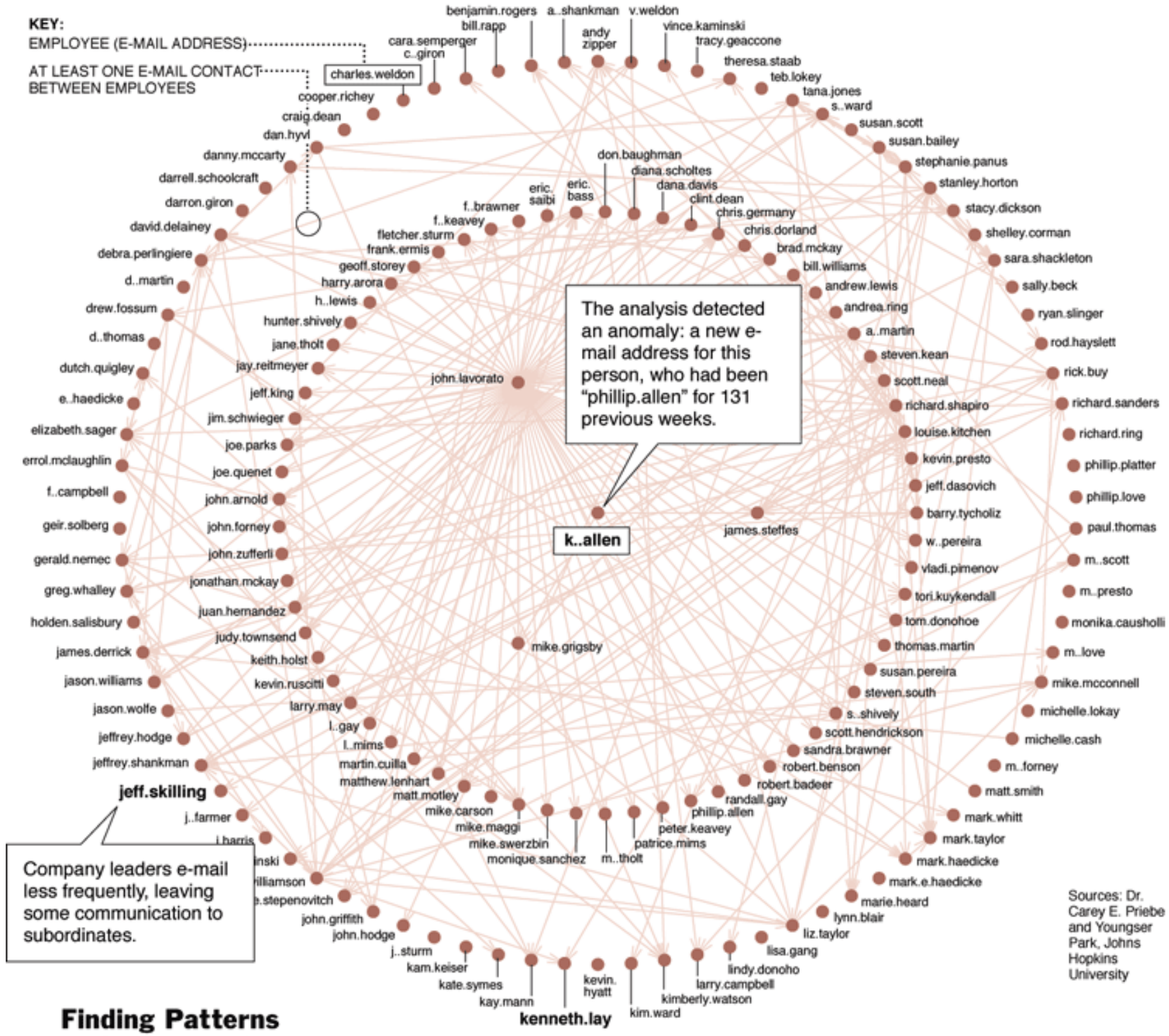
CS630 Graduate Algorithms
December 3, 2024
Dora Erdos and Jeffrey Considine

Random graph models and generation

Slides by Dora Erdos

Enron email network

Analysis of ~2Million emails. The week of May 2001 the email patterns changes



Framingham Heart Study

Longitudinal study maintained from 1947 to present day

- 15K participants
- BU is one of the study owners

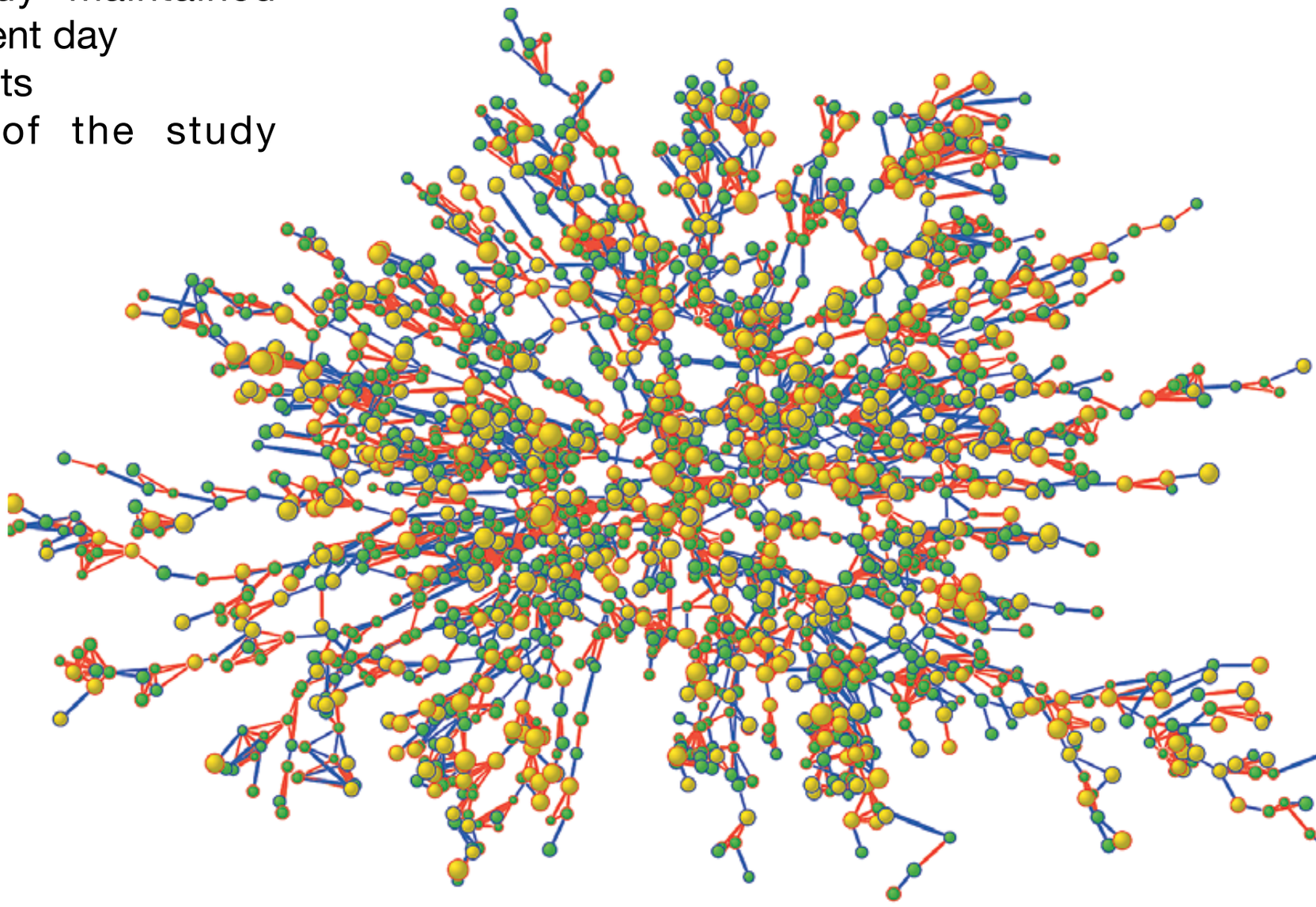
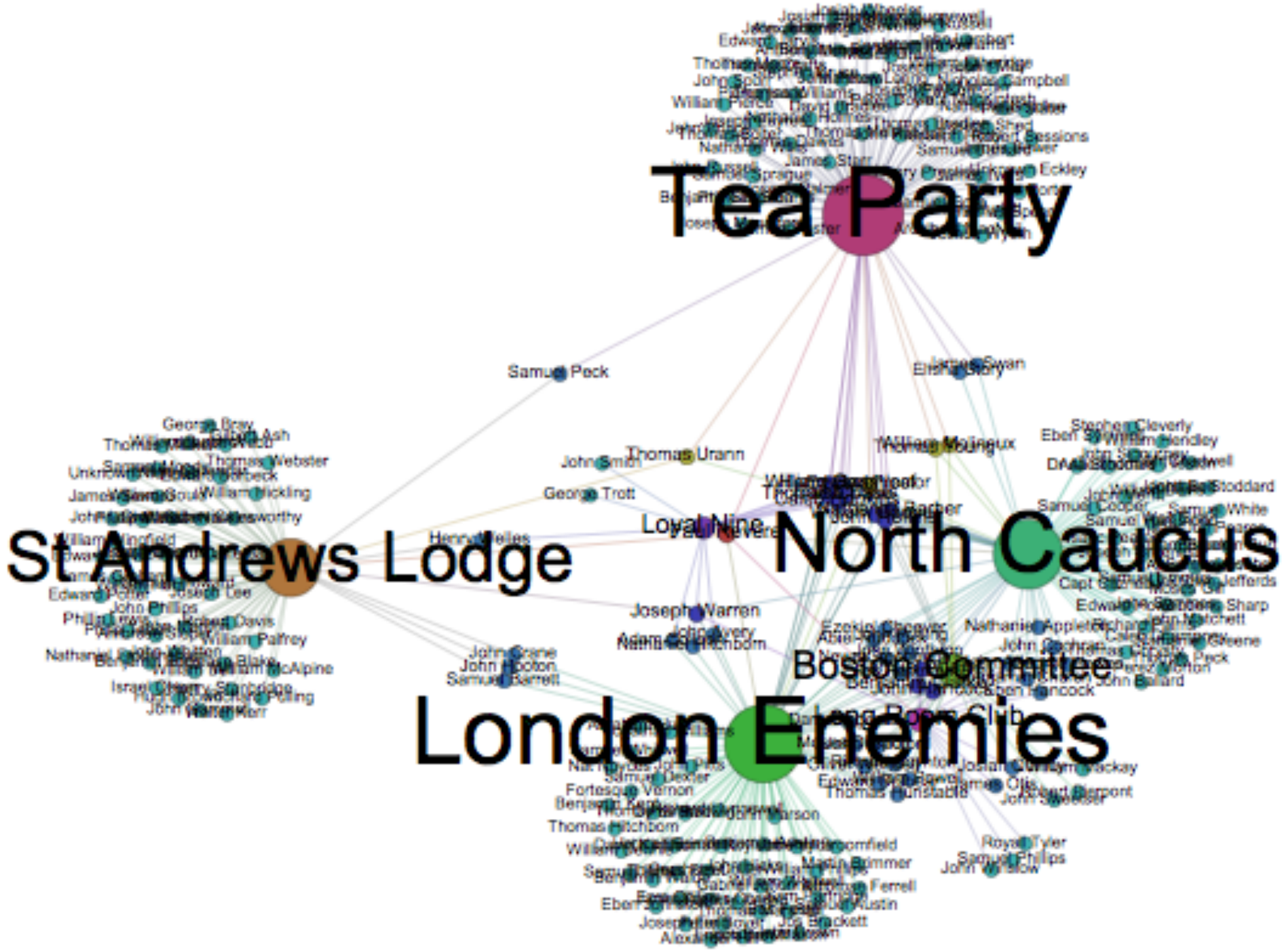


Figure 1. Largest Connected Subcomponent of the Social Network in the Framingham Heart Study in the Year 2000.

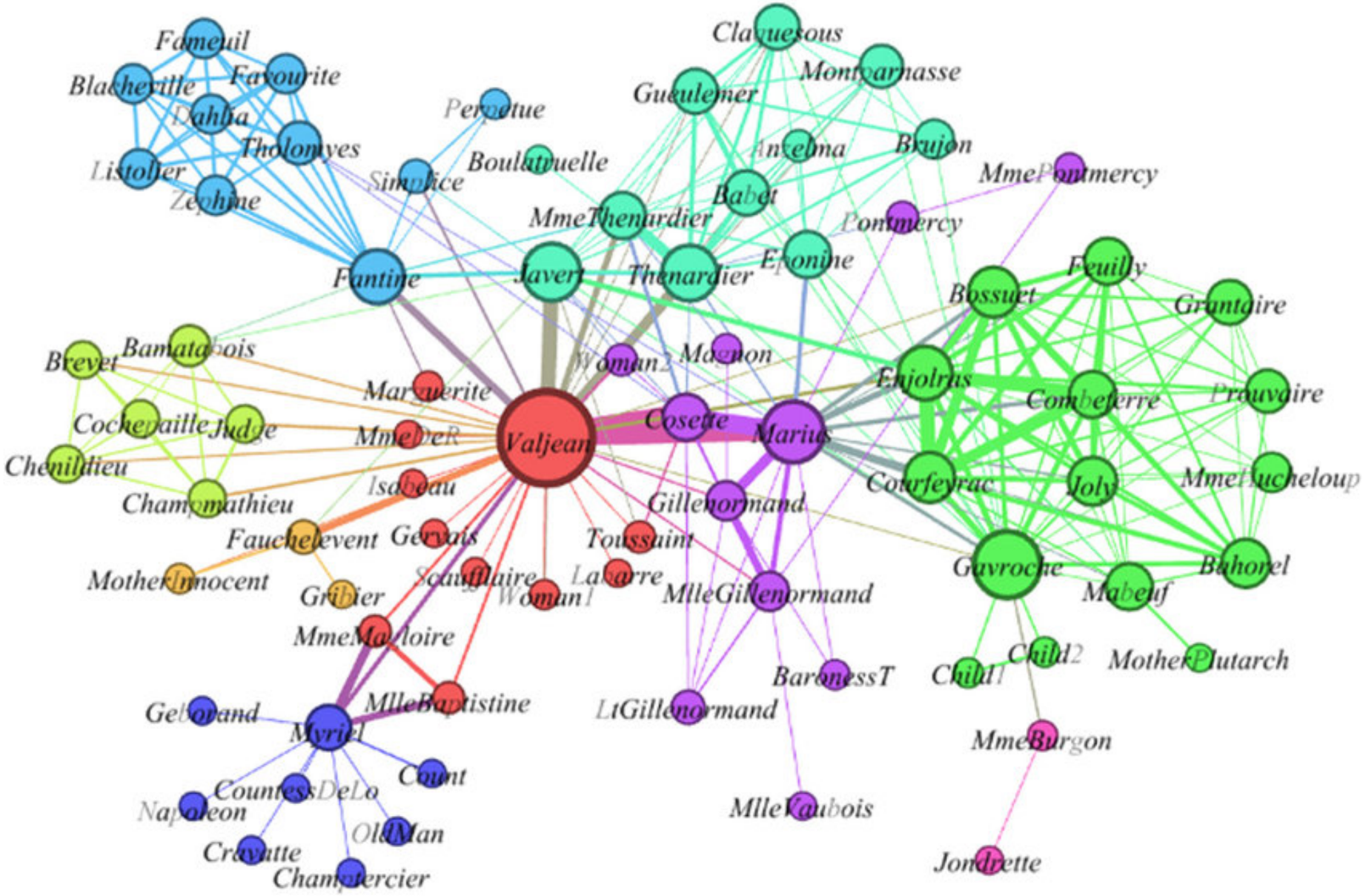
Each circle (node) represents one person in the data set. There are 2200 persons in this subcomponent of the social network. Circles with red borders denote women, and circles with blue borders denote men. The size of each circle is proportional to the person's body-mass index. The interior color of the circles indicates the person's obesity status: yellow denotes an obese person (body-mass index, ≥30) and green denotes a nonobese person. The colors of the ties between the nodes indicate the relationship between them: purple denotes a friendship or marital tie and orange denotes a familial tie.

"The Spread of Obesity in a Large Social Network over 32 Years" by Christakis and Fowler in New England Journal of Medicine, 2007

Affiliations of Colonial Boston Revolutionaries



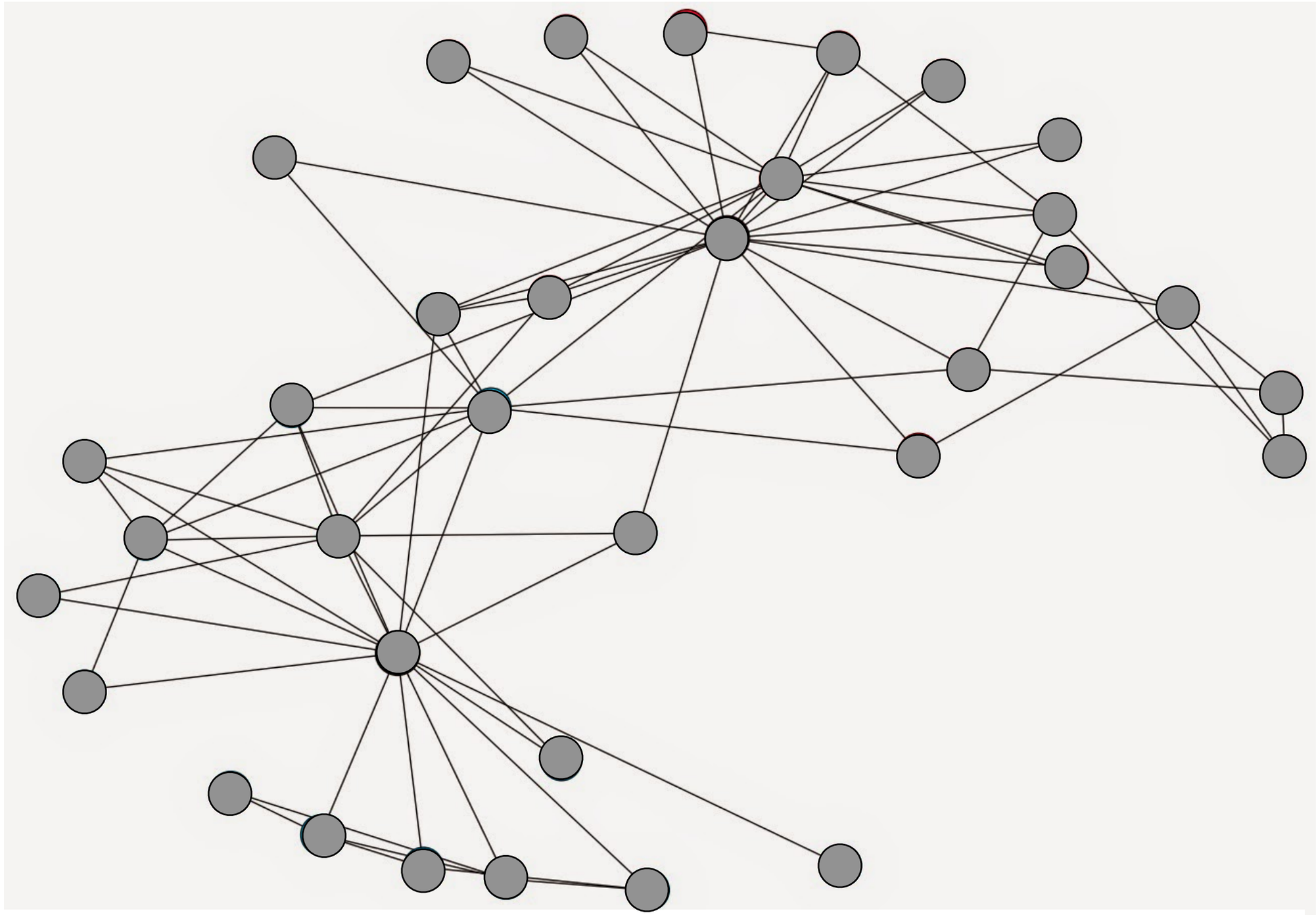
Characters in Les Miserables



Thickness of the connecting edges corresponds to the frequency of interaction between characters — thickness may indicate similarity

Data set originally created by Donald Knuth for the Stanford Graph Base, 1993

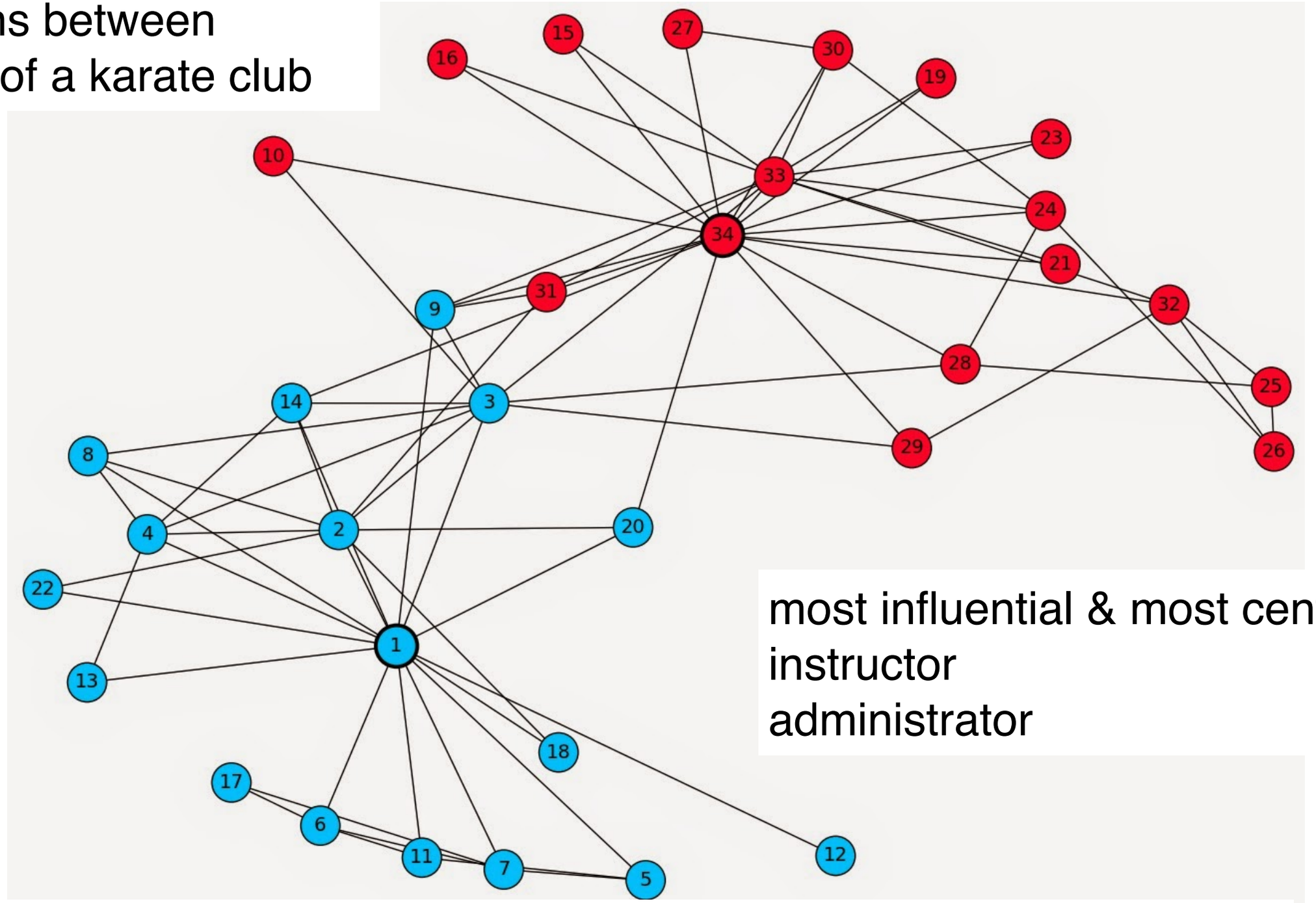
Zachary's Karate Club



WW Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**(4), 452–473 (1977)

Zachary's Karate Club

interactions between
members of a karate club



most influential & most central:
instructor
administrator

WW Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**(4), 452–473 (1977)

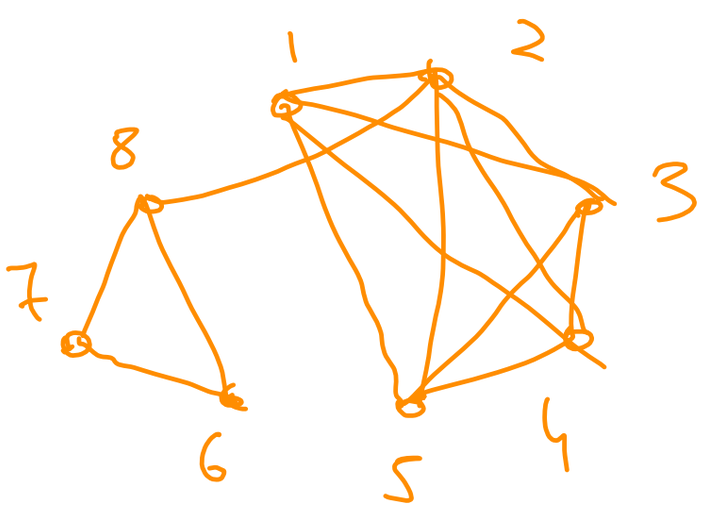
Pattern or not?

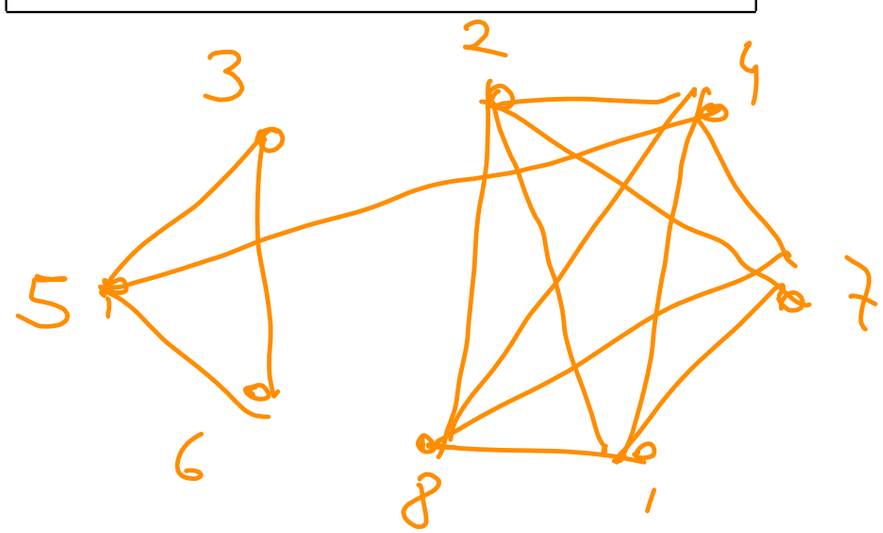
Adjacency mtx

These are two adjacency matrices of graphs. What can you observe about them?

	1	2	3	4	5	6	7	8
1	0	1	1	1	1	0	0	0
2	1	0	1	1	1	0	0	1
3	1	1	0	1	1	0	0	0
4	1	1	1	0	1	0	0	0
5	1	1	1	1	0	0	0	0
6	0	0	0	0	0	0	1	1
7	0	0	0	0	0	1	0	1
8	0	1	0	0	0	1	1	0

0	1	0	1	0	0	1	1
1	0	0	1	0	0	1	1
0	0	0	0	1	1	0	0
1	1	0	0	1	0	1	1
0	0	1	1	0	1	0	0
0	0	1	0	1	0	0	0
1	1	0	1	0	0	0	1
1	1	0	1	0	0	1	0





Random network model

Graph generation:

- pick number of nodes n
- nodes are connected by edges randomly based on some rule
- different rules result in graphs with different properties

Why?

- we can measure some properties of a real-graph and compare to the same measurements on a random graph. This can tell us whether the observed pattern is significant or random.
- we can also use it to generate graphs that are similar to the real life one

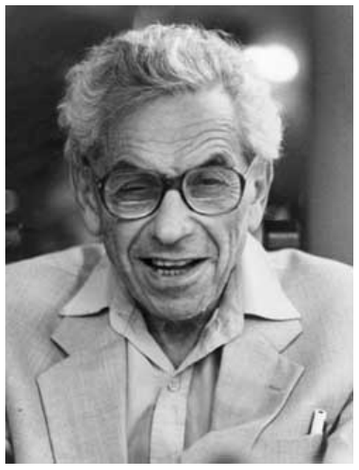
Erdos-Renyi random graph model

Erdős-Rényi random graph G(n,p):

- n nodes
- between each pair of nodes u,v an edge is generated independent at random with probability p.

- most simple model
- we understand its (mathematical) properties very well

Pál Erdős 1913-1996



Alfréd Rényi 1921-1970

Erdos-Renyi random graph model

Erdős-Rényi random graph G(n,p):

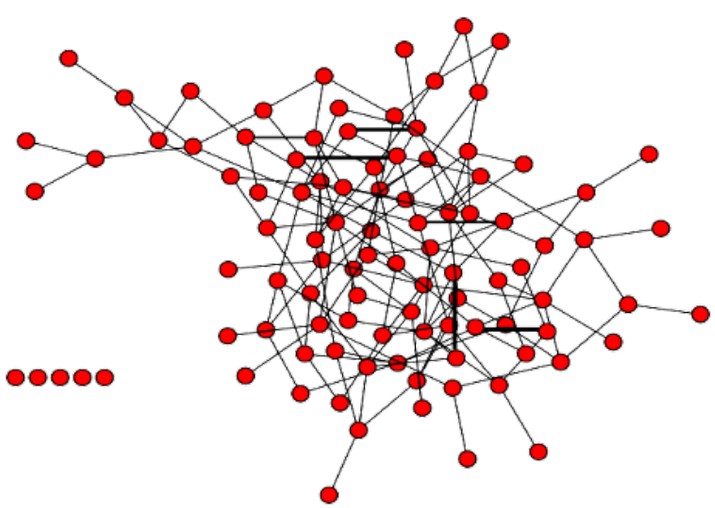
- n nodes
- between each pair of nodes u,v an edge is generated independent at random with probability p.

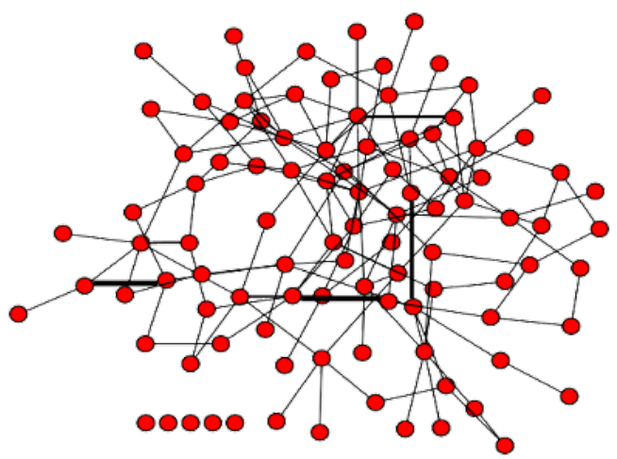
- G(n,p) is a random variable with parameters n and p
- often p is chosen to be $p = \frac{d}{n}$ where d is a constant.
  - expected degree of each node is $(n-1)\frac{d}{n} \approx d$

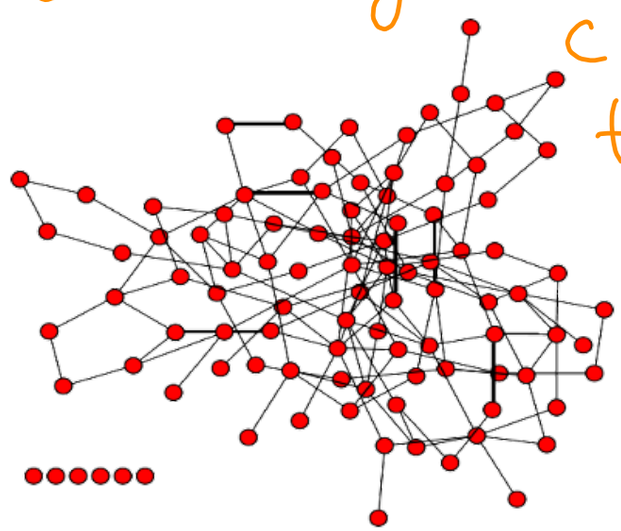
d is small ~ $\frac{d}{n}$ close to 0

$0 \leq \frac{d}{n} \leq 1$

d is large ~ $\frac{d}{n}$ close to 1







pick $p = \frac{d}{n}$ ~ d chosen by us

G(n,p) components

Connectedness of G(n,p):

(connected component = subgraph where each node can be reached through edges)

average degree $d = p(n-1)$ → each vertex has n-1 potential neighbors

- if d<1 then the connected components in G(n,p) are small
  - with high probability the size of each component is less than O(log n)

- if d>1 then there is a so-called giant component that consist of a constant fraction of the vertices
  - with high probability the size of the giant component is O(n)
  - the second largest is O(log n) with high probability

$p = \frac{1}{n}$

- when $d \approx 1$ the transition between G consisting of small components vs a giant component is abrupt
  - called the phase transition

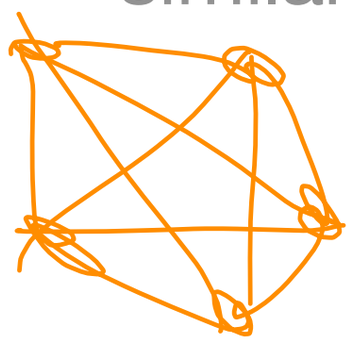
conclusion: Observing a large connected component in a real world graph doesn't imply by default that it's a definite pattern and not random.

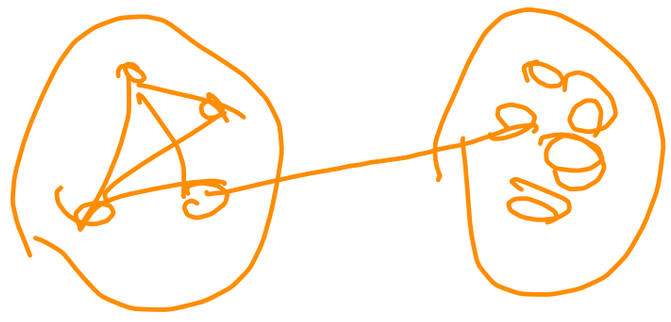
Measuring networks

- Degree distributions
- Small world phenomena
- Clustering coefficient
- Mixing patterns
- Degree correlations
- Communities and clusters

Why?

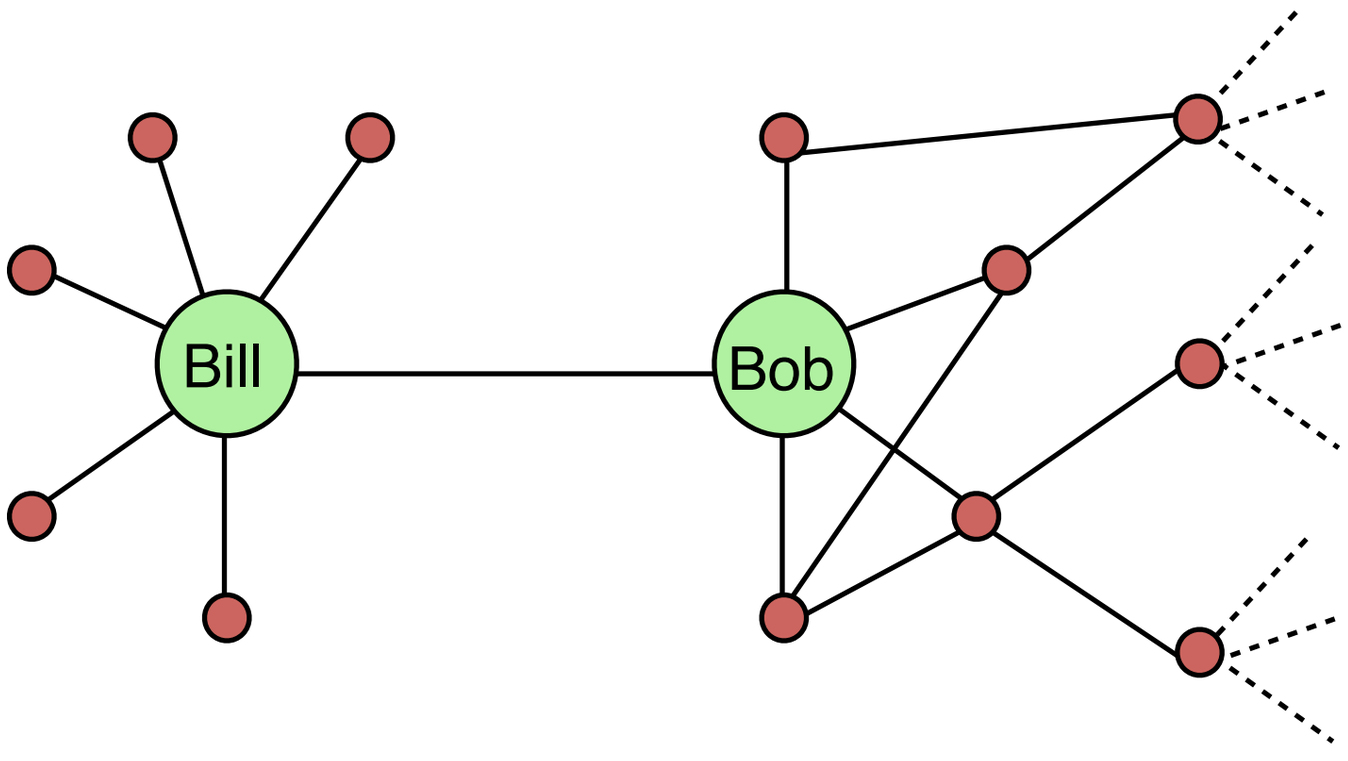
1. Find patterns in data
2. Generate graphs that are similar to real world ones





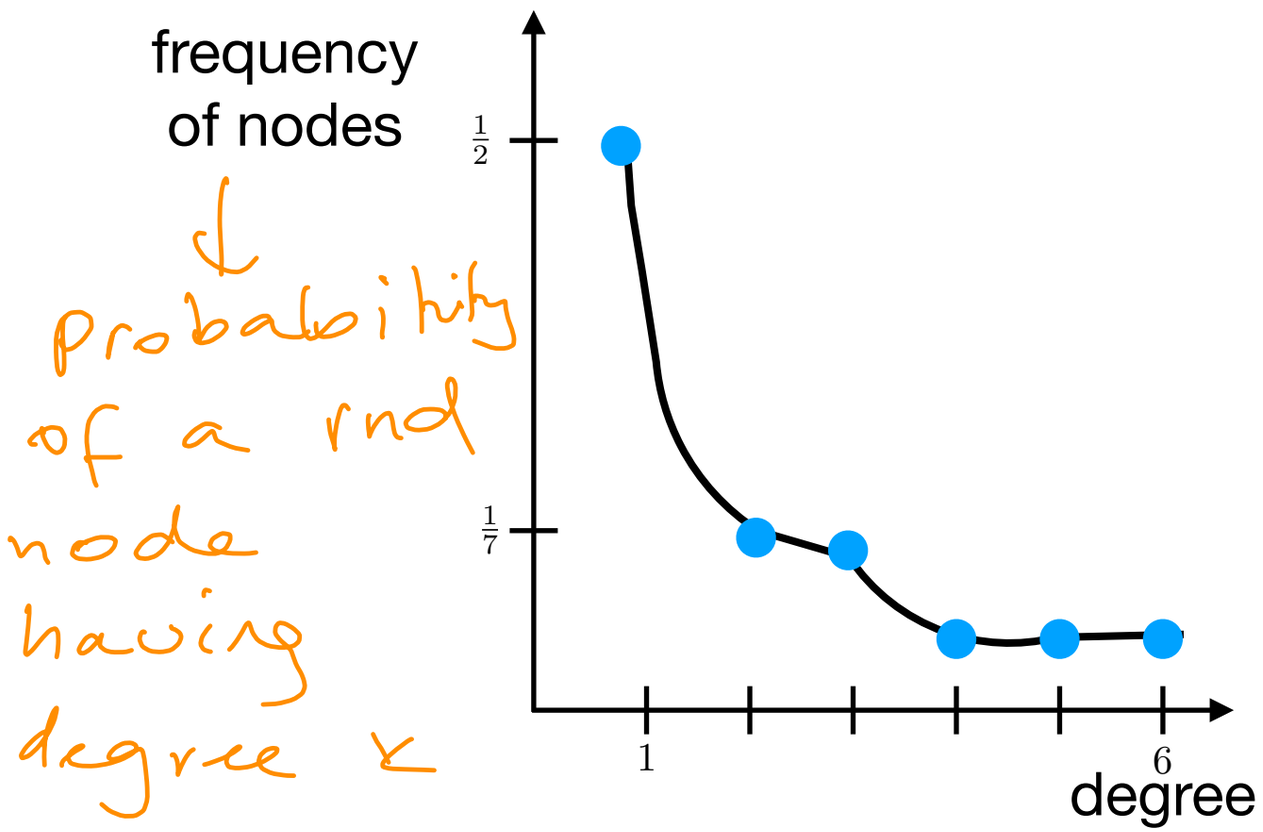
goal: generate graphs that maintain the above properties

Degree distribution



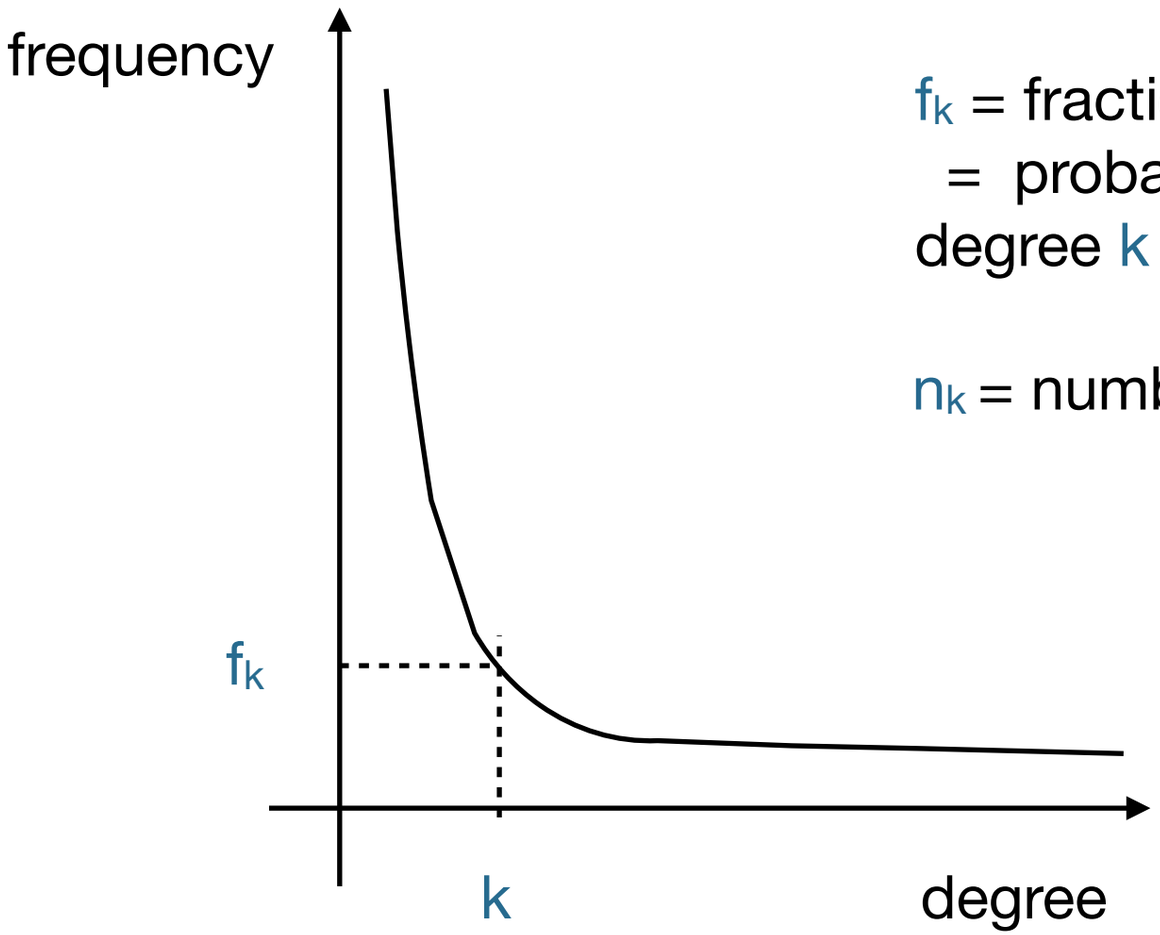
degree distribution

degree	# of nodes
1	7
2	2
3	2
4	1
5	1
6	1



probability of a rnd node having degree x

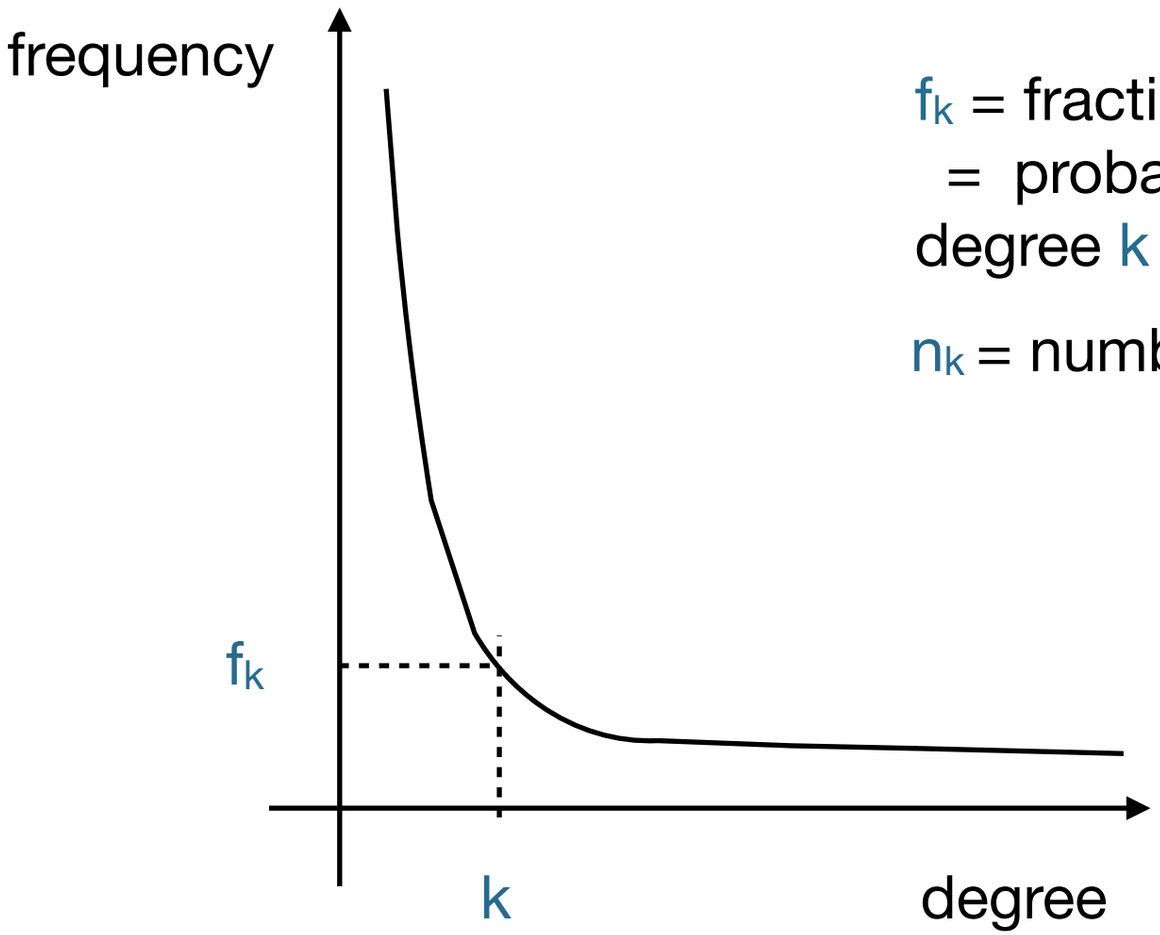
Degree distributions



f_k = fraction of nodes with degree k =
= probability of randomly selected node having degree k

n_k = number of nodes with degree k : $f_k = \frac{n_k}{n}$

Degree distributions



f_k = fraction of nodes with degree k =
= probability of randomly selected node having degree k

n_k = number of nodes with degree k $f_k = \frac{n_k}{n}$

Problem 1: find probability distribution on the node degrees
- can it be described by some known distribution?

Problem 2: how are the edges generated, given the distribution?
- can we generate a similar graph?

E-R Random graph degree distributions

Probability that a randomly selected node has degree k → binomial distribution

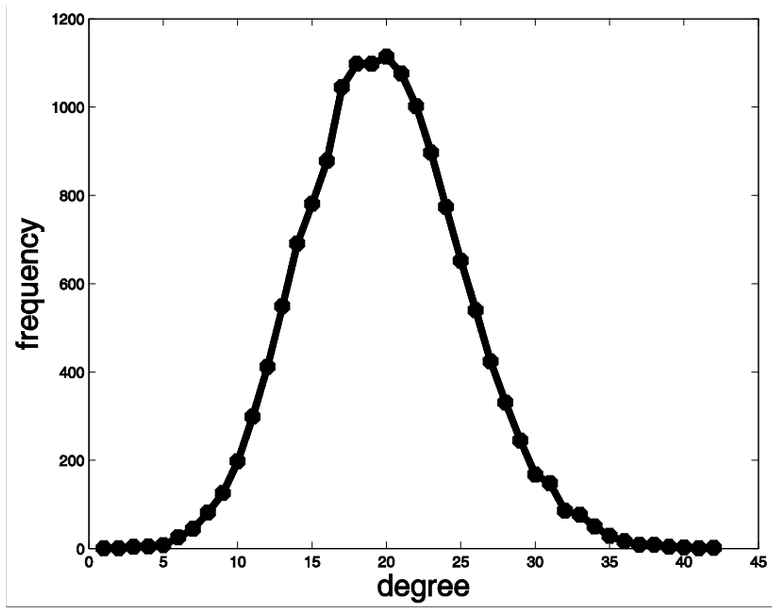
$$P(\text{degree} = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

each edge is gen. with prob p.

Average degree $d = p(n-1)$

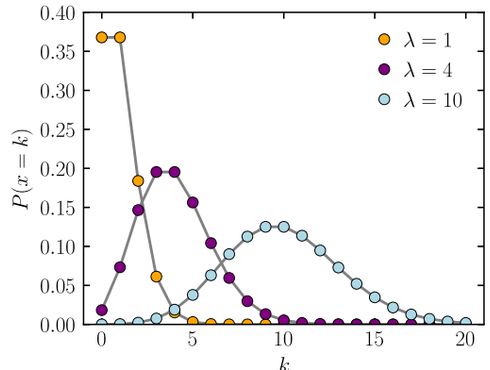
P(k) can be approximated by the Poisson distribution

- reminder: Poisson is the probability of k events happening in unit time
  - here: k number of connections (degrees)
- reminder2: Poisson is the limit of the binomial distribution when $n \to \infty$



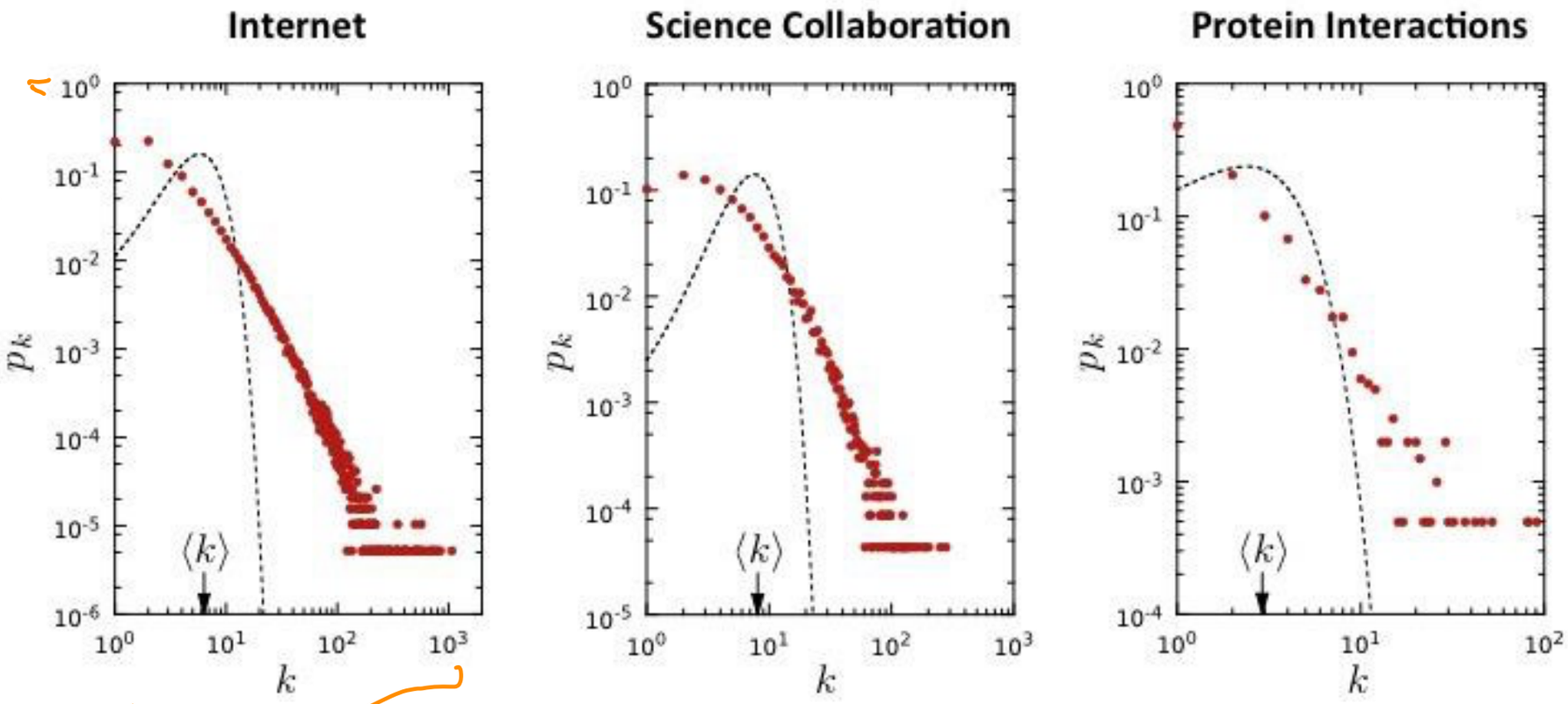
$$P(k) = P(k;d) = \frac{d^k e^{-d}}{k!}$$

- highly concentrated around the mean degree
  - many of the nodes have average degree
- probability (fraction) of high degree nodes is exponentially small



Real networks don't follow Poisson

degree distribution of some networks



from: lecture notes for the Network Science course at Northeastern by Barabasi et al.

(simple) random graphs and real life

- theory studied exhaustively
- random graphs have been used as idealized network models
- unfortunately they don’t capture reality….

Departing from the random graph models

- We need models that better capture the characteristics of real graphs
  - degree sequences
  - short paths
  - clustering coefficient

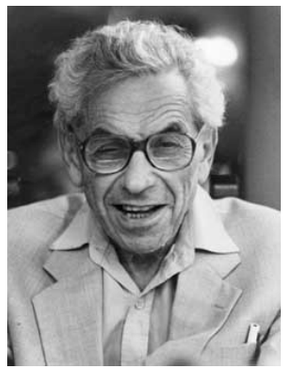
The small world experiment

- Milgram's experiment 1967
  - Picked 300 people at random from Nebraska
  - asked them to get a letter to a stockbroker in Boston. They could pass the letter through friends they know on a first-name basis

- 64 chains completed
  - 6.2 average chain length (thus, six degrees of separation)
- Further observations
  - people who owned stocks had shorter paths to the stockbroker than random people
  - people from the Boston area have even closer ties

Erdős number

- named after Paul Erdos
  - one of the most productive mathematicians of all times
  - 1525 papers during his lifetime, 511 collaborators (math is a social activity)



- collaborators of Paul have number 1
- collaborators of these have number 2, etc.
- 200K mathematicians have an assigned Erdos number, 90% is less than 8 (small world phenomenon!)
- leading mathematicians all tend to have low numbers

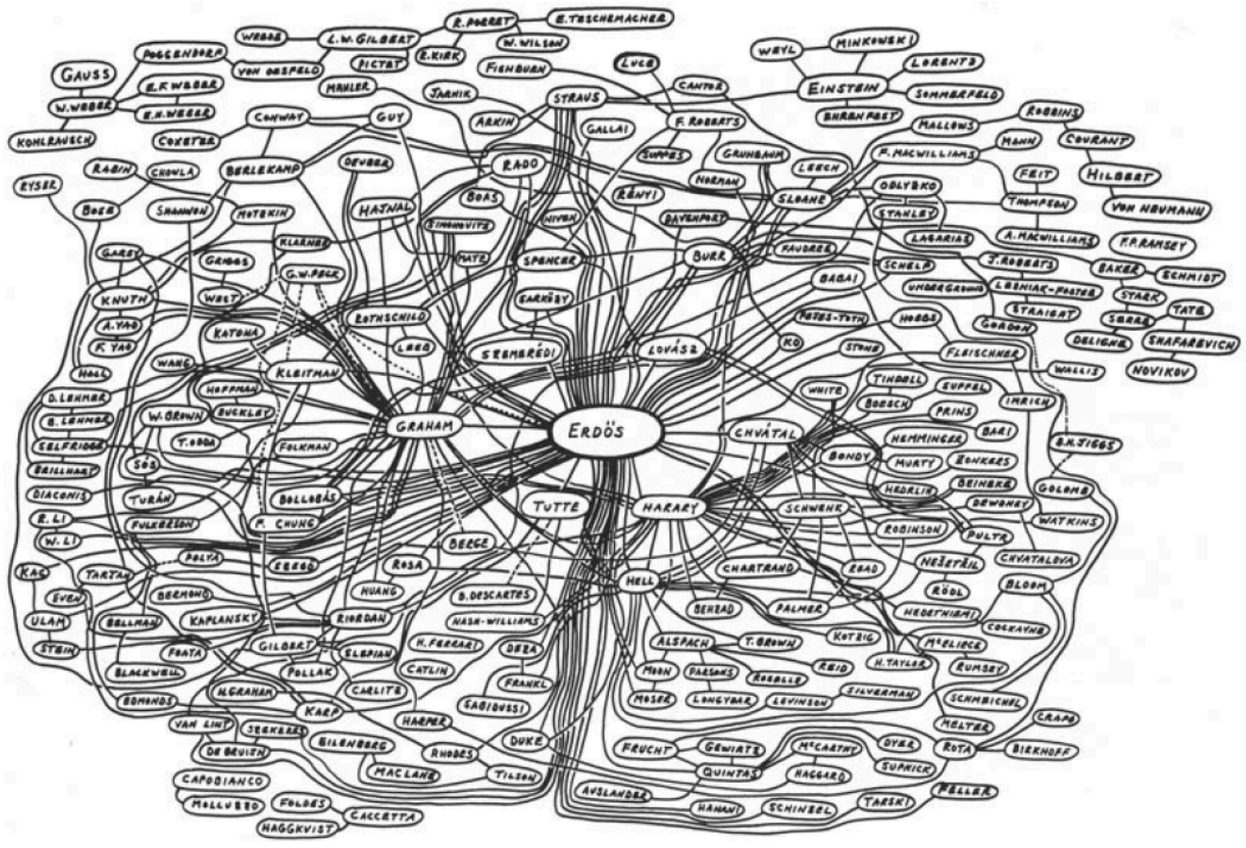
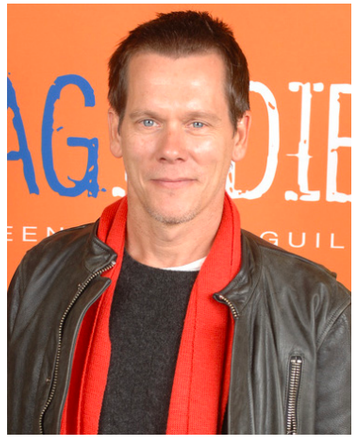


image by Easley and Kleinberg (2010)

Six degrees of Kevin Bacon



- Bacon number:
  - Create a network of Hollywood actors
  - Connect two actors if they play together in a movie
  - Bacon number: number of steps to Kevin Bacon
- As of Dec 2007 the highest (finite) Bacon number is 8
- Only approx 12% of all actors cannot be linked to him
  - What is the Bacon number of Elvis Presley?

- Elvis Presley was in Change of Habit (1969) with Edward Asner
- Edward Asner was in JFK (1991) with Kevin Bacon

Therefore, Asner has a Bacon number of 1, and Presley (who never appeared in a film with Bacon) has a Bacon number of 2.

Measuring the small-world phenomenon

dist(u,v) = length (number of edges) of the shortest paths between u and v

$$\text{diameter} = \max_{u,v \in V} \text{dist}(u, v)$$

- small world graphs have low diameters

$$\text{diameter}_{small\ \ world} \approx constant$$

- E-R graphs have diameter

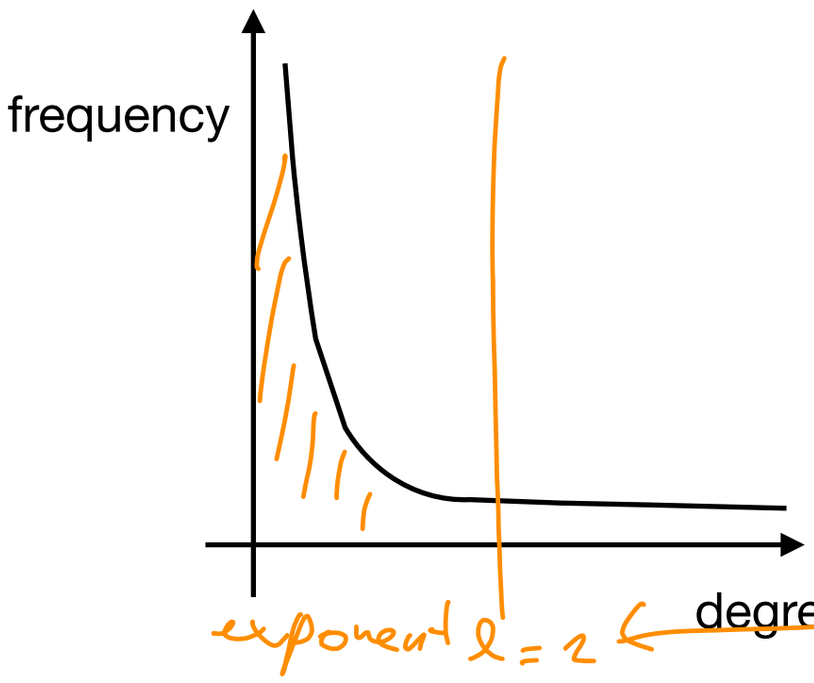
$$\text{diameter}_{E-R} = \frac{\log n}{\log d}$$

Power-law distributions

- Degree distribution of many real-life networks follow a power-law

  Power Law: functions f(x), such that there is a constant c and a constant exponent ~~x~~ ℓ that $f(x) = cx^{\ell}$ $\quad f(x+y) = c(x+y)^{\ell}$

  - functional relationship between quantities. The relative change in one quantity results in a relative change in the other quantity proportional to the change raised to a constant exponent
    - quantity varies as a power of the other
    - properties of the power law are often analyzed in the context of the exponent ℓ
  - right skewed/heavy-tail distribution
    - median is well defined, but mean is not



examples:
- CPU's cache size and number of cache misses
- number of connections per individual in a social network
- Pareto principle - for many outcomes 80% of the consequences comes from 20% of causes (numbers may vary)
- etc.

area of square proportional to the length of its side

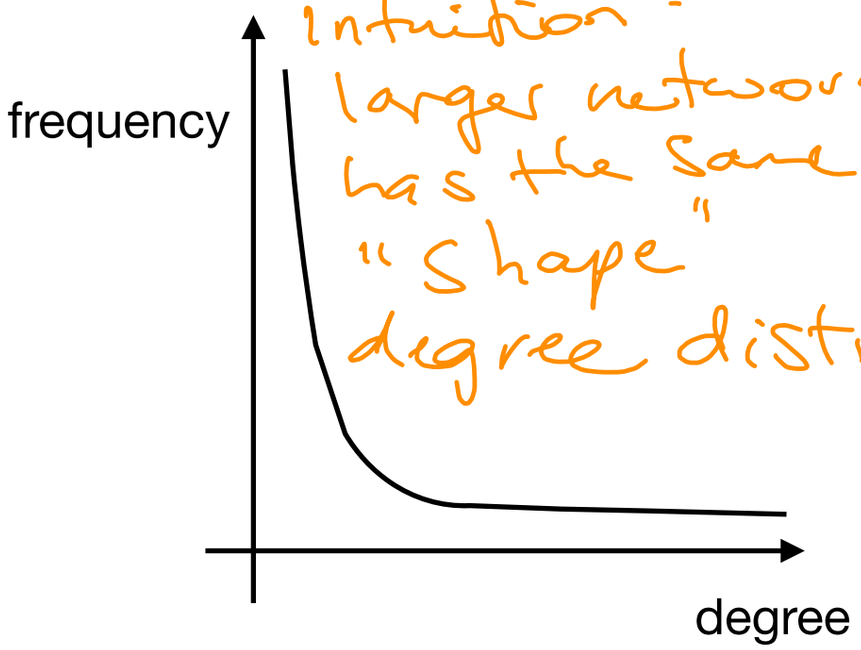
Power-law distributions

- Degree distribution of many real-life networks follow a power-law

probability that random node has degree k $p(k) = Ck^{-\alpha}$

- Right-skewed/heavy tail distribution
  - there is a non-negligible fraction of nodes that has very high degree
    - the mean is not representative of the data

- scale free: the distributions remain the same (up to multiplication by a constant) if the input is scaled.



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

constant

Power-law signature

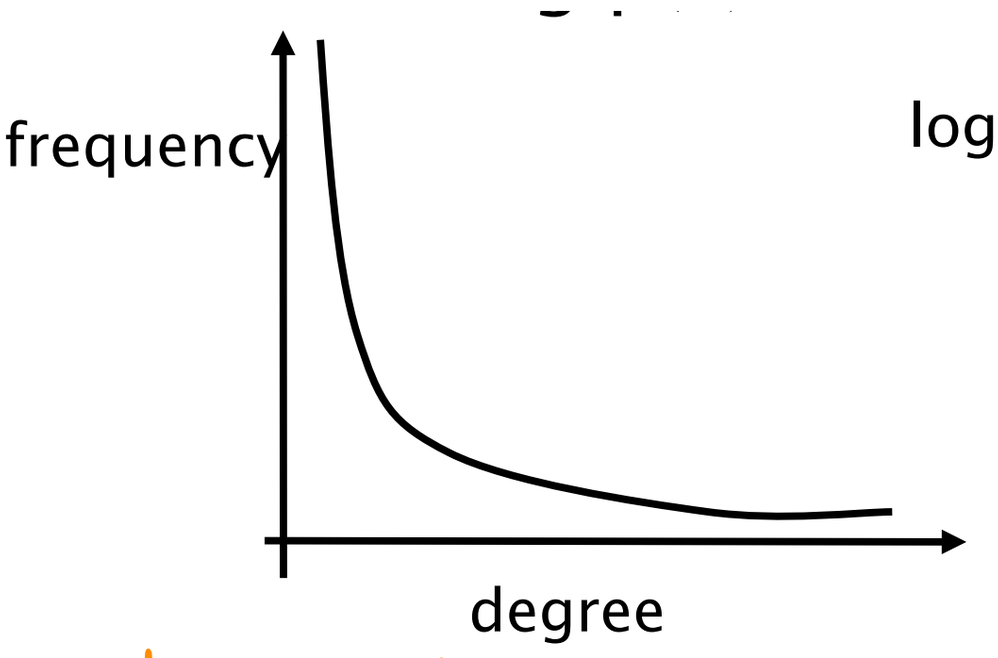
Power-law distribution gives a line in the log-log plot

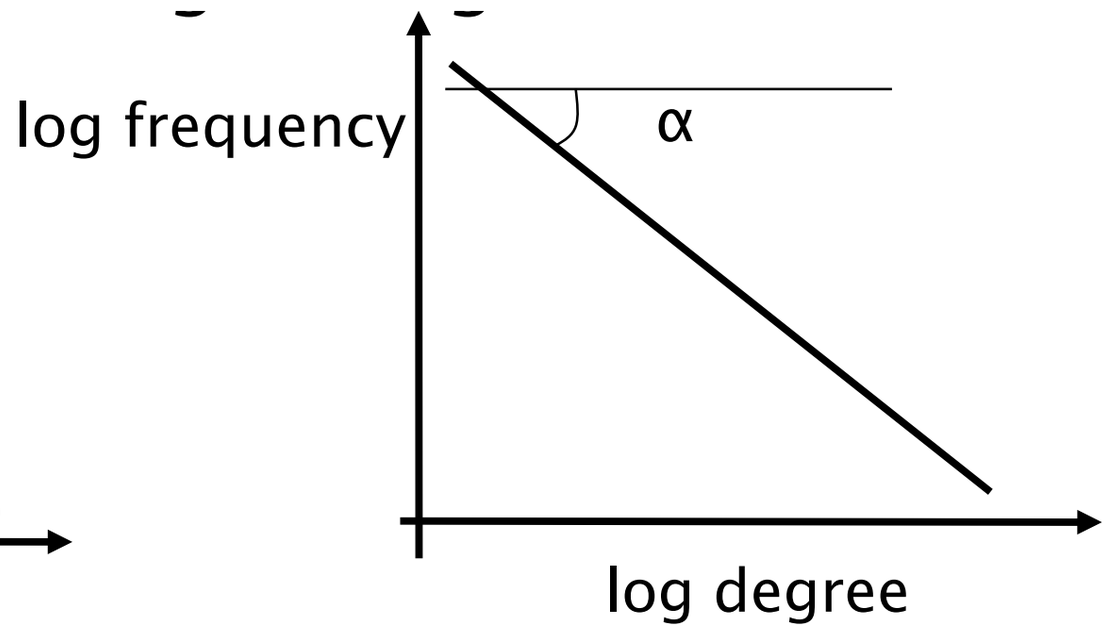
$$\log p(k) = -\alpha \log k + \log C$$

$\log k$ = variable
$\Rightarrow$ equation of a line

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

$\log(p(k)) = \log(C k^{-\alpha})$
$= -\alpha \log k + \log C$





find C and α corresponding to data:
- create log-log data from degree sequencies
- fit a line on this data
  → compute slope

Power-law examples

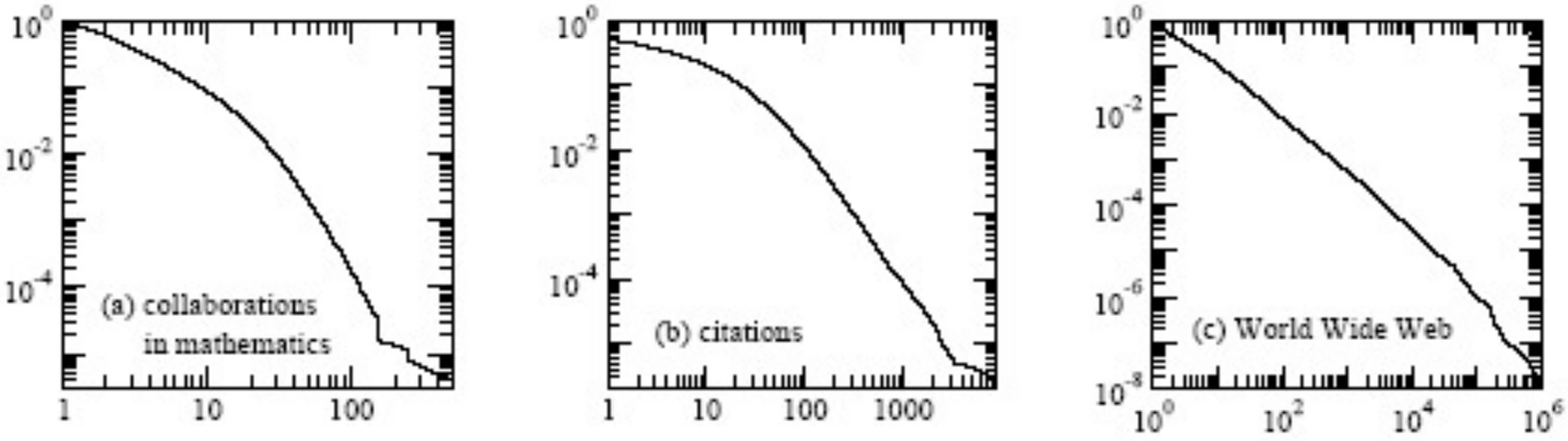


image from Newman [2003]

To check whether data follows a power-law you need to perform a maximum-likelihood estimate.

Preferential attachment in networks

- First considered by Price [1965] as a model of citation networks
- new nodes prefer to link to higher connected nodes
  - each new paper is generated with m citations (this will be the mean degree)
  - new papers cite previous papers with probability proportional to their in-degree (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 $\alpha = 2 + 1/m$

Barabasi-Albert model

- The B-A model (for undirected graph)
  - input: some initial subgraph G_0, and m the number of edges per node
  - the process:
    - nodes arrive one at a time
    - each node selects m other nodes with probability proportional to their degree
    - if $[d_1, d_2, \ldots, d_t]$ is the degree sequence at time t, the node $t+1$ links to node i with probability $\frac{d_i}{\sum_i d_i} = \frac{d_i}{2mt}$

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

goals:
- maintain degree distr.
- small world phenomena → small (const) diam

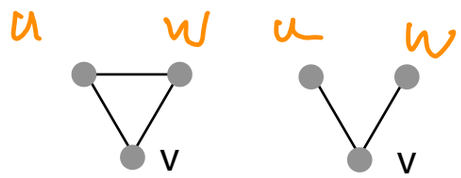
missing:
- clustering coefficient

Clustering Coefficient

- Measures the density of triangles (local clusters) in the graph

$$C_v = \frac{\text{triangles containing } v}{\text{triples centered at } v}$$

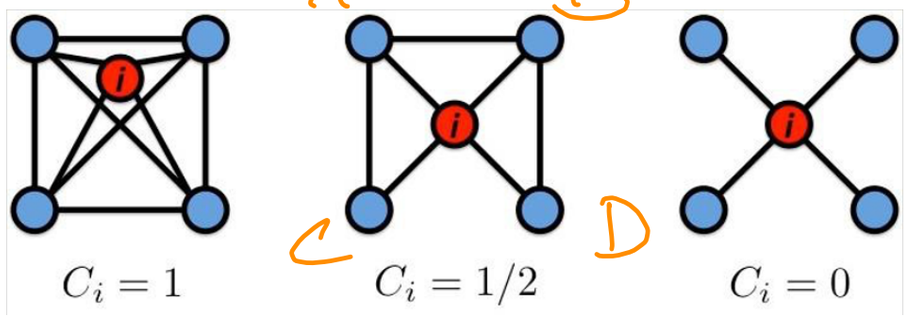
$$C = \frac{1}{n}\sum_v C_v$$



triple centered at v
= pairs of neighbors of v

$0 \le C_v \le 1$

the larger, the more connected your local community



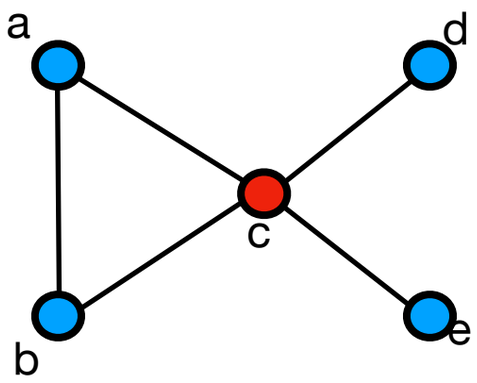
pairs:
A B ✓
A C ✓
A D ✗
B C ✗
B D ✓
C D ✗

Clustering Coefficient - TopHat

- Measures the density of triangles (local clusters) in the graph

$$C_v = \frac{\text{triangles containing } v}{\text{triples centered at } v}$$

$$C = \frac{1}{n}\sum_v C_v$$



Select which clustering coefficients are correct

A. $C_a = 1$ ✓

B. $C_b = 1/2$ ✗

C. $C_c = 1/4$ ~ 6 combinations of neighbors of C_1 only one (a,b) form triange

D. $C_d = 1$

E. $C_e = 0$ ✓

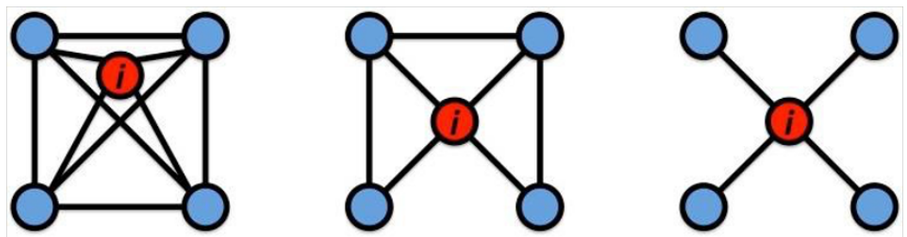
$C_c = \frac{1}{6}$

Clustering Coefficient

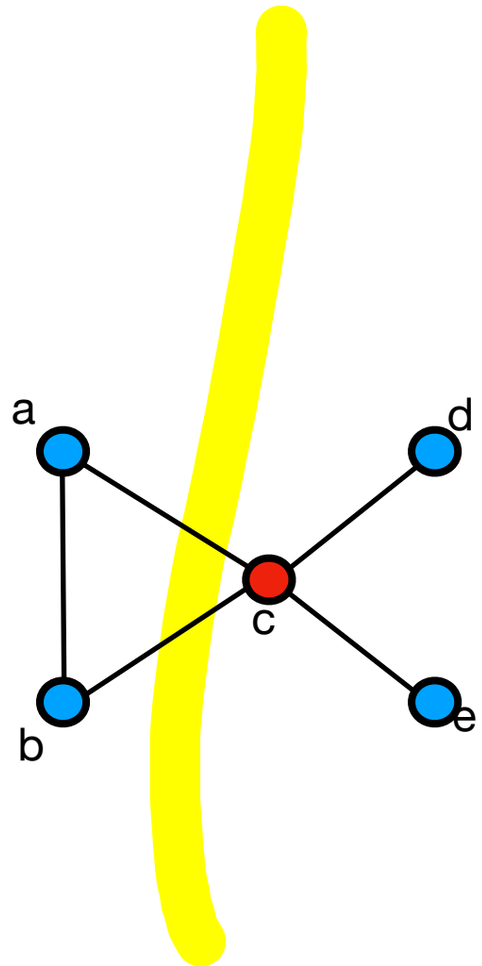
- Measures the density of triangles (local clusters) in the graph

$$C_v = \frac{\text{triangles containing } v}{\text{triples centered at } v}$$

$$C = \frac{1}{n}\sum_v C_v$$



$C_i = 1$ $C_i = 1/2$ $C_i = 0$



$$C_a = C_b = \frac{1}{1}$$

$$C_c = \frac{1}{6}$$

$$C_d = C_e = 0$$

$$C = \frac{1}{5}(1 + 1 + \frac{1}{6}) = \frac{13}{30}$$

Clustering coefficient for E-R random graphs

- The probability of two of your neighbors also being neighbors is p independent of local structure
  - clustering coefficient C=p
  - when z is fixed C = z/n = O(1/n)

compare real-world graphs to end counterparts

clustering coefficient

E-R graph with the same "density"

Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

Small-world phenomenon

- So far we focused on obtaining graphs with power-law distributions on the degrees. What about other properties?
  - Clustering coefficient: real-life networks tend to have large clustering coefficients
  - short parts: real-life networks are “small worlds”
    - this property is easy to generate
  - Can we combine these two properties?

Small-world graphs

- According to Watts [1999]
  - large networks ($n >> 1$)
  - sparse connectivity (average degree $z << n$)
  - no central node ($k_{max} << n$)
  - large clustering coefficient (larger than in random graphs of same size)
  - short average paths (~$\log n$, close to those of random graphs of the same size)

Watts and Strogatz model [1998]

- start with a ring, where every node is connected to the next z nodes
- with probability p, rewire every edge (or, add a shortcut) to a uniformly chosen destination