 = 6$$

3 connections present

$$C_i = 3 / 6 = 0.5$$

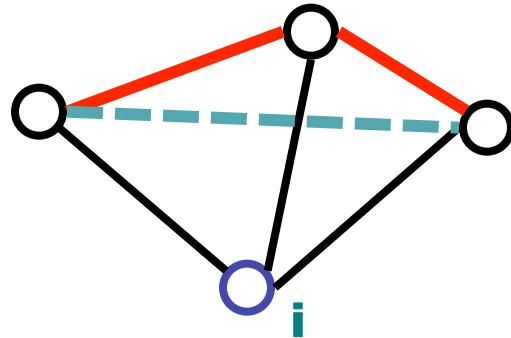
Quiz Q:

- ❑ The clustering coefficient for vertex A is:



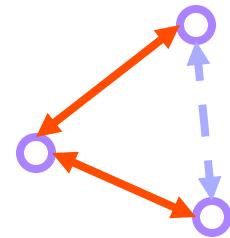
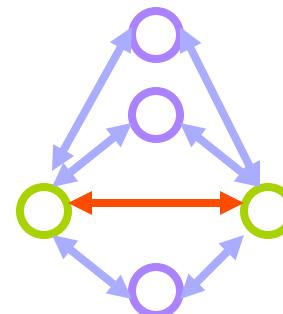
Explanation

- ❑ $n_i = 3$
- ❑ there are 2 connections present out of max of 3 possible
- ❑ $C_i = 2/3$



Are strong ties “local”?

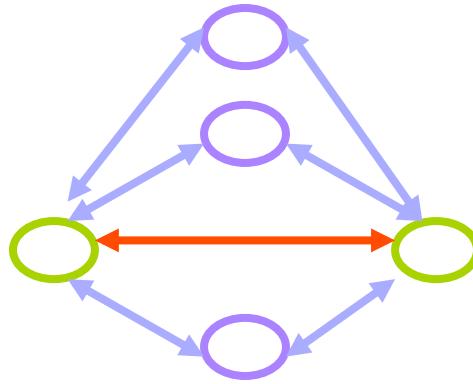
- ❑ A strong tie
 - ❑ frequent contact
 - ❑ affinity
 - ❑ many mutual contacts



“forbidden triad”:
strong ties are likely to “close”

edge embeddeness

- embeddeness: number of common neighbors the two endpoints have



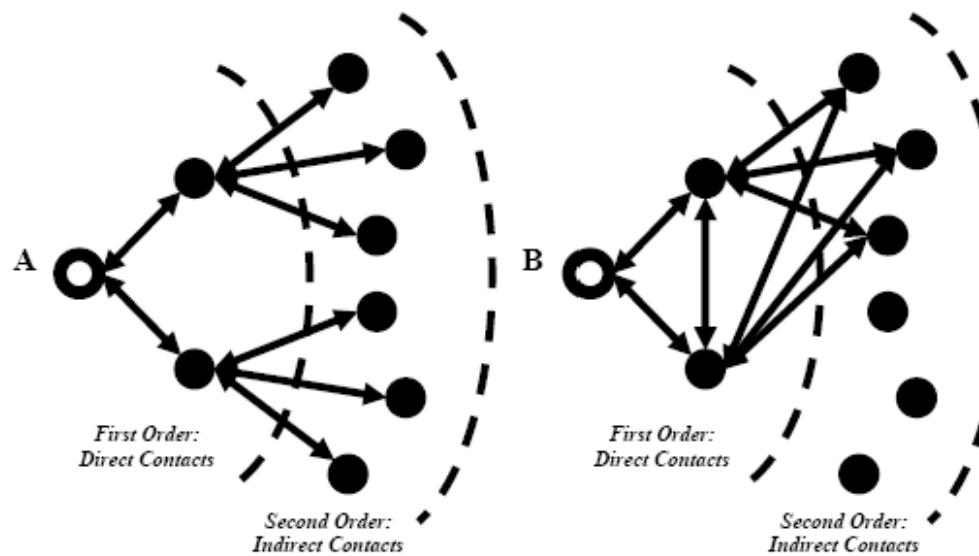
- neighborhood overlap:

$$\frac{\text{number of nodes who are neighbors of } both A \text{ and } B}{\text{number of nodes who are neighbors of } at \text{ least one of } A \text{ or } B}$$

school kids and 1st through 8th choices of friends

■ snowball sampling:

- will you reach more different kids by asking each kid to name their 2 best friends, or their 7th & 8th closest friend?



Source: M. van Alstyne, S. Aral. Networks, Information & Social Capital, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=958158

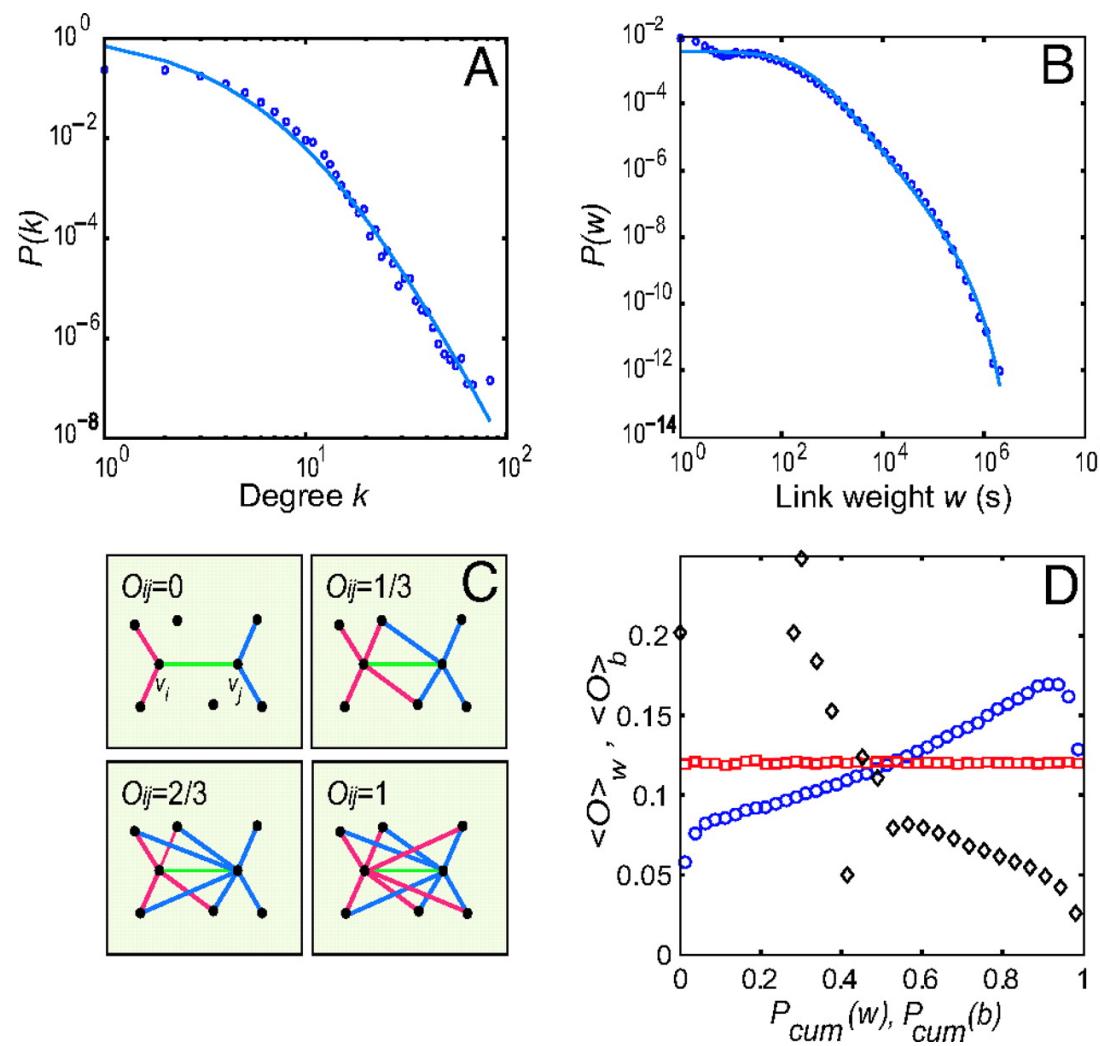
is it good to be embedded?

- ❑ What are the advantages of occupying an embedded position in the network?
- ❑ What are the disadvantages of being embedded?
- ❑ Advantages of being a broker (spanning structural holes)?
- ❑ Disadvantages of being a broker?

the strength of intermediate ties

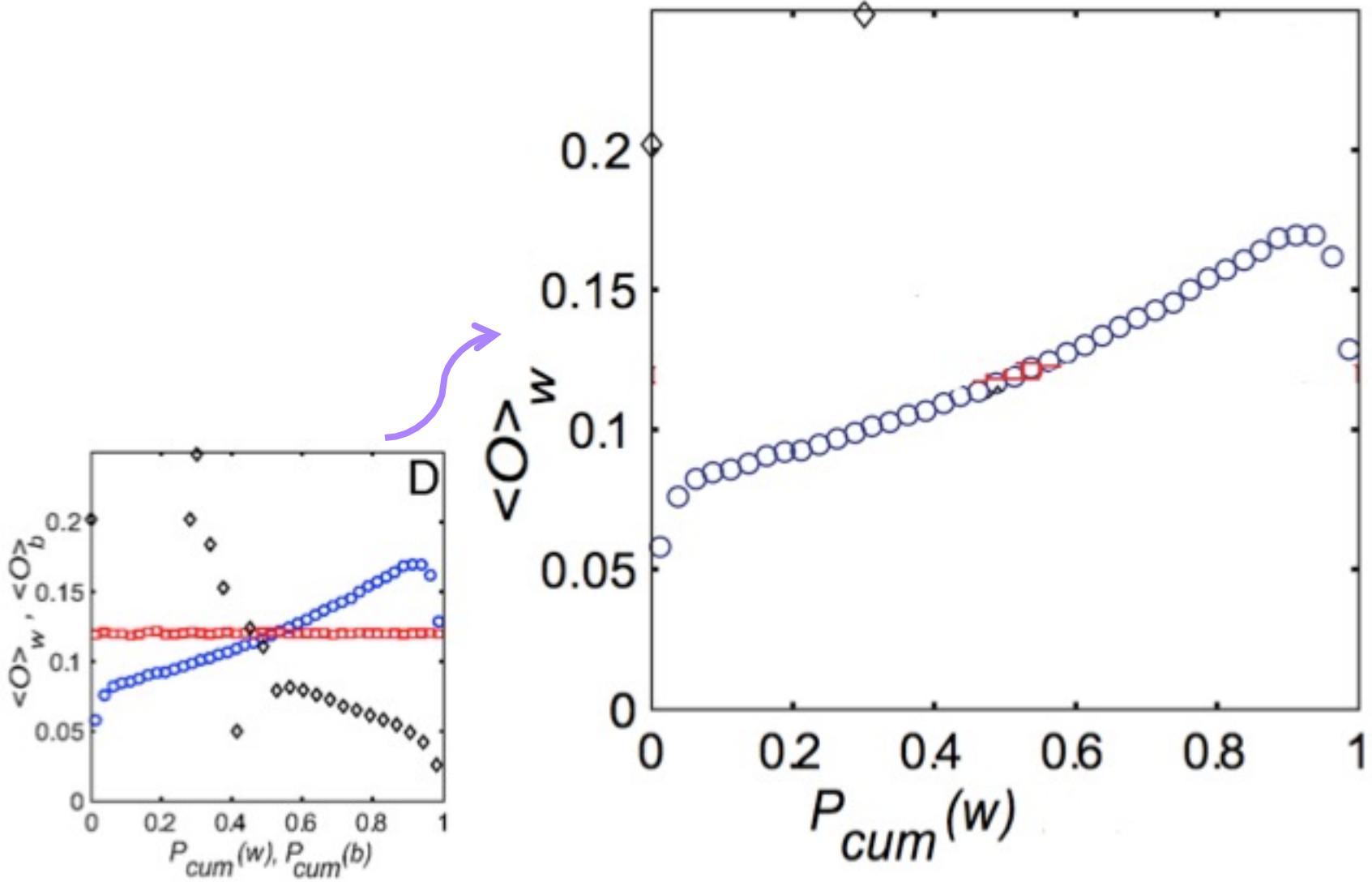
- ❑ study of a large call graph
- ❑ strong ties
 - ❑ frequent communication, but ties are redundant due to high clustering
- ❑ weak ties
 - ❑ reach far across network, but communication is infrequent...
- ❑ Onnela J. et.al. PNAS 2007;104:7332-7336
 - ❑ use nation-wide cellphone call records and simulate diffusion using actual call timing
 - ❑ in simulation, individuals are most likely to obtain novel information through ties of intermediate strength

Characterizing the large-scale structure and the tie strengths of the mobile call graph

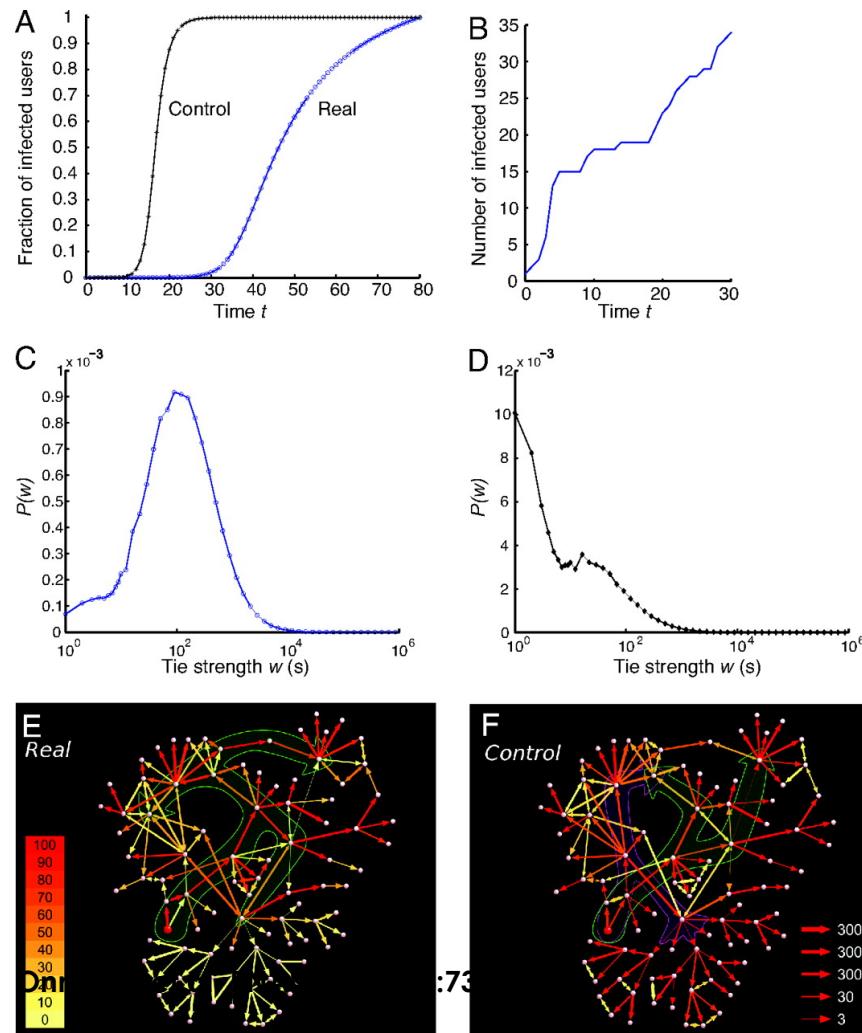


Onnela J et al. PNAS 2007;104:7332-7336

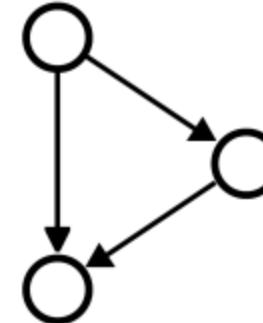
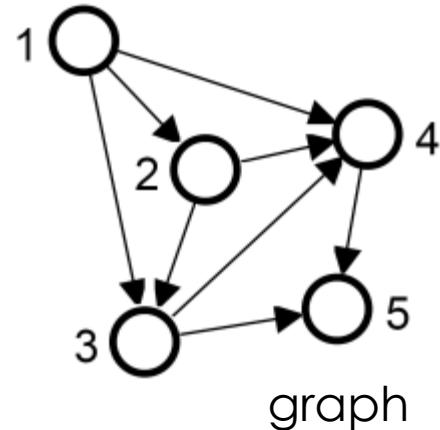
Edge neighborhood overlap as a function of tie strength



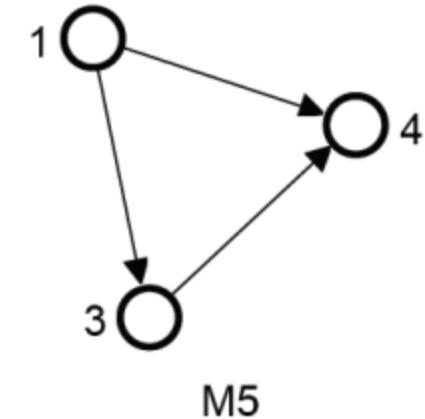
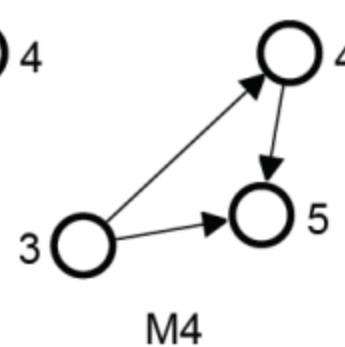
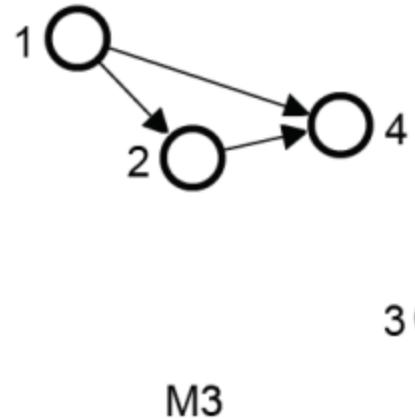
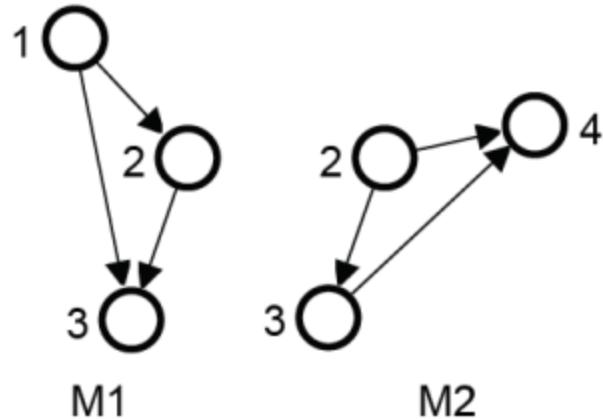
The dynamics of spreading on the weighted mobile call graph, assuming that the probability for a node v_i to pass on the information to its neighbor v_j in one time step is given by $P_{ij} = xw_{ij}$, with $x = 2.59 \times 10^{-4}$



Resolving local structure: network motifs

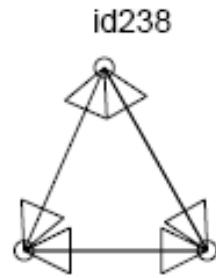
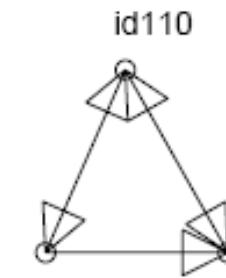
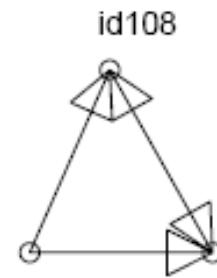
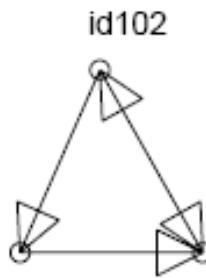
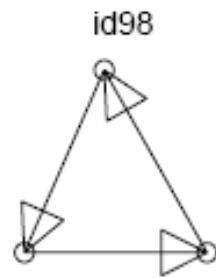
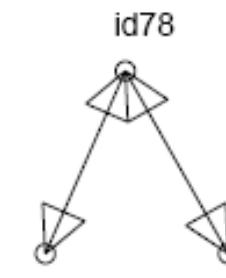
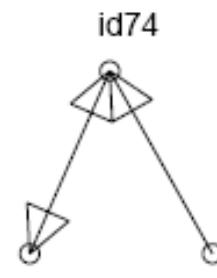
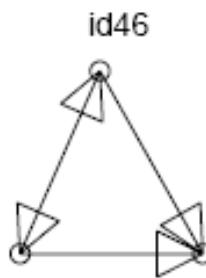
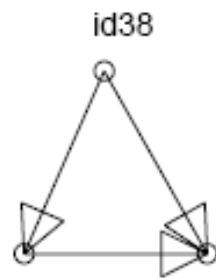
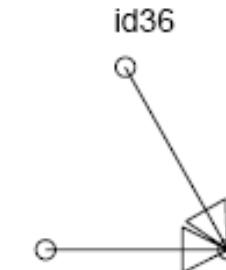
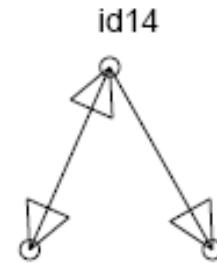
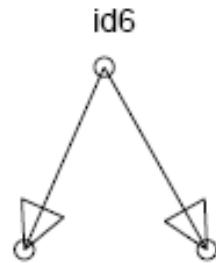


motif to be found



motif matches in the target graph

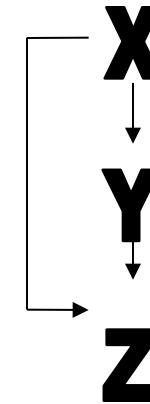
All 3 node motifs



Examples of network motifs (3 nodes)

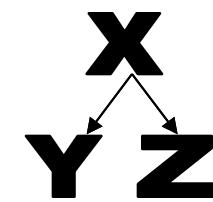
❑ Feed forward loop

- ❑ Found in neural networks
- ❑ Seems to be used to neutralize “biological noise”



❑ Single-Input Module

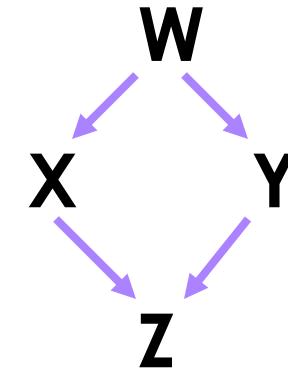
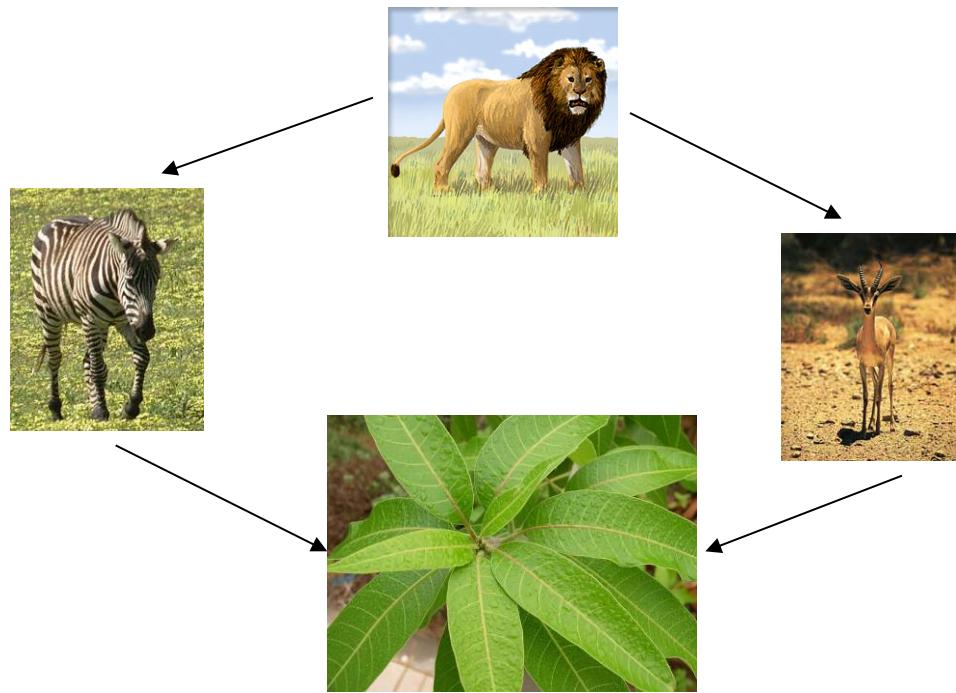
- ❑ e.g. gene control networks



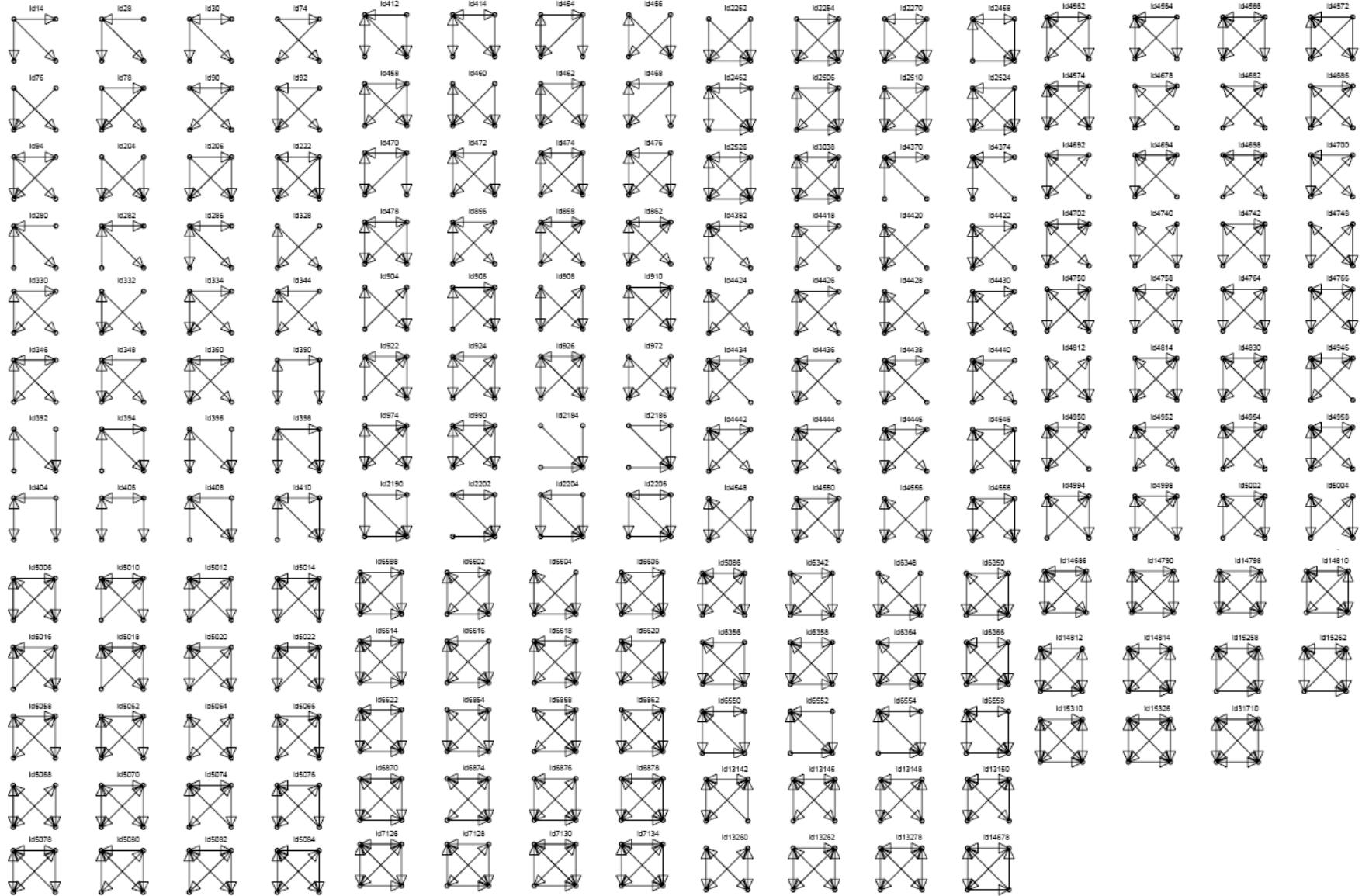
Examples of network motifs (4 nodes)

□ Parallel paths

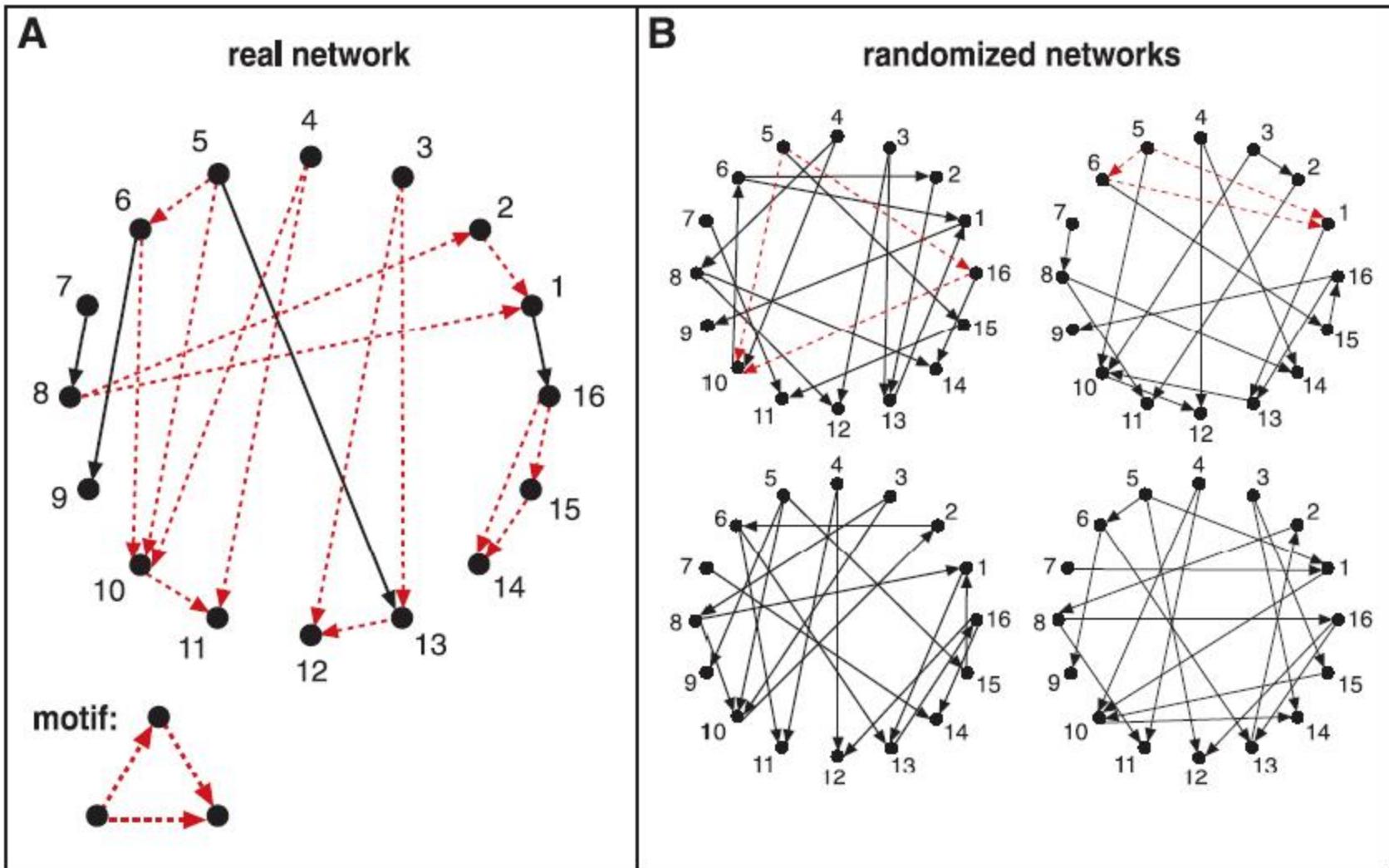
- Found in neural networks
- Food webs



4 node subgraphs (computational expense increases with the size of the graph!)



Compare to “equivalent” random graph



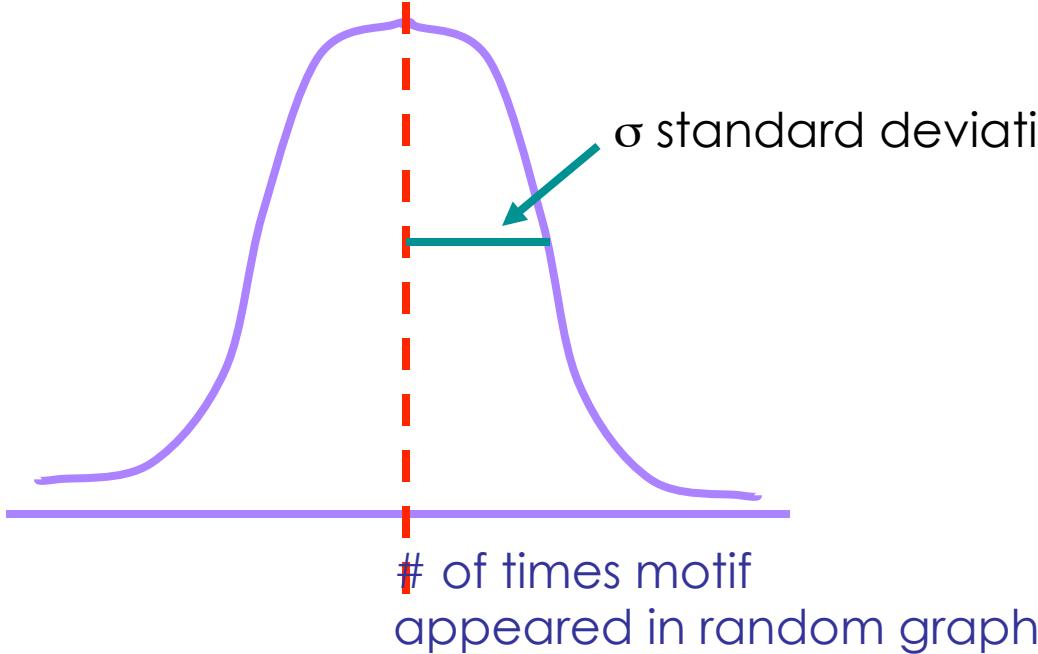
Milo et al., Network motifs: Simple building blocks of complex networks, Science 298:824-827, 2002

Network motif detection

- Some motifs will occur more often in real world networks than random networks
- Technique:
 - construct many random graphs with the same number of nodes and edges (same node degree distribution?)
 - count the number of motifs in those graphs
 - calculate the Z score: the probability that the given number of motifs in the real world network could have occurred by chance
- Software available:
 - <http://www.weizmann.ac.il/mcb/UriAlon/> (the original)
 - <http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html>
(faster and more user friendly)

What the Z score means

μ = mean number of times the motif appeared in the random graph



the probability observing a Z score of 2 is 0.02275

In the context of motifs:
 $Z > 0$, motif occurs more often than for random graphs
 $Z < 0$, motif occurs less often than in random graphs

$$z_x = \frac{x - \mu_x}{\sigma_x}$$

$|Z| > 1.65$, only a 5% chance of random occurrence

software: FANMOD (also igraph)

- http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html

FANMOD a tool for fast network motif detection

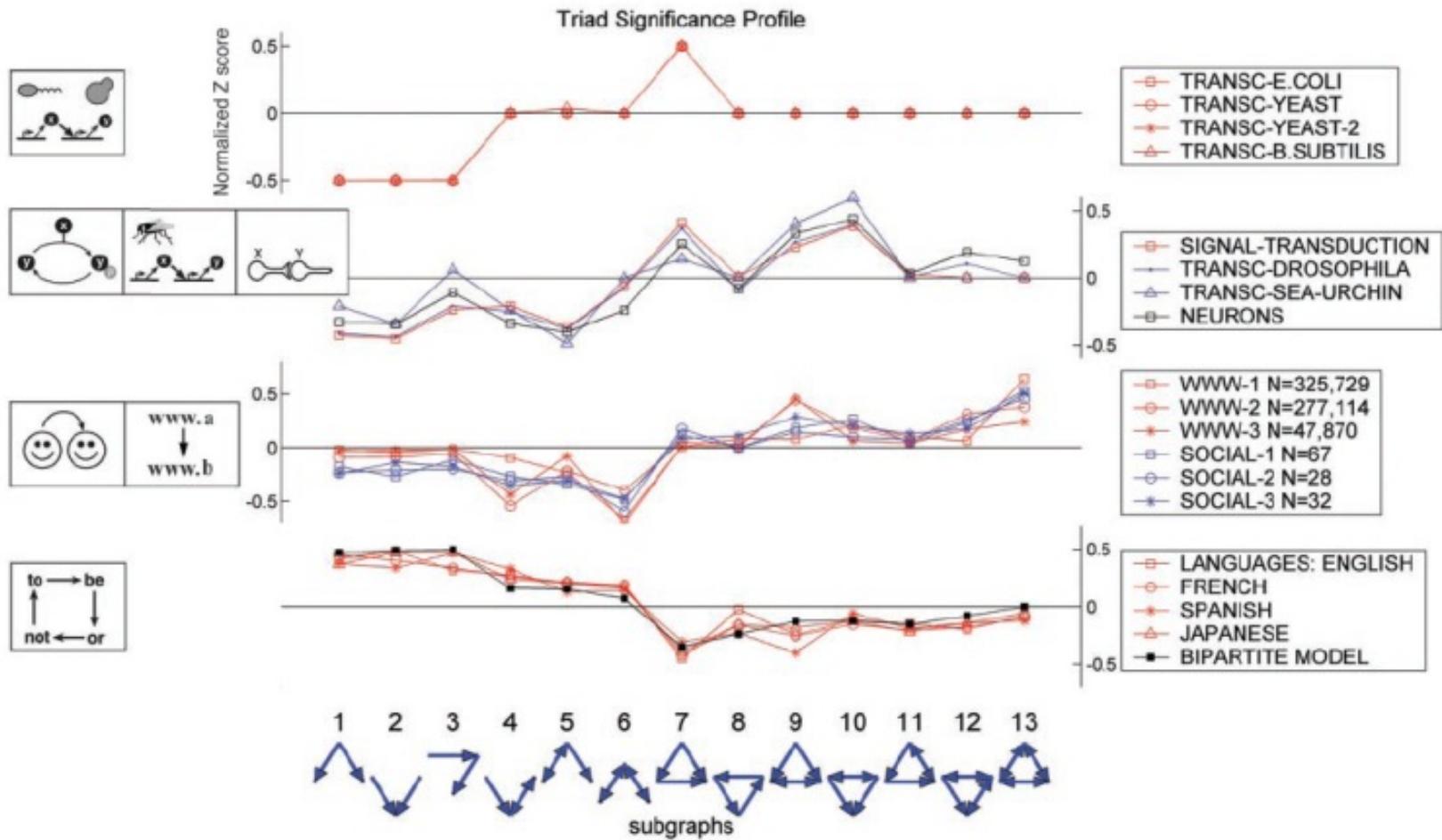
The screenshot displays the FANMOD software interface across three windows:

- MAIN WINDOW (FANMOD):** Shows the "SETUP" tab with parameters: Subgraph size: 4, Samples to estimate: 100000, Input graph: "coll1_1Inter_st.txt", Output: "coll1_1Inter_st.out". It also shows "ALGORITHM" progress: Total Progress: 1000000000 of 1000000000, Time elapsed: 00:00:04.
- RESULTS WINDOW:** Shows a table of Size-4 Network Motifs from "coll1_1Inter_st.txt". The table includes columns: ID, Atj, Frequency [Original], Mean-Freq [Random], Standard-Dev [Random], t-Score, p-Value. The data is as follows:

ID	Atj	Frequency [Original]	Mean-Freq [Random]	Standard-Dev [Random]	t-Score	p-Value
206	(1,2,3,4)	0.007152%	2.352e-005%	5.2369e-007	135	0
206	(1,2,3,4)	0.004768%	2.3727e-005%	5.3029e-007	89.87	0
2188	(1,2,3,4)	0.002994%	1.1625e-005%	2.7394e-007	69.739	0
205	(1,2,3,4)	0.003576%	3.5635e-005%	6.4999e-007	54.981	0

- HTML EXPORT DIALOG:** Shows "HTML Export options" for "coll1_1Inter_st.out". It includes "HTML-Output-File(s)": "coll1_1Inter_st.html", "Motif per file": 20, and "Create pictures" checked. It also has "Filters" for "t-Score" (greater than 2, 0.05), "p-Value" (less than 0.05), "Frequency" (greater than 0.01%), and "Motif found in random networks" (checkboxes for "put at top of list" and "put at bottom of list").

Superfamilies of networks

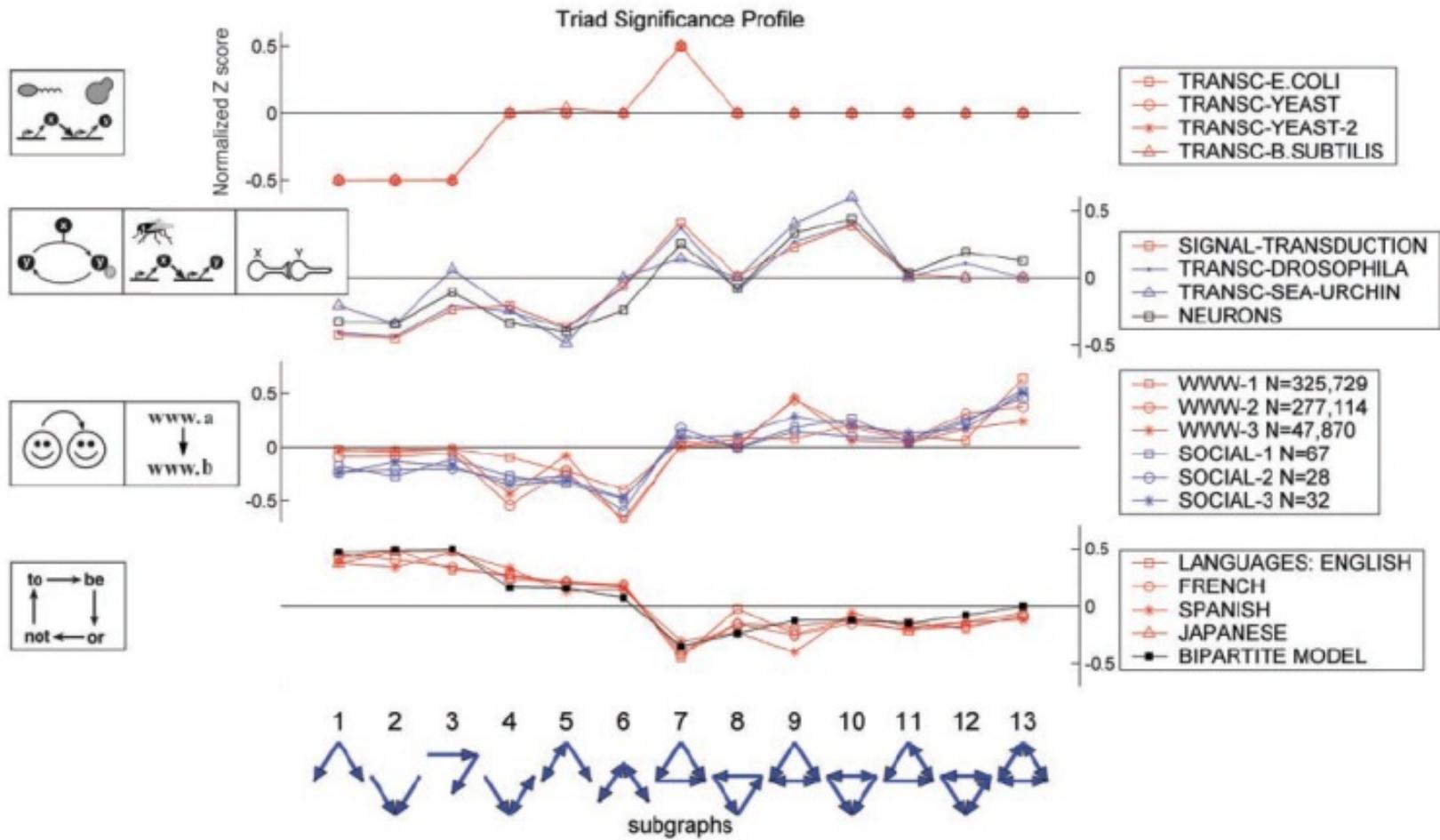


source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004

Quiz Q:

- ❑ Based on their triad census profiles, which two kinds of networks exhibit similar structure?

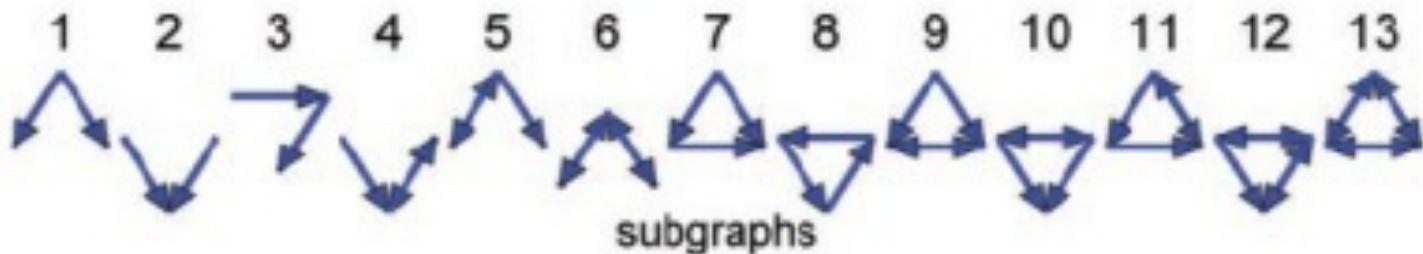
Superfamilies of networks



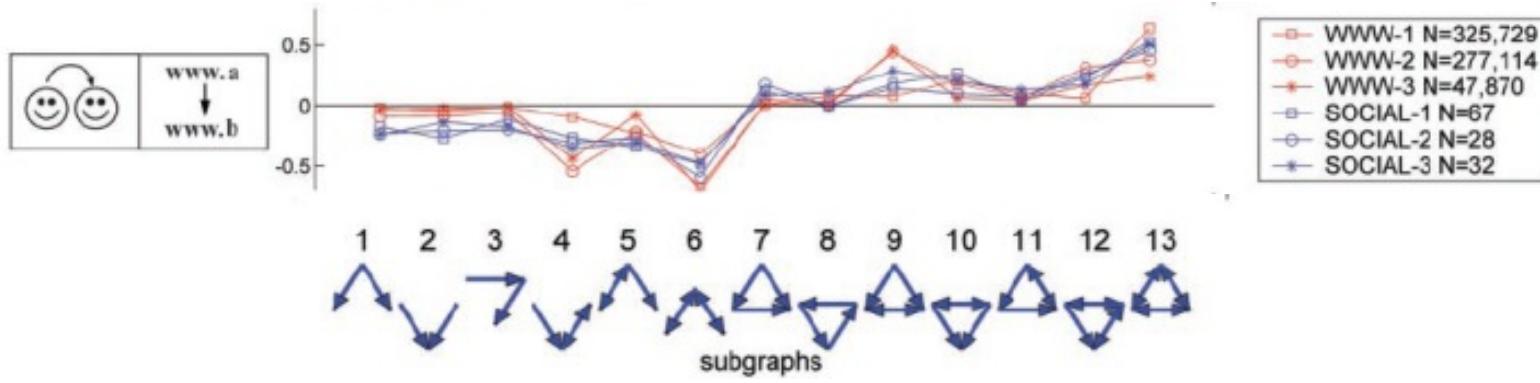
source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004

Quiz Q:

- Which of the following triads is underrepresented in social networks?

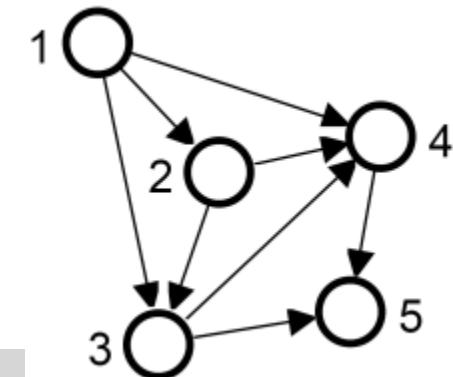


Superfamilies of networks



Motifs: recap

- Given a particular structure, search for it in the network, e.g. complete triads
- advantage: motifs can correspond to particular functions, e.g. in biological networks
- disadvantage: don't know if motif is part of a larger cohesive community



beyond social networks

Small world phenomenon:

high clustering

$$C_{\text{network}} \gg C_{\text{random graph}}$$

low average shortest path

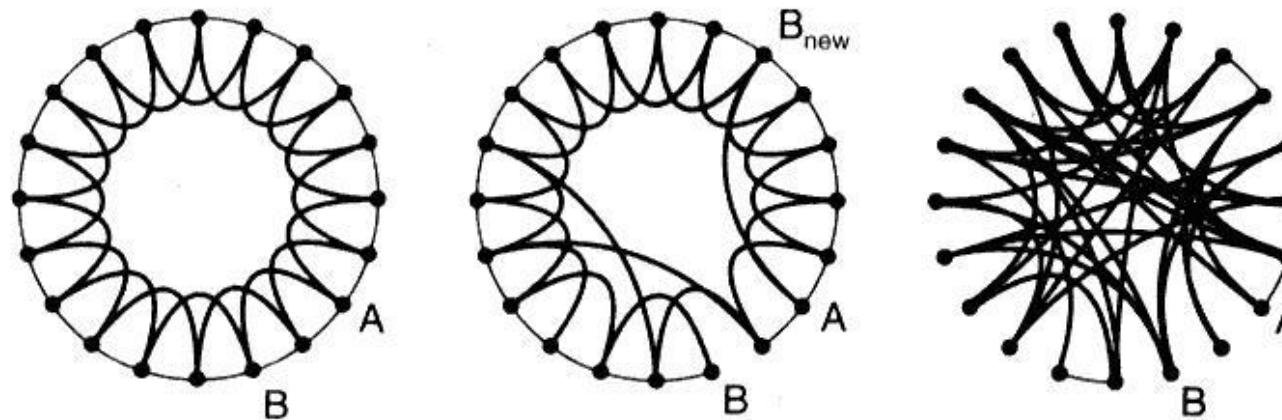
$$l_{\text{network}} \approx \ln(N)$$

- neural network of *C. elegans*,
- semantic networks of languages,
- actor collaboration graph
- food webs

Small world phenomenon: Watts/Strogatz model

Reconciling two observations:

- **High clustering:** my friends' friends tend to be my friends
- **Short average paths**



Watts-Strogatz model: Generating small world graphs



Select a fraction p of edges
Reposition one of their endpoints

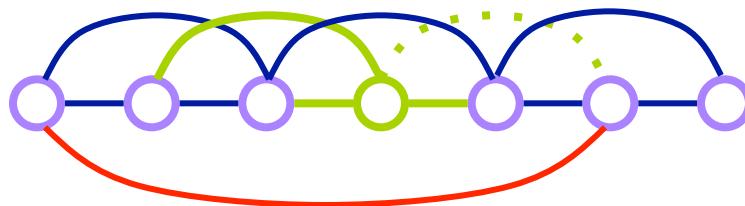


Add a fraction p of additional edges leaving underlying lattice intact

- As in many network generating algorithms
 - Disallow self-edges
 - Disallow multiple edges

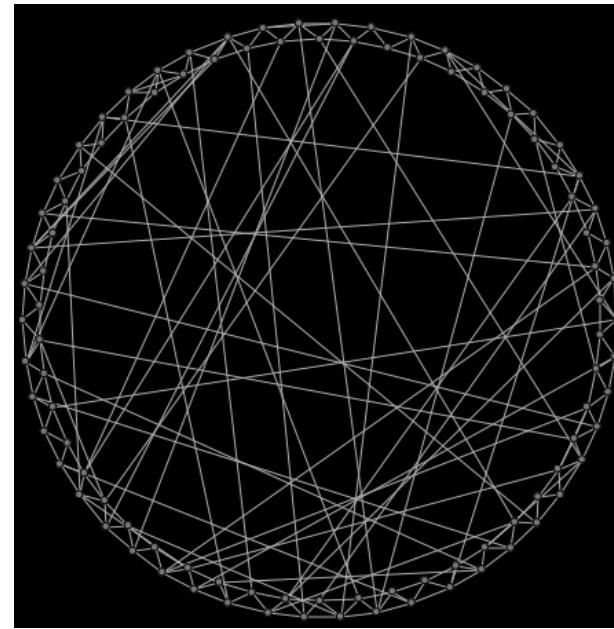
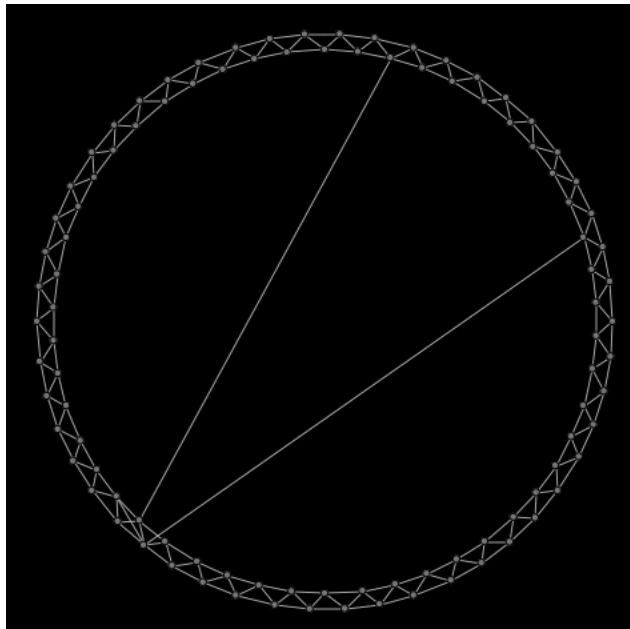
Watts-Strogatz model: Generating small world graphs

- Each node has $K \geq 4$ nearest neighbors (local)
- tunable: vary the probability p of rewiring any given edge
- small p : regular lattice
- large p : classical random graph



Quiz question:

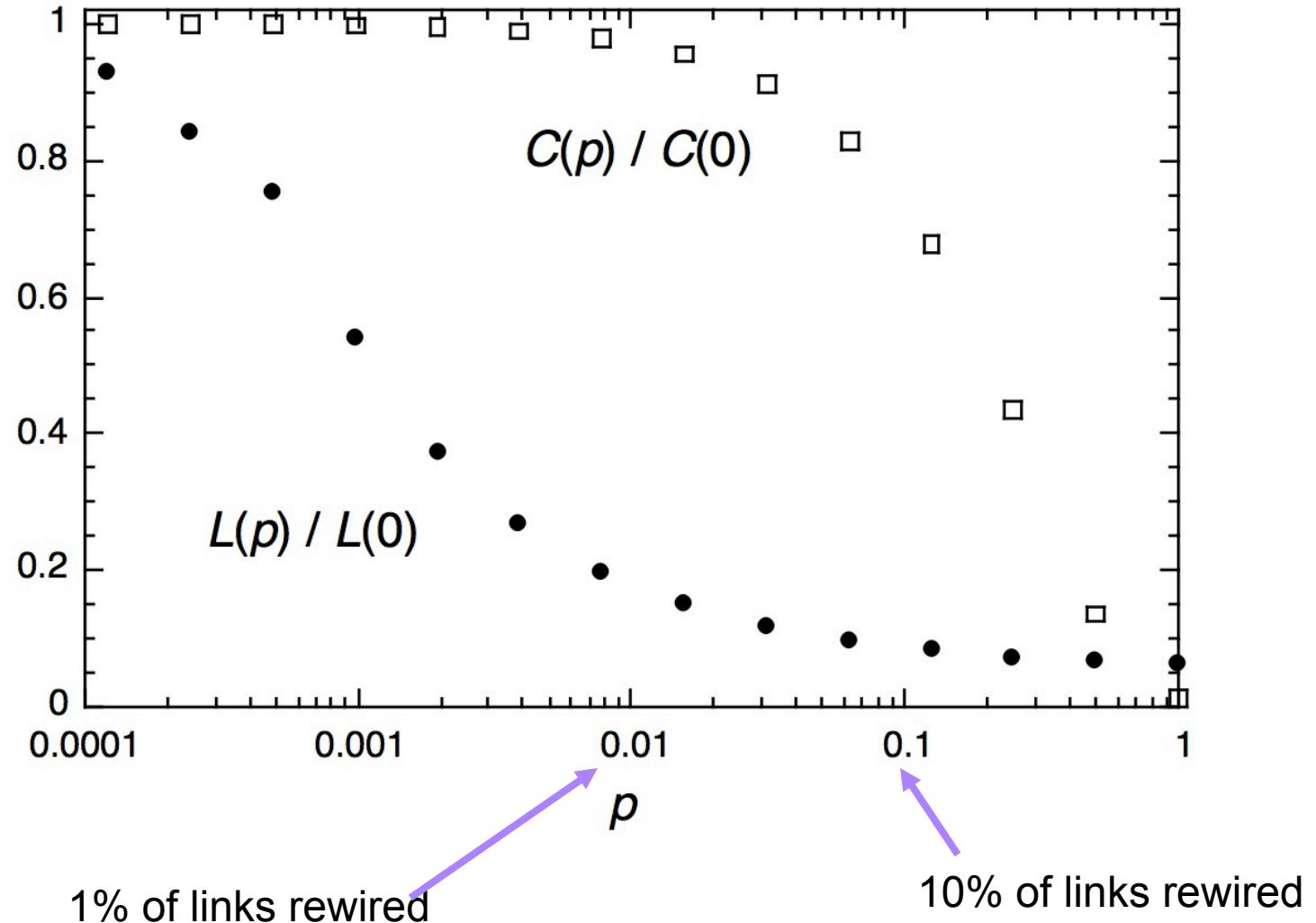
- Which of the following is a result of a higher rewiring probability?



What happens in between?

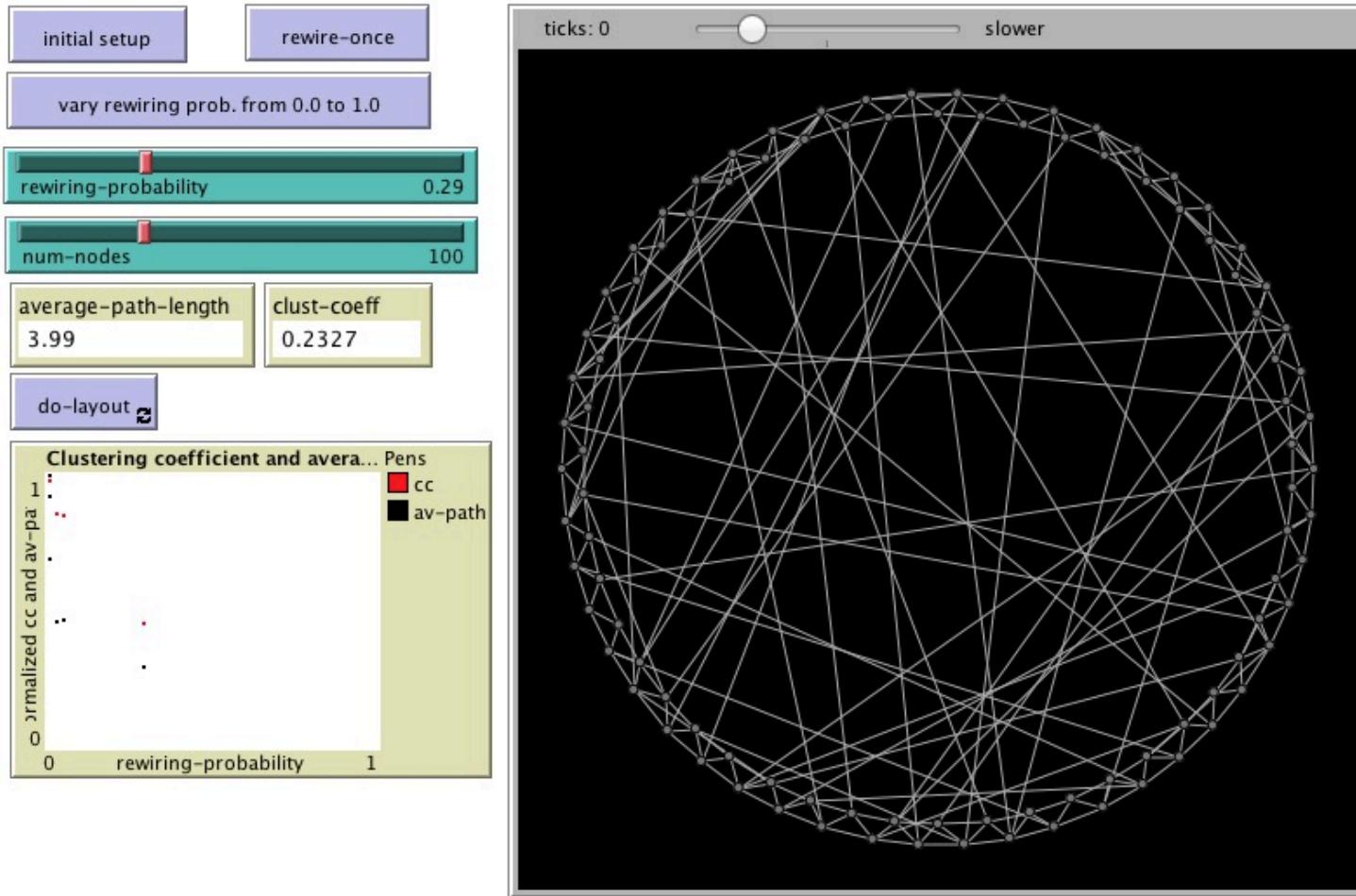
- ❑ Small shortest path means low clustering?
- ❑ Large shortest path means high clustering?
- ❑ Through numerical simulation
 - ❑ As we increase p from 0 to 1
 - ❑ Fast decrease of mean distance
 - ❑ Slow decrease in clustering

Clust coeff. and ASP as rewiring increases



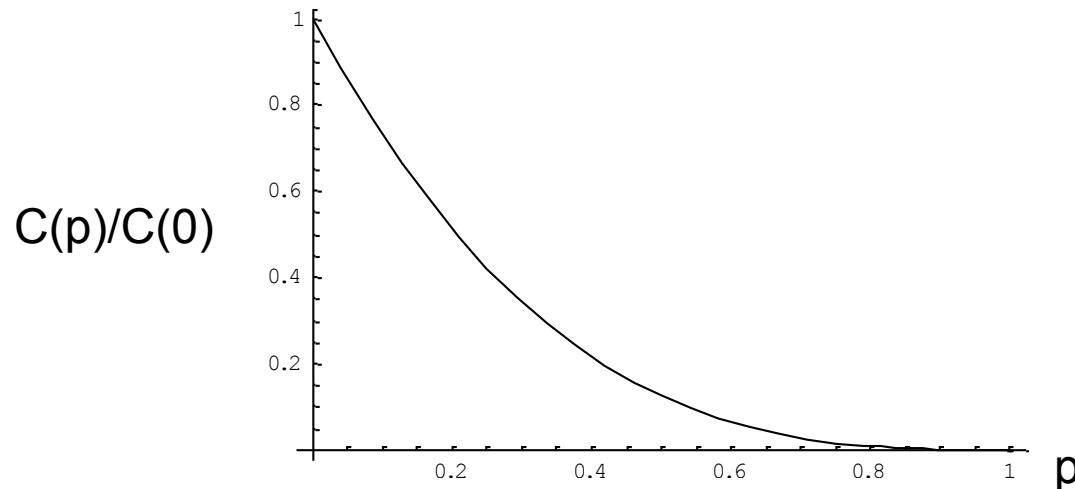
Trying this with NetLogo

<http://www.ladamic.com/netlearn/NetLogo4/SmallWorldWS.html>



WS model clustering coefficient

- The probability that a connected triple stays connected after rewiring
 - probability that none of the 3 edges were rewired $(1-p)^3$
 - probability that edges were rewired back to each other very small, can ignore
- Clustering coefficient = $C(p) = C(p=0) * (1-p)^3$



Source: Watts, D.J., Strogatz, S.H.(1998) Collective dynamics of 'small-world' networks. Nature 393:440-442.

Comparison with “random graph” used to determine whether real-world network is “small world”

Network	size	av. shortest path	Shortest path in fitted random graph	Clustering (averaged over vertices)	Clustering in random graph
Film actors	225,226	3.65	2.99	0.79	0.00027
MEDLINE co-authorship	1,520,251	4.6	4.91	0.56	1.8×10^{-4}
E.Coli substrate graph	282	2.9	3.04	0.32	0.026
C.Elegans	282	2.65	2.25	0.28	0.05

Quiz Q

- Which of the following is a description matching a small-world network?

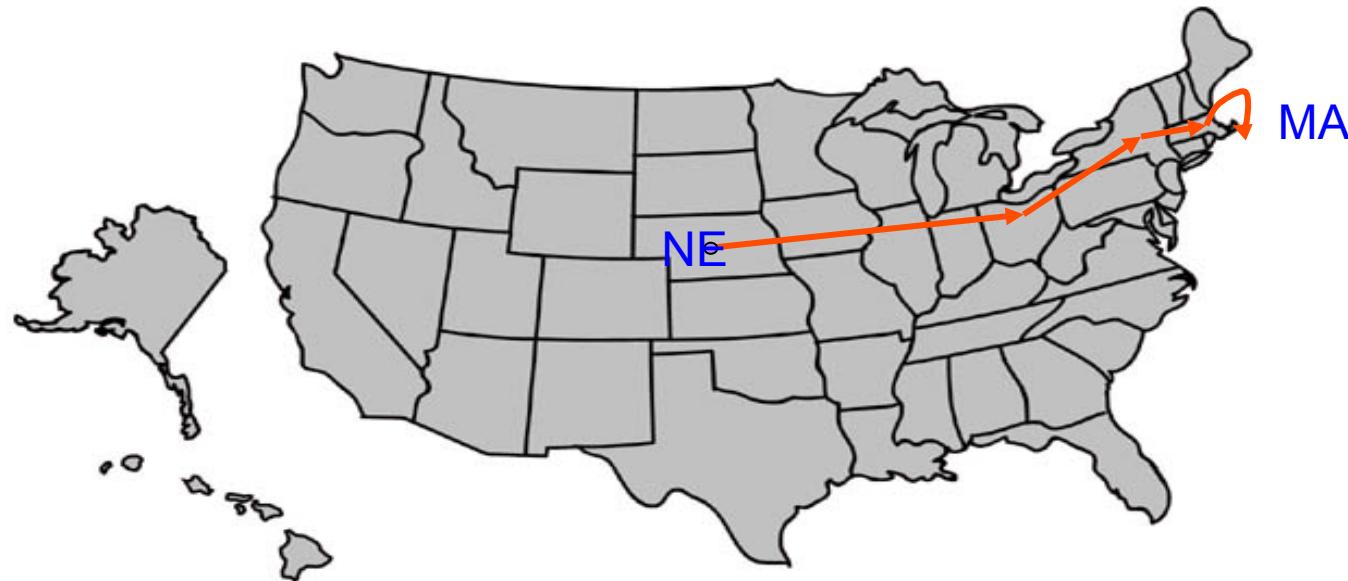
WS Model: What's missing?

- Long range links not as likely as short range ones
- Hierarchical structure / groups
- Hubs

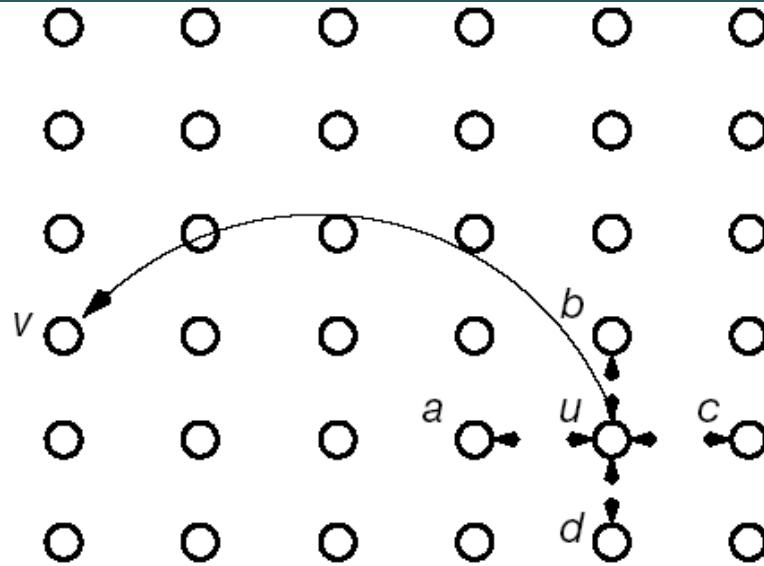
Ties and geography

“The geographic movement of the [message] from Nebraska to Massachusetts is striking. There is a progressive closing in on the target area as each new person is added to the chain”

S.Milgram ‘The small world problem’ , Psychology Today 1,61,1967



Kleinberg's geographical small world model



nodes are placed on a lattice and connect to nearest neighbors

exponent that will determine navigability

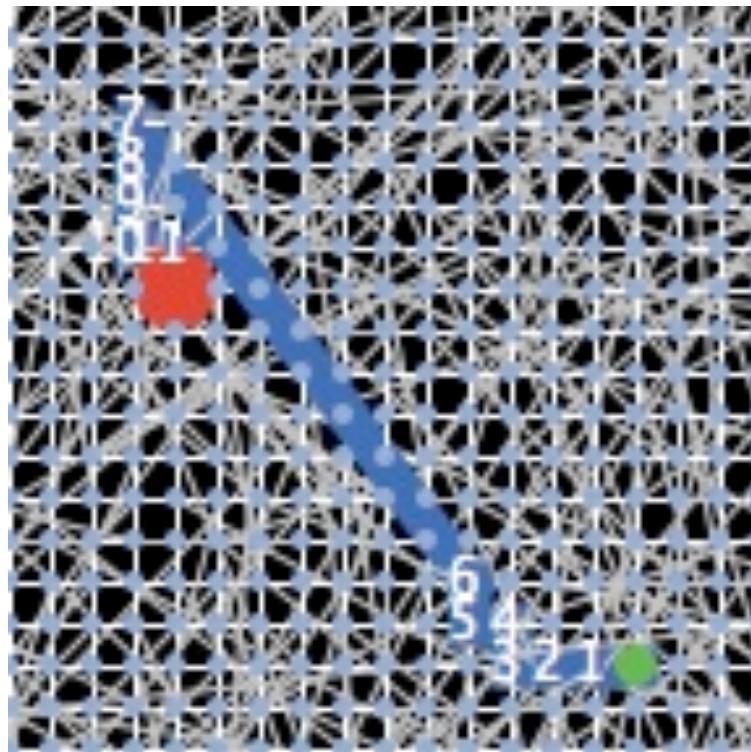
additional links placed with

$$p(\text{link between } u \text{ and } v) = (\text{distance}(u,v))^{-r}$$



NetLogo demo

- how does the probability of long-range links affect search?

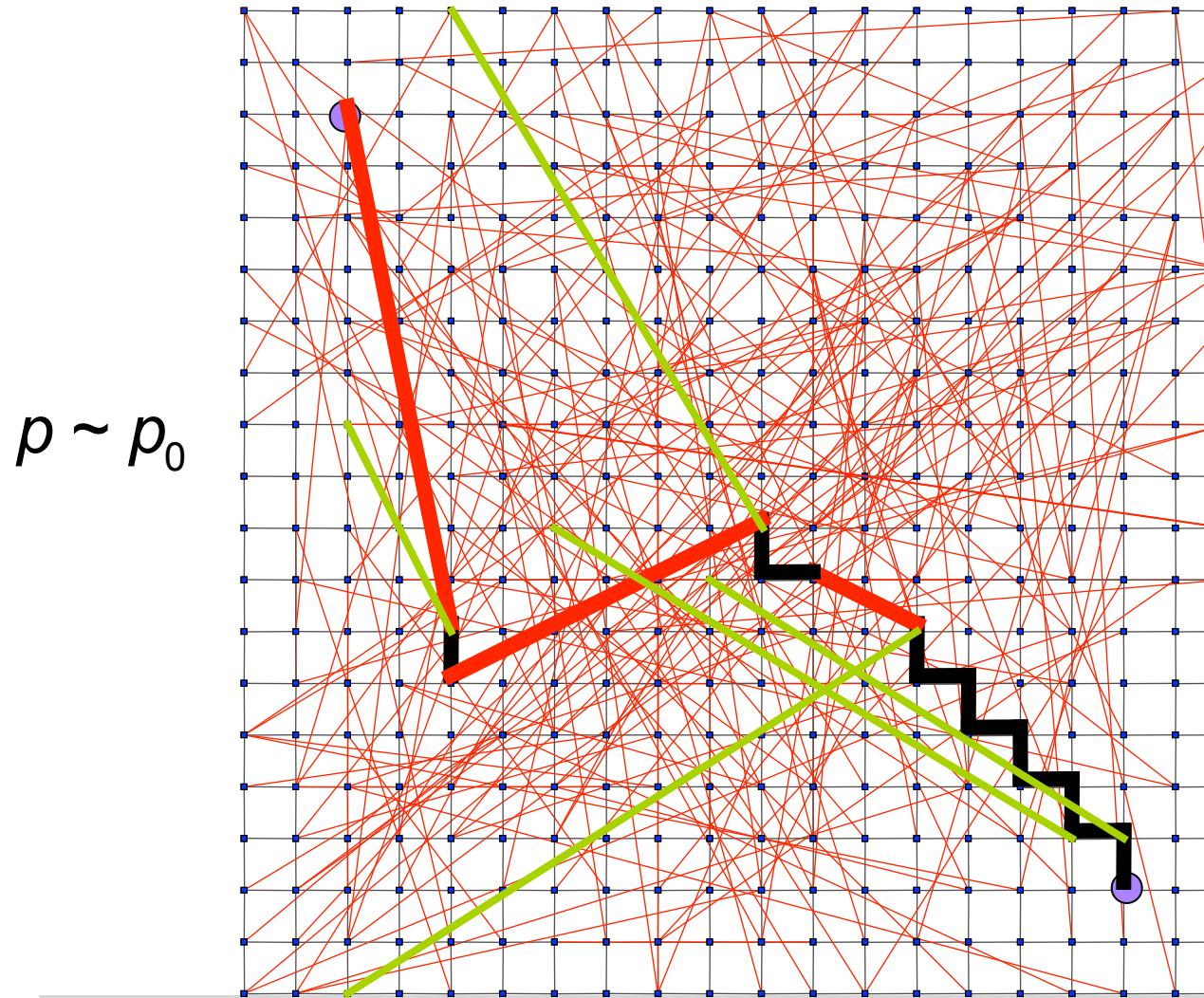


[http://www.ladamic.com/netlearn/
NetLogo4/SmallWorldSearch.html](http://www.ladamic.com/netlearn/NetLogo4/SmallWorldSearch.html)

geographical search when network lacks locality

When $r=0$, links are randomly distributed, $\text{ASP} \sim \log(n)$, n size of grid

When $r=0$, any decentralized algorithm is at least $a_0 n^{2/3}$

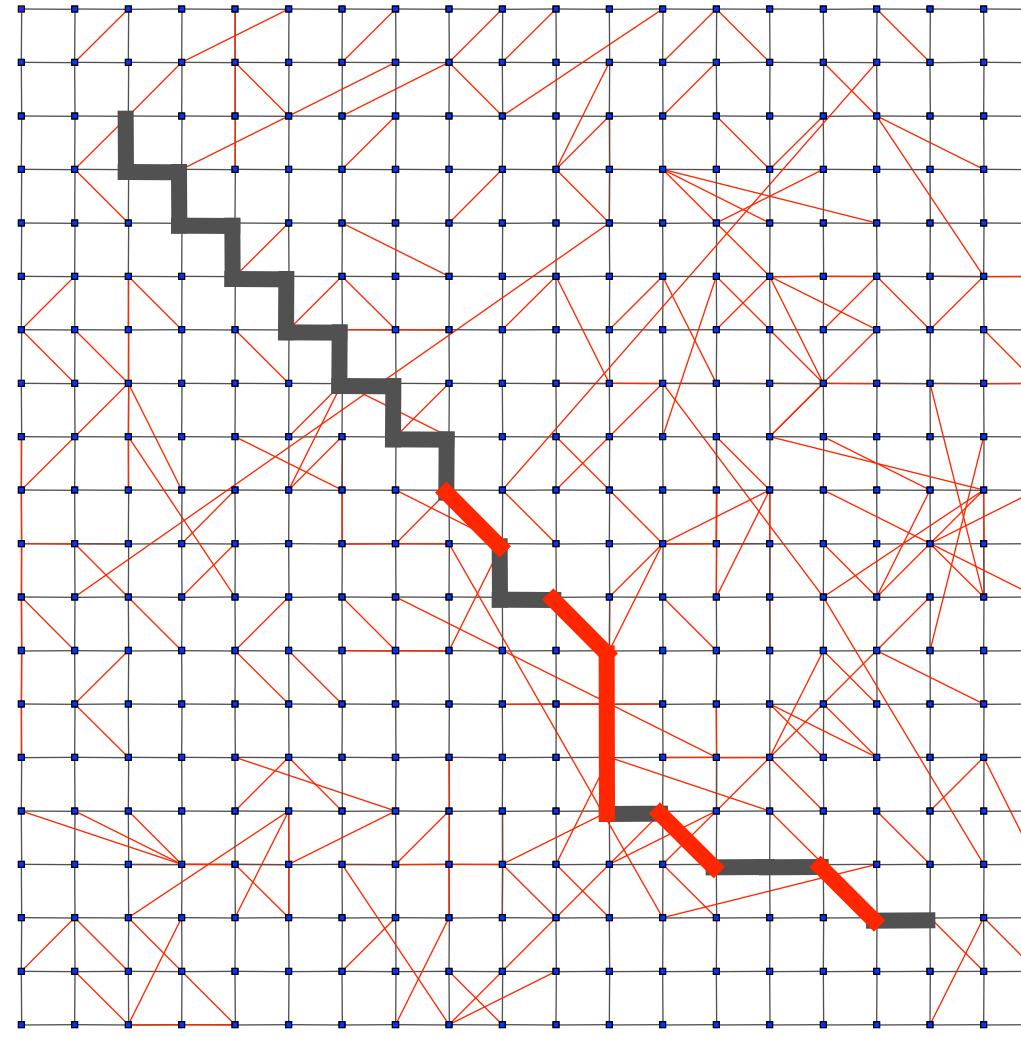


When $r < 2$,
expected
time at
least $\alpha_r n^{(2-r)/3}$

Overly localized links on a lattice

When $r > 2$ expected search time $\sim N^{(r-2)/(r-1)}$

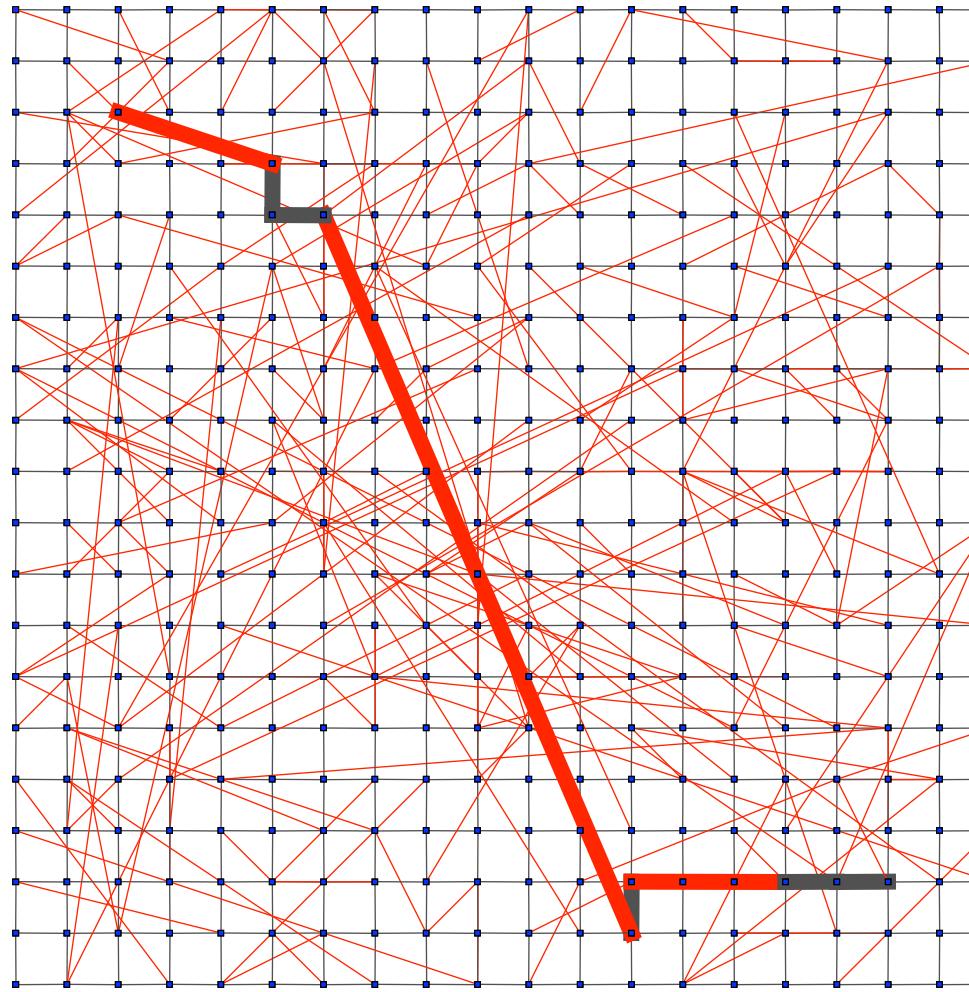
$$p \sim \frac{1}{d^4}$$



Just the right balance

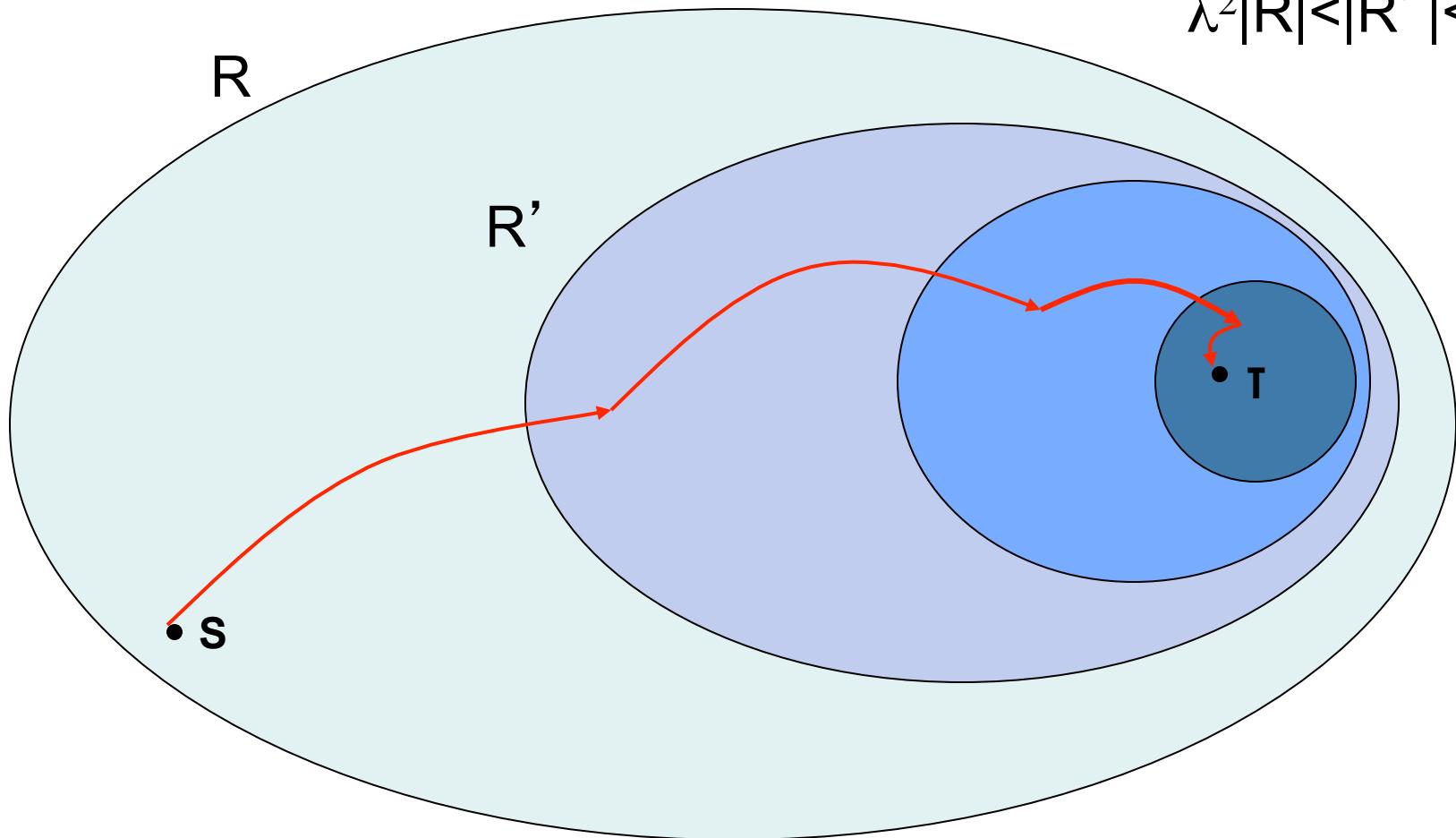
When $r=2$, expected time of a DA is at most $C (\log N)^2$

$$p \sim \frac{1}{d^2}$$



Navigability

$$\lambda^2|R| < |R'| < \lambda|R|$$



$$k = c \log^2 n$$

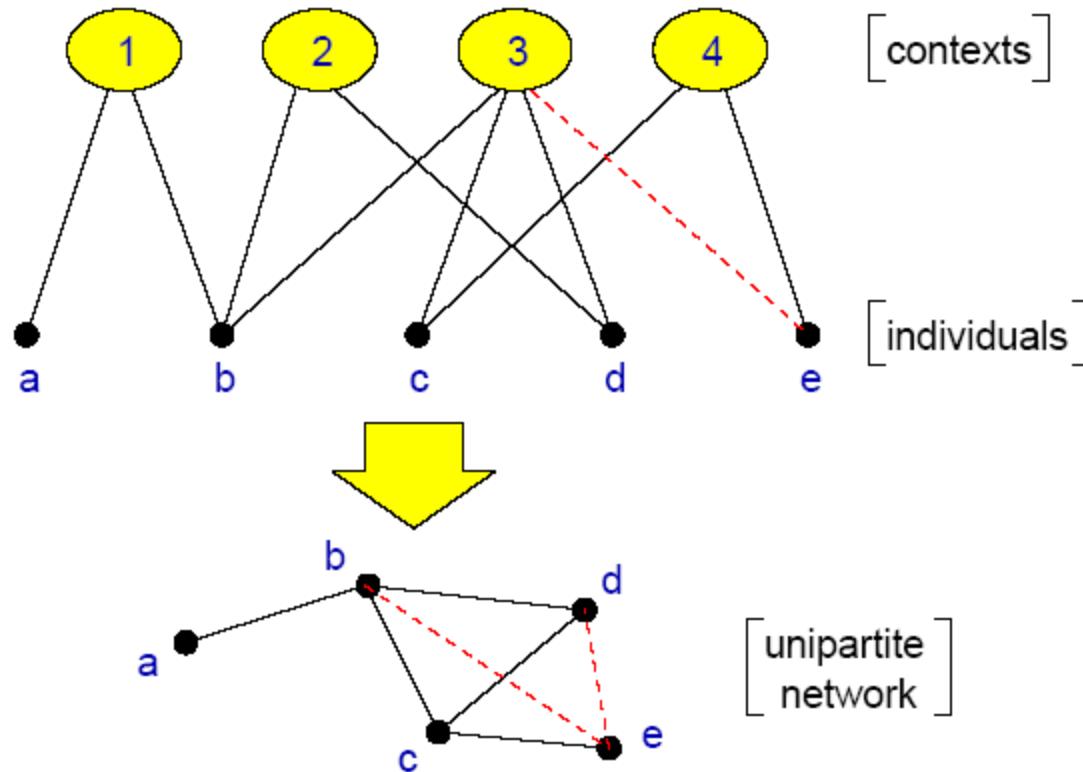
calculate probability that s fails to have a link in R'

Quiz Q:

- ❑ What is true about a network where the probability of a tie falls off as distance^{-2}

Origins of small worlds: group affiliations

Social distance—Bipartite networks:



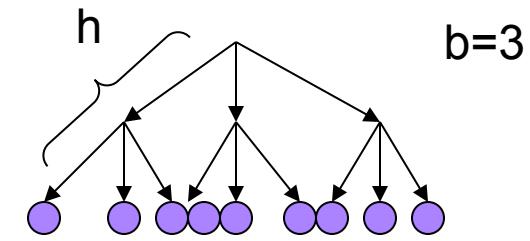
hierarchical small-world models: Kleinberg

Hierarchical network models:

Individuals classified into a hierarchy,
 h_{ij} = height of the least common ancestor.

$$p_{ij} : b^{-\alpha h_{ij}}$$

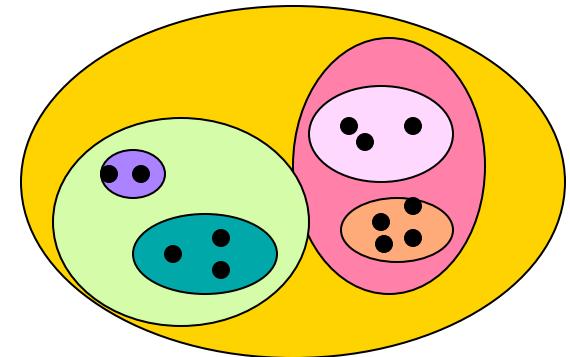
e.g. state-county-city-neighborhood
industry-corporation-division-group



Group structure models:

Individuals belong to nested groups
 q = size of smallest group that v,w belong to

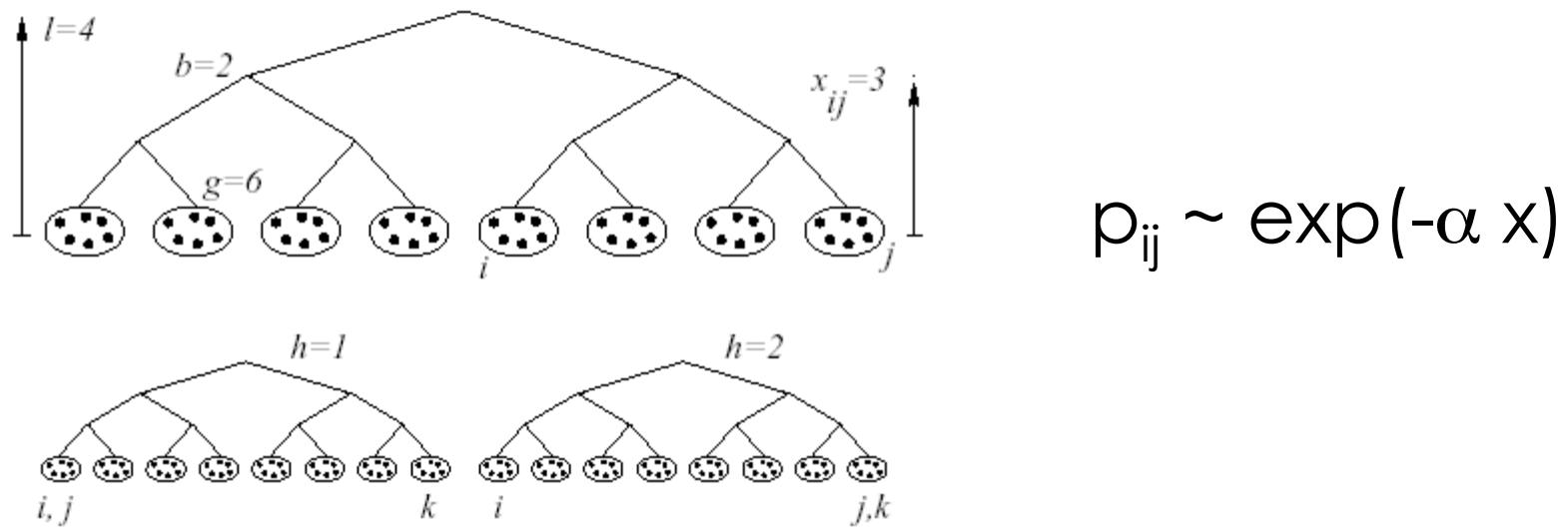
$$f(q) \sim q^{-\alpha}$$



Source: [Kleinberg, ‘Small-World Phenomena and the Dynamics of Information’ NIPS 14, 2001.](#)

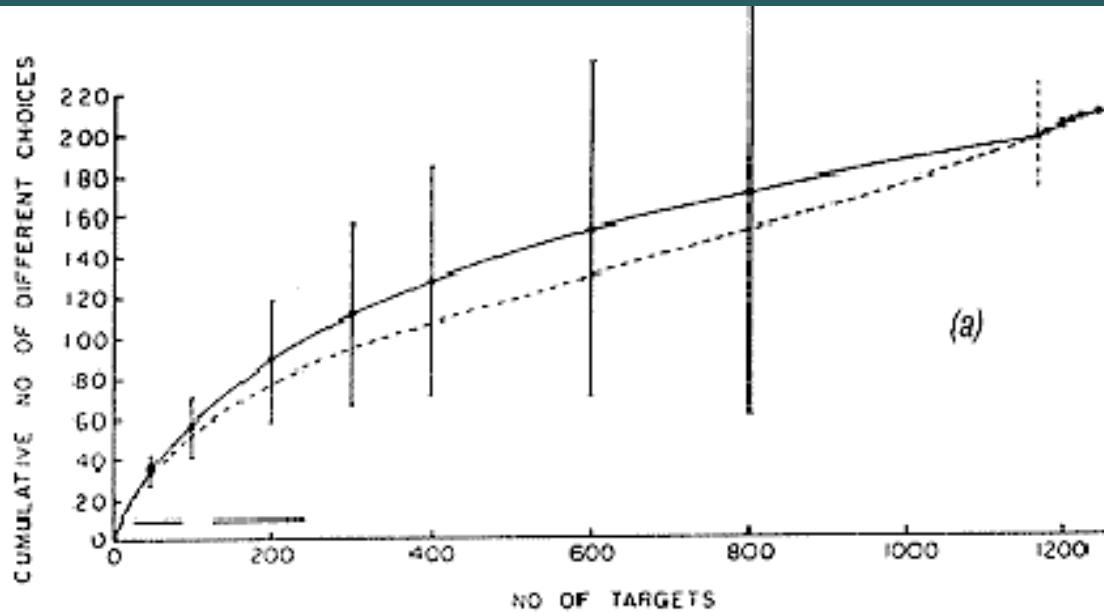
hierarchical small-world models: WDN

Watts, Dodds, Newman (Science, 2001)
individuals belong to hierarchically nested groups



multiple independent hierarchies $h=1, 2, \dots, H$
coexist corresponding to occupation,
geography, hobbies, religion...

Navigability and search strategy: Reverse small world experiment



- Killworth & Bernard (1978):
- Given hypothetical targets (name, occupation, location, hobbies, religion...) participants choose an acquaintance for each target
- based on (most often) occupation, geography
- only 7% because they “know a lot of people”
- Simple greedy algorithm: most similar acquaintance
- two-step strategy rare

Navigability and search strategy: Small world experiment @ Columbia

Successful chains disproportionately used

- weak ties (Granovetter)
- professional ties (34% vs. 13%)
- ties originating at work/college
- target's work (65% vs. 40%)

. . . and disproportionately avoided

- hubs (8% vs. 1%) (+ no evidence of funnels)
- family/friendship ties (60% vs. 83%)

Strategy: Geography -> Work

Generating small-world networks

- Assign properties to nodes (e.g. spatial location, group membership)
- Add or rewire links according to some rule
 - optimize for a particular property (simulated annealing)
 - add links with probability depending on property of existing nodes, edges (preferential attachment, link copying)
 - simulate nodes as agents ‘deciding’ whether to rewire or add links

Origins of small worlds: efficient network example trade-off between wiring and connectivity

Small worlds: How and Why, Nisha Mathias and Venkatesh Gopal

$$E = \lambda L + (1 - \lambda)W$$

$$L = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$$

$$W = \sum_{e_{ij}} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

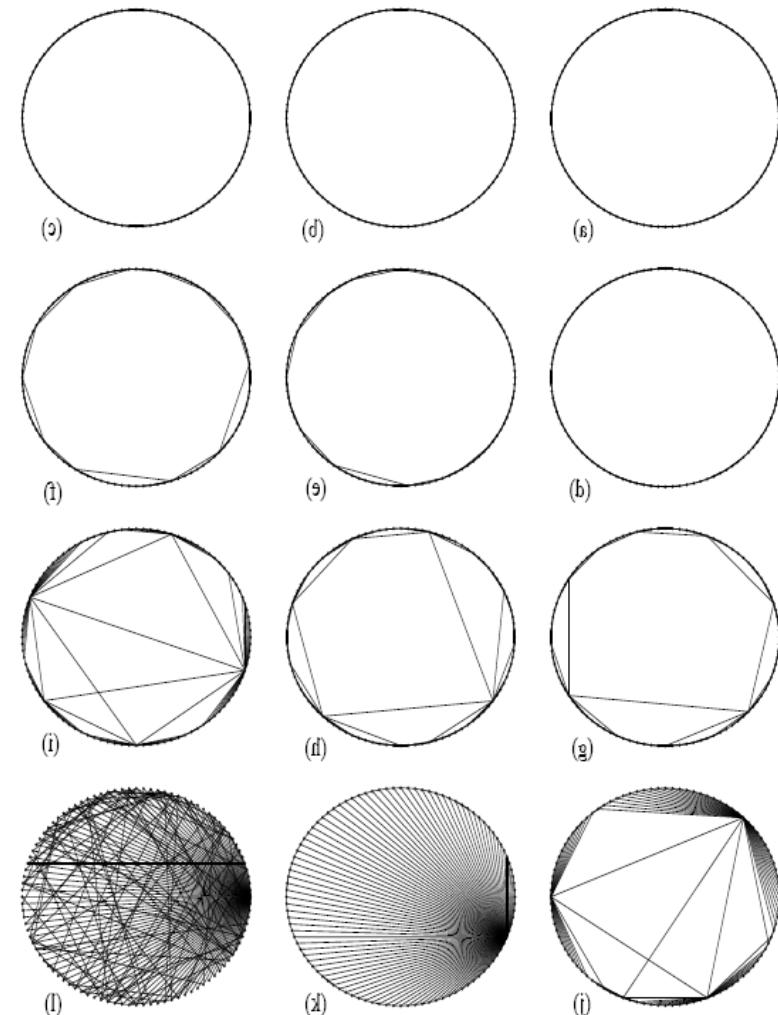
- E is the ‘energy’ cost we are trying to minimize
- L is the average shortest path in ‘hops’
- W is the total length of wire used

Quiz Q:

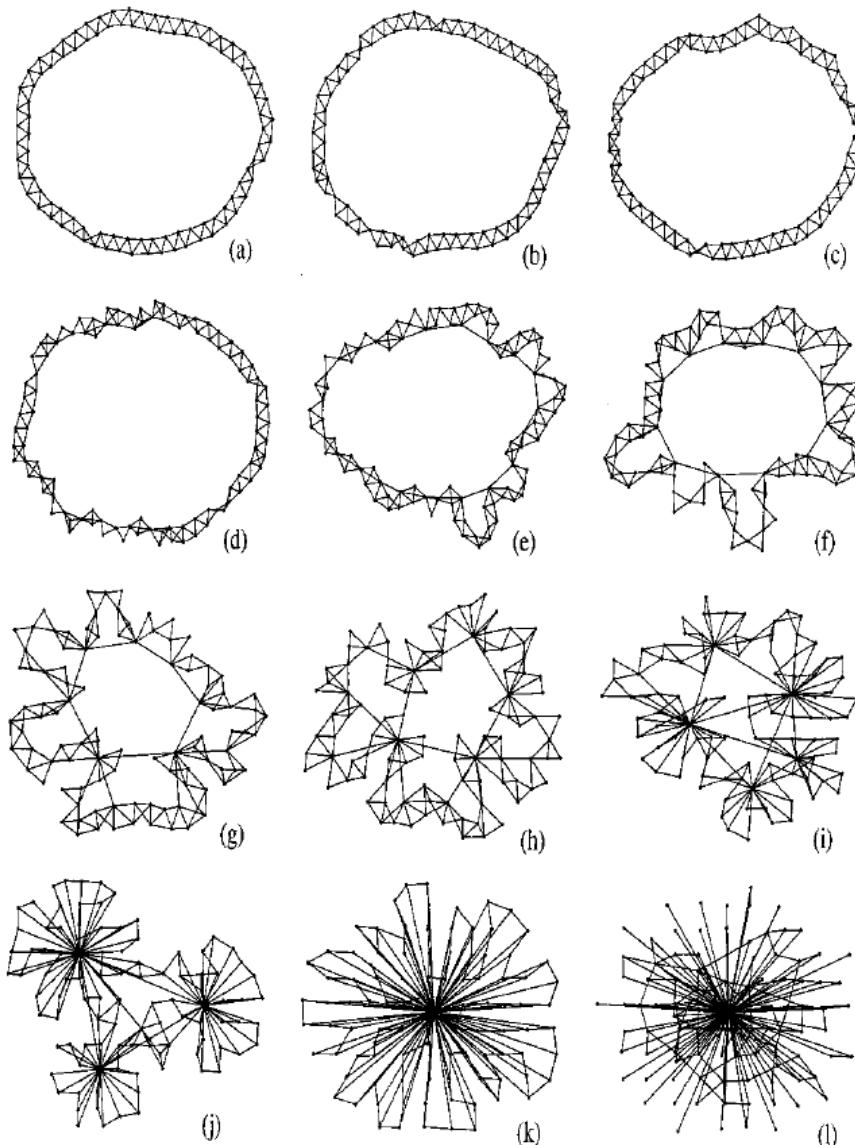
- ❑ Incorporates a person's preference for short distances or a small number of hops
- ❑ Relative to setting up the optimization for a road network, the optimization for an airline transportation network, from the passengers' point of view, should:

optimized networks

- ❑ rewire using simulated annealing
- ❑ sequence is shown in order of increasing λ

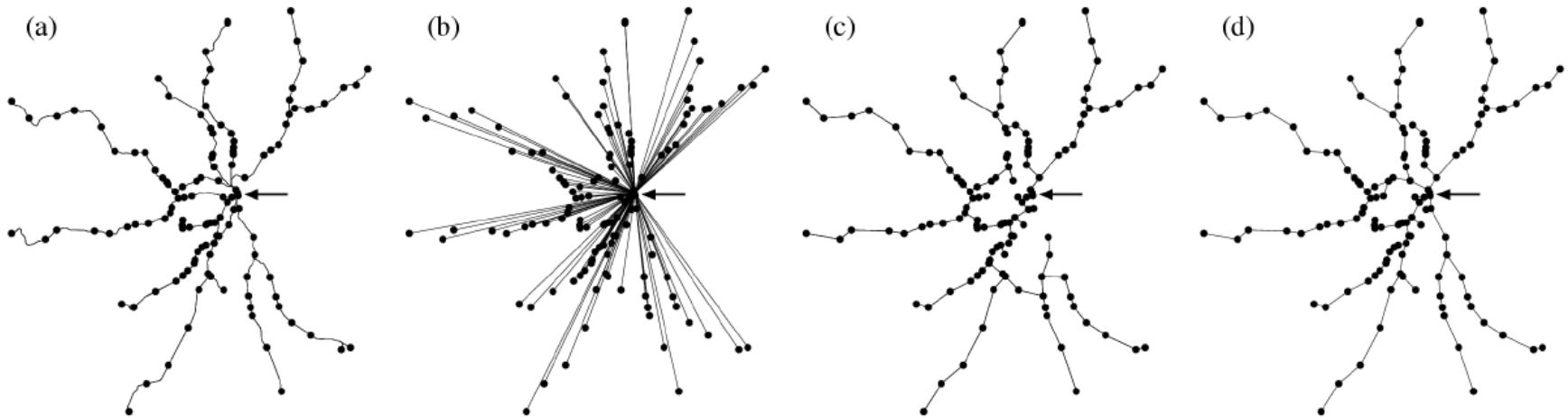


another view of optimized networks



- same networks, but the vertices are allowed to move using a spring layout algorithm
- wiring cost associated with the physical distance between nodes

optimizing from scratch

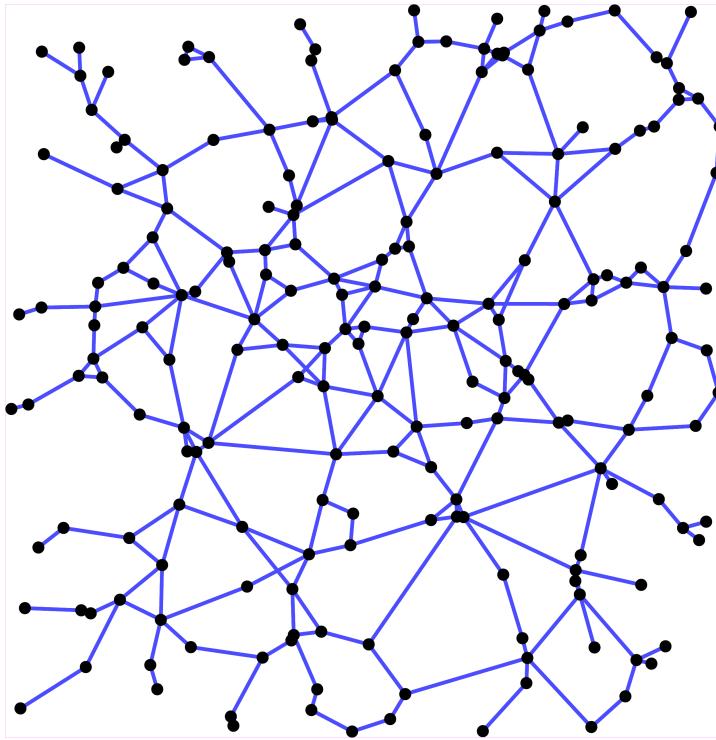


- (a) Commuter rail network in the Boston area. The arrow marks the assumed root of the network.
- (b) Star graph.
- (c) Minimum spanning tree.
- (d) The model applied to the same set of stations.

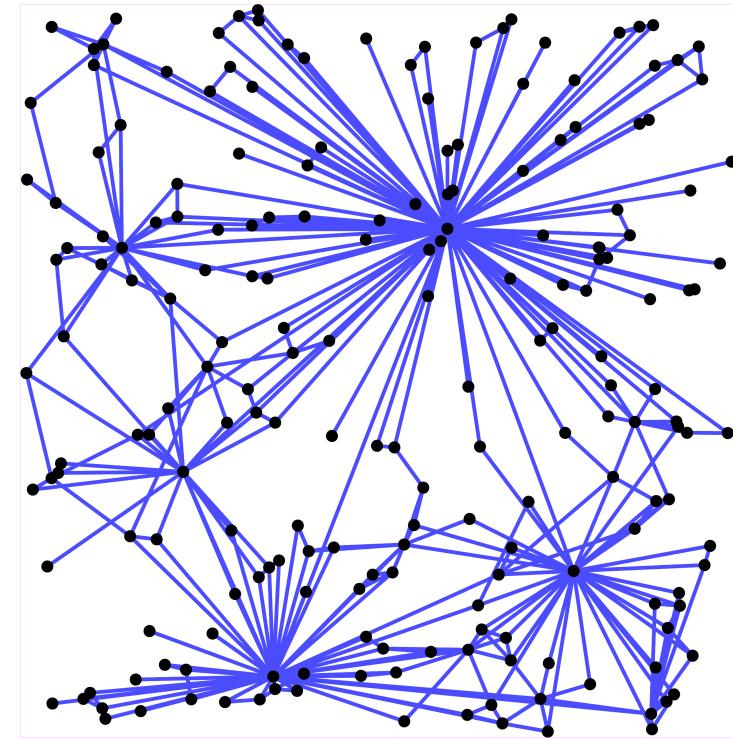
add edge with smallest weight $w'_{ij} = d_{ij} + \beta l_{j0}$

hops to root node
Euclidean distance between i and j

reminiscent of



Roads



Air routes

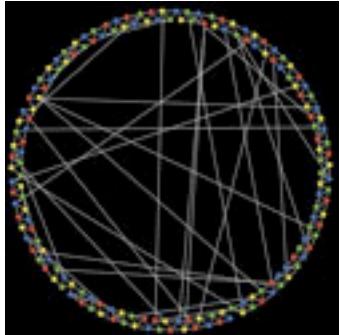
Quiz Q:

- ❑ A network that contains many hubs with far reaching edges is indicative of (check all that apply)
 - ❑ high cost of distance traveled
 - ❑ low cost of distance traveled
 - ❑ high cost of making many hops
 - ❑ low cost of making many hops

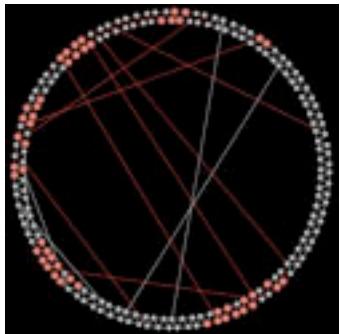
recap

- ❑ The world is small if you look at it as a network
- ❑ Yet it has lots of interesting local structure
- ❑ Watts & Strogatz came up with a simple model to incorporate the two
- ❑ Other models incorporate geography and hierarchical social structure
- ❑ Small worlds may evolve from different constraints (navigation, constraint optimization, group affiliation)

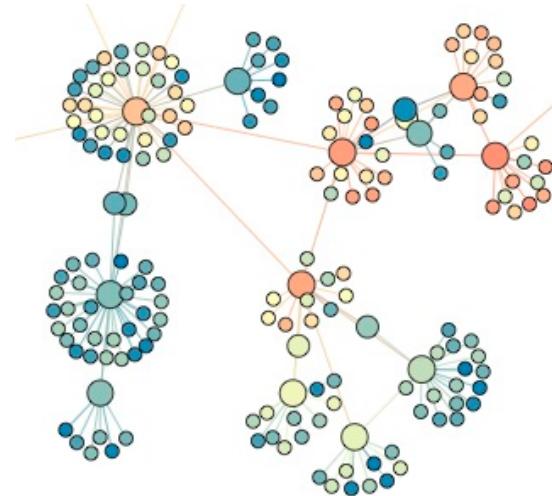
Next week: learning and diffusing on small world (and other topologies)



Graph coloring



Diffusion



SNA 6: processes on networks

Lada Adamic

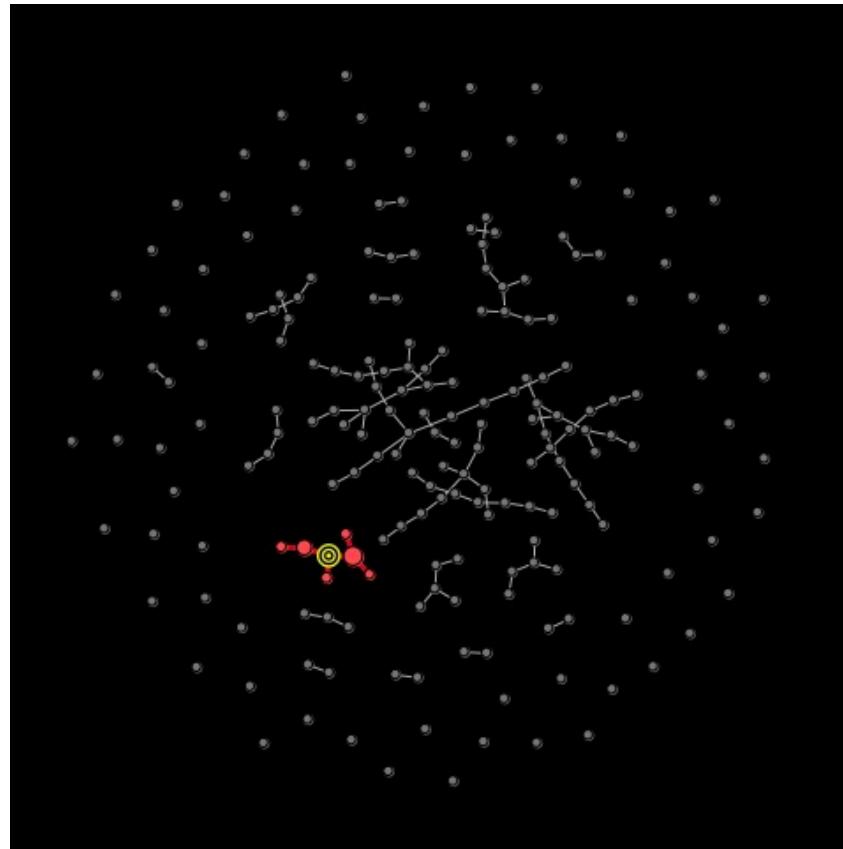


Processes on networks

- ❑ Diffusion (simple)
 - ❑ ER graphs
 - ❑ Scale-free graphs
 - ❑ Small-world topologies
- ❑ Complex contagion/thresholds
- ❑ Collective action
- ❑ Innovation
- ❑ Problem solving

Diffusion in networks: ER graphs

- ❑ review: diffusion in ER graphs

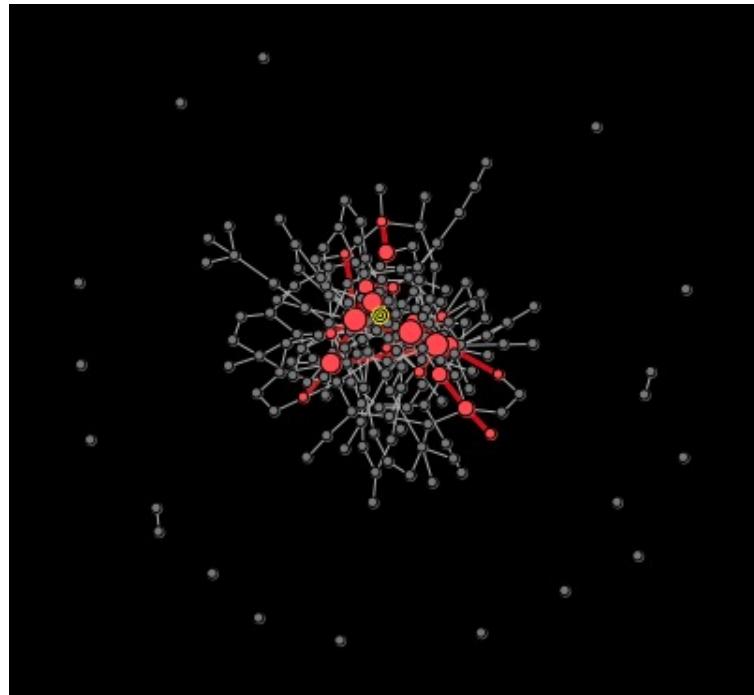


<http://www.ladamic.com/netlearn/NetLogo501/ERDiffusion.html>

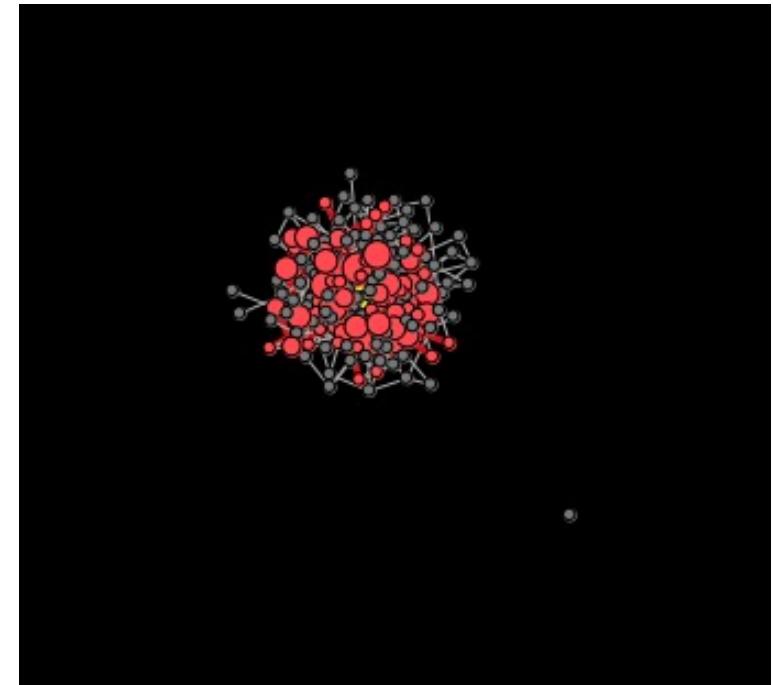
ER graphs: connectivity and density

nodes infected after 10 steps, infection rate = 0.15

average degree = 2.5



average degree = 10

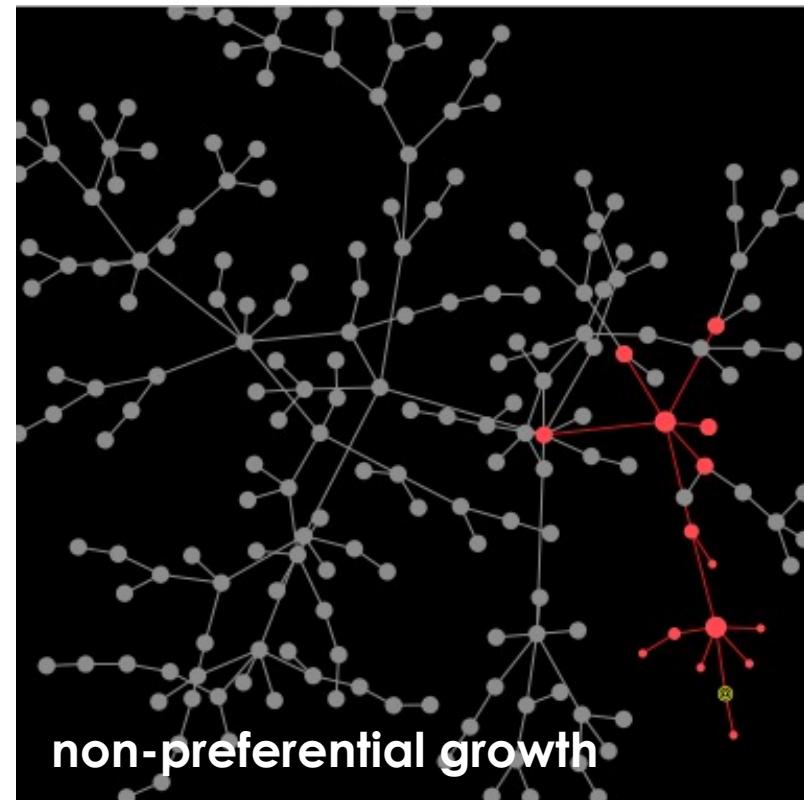
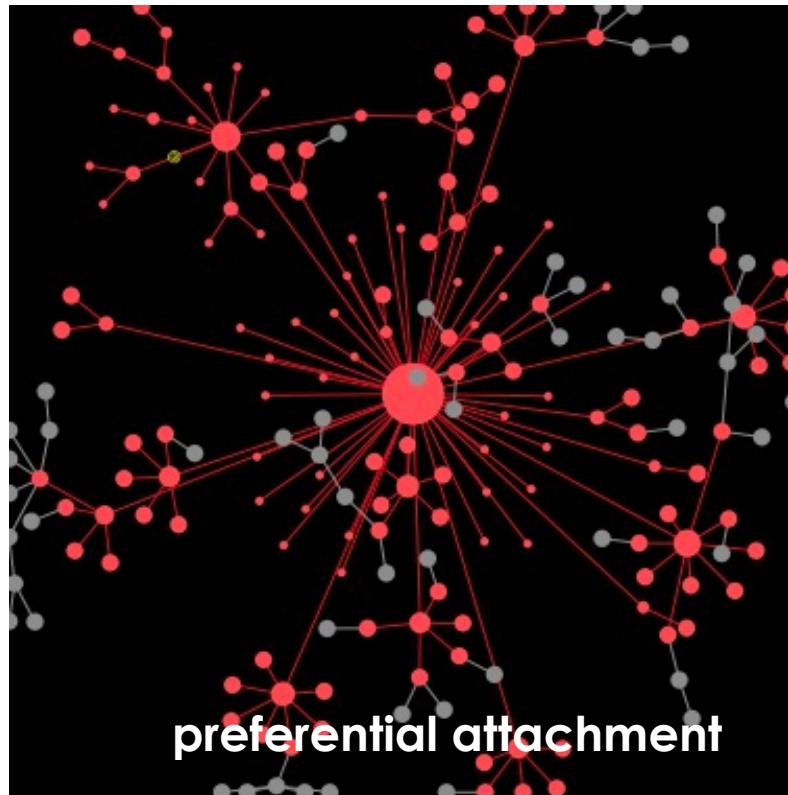


Quiz Q:

- ❑ When the density of the network increases, diffusion in the network is
 - ❑ faster
 - ❑ slower
 - ❑ unaffected

Diffusion in “grown networks”

- nodes infected after 4 steps, infection rate = 1



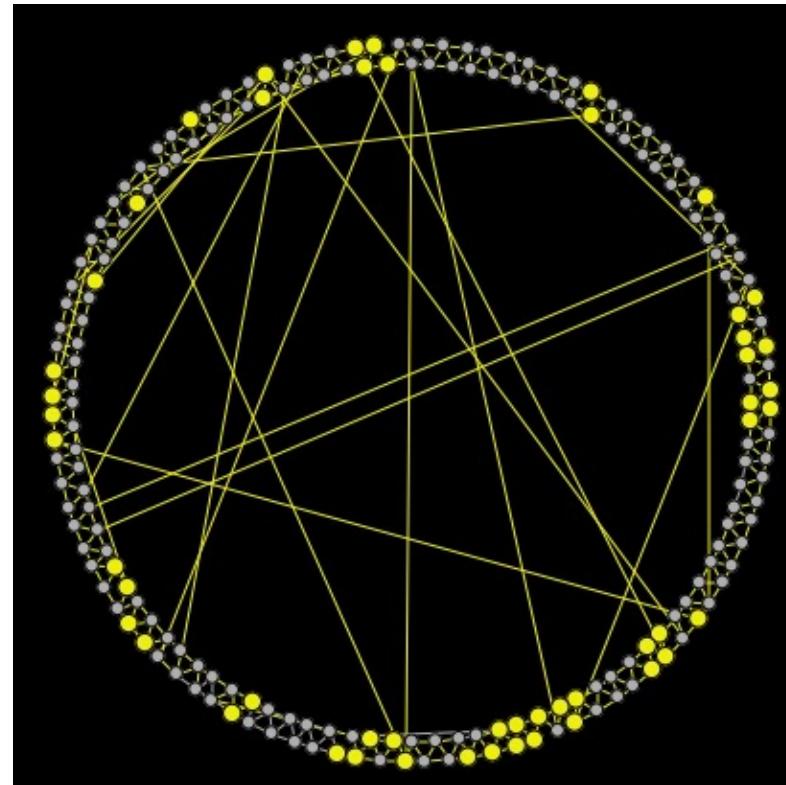
<http://www.ladamic.com/netlearn/NetLogo501/BADiffusion.html>

Quiz Q:

- ❑ When nodes preferentially attach to high degree nodes, the diffusion over the network is
 - ❑ faster
 - ❑ slower
 - ❑ unaffected

Diffusion in small worlds

- What is the role of the long-range links in diffusion over small world topologies?



<http://www.ladamic.com/netlearn/NetLogo4/SmallWorldDiffusionSIS.html>

Quiz Q:

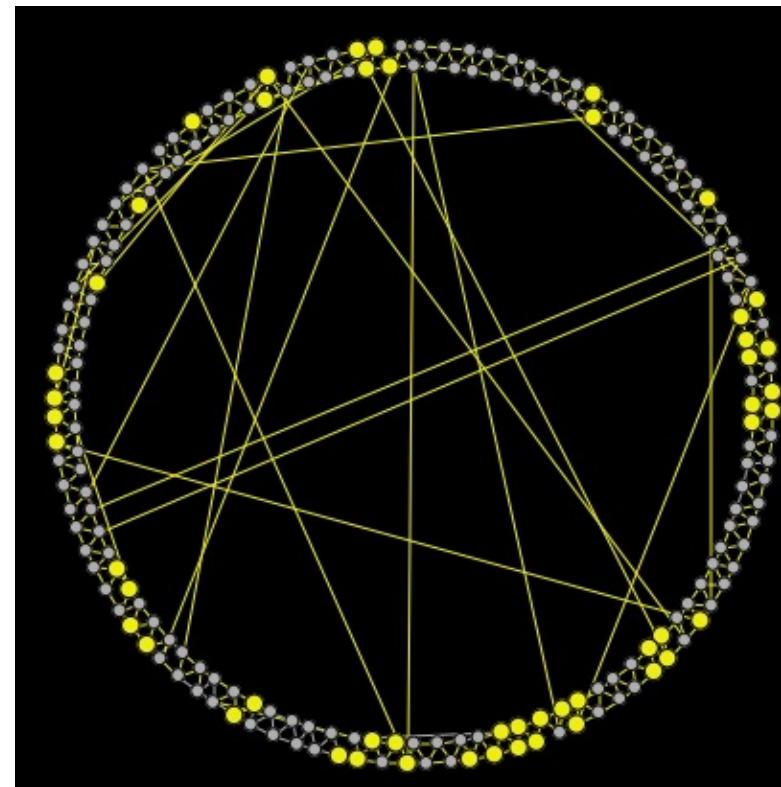
- ❑ As the probability of rewiring increases, the speed with which the infection spreads
 - ❑ increases
 - ❑ decreases
 - ❑ remains the same

Simple vs. complex contagion

- ❑ Simple contagion: each friend infects you with some probability for each unit of time
- ❑ Complex contagion: you will only take action if a certain number or fraction of your neighbors do

What is the role of the shortcuts?

- long range links unlikely to coincide in influence



Quiz Q:

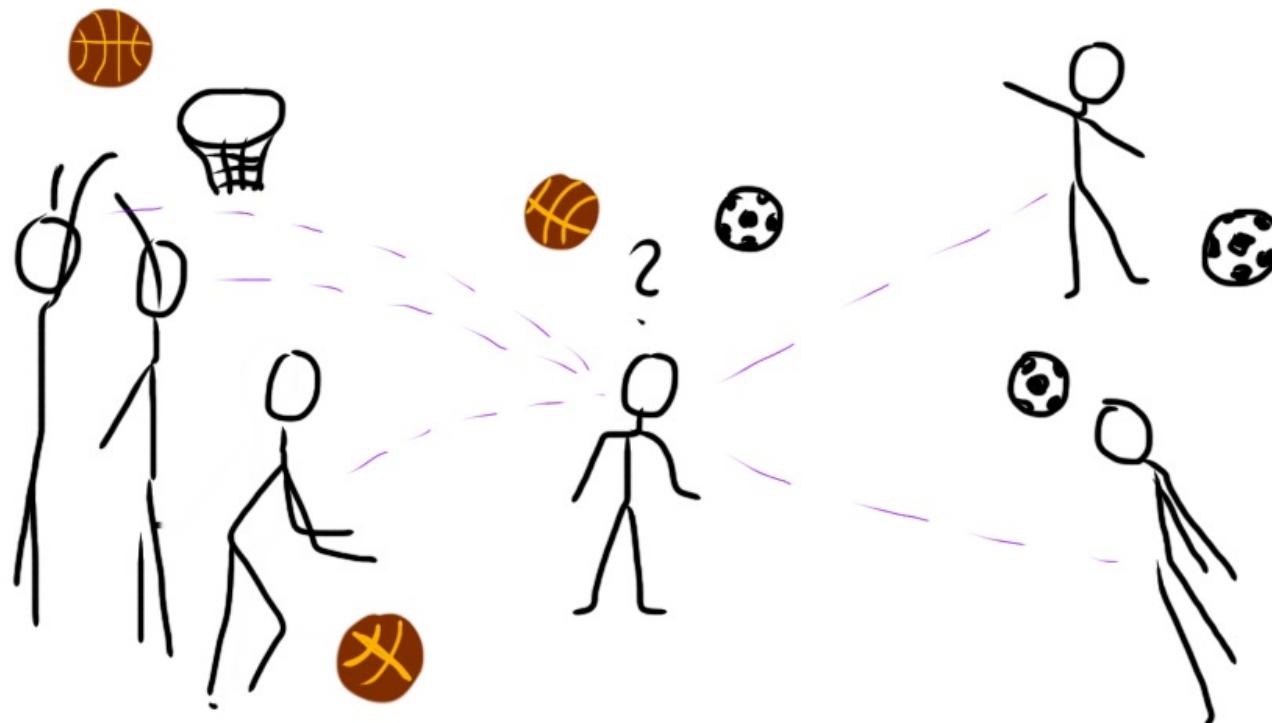
- ❑ Relative to the simple contagion process the complex contagion process:
 - ❑ is better able to use shortcuts
 - ❑ advances more rapidly through the network
 - ❑ infects a greater number of nodes

networked coordination game

- ❑ choice between two things, A and B (e.g. basketball and soccer)
- ❑ if friends choose A, they get payoff a
- ❑ if friends choose B, they get payoff b
- ❑ if one chooses A while the other chooses B, their payoff is 0

coordinating with one's friends

Let A = basketball, B = soccer. Which one should you learn to play?

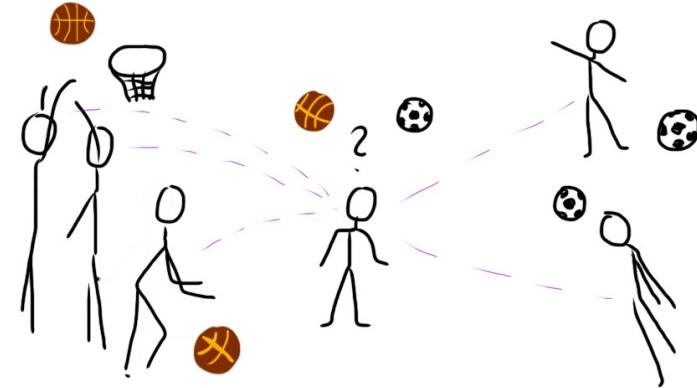


fraction p = 3/5 play basketball

fraction p = 2/5 play soccer

which choice has higher payoff?

- ❑ d neighbors
- ❑ p fraction play basketball (A)
- ❑ $(1-p)$ fraction play soccer (B)
- ❑ if choose A, get payoff
 $p * d * a$
- ❑ if choose B, get payoff
 $(1-p) * d * b$
- ❑ so should choose A if
 - ❑ $p d a \geq (1-p) d b$
 - ❑ or
 - ❑ $p \geq b / (a + b)$



two equilibria

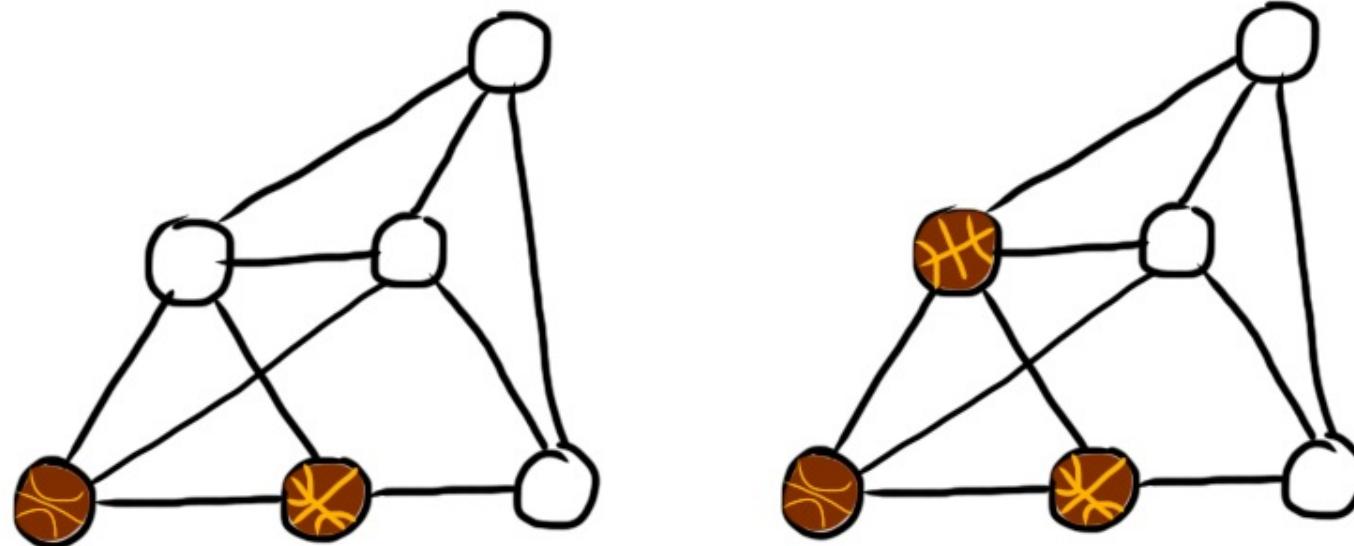
- ❑ everyone adopts A
- ❑ everyone adopts B

what happens in between?

- ❑ What if two nodes switch at random? Will a cascade occur?
- ❑ example:
 - ❑ $a = 3, b = 2$
 - ❑ payoff for nodes interaction using behavior A is $3/2$ as large as what they get if they both choose B
 - ❑ nodes will switch from B to A if at least $q = 2/(3+2) = 2/5$ of their neighbors are using A

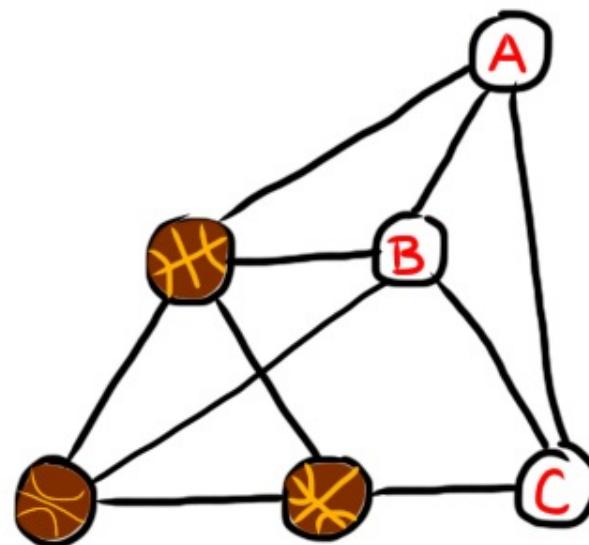
how does a cascade occur

- ❑ suppose 2 nodes start playing basketball due to external factors (e.g. they are bribed with a free pair of shoes by some devious corporation)

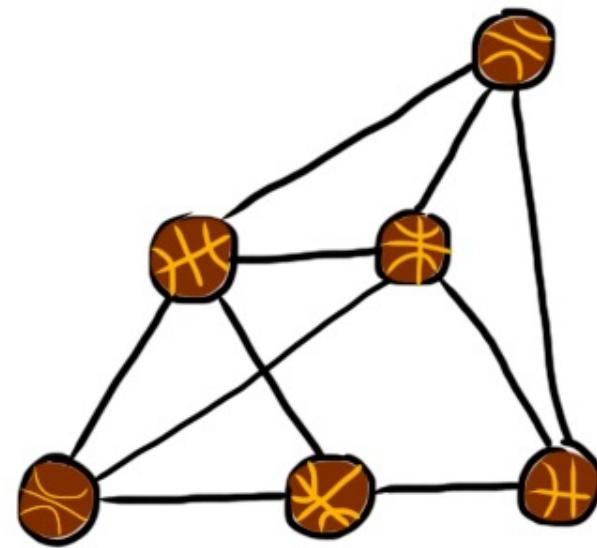
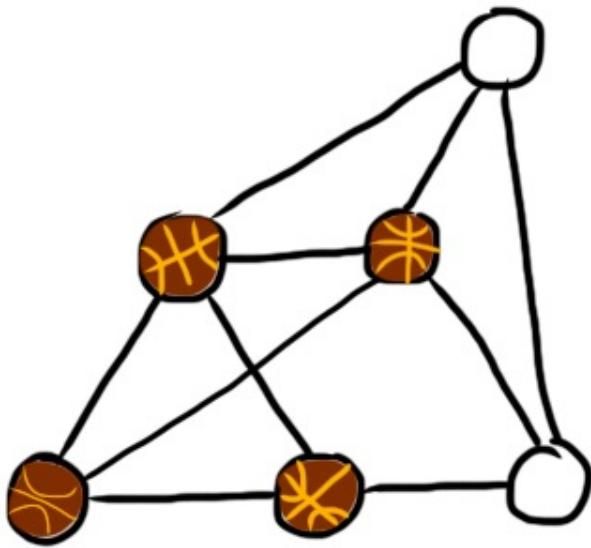


Quiz Q:

Which node(s) will switch to playing basketball next?

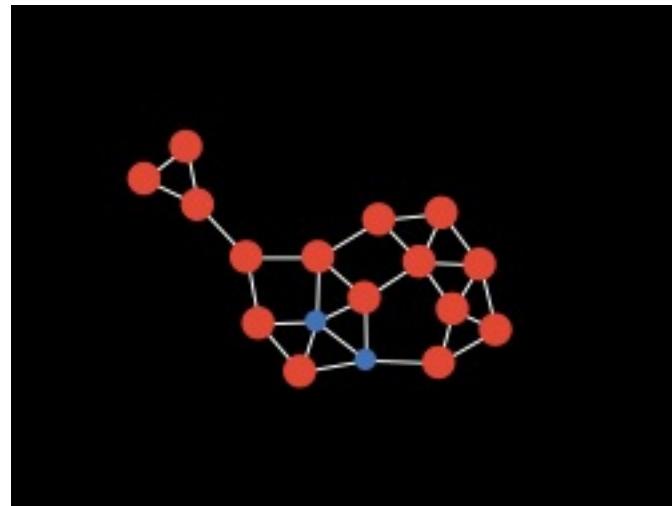


the complete cascade



you pick the initial 2 nodes

- ❑ A larger example (Easley/Kleinberg Ch. 19)
- ❑ does the cascade spread throughout the network?



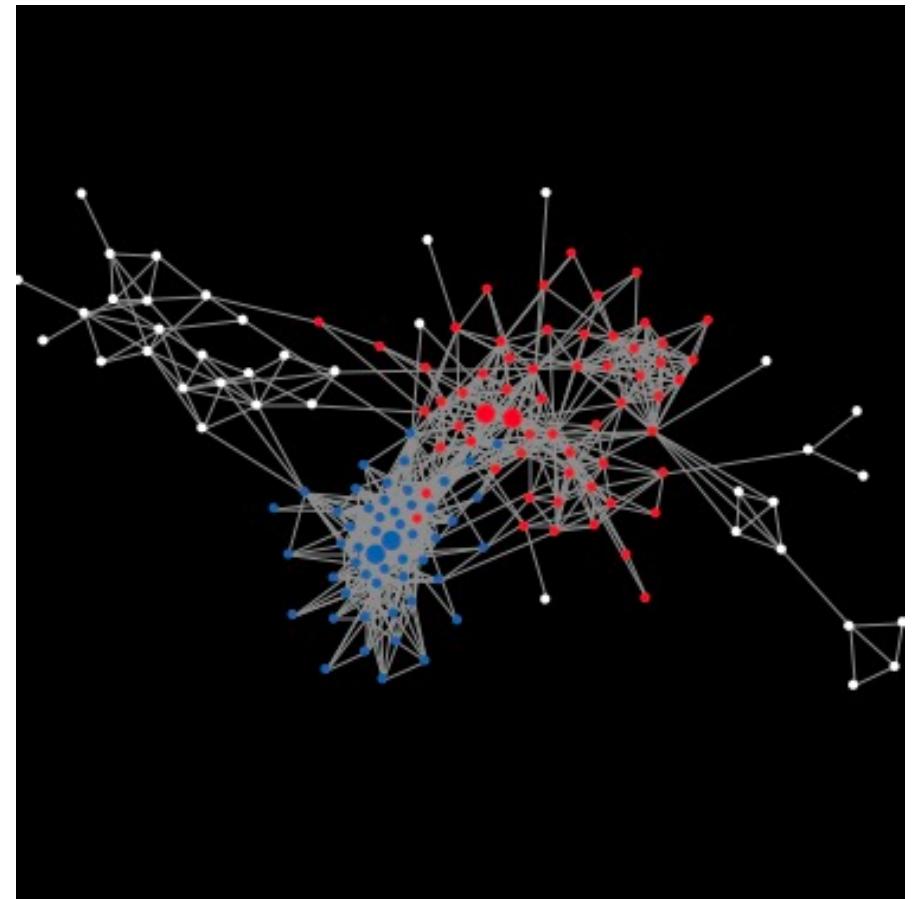
<http://www.ladamic.com/netlearn/NetLogo412/CascadeModel.html>

implications for viral marketing

- ❑ if you could pay a small number of individuals to use your product, which individuals would you pick?

try it on Lada's Facebook network

- you can play with a partner
- each person gets to pick 2 nodes
 - first person picks one blue
 - second person picks one red
 - first person picks an additional blue
 - second person picks an additional red



Quiz question:

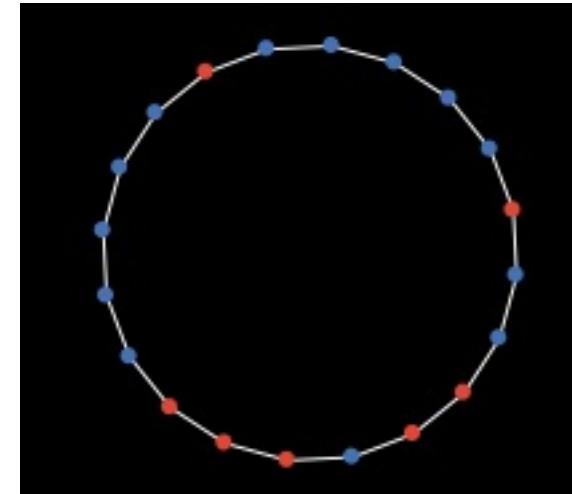
- ❑ What is the role of communities in complex contagion
 - ❑ enabling ideas to spread in the presence of thresholds
 - ❑ creating isolated pockets impervious to outside ideas
 - ❑ allowing different opinions to take hold in different parts of the network

bilingual nodes

- ❑ so far nodes could only choose between A and B
- ❑ what if you can play both A and B, but pay an additional cost c ?

try it on a line

- ❑ Increase the cost of being bilingual so that no node chooses to do so. Let the cascade run
- ❑ Now lower the cost.
 - ❑ What happens?



Quiz Q:

- ❑ The presence of bilingual nodes
 - ❑ helps the superior solution to spread throughout the network
 - ❑ helps inferior options to persist in the network
 - ❑ causes everyone in the network to become bilingual

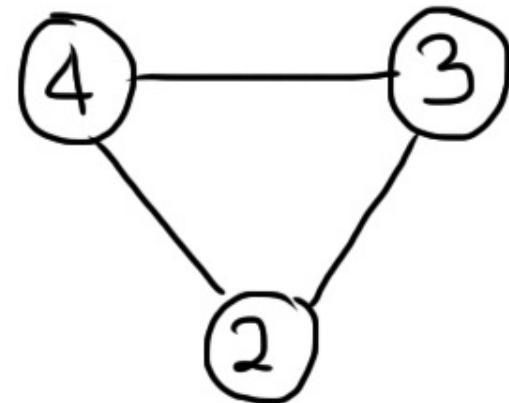
knowledge, thresholds, and collective action

- ❑ nodes need to coordinate across a network, but have limited horizons



can individuals coordinate?

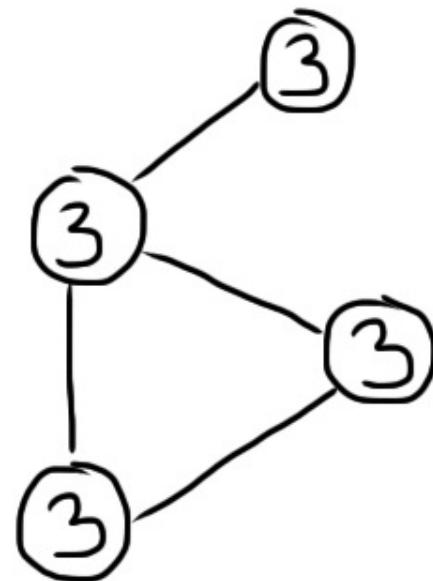
- ❑ each node will act if at least x people (including itself) mobilize



nodes will not mobilize

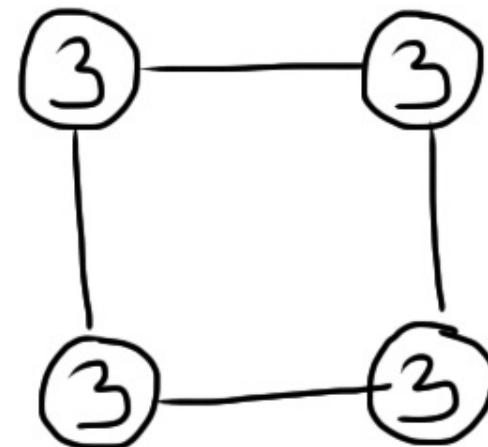
mobilization

- there will be some turnout



Quiz Q:

- ☐ will this network mobilize (at least some fraction of the nodes will protest)?



innovation in networks

- ❑ network topology influences who talks to whom
- ❑ who talks to whom has important implications for innovation and learning

better to innovate or imitate?

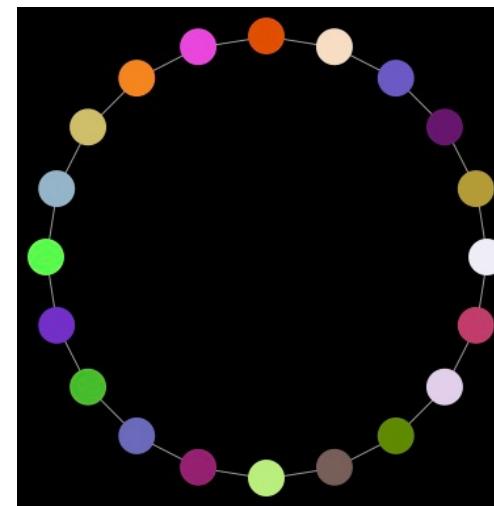
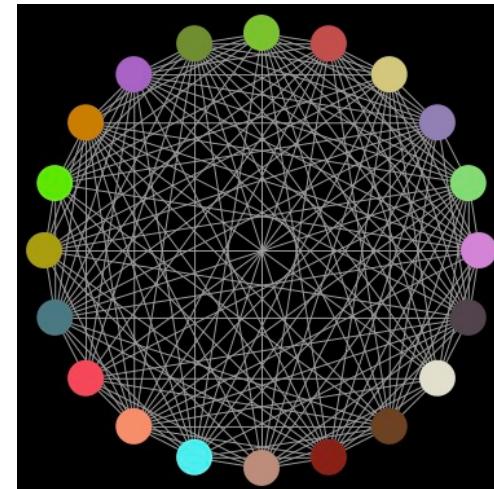
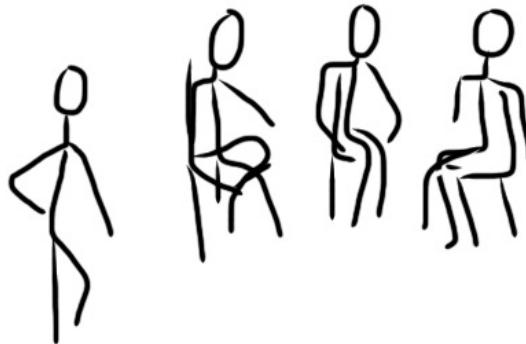


brainstorming:
more minds together,
but also danger of groupthink

working in isolation:
more independence
slower progress



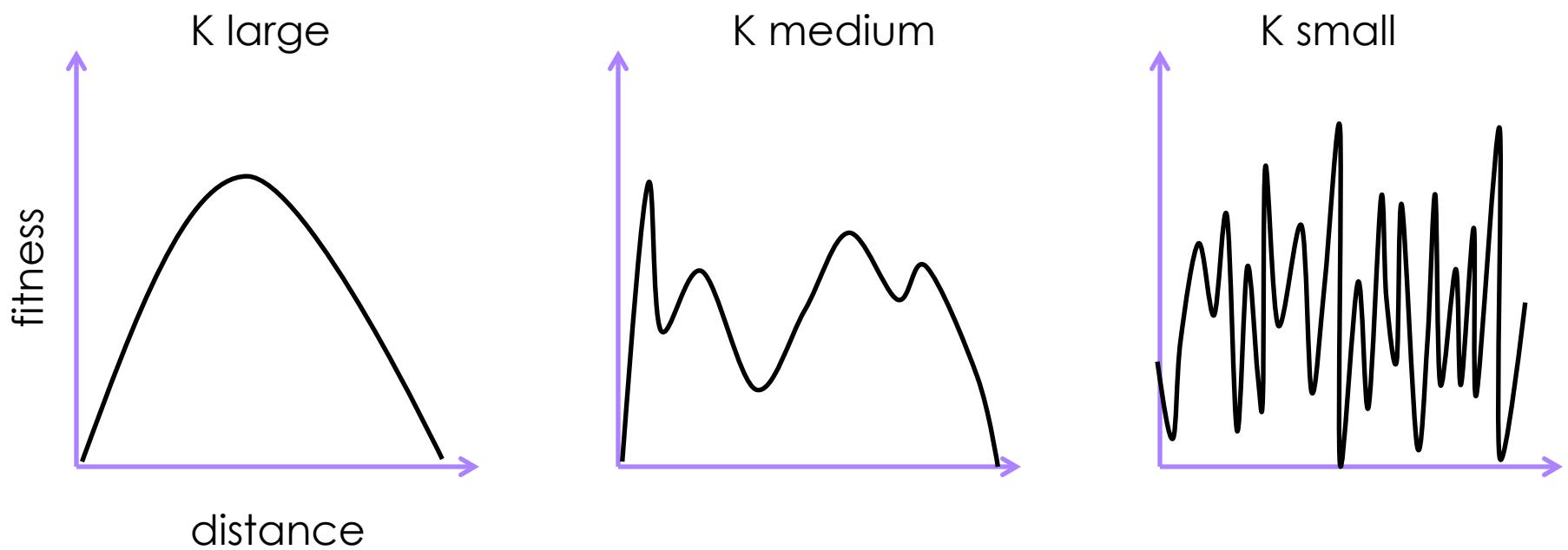
in a network context



modeling the problem space

- ❑ Kauffman's NK model
- ❑ N dimensional problem space
 - ❑ N bits, each can be 0 or 1
- ❑ K describes the smoothness of the fitness landscape
 - ❑ how similar is the fitness of sequences with only 1-2 bits flipped (K = 0, no similarity, K large, smooth fitness)

Kauffman's NK model



Update rules

- ❑ As a node, you start out with a random bit string
- ❑ At each iteration
 - ❑ If one of your neighbors has a solution that is more fit than yours, imitate (copy their solution)
 - ❑ Otherwise innovate by flipping one of your bits

Quiz Q:

- ❑ Relative to the regular lattice, the network with many additional, random connections has on average:
 - ❑ slower convergence to a local optimum
 - ❑ smaller improvement in the best solution relative to the initial maximum
 - ❑ more oscillations between solutions

Coordination: graph coloring

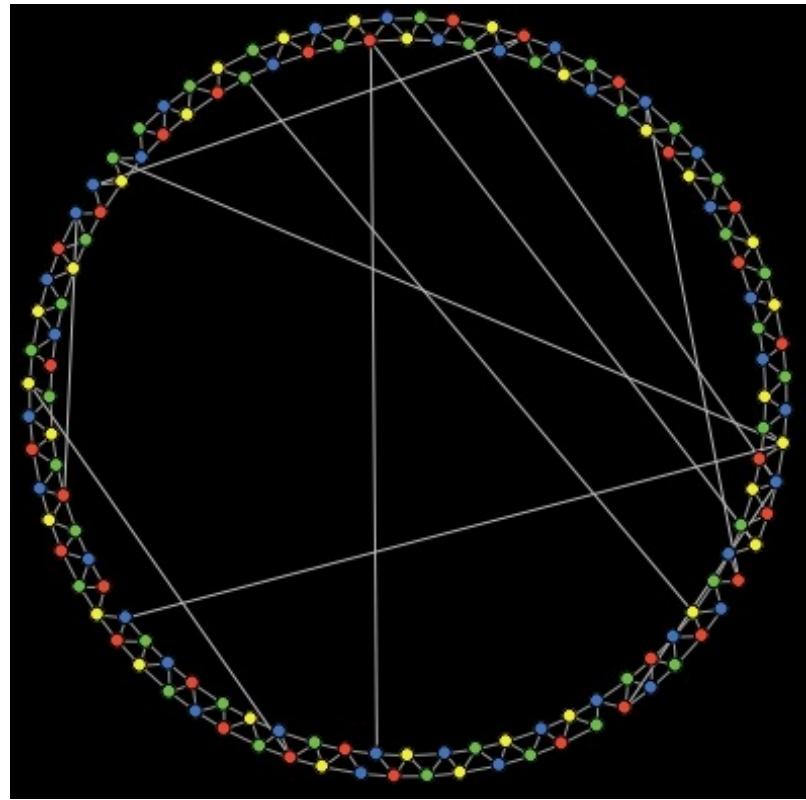
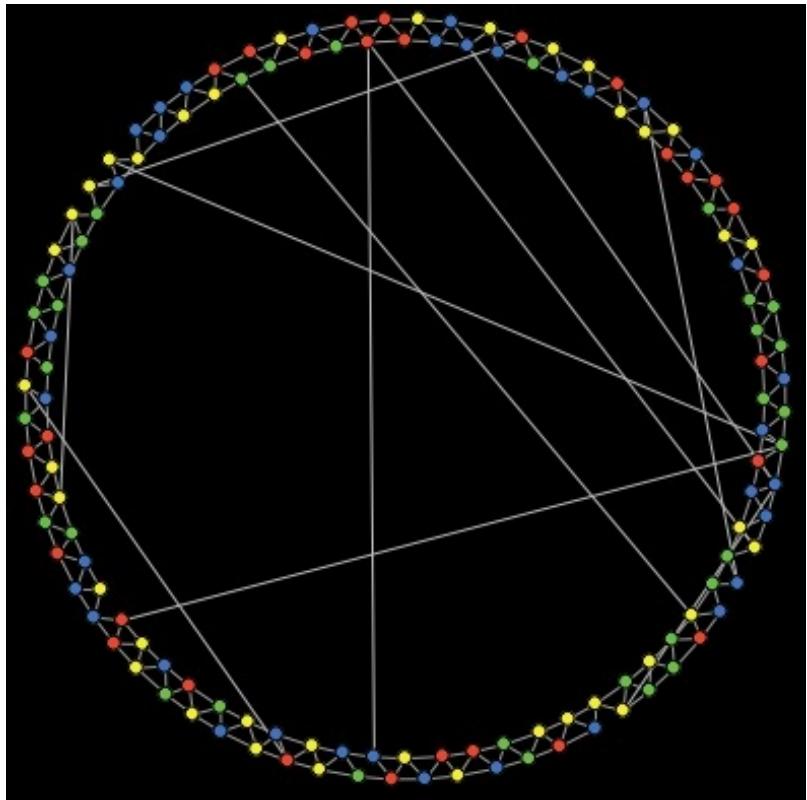
- Application: coloring a map: limited set of colors, no two adjacent countries should have the same color



graph coloring on a network

- ❑ Each node is a human subject. Different experimental conditions:
 - ❑ knowledge of neighbors' color
 - ❑ knowledge of entire network
- ❑ Compare:
 - ❑ regular ring lattice
 - ❑ small-world topology
 - ❑ scale-free networks

simulation



Quiz Q:

- ❑ As the rewiring probability is increased from 0 to 1 the following happens:
 - ❑ the solution time decreases
 - ❑ the solution time increases
 - ❑ the solution time initially decreases then increases again

recap

- ❑ network topology influences processes occurring on networks
 - ❑ what state the nodes converge to
 - ❑ how quickly they get there
- ❑ process mechanism matters:
 - ❑ simple vs. complex contagion
 - ❑ coordination
 - ❑ learning

Recipe recommendation using ingredient networks

Chun-Yuen Teng¹, Yu-Ru Lin^{2,3}, Lada A. Adamic¹

¹School of Information, University of Michigan

²IQSS, Harvard University

³CCS, Northeastern University

Online recipes

Example: cupcakes More searches: Ingredient | Nutrition | Advanced

new at  recipes » videos » menus » holidays »



Spicy Corn Salad Corn, sweet onions, and jalapenos combine into a spicy summer salad. »



Top Grilling Recipes The best of burgers, chicken, and grilled sides. »

Summer Dinners Seasonal stir-fries, salads, and wraps. »

Menu Planner A year's worth of menus costs less than one take-out dinner. »

More Recipes Like This
[Sweet and Spicy Baked](#)



Crispy Herb Baked Chicken

By: DCANTER
"The secret ingredient is instant mashed potatoes, used to make the crispy coating."

★★★★★ Rate/Review | Read Reviews (438)

 133  0 

1 of 23 Photos

Prep Time: 15 Min Cook Time: 45 Min Ready In: 1 Hr

Servings (Help)
4 US Metric

Original Recipe Yield 4 - 5 servings

Ingredients

2/3 cup dry potato flakes
1/3 cup grated Parmesan cheese
1 teaspoon garlic salt
1 (3 pound) chicken, skin removed, cut into pieces
1/3 cup butter, melted

Directions

kitchenapproved

Add to Recipe Box
 Add to Shopping List
 Print this Recipe

supportingmembers

Create Menu
 Customize Recipe
 Kitchen-friendly View

What to Drink?

[Somewhere Blue](#)

Our online recipes



- 46,337 recipes
- Each recipe includes directions, ingredients, nutrition info, cooking time, and regional information
- 1,976,920 reviews include ratings and text

Research questions

- ❑ What patterns emerge from the collective cooking knowledge aggregated in recipes?
- ❑ How can ingredient networks be used for predicting recipe ratings?

Recipe mining

- ❑ Cooking methods
 - ❑ Regional preferences
- ❑ Ingredients
 - ❑ Combination of ingredients
 - ❑ Modification of ingredients
- ❑ Predicting ratings:



OR



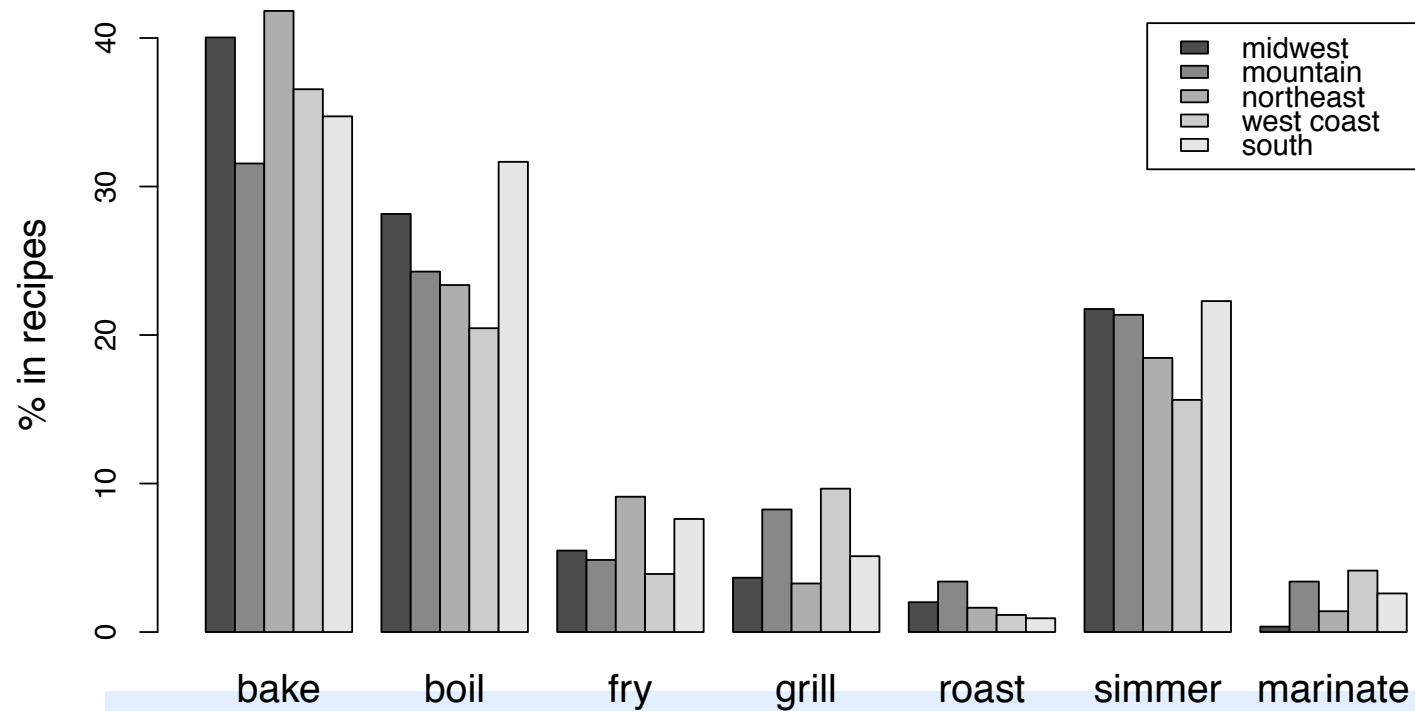
Credit: sonomaorganics.com

Cooking methods



US Regional Preferences

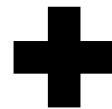
- Significantly varying preferences
- In the West, 42% of grilling recipes involve seafood relative to other regions (6%)



Ingredients



Combining ingredients

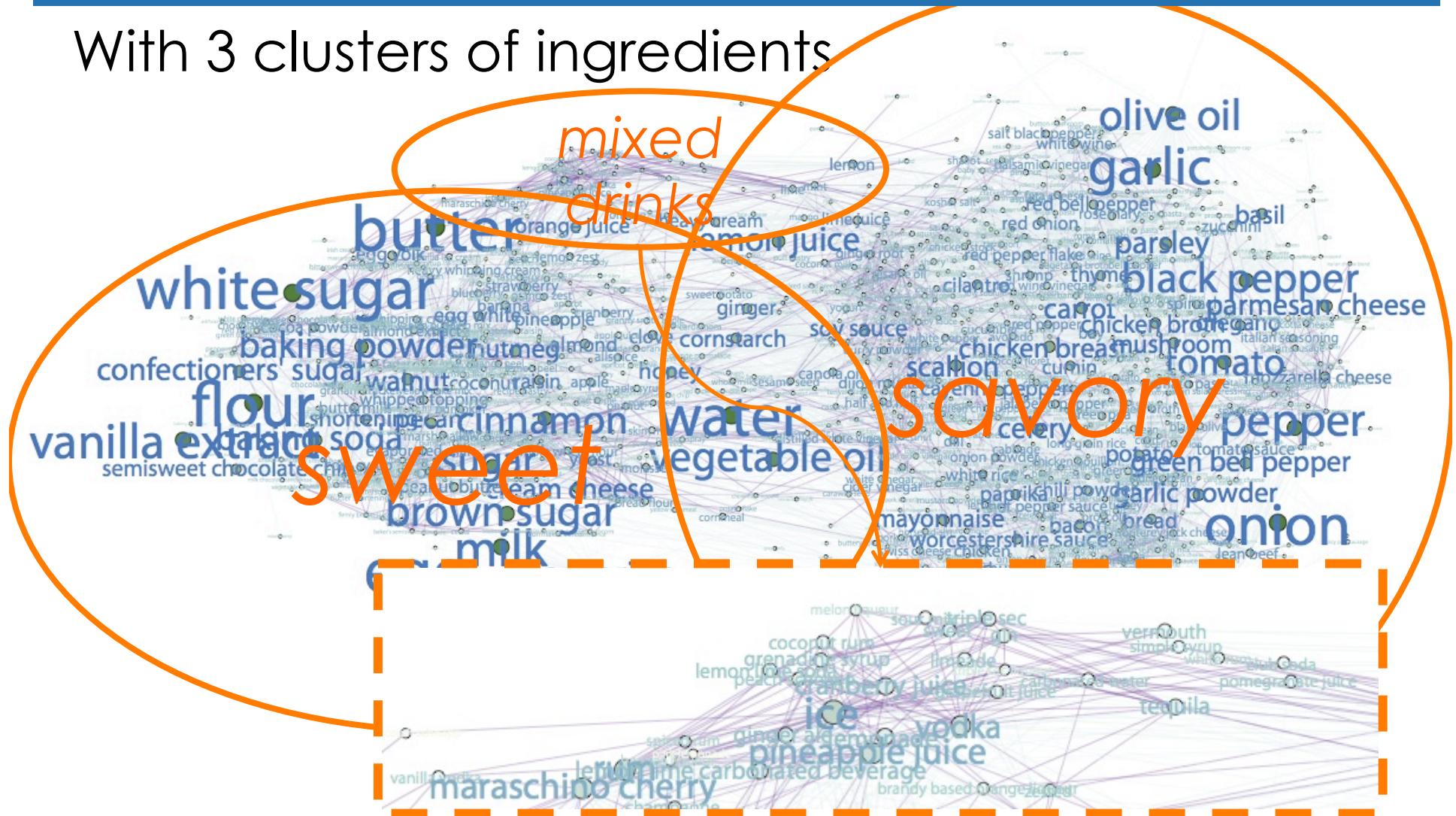


Complement network

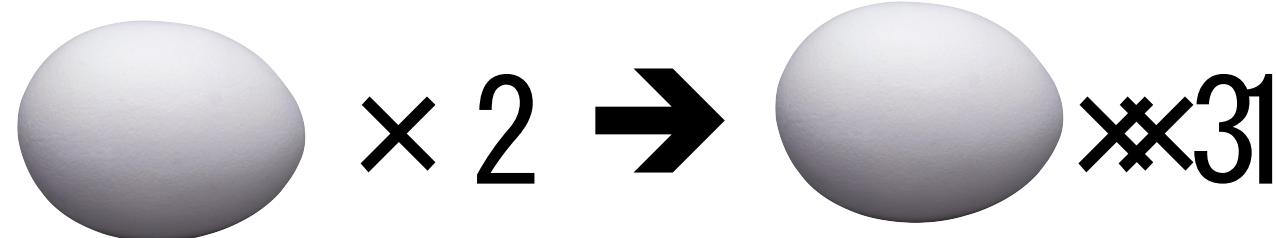
- Nodes: ingredients
- Undirected edges:
 - Weighted by pointwise mutual information
 $PMI = \log \left(P(a,b) / P(a)P(b) \right)$
 - $P(a,b) = (\# \text{ of recipes containing } a \text{ and } b) / (\# \text{ of recipes})$
 - $P(a) = (\# \text{ of recipes containing } a) / (\# \text{ of recipes})$
 - $P(b) = (\# \text{ of recipes containing } b) / (\# \text{ of recipes})$
 - Recipe rating and PMI of its ingredient pairs
 - Mean and minimum of PMI (no correlation with rating)
 - Max of PMI ($\rho=0.09$, $p < 0.001$)

Complement network

With 3 clusters of ingredients



Recipe modification

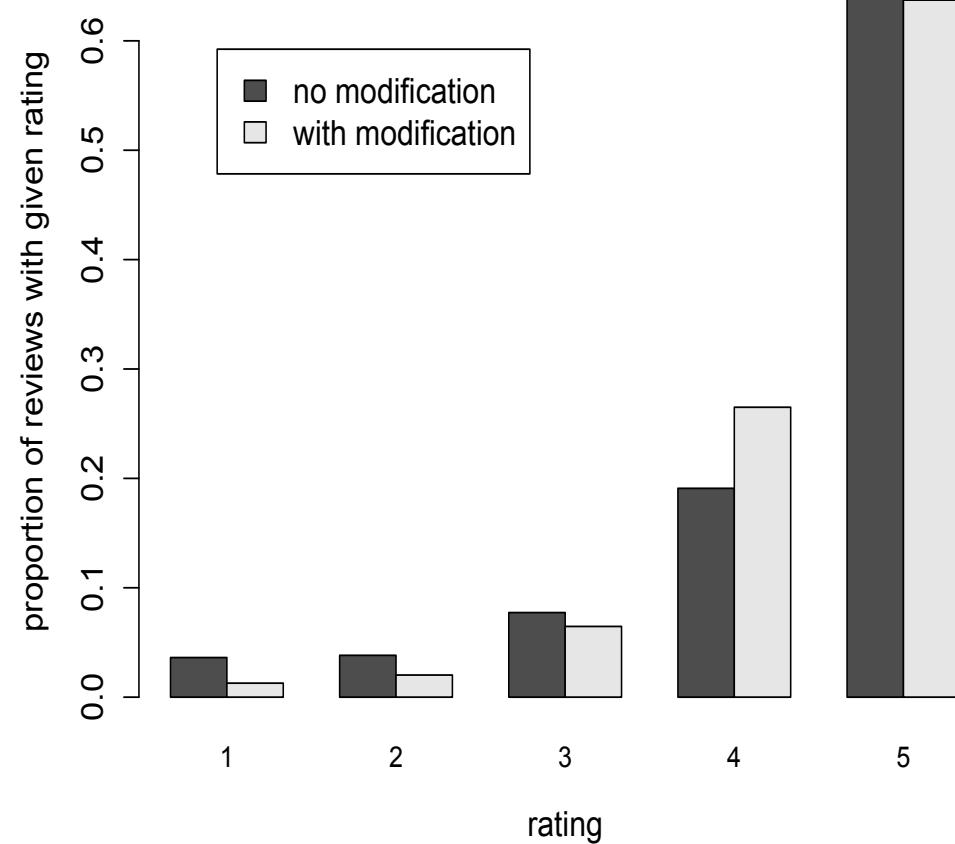


+

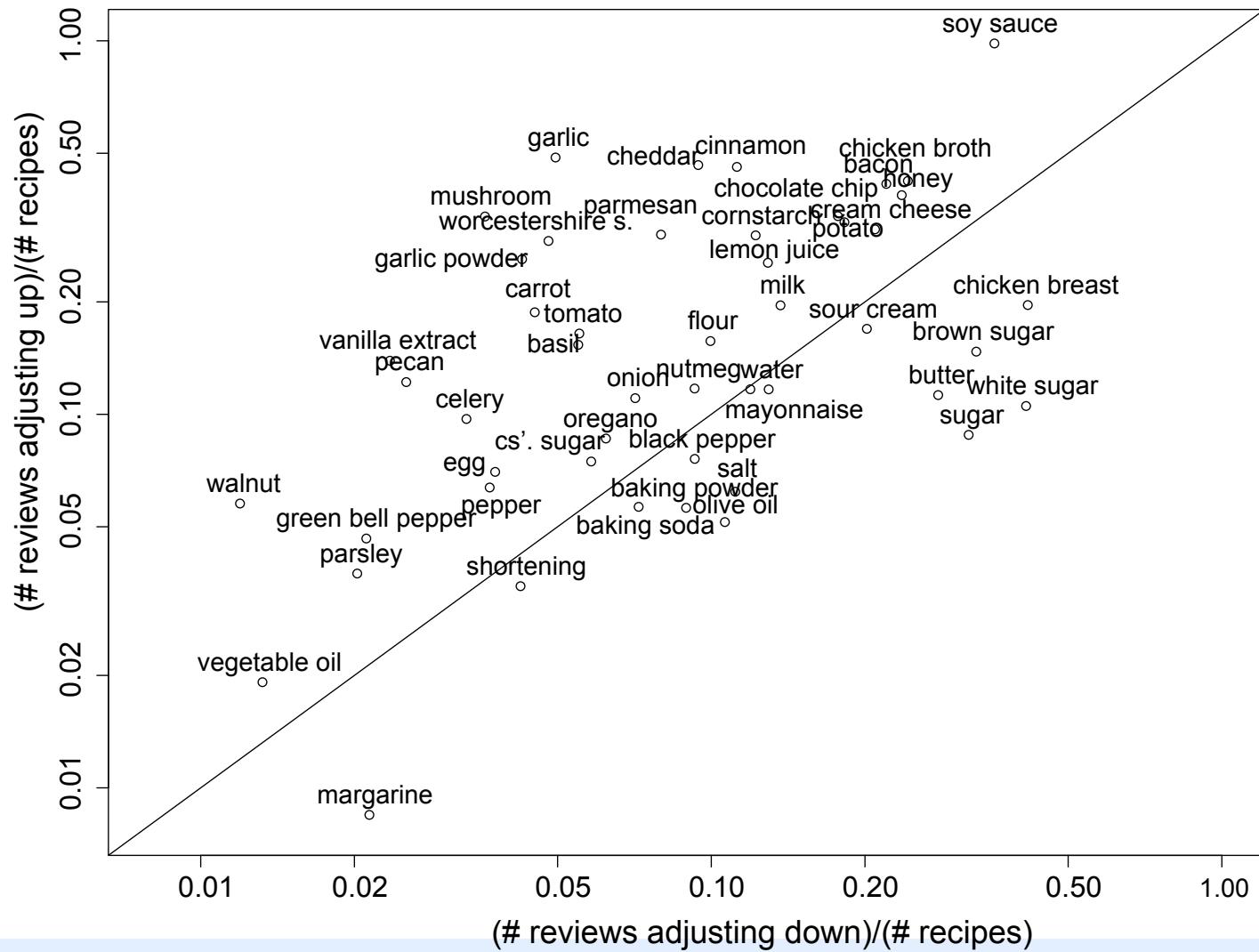


Recipe modification

- 60% of reviews contain “add”, “omit”, “instead”, “extra”, and 14 others.
- Reviews that include changes assign higher star ratings (4.49 vs. 4.39, $p<10^{-10}$)
- almost perfect but not quite (4 star) reviews often suggest modifications



Suggested modifications of quantities



Correlations between ingredient modifications

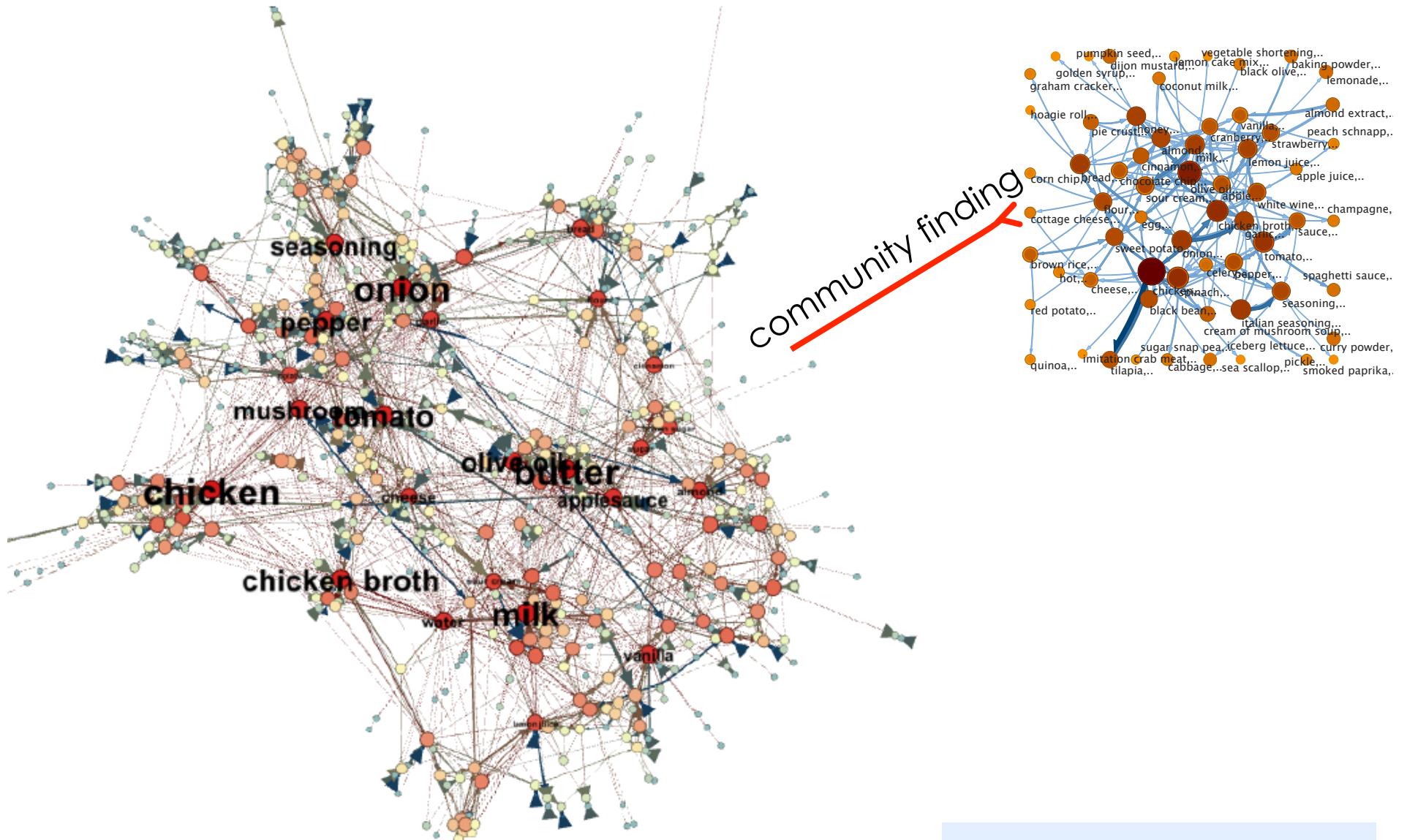
- Recipe freq. vs. deletion/recipe ($\rho = -0.22$)
- Recipe freq. vs. addition/recipe ($\rho = -0.25$)
- Recipe freq. vs. increase/recipe ($\rho = -0.26$)
- Correlations between ingredient modifications

	addition	deletion	increase	decrease
# recipes	0.41	0.22	0.61	0.68
addition		-0.15	0.79	0.11
deletion			0.09	0.58
increase				0.39

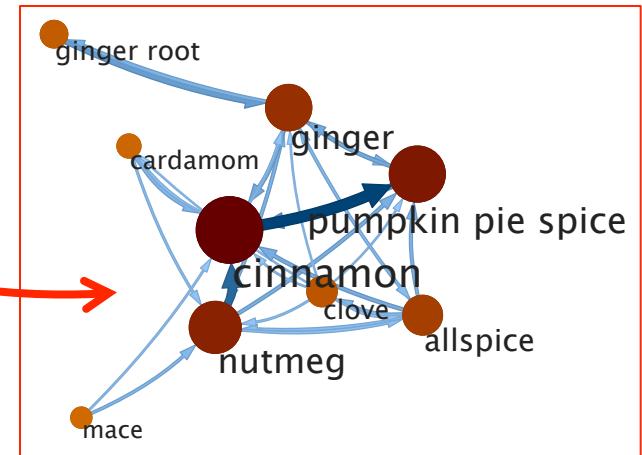
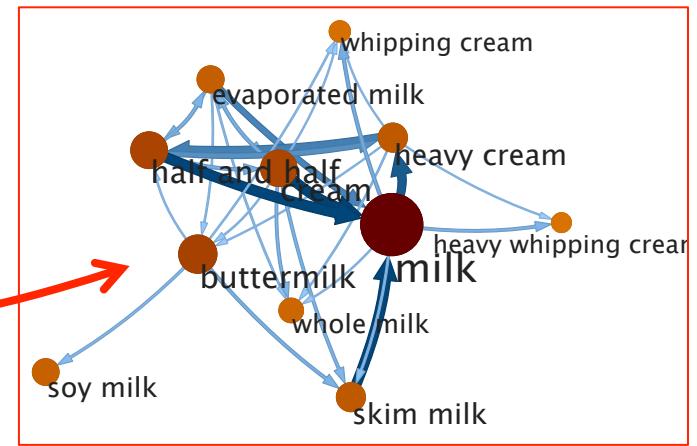
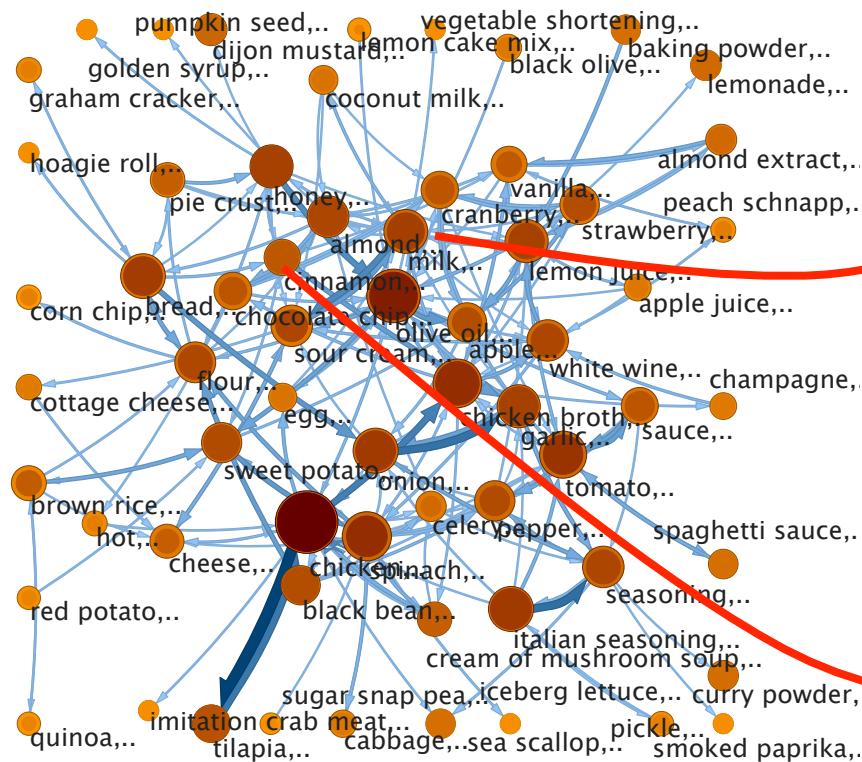
Substitution network

- Extract substitution relationships from comments
 - e.g. “I replaced the butter with sour cream”
 - “replace **a** with **b**”, “substitute **b** for **a**”, “**b** instead of **a**”
- Nodes: ingredients
- Edge weights = $p(\mathbf{b} \mid \mathbf{a})$, which is the proportion of substitutions of ingredient **a** that suggests ingredient **b**

Substitution network



Substitution network: communities



Examples of substitution

main	other ingredients
chicken	turkey, beef, sausage, chicken breast, bacon
olive oil	butter, apple sauce, oil, banana, margarine
sweet potato	yam, potato, pumpkin, butternut squash, parsnip
baking powder	baking soda, cream of tartar
almond	pecan, walnut, cashew, peanut, sunflower s.
apple	peach, pineapple, pear, mango, pie filling
egg	egg white, egg substitute, egg yolk
tilapia	cod, catfish, flounder, halibut, orange roughy
spinach	mushroom, broccoli, kale, carrot, zucchini
italian seasoning	basil, cilantro, oregano, parsley, dill
cabbage	coleslaw mix, sauerkraut, bok choy napa cabbage

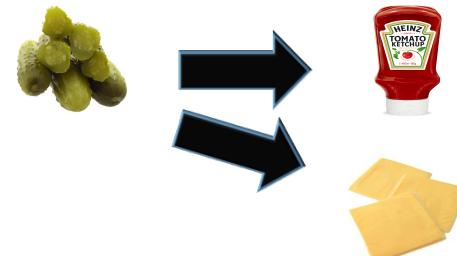
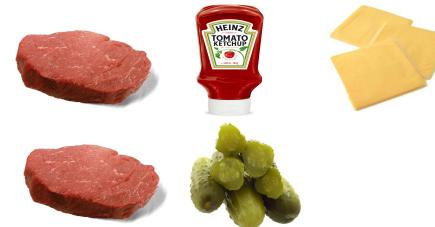
Substitution network and users' preference

Preference network

- Create an edge from ingredient **a** to **b** if $\text{rating}(a) < \text{rating}(b)$

- ex:

- Recipe X contains
- Recipe Y contains
- $\text{Rating}(X) > \text{Rating}(Y)$



Substitute network and users' preference

- Weight of preference network
 - $\text{PMI}(a \rightarrow b) = \log(p(a \rightarrow b) / p(a)p(b))$
 - where $p(a \rightarrow b) = (\# \text{ of recipe pairs from } a \text{ to } b) / (\# \text{ of recipe pairs})$
- Correlations between preference network and substitute network ($\rho = 0.72$, $p < 0.001$)

Prediction task

- Given a recipe pair with overlapped ingredients, determine which one has the higher rating



Prediction task

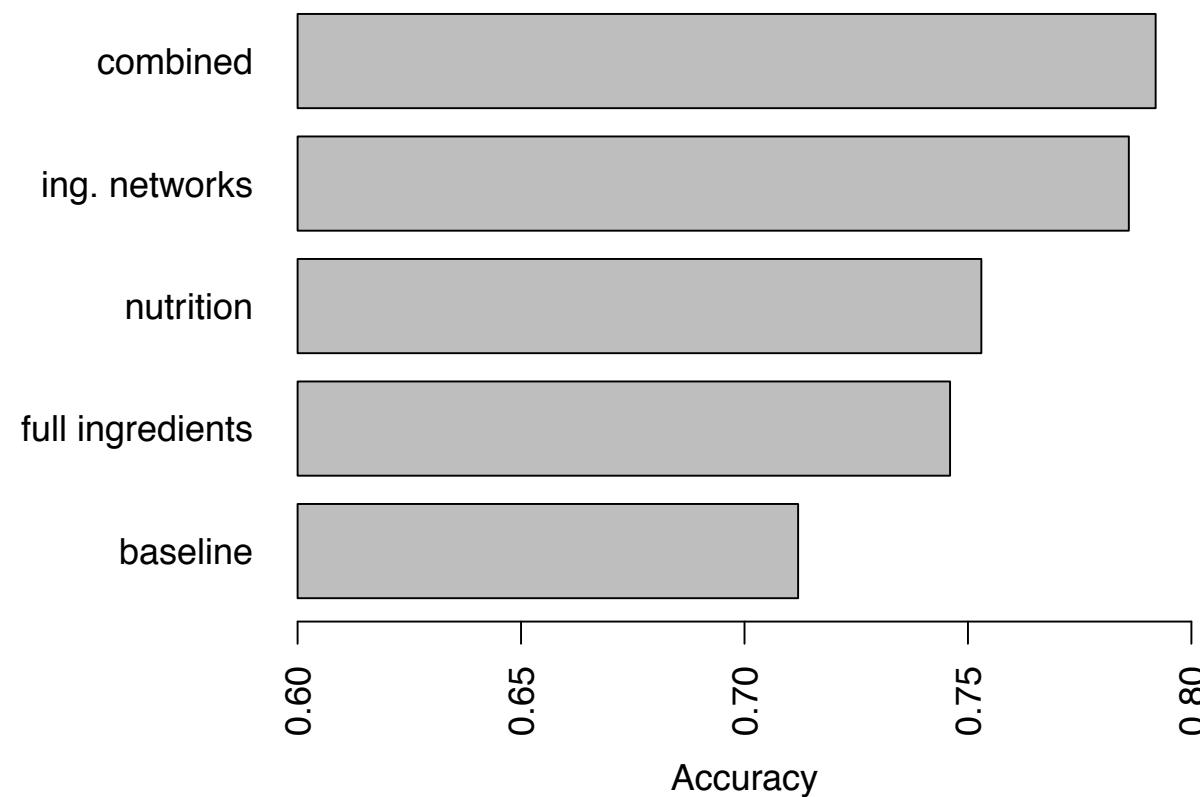
- Features
 - Baseline
 - Cooking methods, preparation time, the number of servings
 - 1000 popular ingredient list
 - Binary vector indicating the occurrence of ingredients
 - Nutrition
 - Calories, carbohydrates, fat, etc.
 - Ingredient networks
 - Network positions (centrality) and communities (SVD)
 - Combined set
 - Everything listed above

Prediction task

- 62,031 recipe pairs (X,Y)
 - where rating(X) > rating(Y)
 - ≥ 10 user reviews
 - $\geq 50\%$ users have rated both recipes
 - Cosine similarity of ingredients $(X,Y) > 0.2$
- Train with gradient boosting tree
 - balanced dataset
 - 2/3 for training, 1/3 for testing
 - Evaluate based on accuracy

Prediction performance

- Ingredient network features lead to impressive performance

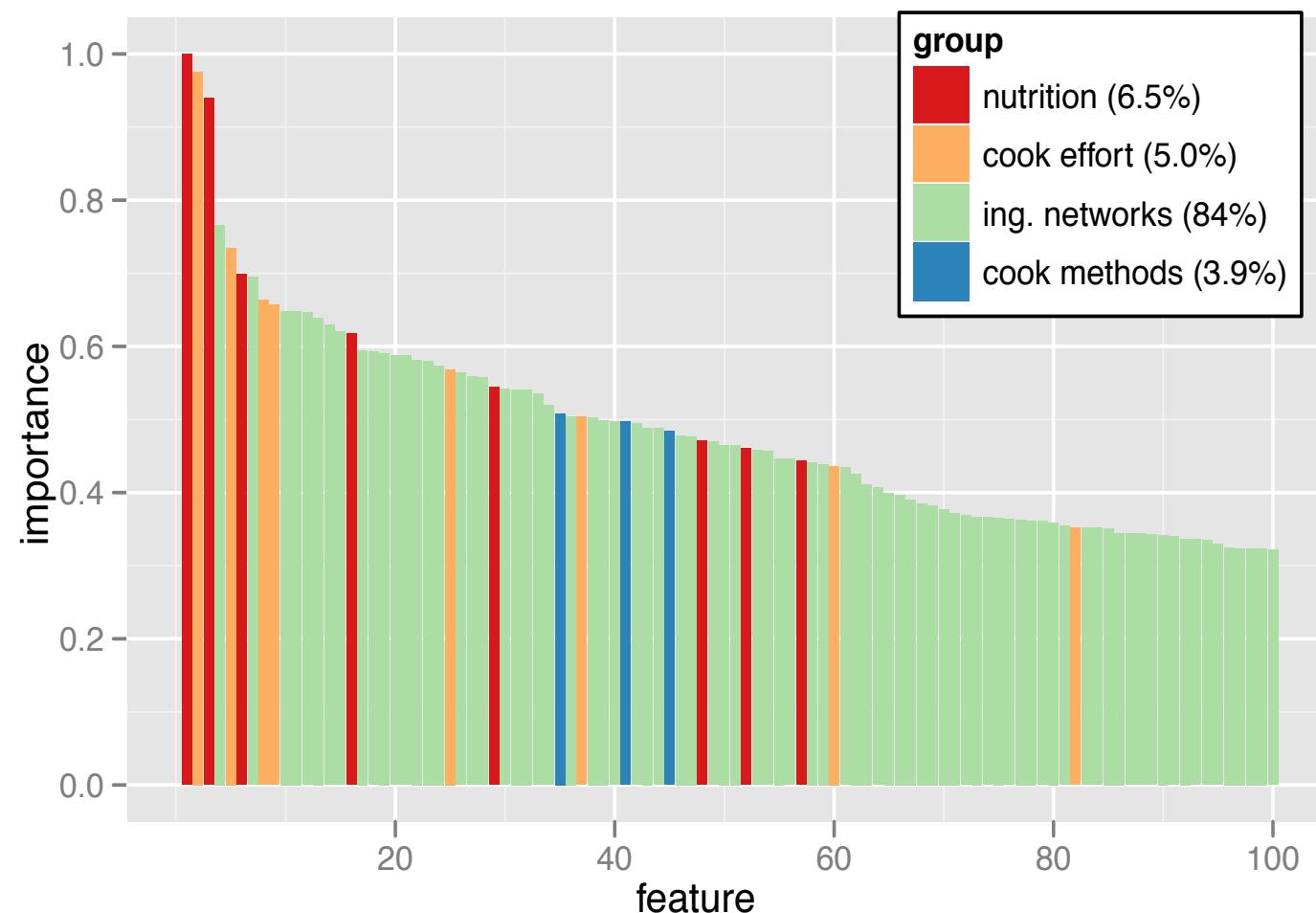


Relative importance of features

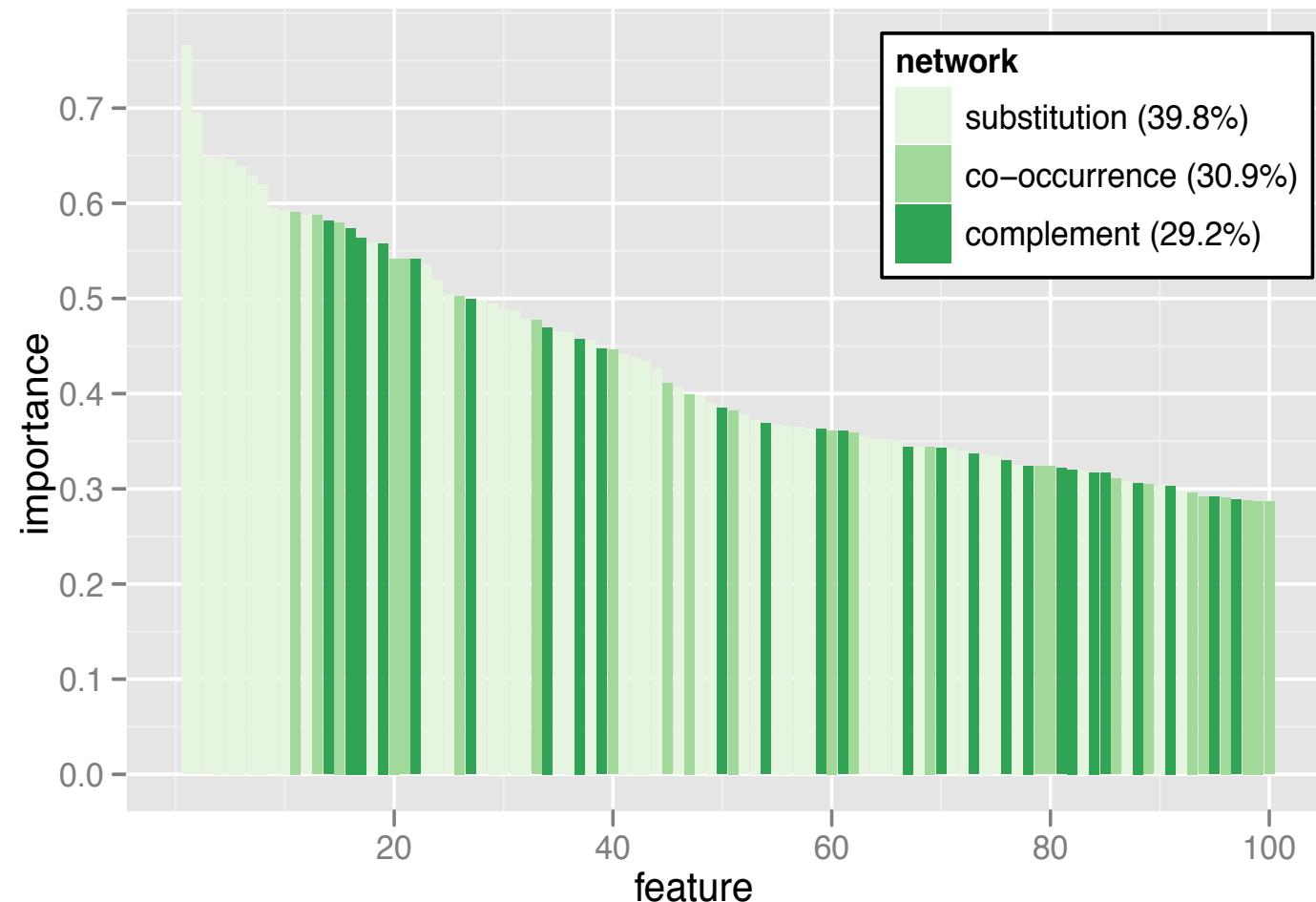
- Relative importance of feature x^j is the sum of squared improvement over all internal nodes
- i^k is the empirical improvement by the k-th node splitting on x^j

$$imp(j) = \sum i_k^2 I(\text{splits on } x^j)$$

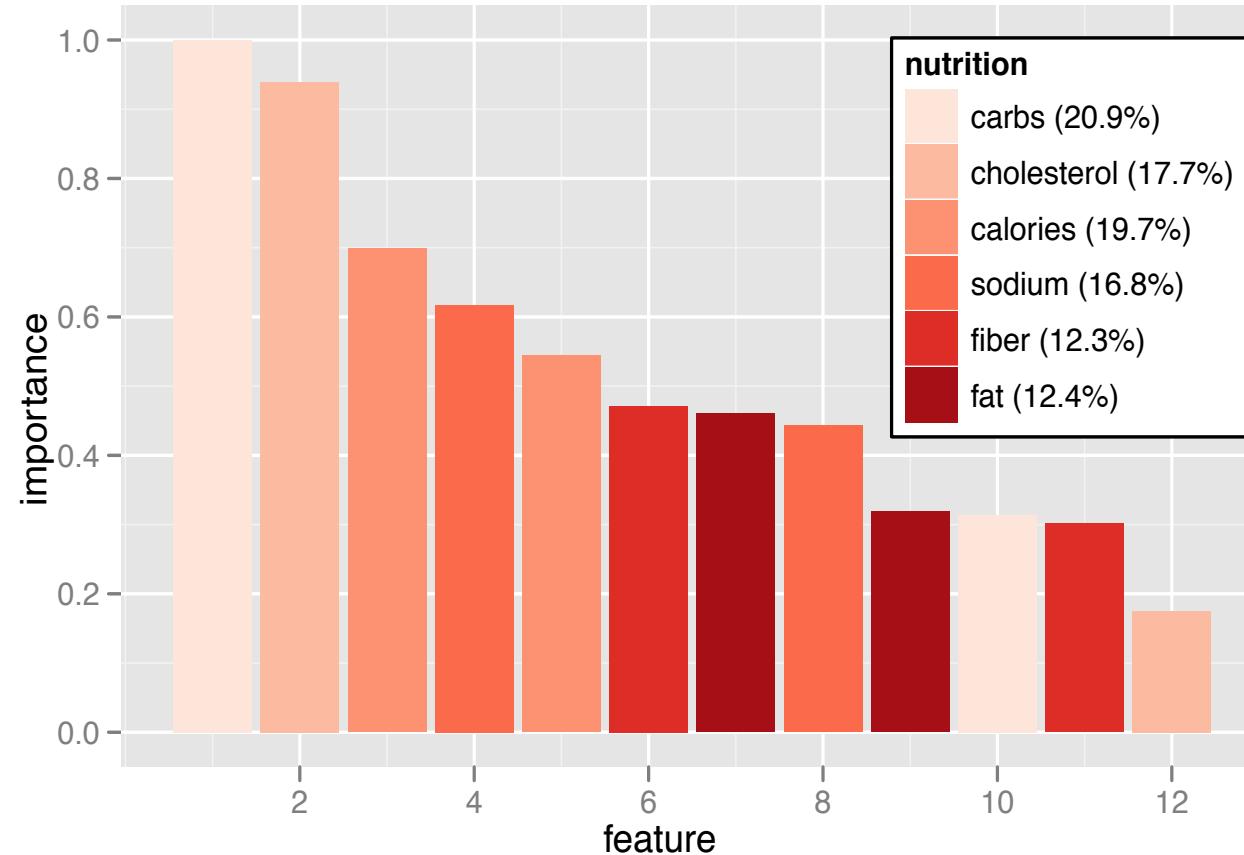
Relative importance of features



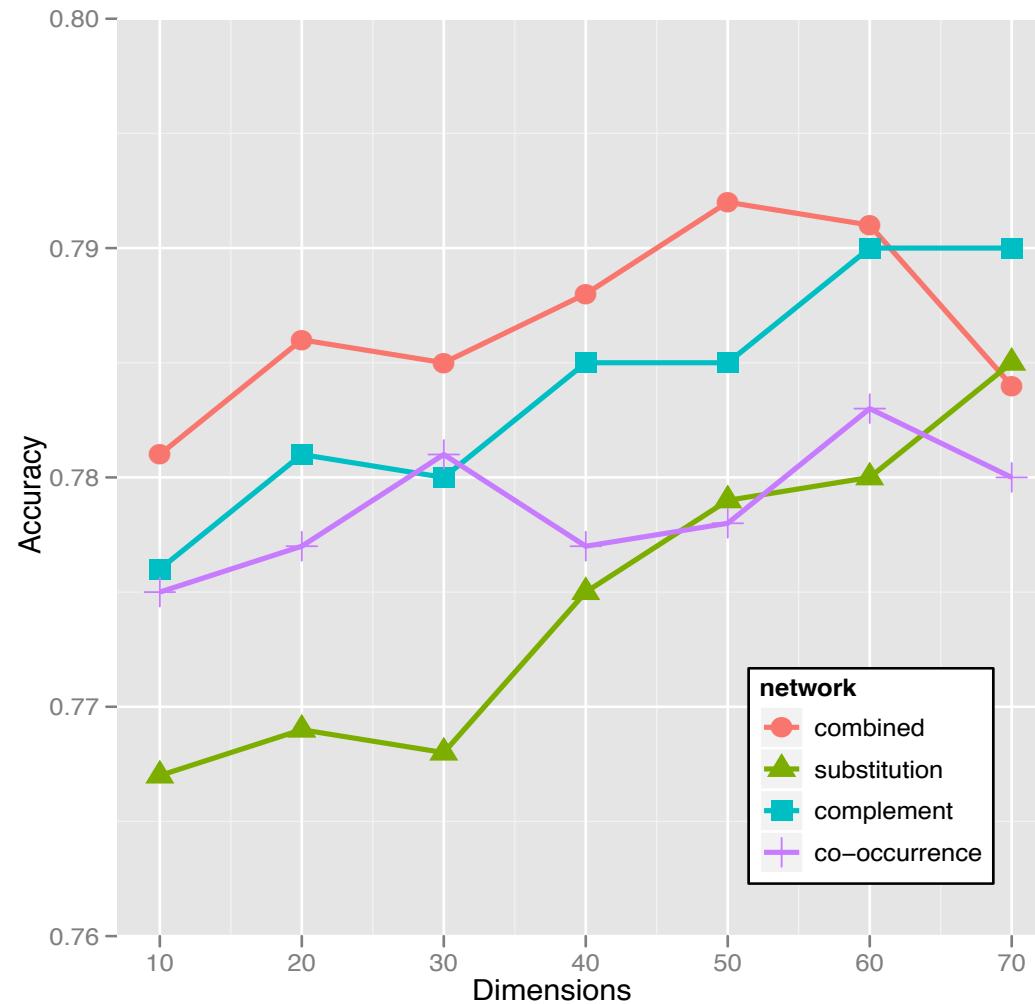
Relative importance of network features



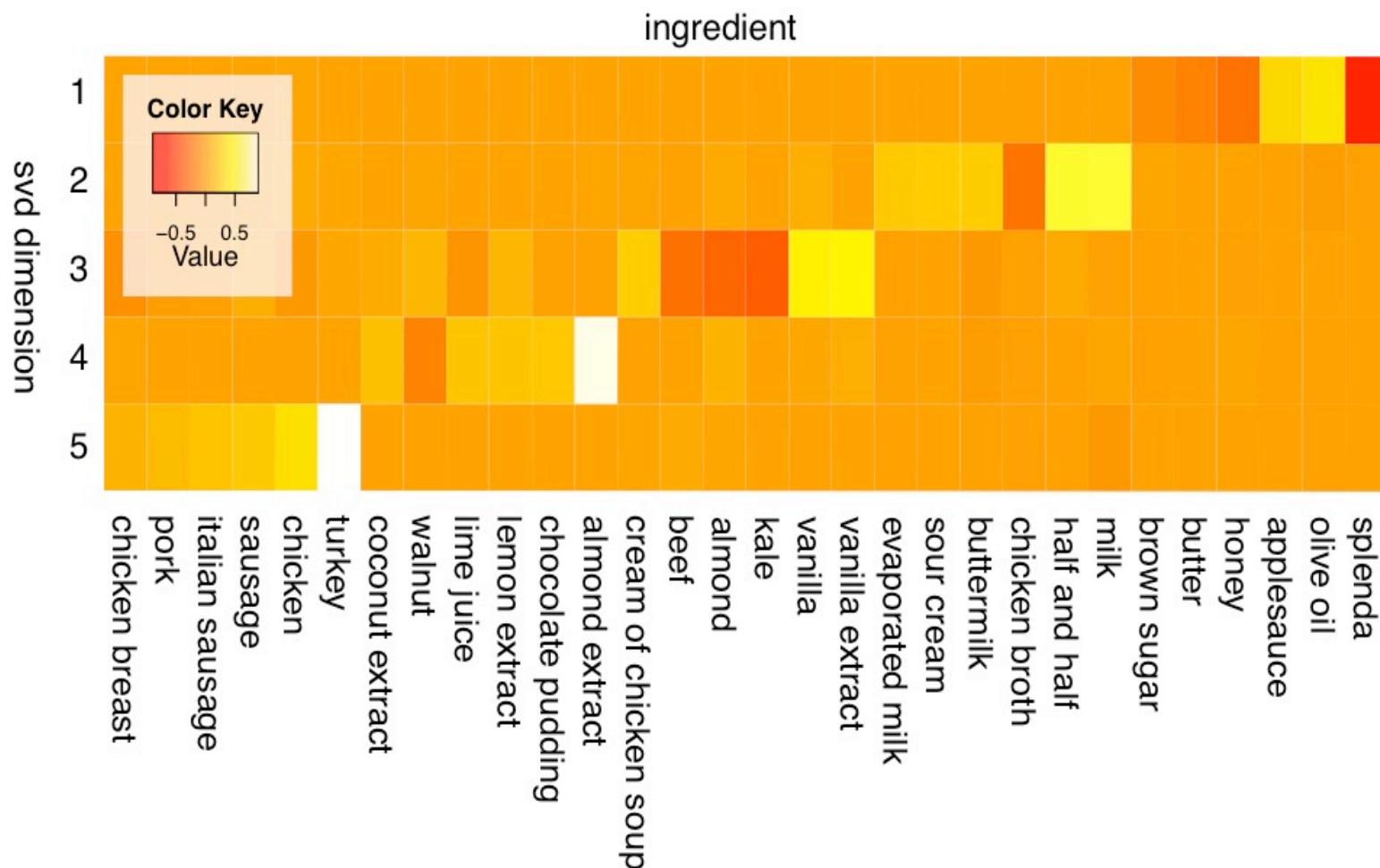
Relative importance of features from nutrition



Prediction performance vs. dimensionality



Influential substitution communities

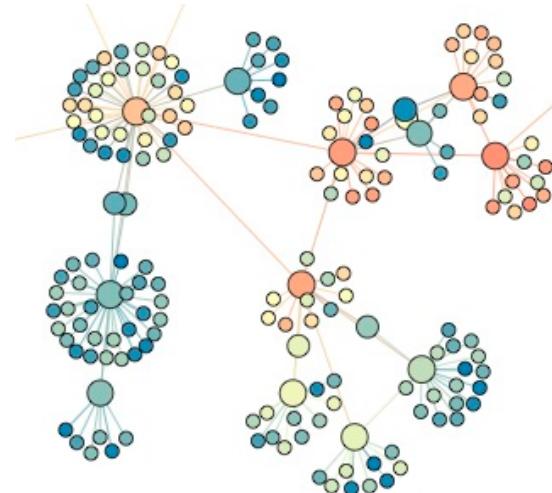


Conclusion: recipes encode our collective cooking knowledge

- regional preferences for cooking methods
- complementarity of ingredients
- substitutability of ingredients
- complement and substitute networks encode users preferences and can be used to effectively predict recipe ratings

Future work

- ❑ Extend ingredient networks by incorporating cooking methods
- ❑ Build a recommender system that incorporates users' background info (ex: region & diet)
- ❑ More info
 - ❑ <http://netsi.org>



SNA 8: network resilience

Lada Adamic

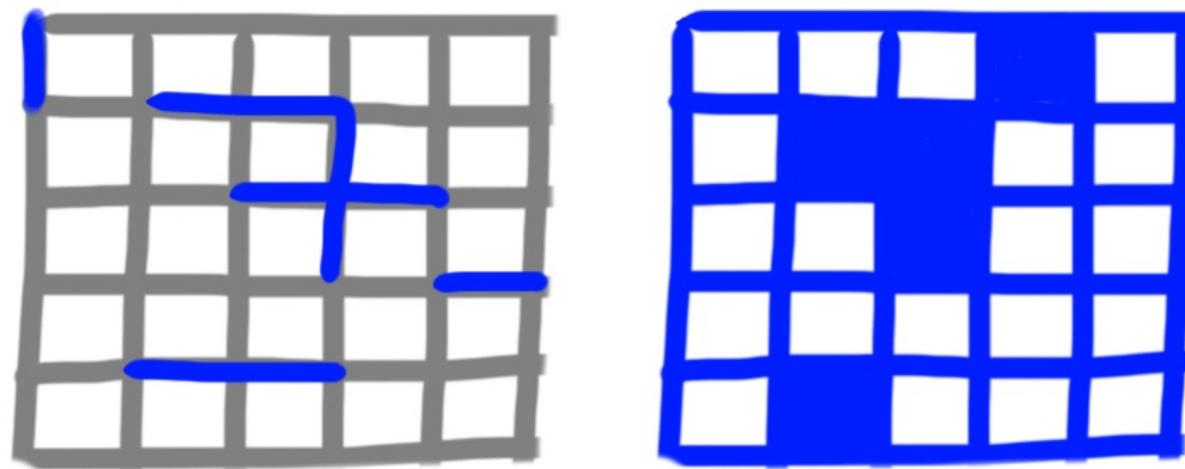


Outline

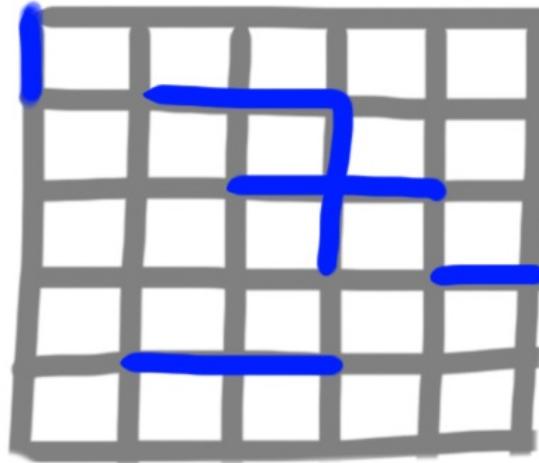
- ❑ Node vs. edge percolation
- ❑ Resilience of randomly vs. preferentially grown networks
- ❑ Resilience in real-world networks

network resilience

- Q: If a given fraction of nodes or edges are removed...
 - how large are the connected components?
 - what is the average distance between nodes in the components
- Related to percolation (previously studied on lattices):



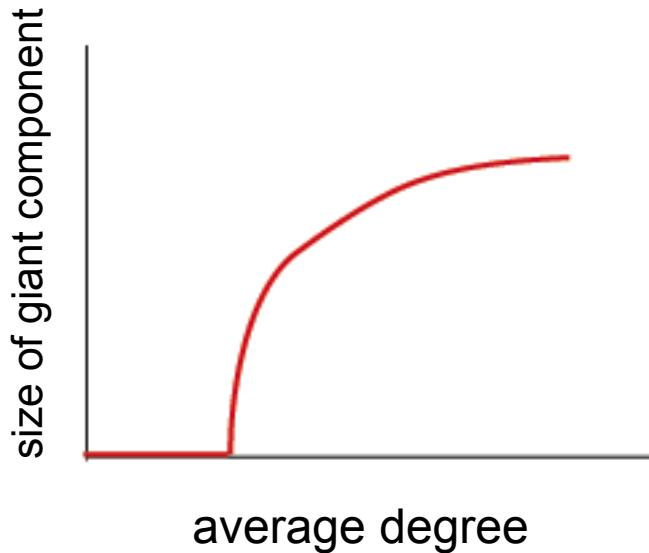
edge percolation



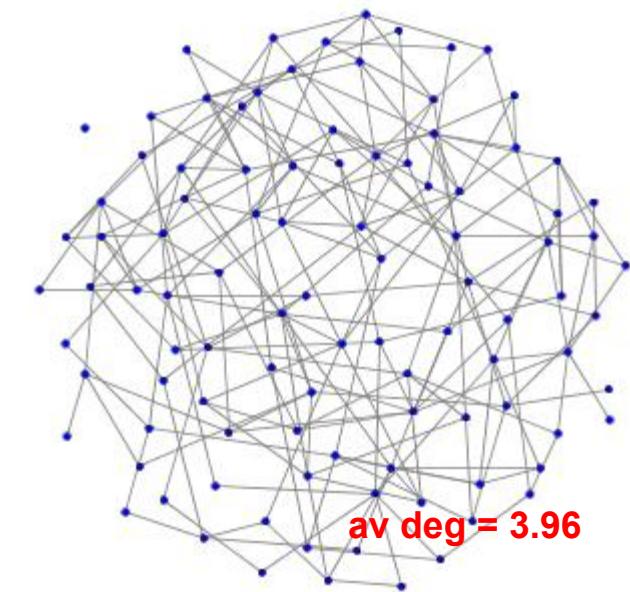
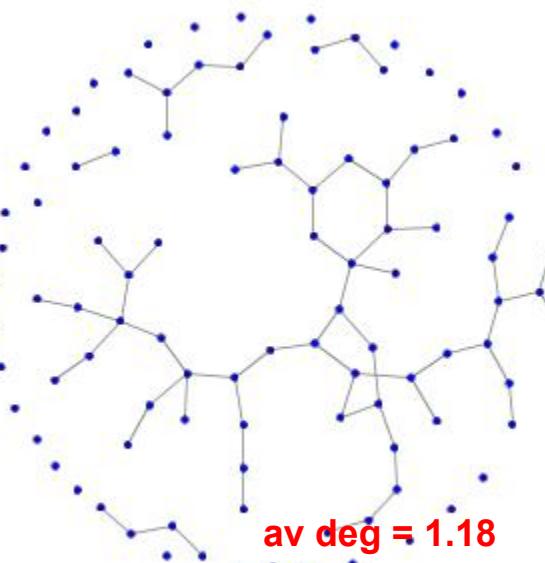
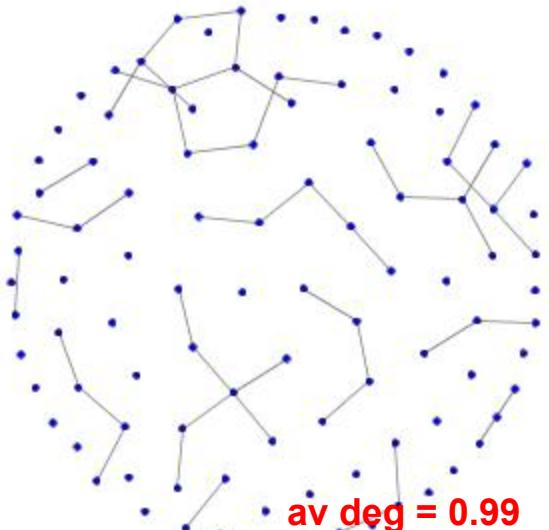
□ Edge removal

- bond percolation: each edge is removed with probability $(1-p)$
 - corresponds to random failure of links
- targeted attack: causing the most damage to the network with the removal of the fewest edges
 - strategies: remove edges that are most likely to break apart the network or lengthen the average shortest path
 - e.g. usually edges with high betweenness

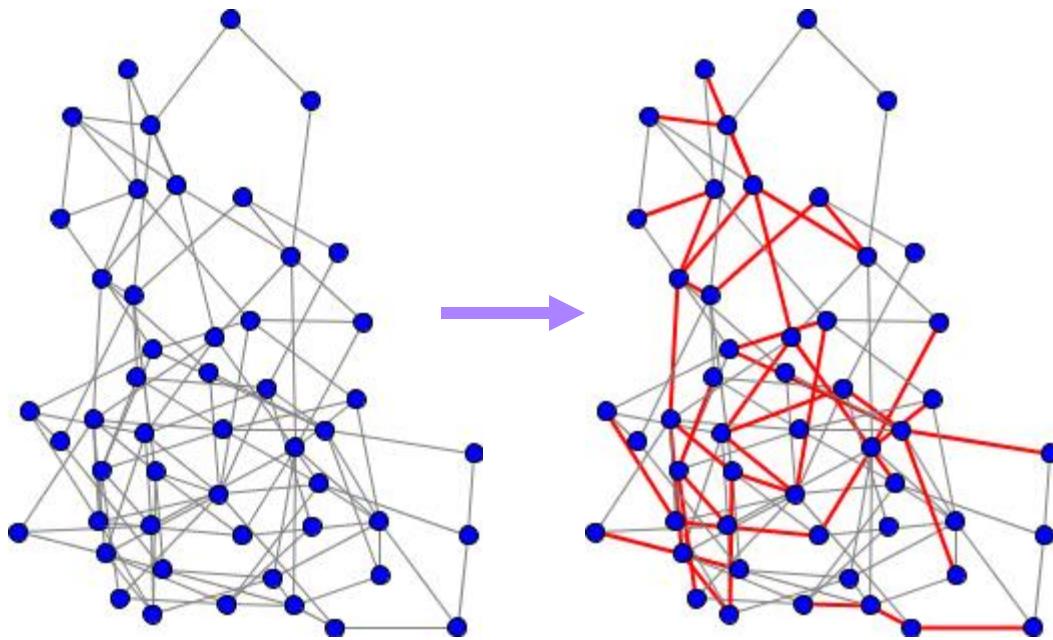
reminder: percolation in ER graphs



- As the average degree increases to $z = 1$, a giant component suddenly appears
- Edge removal is the opposite process – at some point the average degree drops below 1 and the network becomes disconnected

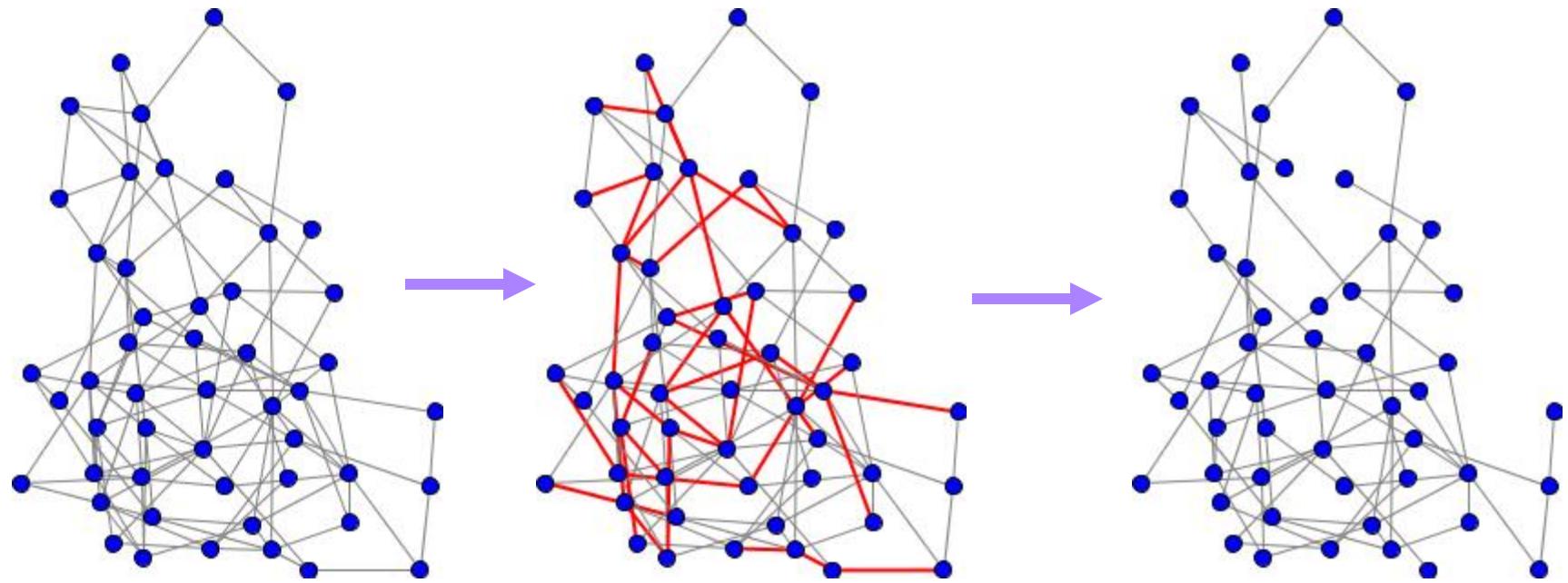


Quiz Q:



In this network each node has average degree 4.64, if you removed 25% of the edges, by how much would you reduce the giant component?

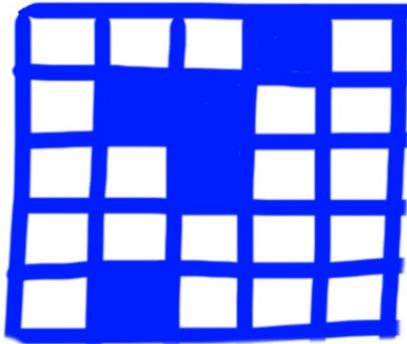
edge percolation



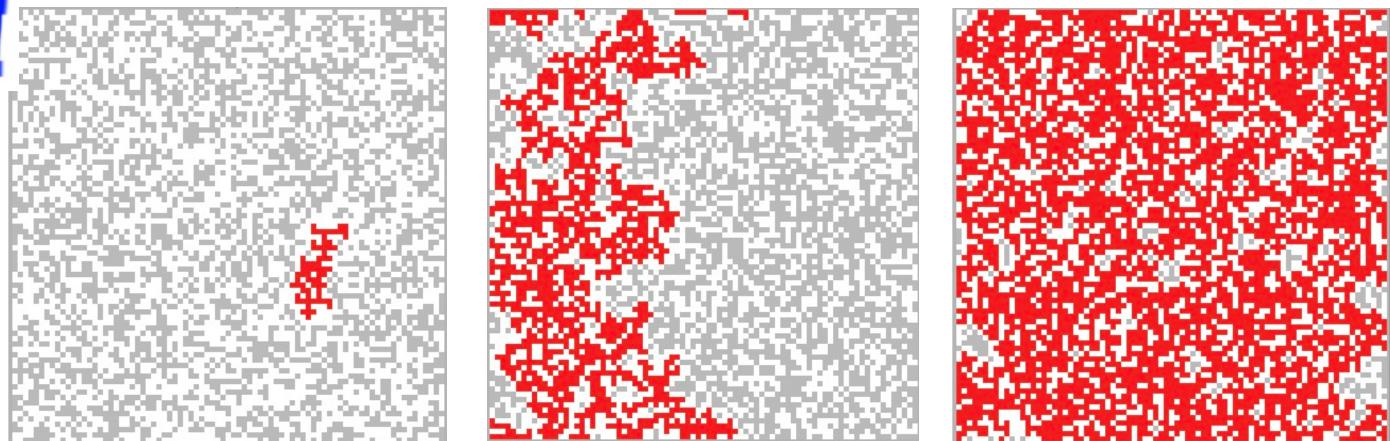
50 nodes, 116 edges, average degree 4.64
after 25 % edge removal

76 edges, average degree 3.04 – still well above
percolation threshold

node removal and site percolation



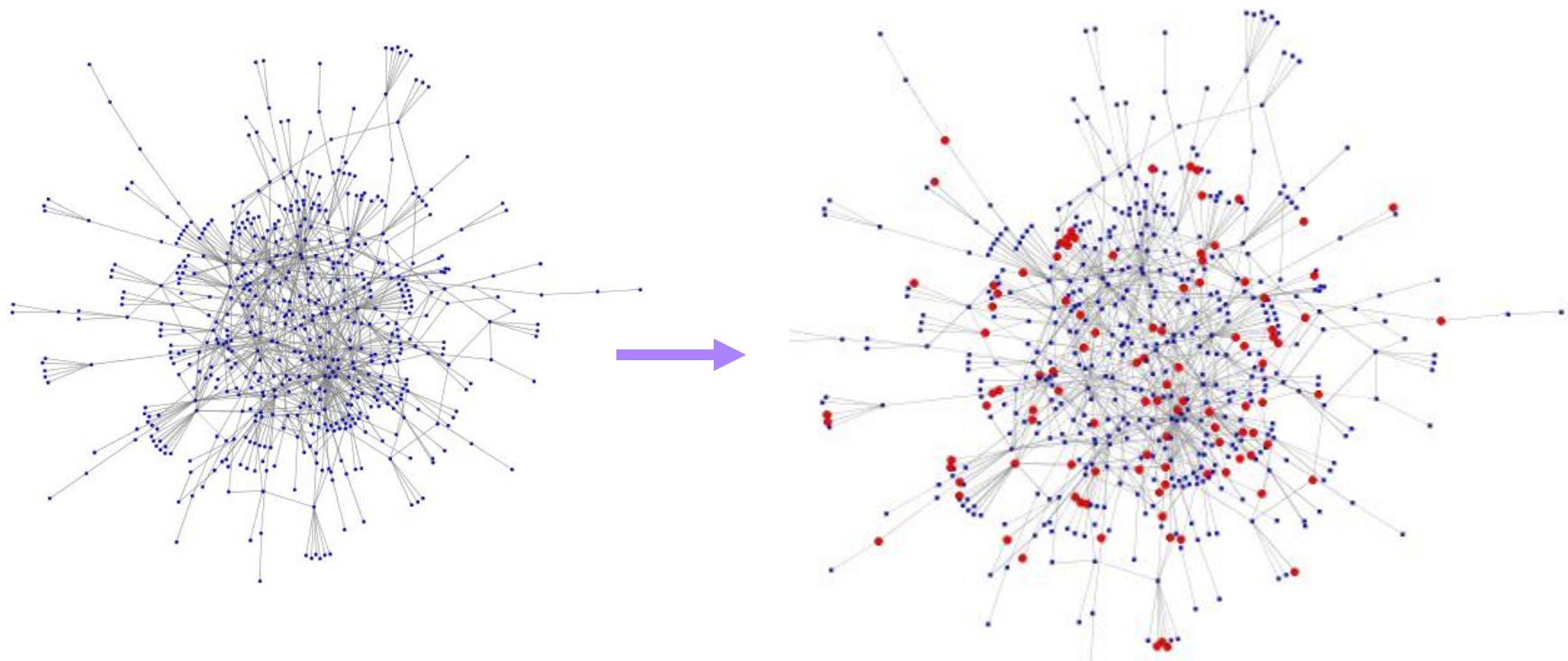
Ordinary Site Percolation on Lattices:
Fill in each site (site percolation) with probability p



- **low p :** small islands
- **p critical:** giant component forms, occupying finite fraction of infinite lattice.
- **p above critical value:** giant component occupies an increasingly larger portion of the graph

<http://www.ladamic.com/netlearn/NetLogo501/LatticePercolation.html>

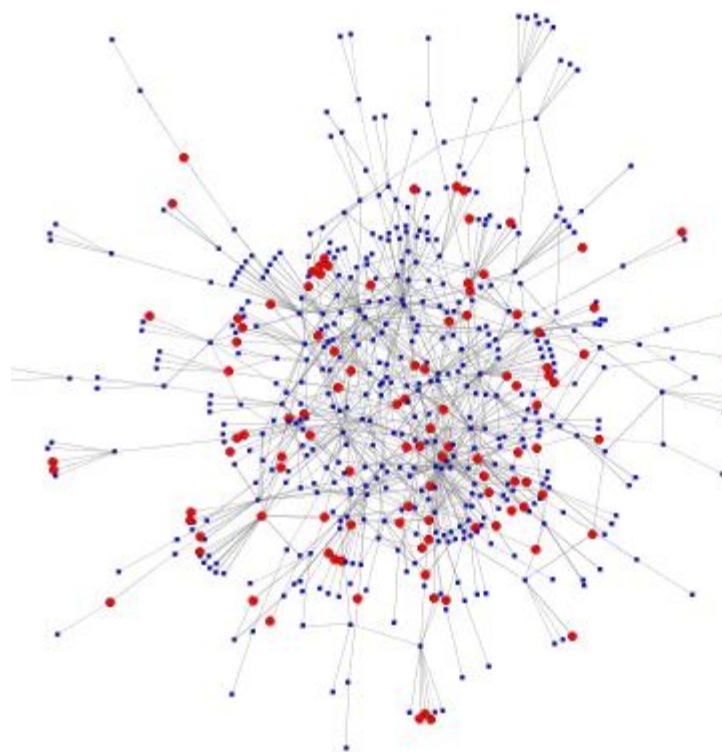
Percolation on networks



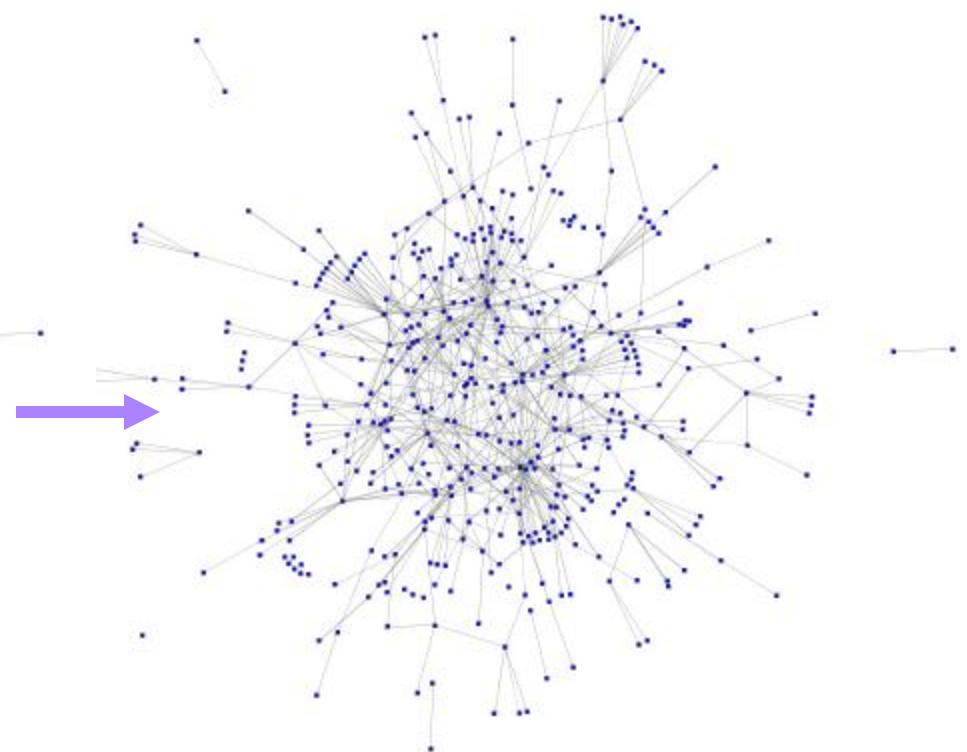
- Percolation can be extended to networks of arbitrary topology.
- We say the network percolates when a giant component forms.

Random attack on scale-free networks

- Example: gnutella filesharing network, 20% of nodes removed at random



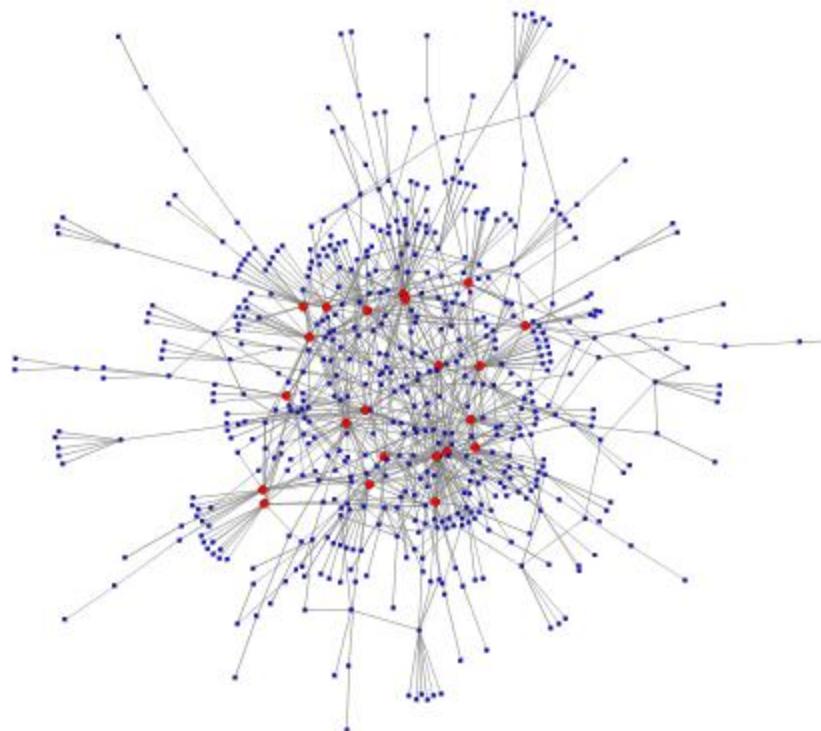
574 nodes in giant component



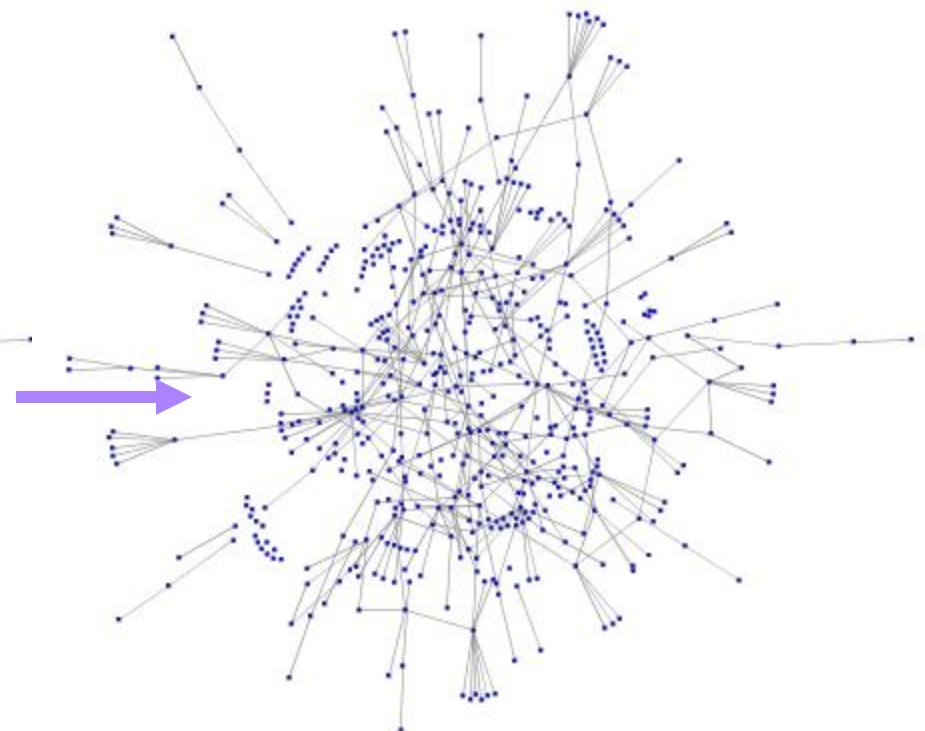
427 nodes in giant component

Targeted attacks on power-law networks

- Power-law networks are vulnerable to targeted attack
- Example: same gnutella network, 22 most connected nodes removed (2.8% of the nodes)



574 nodes in giant component

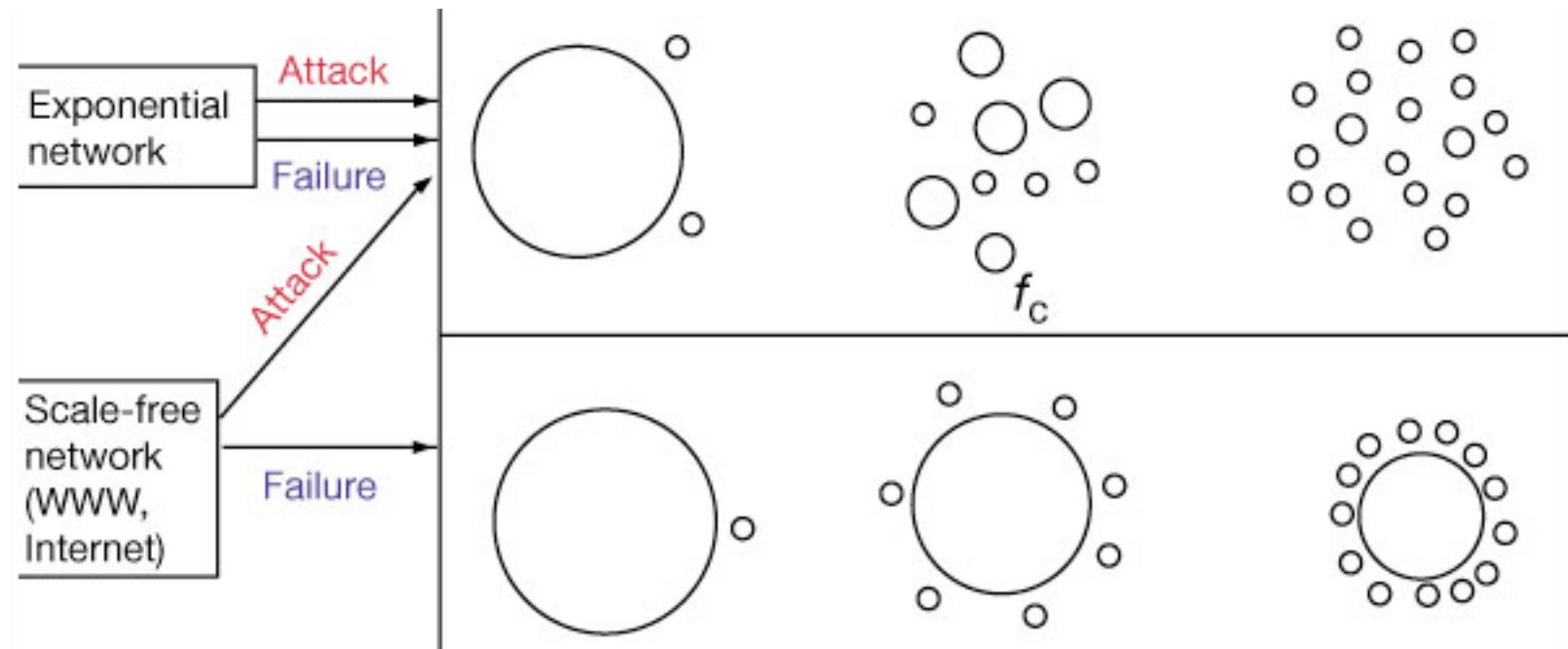


301 nodes in giant component

Quiz Q:

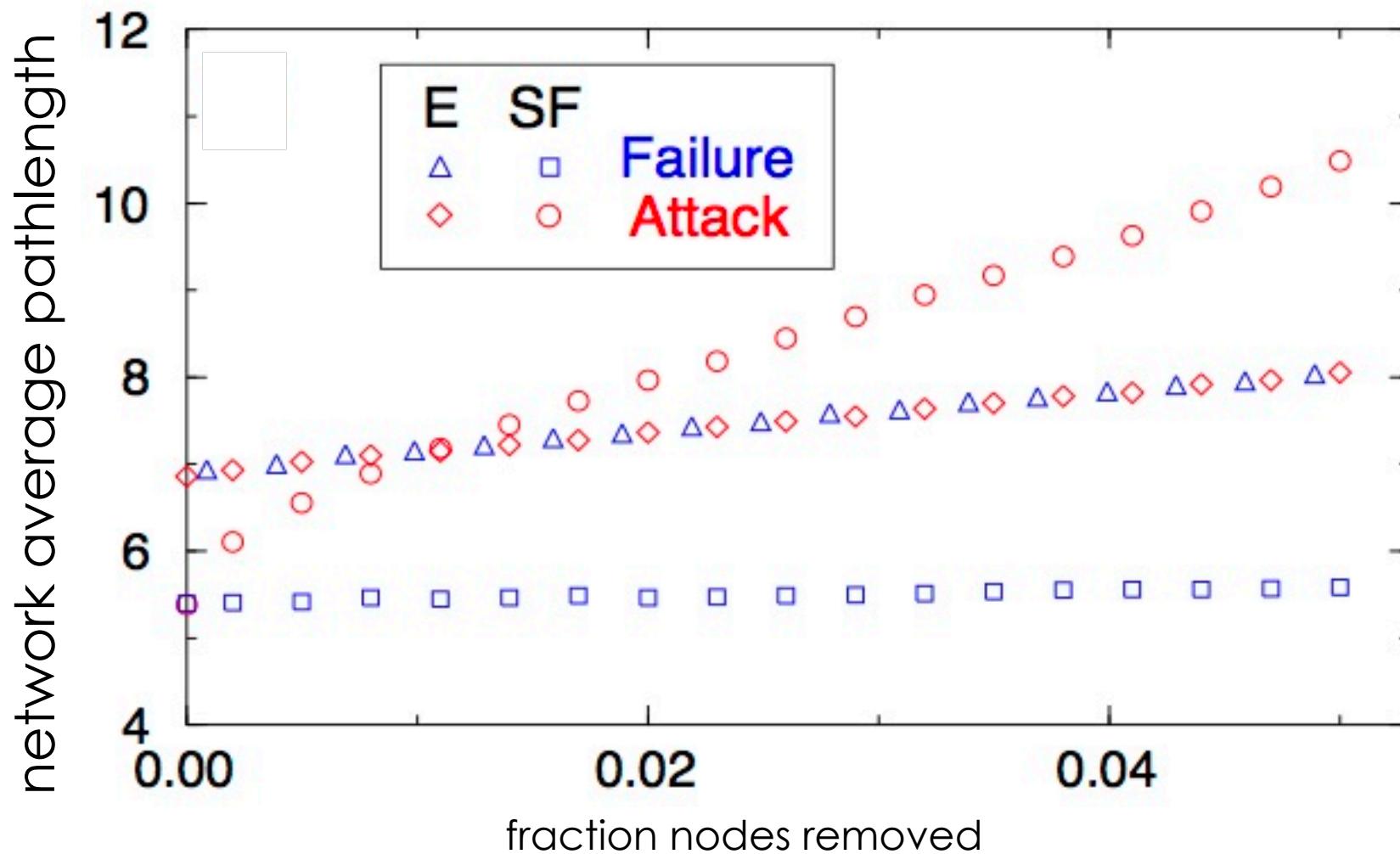
- ❑ Why is removing high-degree nodes more effective?
 - ❑ it removes more nodes
 - ❑ it removes more edges
 - ❑ it targets the periphery of the network

random failures vs. attacks



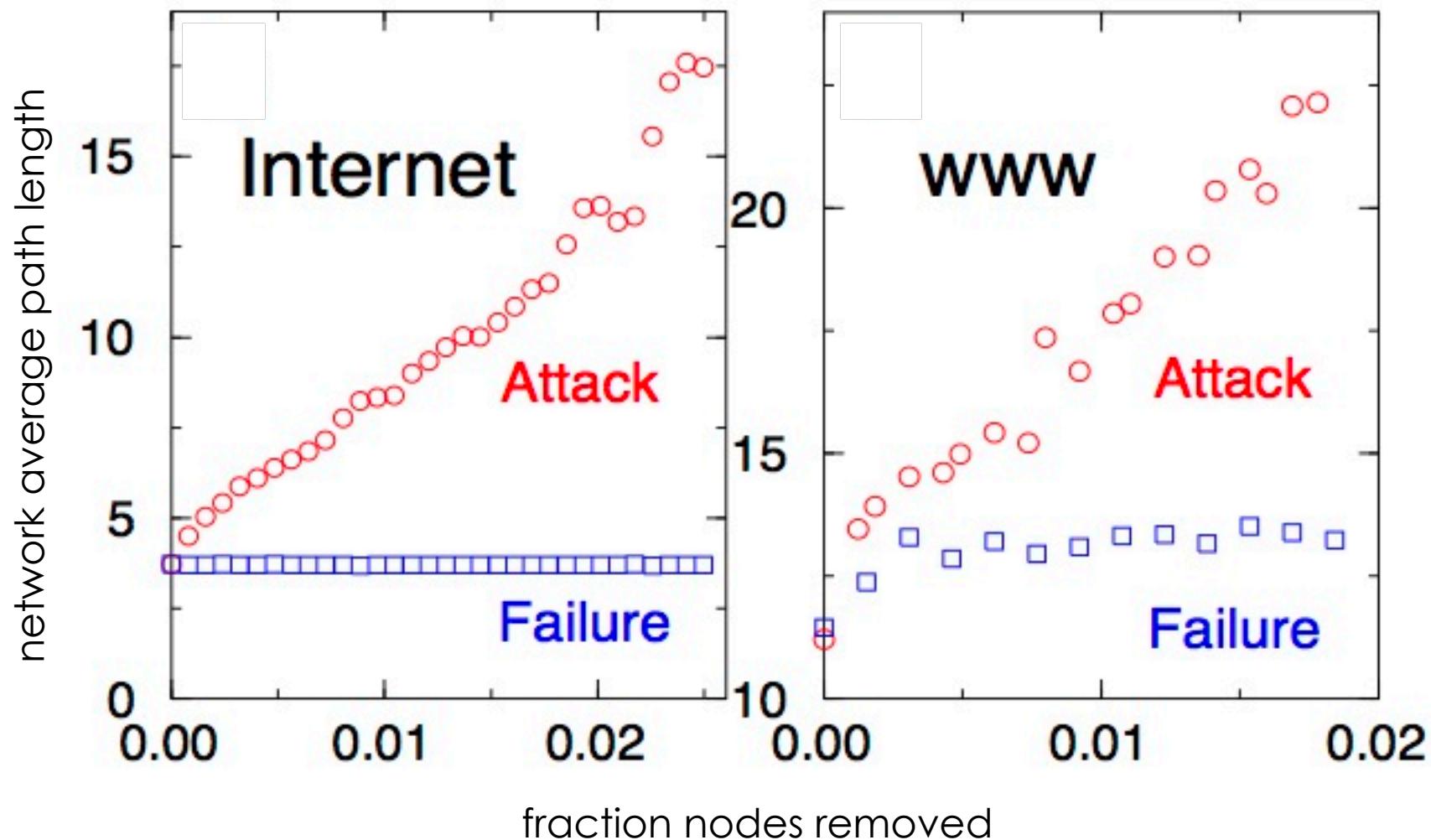
Source: Error and attack tolerance of complex networks. Réka Albert, Hawoong Jeong and Albert-László Barabási. Nature 406, 378-382(27 July 2000); <http://www.nature.com/nature/journal/v406/n6794/abs/406378A0.html>

effect on path length



Source: Error and attack tolerance of complex networks. Réka Albert, Hawoong Jeong and Albert-László Barabási. Nature 406, 378-382(27 July 2000); <http://www.nature.com/nature/journal/v406/n6794/abs/406378A0.html>

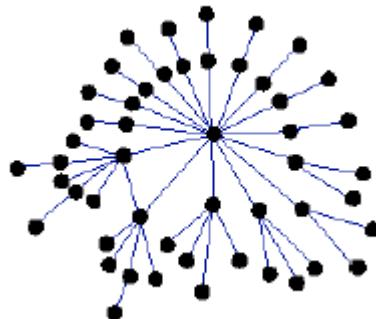
applied to empirical networks



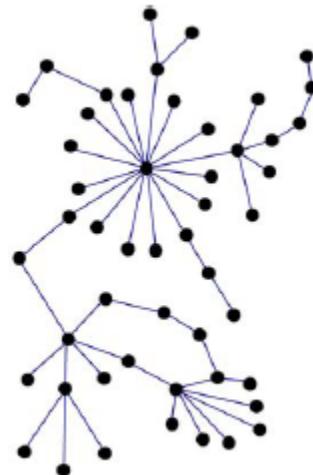
Source: Error and attack tolerance of complex networks. Réka Albert, Hawoong Jeong and Albert-László Barabási. Nature 406, 378-382(27 July 2000); <http://www.nature.com/nature/journal/v406/n6794/abs/406378A0.html>

Assortativity

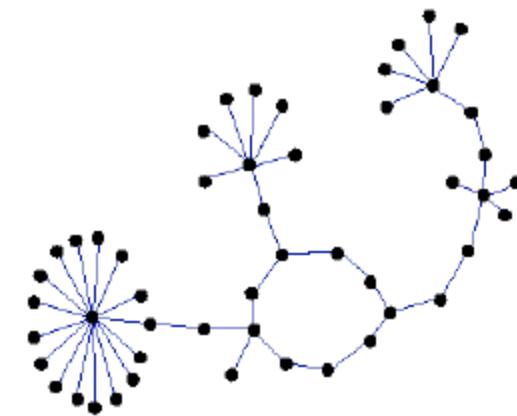
- Social networks are assortative:
 - the gregarious people associate with other gregarious people
 - the loners associate with other loners
- The Internet is disassortative:



Assortative:
hubs connect to hubs



Random

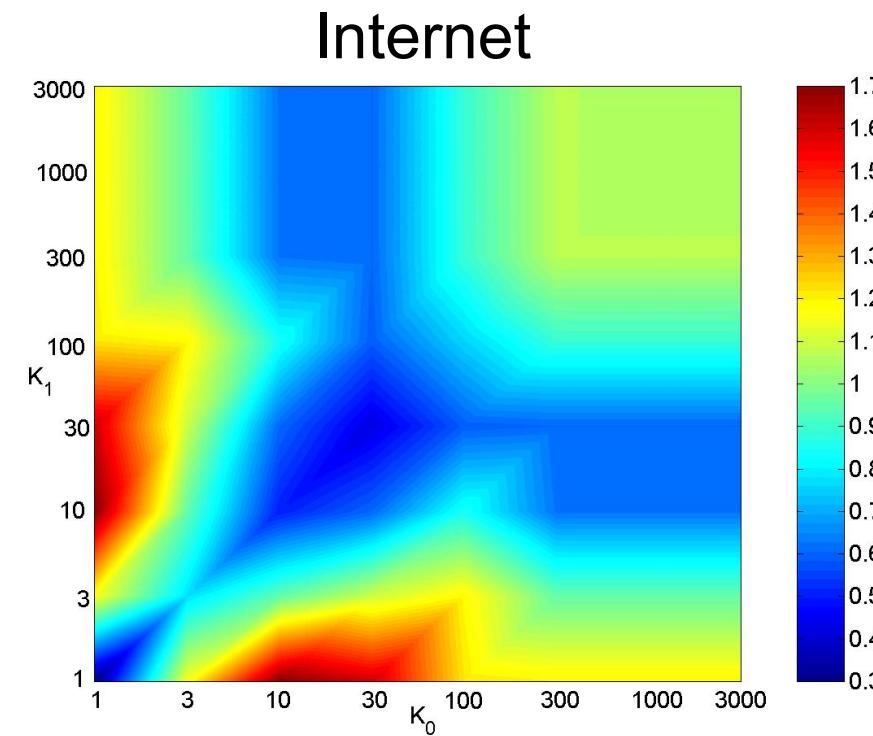


Disassortative:
hubs are in the
periphery

Correlation profile of a network

- ❑ Detects preferences in linking of nodes to each other based on their connectivity
- ❑ Measure $N(k_0, k_1)$ – the number of edges between nodes with connectivities k_0 and k_1
- ❑ Compare it to $N_r(k_0, k_1)$ – the same property in a properly randomized network
- ❑ Very noise-tolerant with respect to both false positives and negatives

Degree correlation profiles: 2D

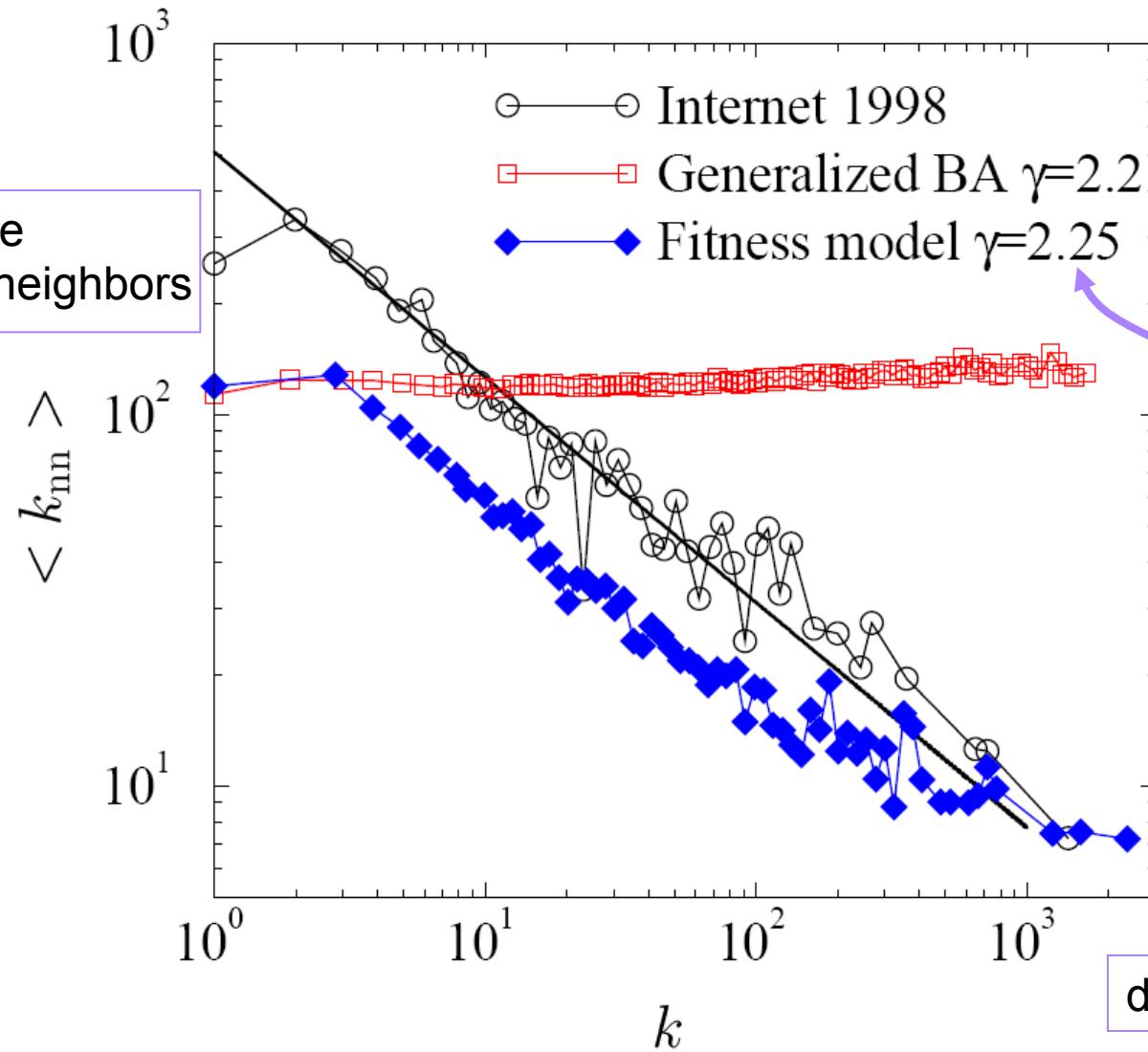


source: Sergei Maslov

Average degree of neighbors

□ Pastor-Satorras and Vespignani: 2D plot

average degree
of the node's neighbors



degree of node

probability
of aquiring
edges is
dependent
on 'fitness'
+ degree
Bianconi &
Barabasi

Single number

- $\text{cor}(\text{deg}(i), \text{deg}(j))$ over all edges $\{ij\}$

$$\rho_{\text{internet}} = -0.189$$

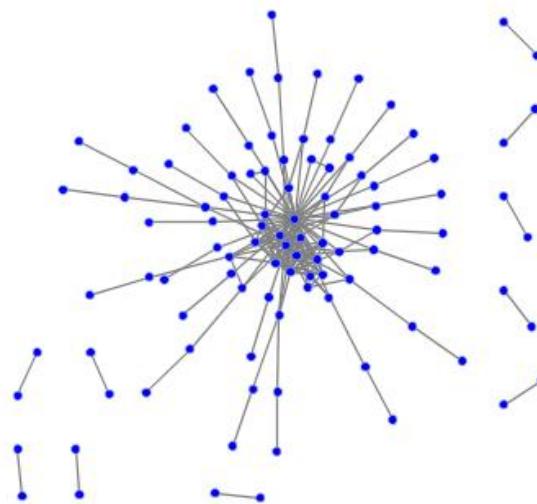
The Pearson correlation coefficient of nodes on each side on an edge

assortative mixing more generally

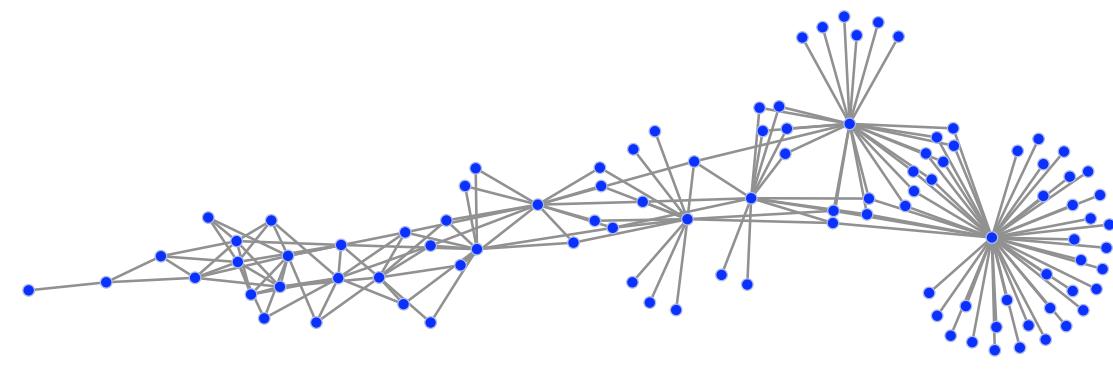
- ❑ Assortativity is not limited to degree-degree correlations other attributes
 - ❑ social networks: race, income, gender, age
 - ❑ food webs: herbivores, carnivores
 - ❑ internet: high level connectivity providers, ISPs, consumers
- ❑ Tendency of like individuals to associate = ‘homophily’

Quiz Q:

will a network with positive or negative degree assortativity be more resilient to attack?



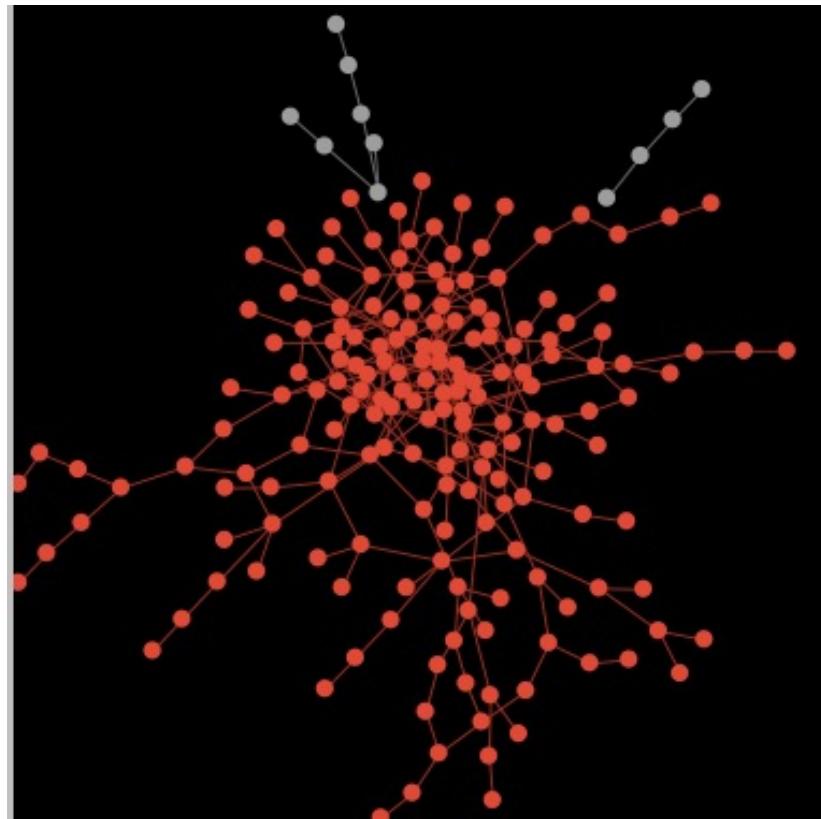
assortative



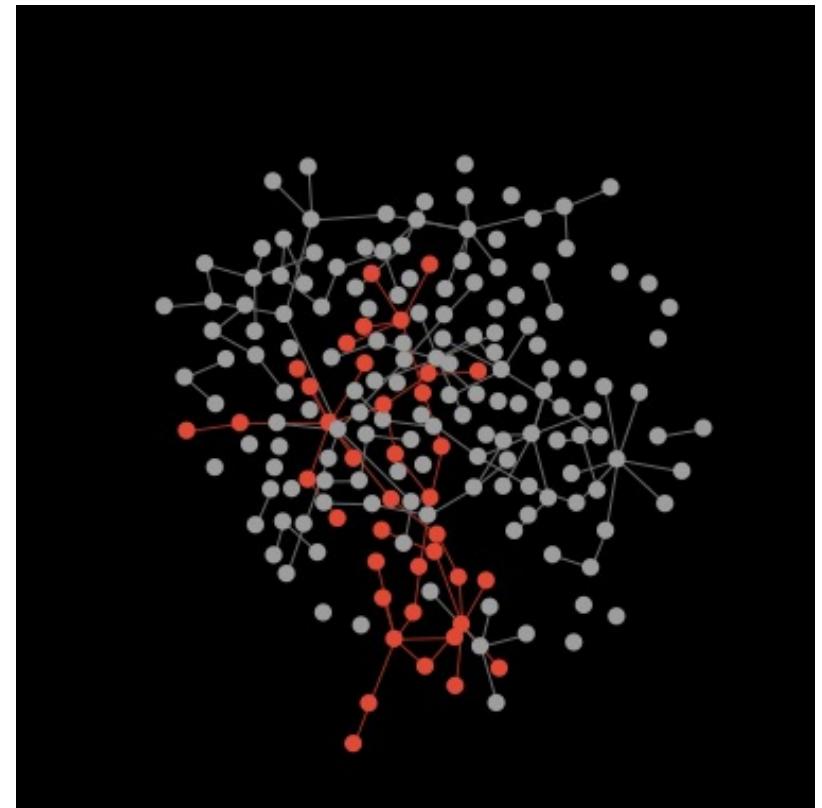
disassortative

Assortativity and resilience

assortative



disassortative

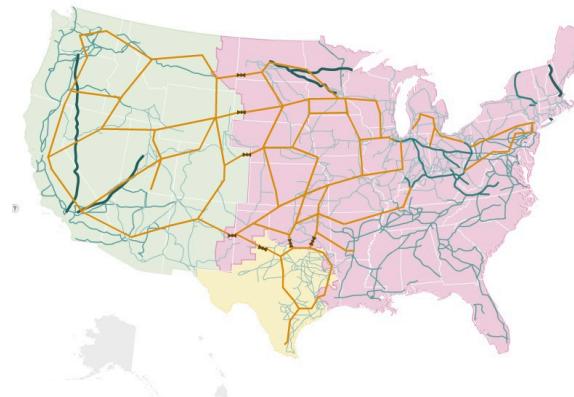


Is it really that simple?

- ❑ Internet?
- ❑ terrorist/criminal networks?

Power grid

- Electric power flows simultaneously through multiple paths in the network.
- For visualization of the power grid, check out NPR's interactive visualization:
[http://www.npr.org/templates/story/story.php?
storyId=110997398](http://www.npr.org/templates/story/story.php?storyId=110997398)



Cascading failures

- ❑ Each node has a **load** and a **capacity** that says how much load it can tolerate.
- ❑ When a node is removed from the network its load is redistributed to the remaining nodes.
- ❑ If the load of a node exceeds its capacity, then the node fails

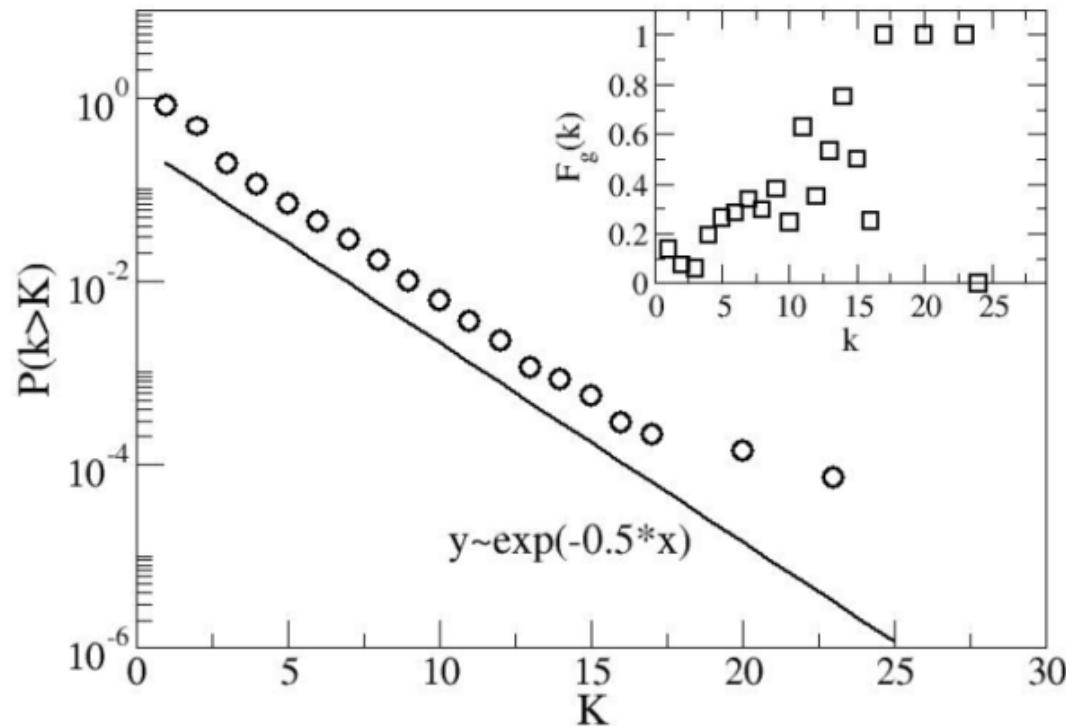
Case study: US power grid

Modeling cascading failures in the North American power grid

R. Kinney, P. Crucitti, R. Albert, and V. Latora, Eur. Phys. B, 2005

- ❑ Nodes: generators, transmission substations, distribution substations
- ❑ Edges: high-voltage transmission lines
- ❑ 14099 substations:
 - ❑ N_G 1633 generators,
 - ❑ N_D 2179 distribution substations
 - ❑ N_T the rest transmission substations
- ❑ 19,657 edges

Degree distribution is exponential



$$P(k > K) \approx \exp(-0.5K)$$

Efficiency of a path

- efficiency $e \in [0, 1]$, 0 if no electricity flows between two endpoints, 1 if the transmission lines are working perfectly
- harmonic composition for a path

$$e_{path} = \left[\sum_{edges} \frac{1}{e_{edge}} \right]^{-1}$$

- path A, 2 edges, each with $e=0.5$, $e_{path} = 1/4$
- path B, 3 edges, each with $e=0.5$ $e_{path} = 1/6$
- path C, 2 edges, one with $e=0$ the other with $e=1$, $e_{path} = 0$
- simplifying assumption: electricity flows along most efficient path

Efficiency of the network

- ☐ Efficiency of the network:
 - ☐ average over the most efficient paths from each generator to each distribution station

$$E = \frac{1}{N_G N_D} \sum_{i \in G_G} \sum_{j \in G_D} \epsilon_{ij}$$

ϵ_{ij} is the efficiency of the most efficient path between i and j

capacity and node failure

- Assume capacity of each node is proportional to initial load

$$C_i = \alpha L_i(0) \quad i = 1, 2..N$$

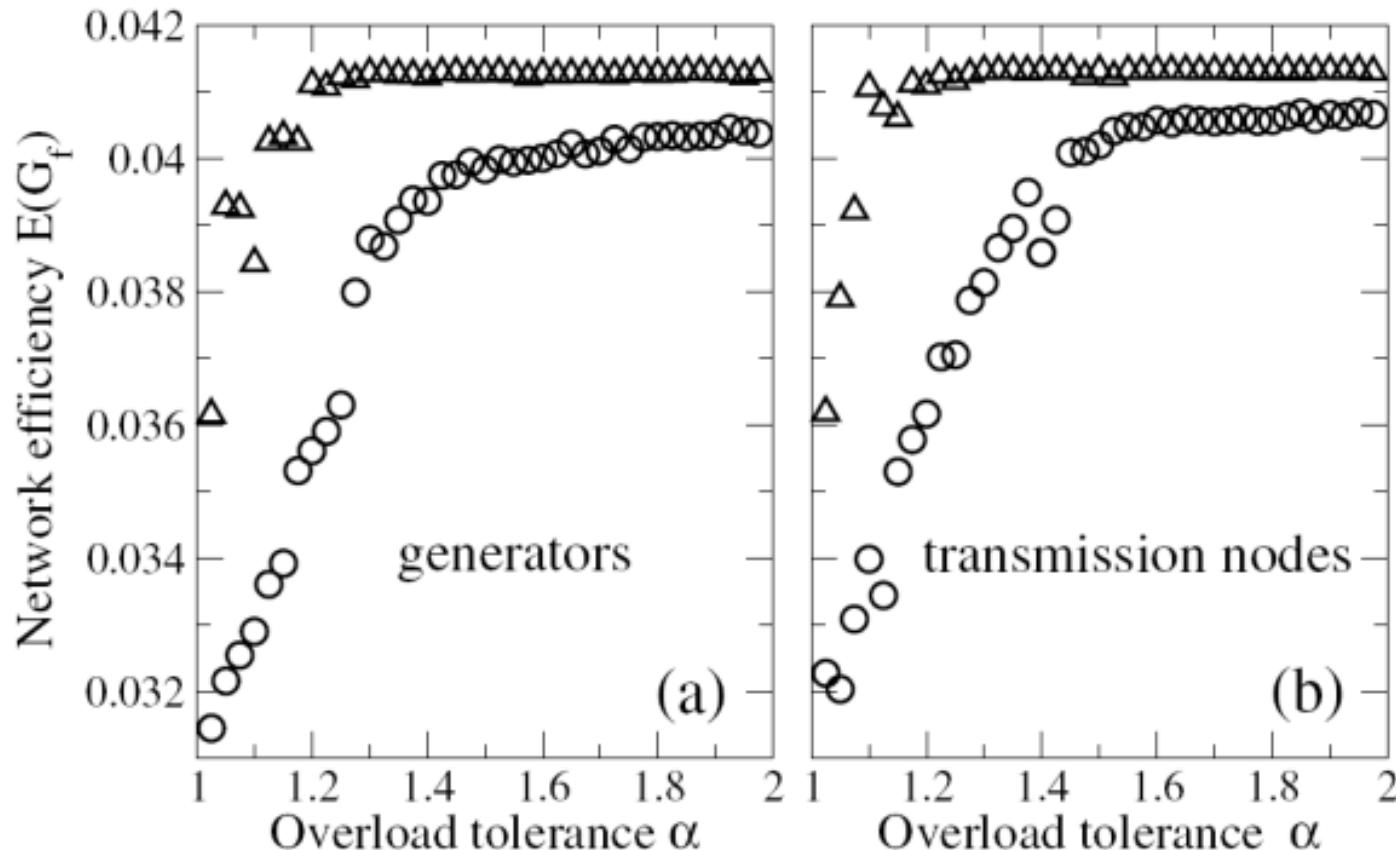
- L represents the weighted betweenness of a node
- Each neighbor of a node is impacted as follows

$$e_{ij}(t+1) = \begin{cases} e_{ij}(0)/\frac{L_i(t)}{C_i} & \text{if } L_i(t) > C_i \\ e_{ij}(0) & \text{if } L_i(t) \leq C_i \end{cases} \quad \text{load exceeds capacity}$$

- Load is distributed to other nodes/edges
- The greater a (reserve capacity), the less susceptible the network to cascading failures due to node failure

power grid structural resilience

- efficiency is impacted the most if the node removed is the one with the highest load



○ highest load generator/transmission station removed

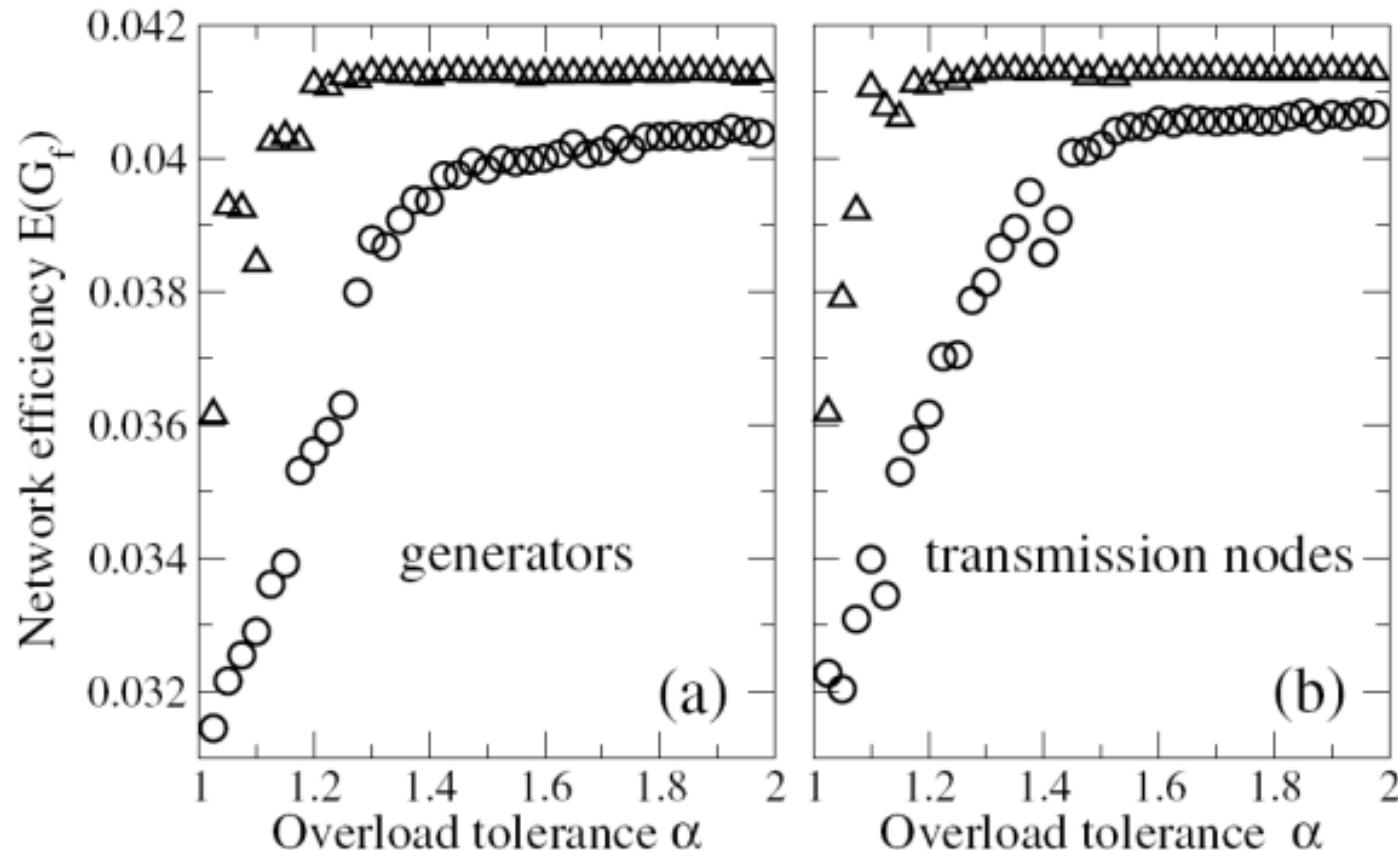
Source: Modeling cascading failures in the North American power grid; R. Kinney, P. Crucitti, R. Albert, V. Latora, Eur. Phys. B, 2005

Quiz Q:

- ❑ Approx. how much higher would the capacity of a node need to be relative to the initial load in order for the network to be efficient? (remember capacity $C = \alpha * L(0)$, the initial load).

power grid structural resilience

- efficiency is impacted the most if the node removed is the one with the highest load



○ highest load generator/transmission station removed

Source: Modeling cascading failures in the North American power grid; R. Kinney, P. Crucitti, R. Albert, V. Latora, Eur. Phys. B, 2005

recap: network resilience

- ❑ resilience depends on topology
- ❑ also depends on what happens when a node fails
 - ❑ e.g. in power grid load is redistributed