

Analysis of Social Media

MLD 10-802, LTI 11-772

William Cohen

9-20-12

Administrivia

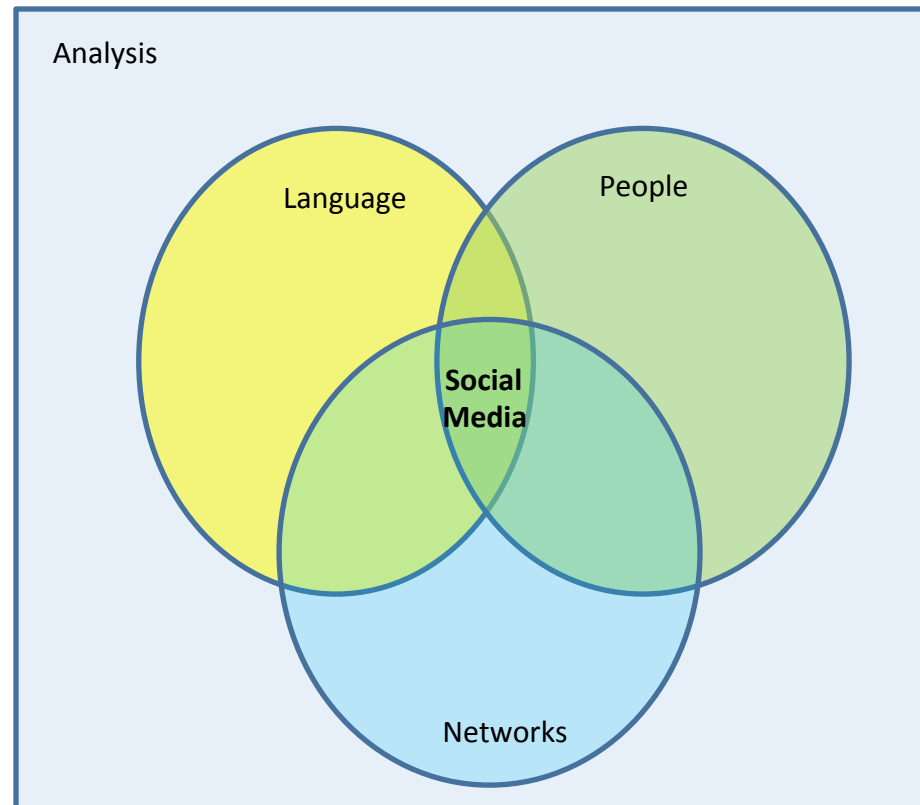
- William's office hours: **1:00-2:00 Friday** or by appointment
 - but not this week, 9/21
- Wiki access: send email to Katie Rivard (krivard@andrew)
- First assignment....

Assignments

- Monday 9/24, by 10am:
 - send us 3 papers you plan to summarize
 - title, authors, ptr to on-line version
 - any not-yet-summarized papers on this years or last year's syllabus are pre-approved
 - search the wiki to make sure it's new!
 - also, don't propose papers I've presented in class
 - or, you can propose another paper I don't know about
 - you can work in groups if you like - but you still need to do 3 papers per person
 - eg if this will be a group for the project
 - hint: short papers are not necessarily easier to read
- Thus 9/27 at 10:30am:
 - first summary is due
 - structured summary + the study plan you followed
 - Example of a structured summary:
http://malt.ml.cmu.edu/mw/index.php/Recent_or_influential_technical_papers_in_Analysis_of_Social_Media
 - Example of a study plan: to be posted
 - if the paper contains a novel Problem, Method, or Dataset then also add a page for it to the wiki
- Tues 10/2:
 - three summaries per student on the wiki
- Thus 10/4:
 - project proposal - first draft - one per student
 - we will allow lit review projects - more on this later
- Tues 10/9:
 - form project teams of 2-3 people

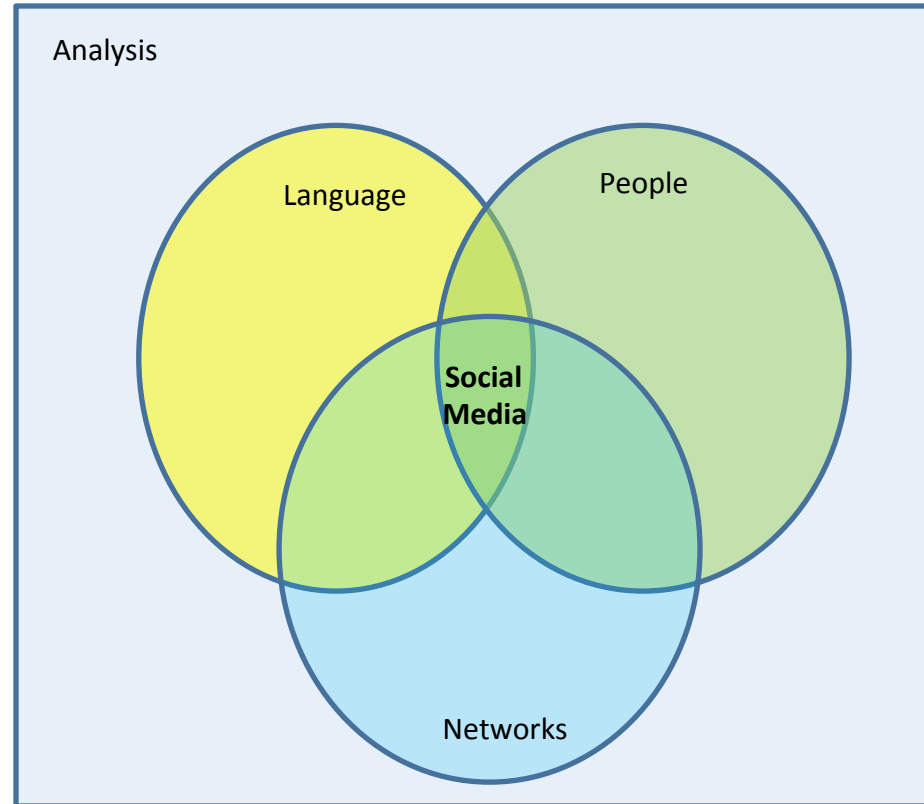
- 1. Background(6-8 wks, mostly me):
 - Opinion mining and sentiment analysis. Pang & Li, FnTIR 2008.
 - **Properties of social networks. Easley & Kleinberg, ch 1-5, 13-14; plus some other stuff.**
 - Stochastic graph models. Goldenberg et al, 2009.
 - Models for graphs and text. [Recent papers]

Syllabus



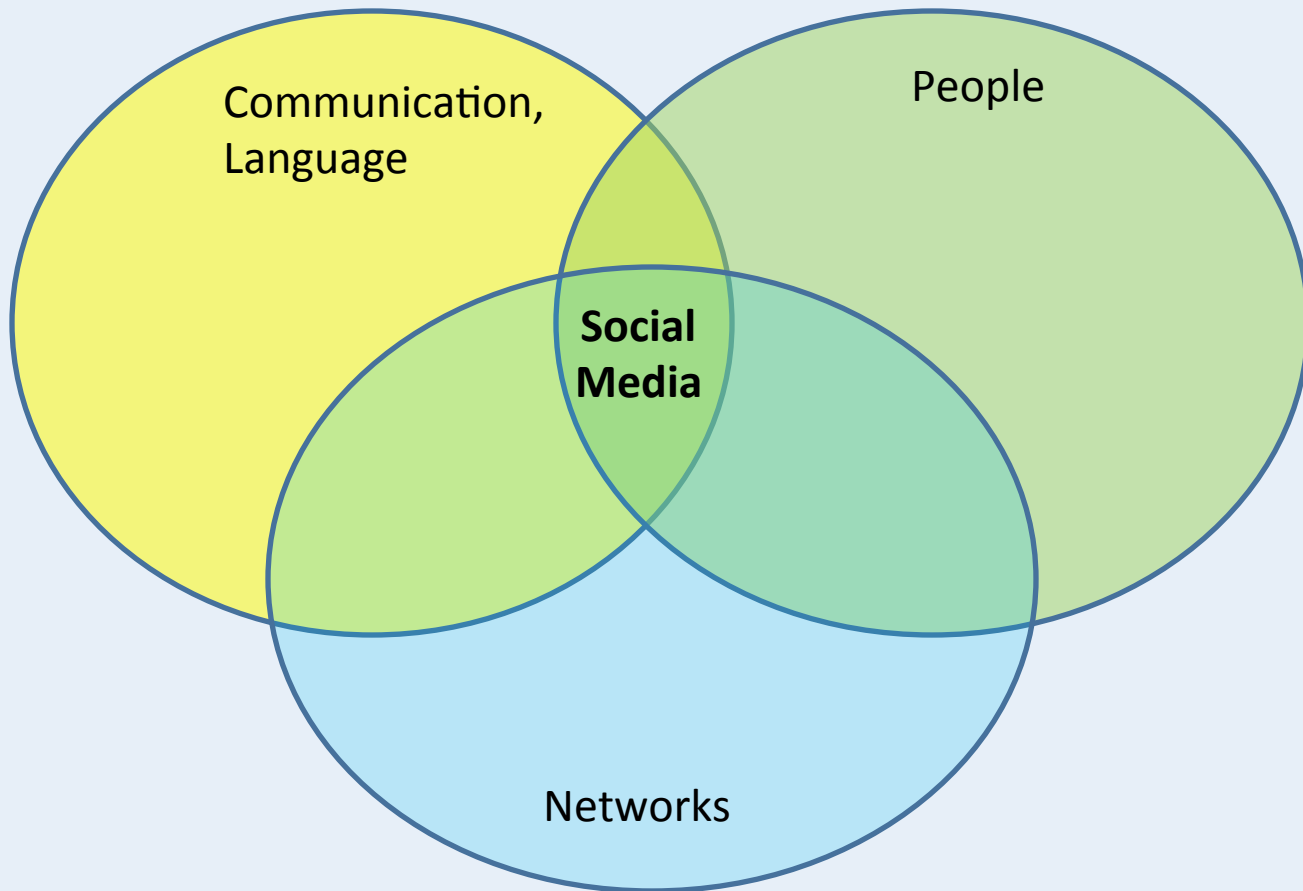
- 2. Review of current literature (mostly you, via assignments).
- Asynchronously: 3-5 guest talks from researchers & industry on sentiment, behavior, etc.

Syllabus



Today's topic: networks

Analysis : modeling & learning



Additional reference

The structure and function of complex networks

Authors: [M. E. J. Newman](#)

(Submitted on 25 Mar 2003)

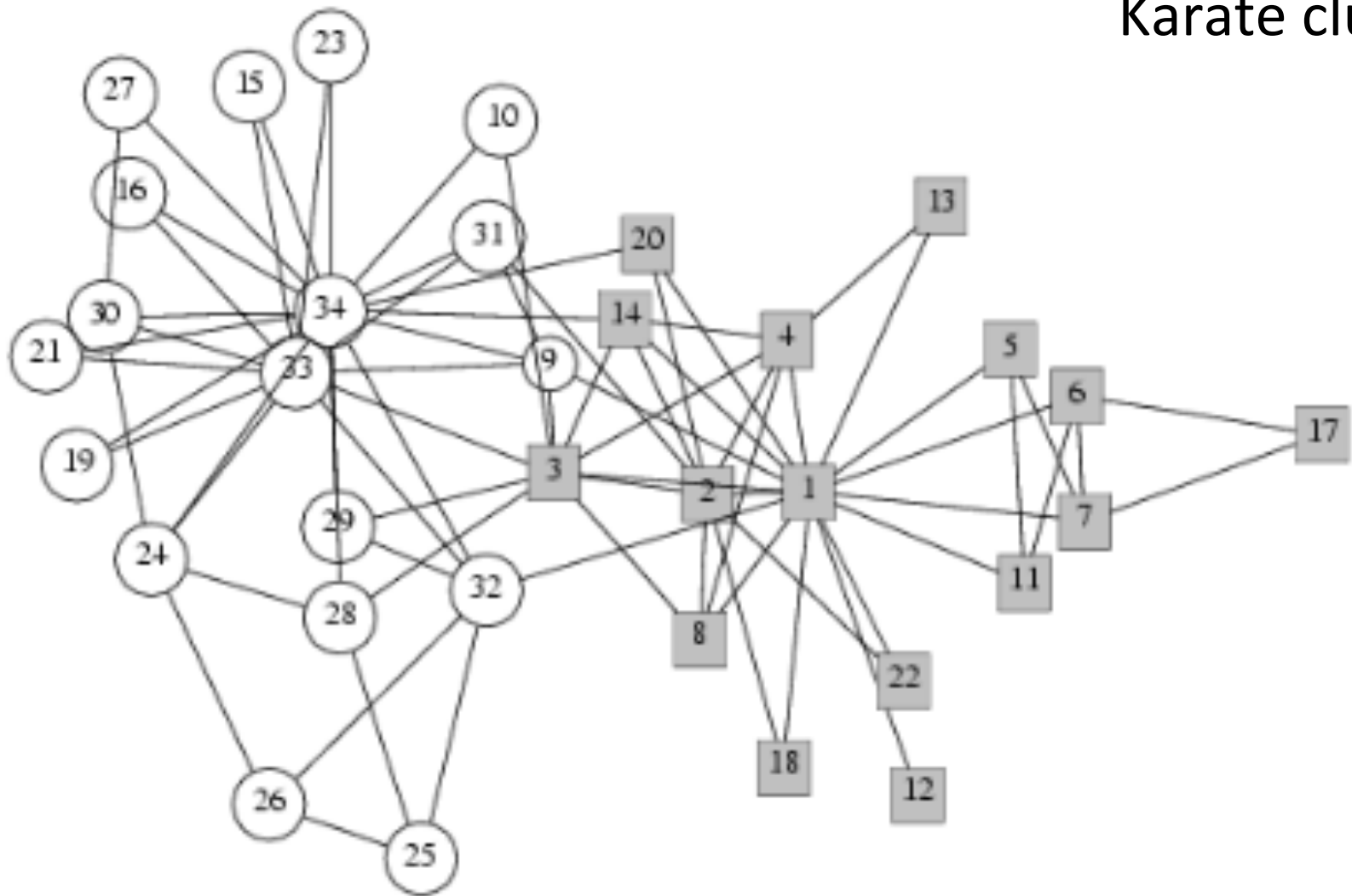
Abstract: Inspired by empirical studies of networked systems such as the Internet, social networks, and biological networks, researchers have in recent years developed a variety of techniques and models to help us understand or predict the behavior of these systems. Here we review developments in this field, including such concepts as the small-world effect, degree distributions, clustering, network correlations, random graph models, models of network growth and preferential attachment, and dynamical processes taking place on networks. Comments: Review article, 58 pages, 16 figures, 3 tables, 429 references, published in SIAM Review (2003)

Subjects: Statistical Mechanics (cond-mat.stat-mech); Disordered Systems and Neural Networks (cond-mat.dis-nn) J

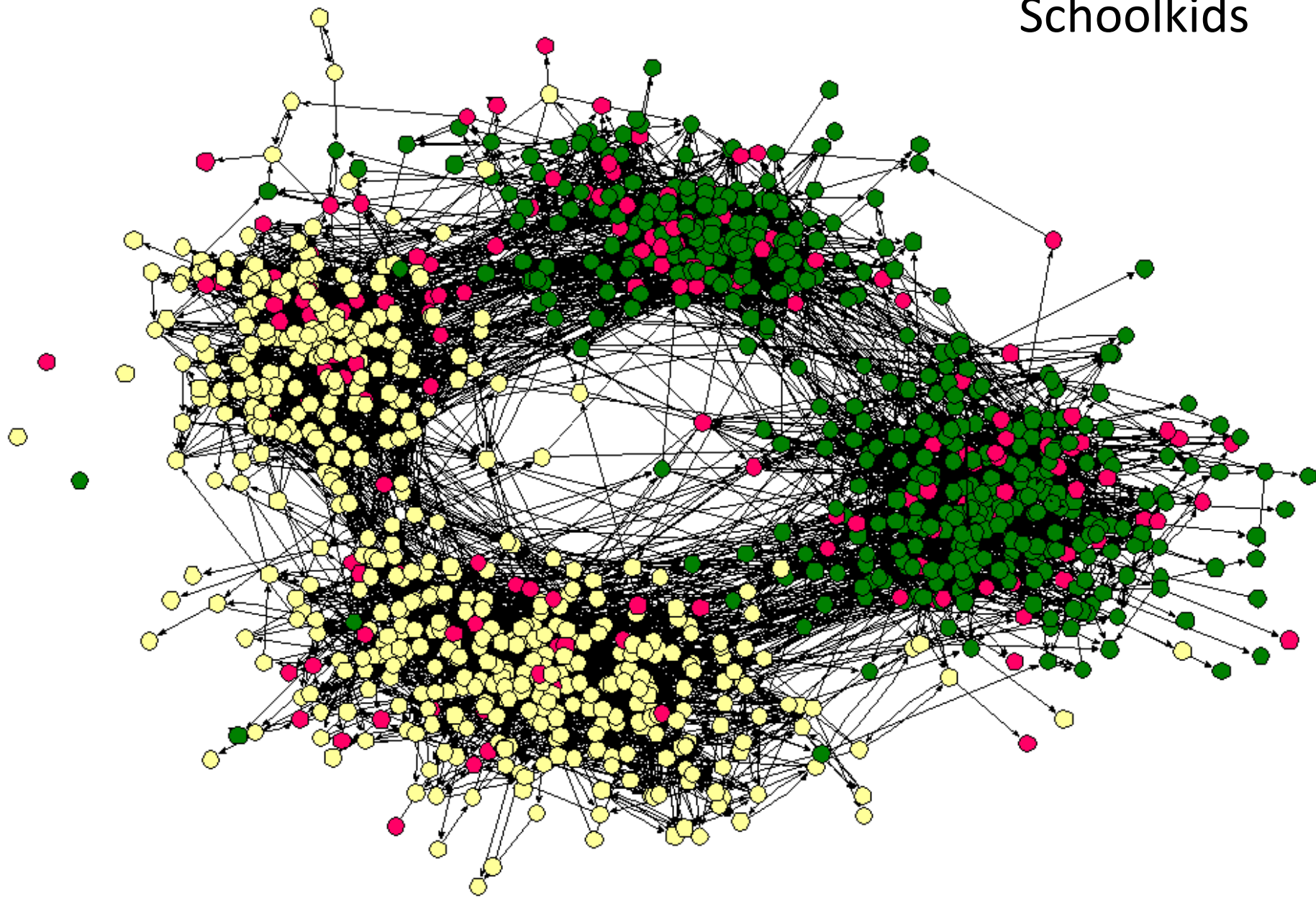
Journal reference: SIAM Review 45, 167-256 (2003)

Cite as: [arXiv:cond-mat/0303516v1](#) [cond-mat.stat-mech]

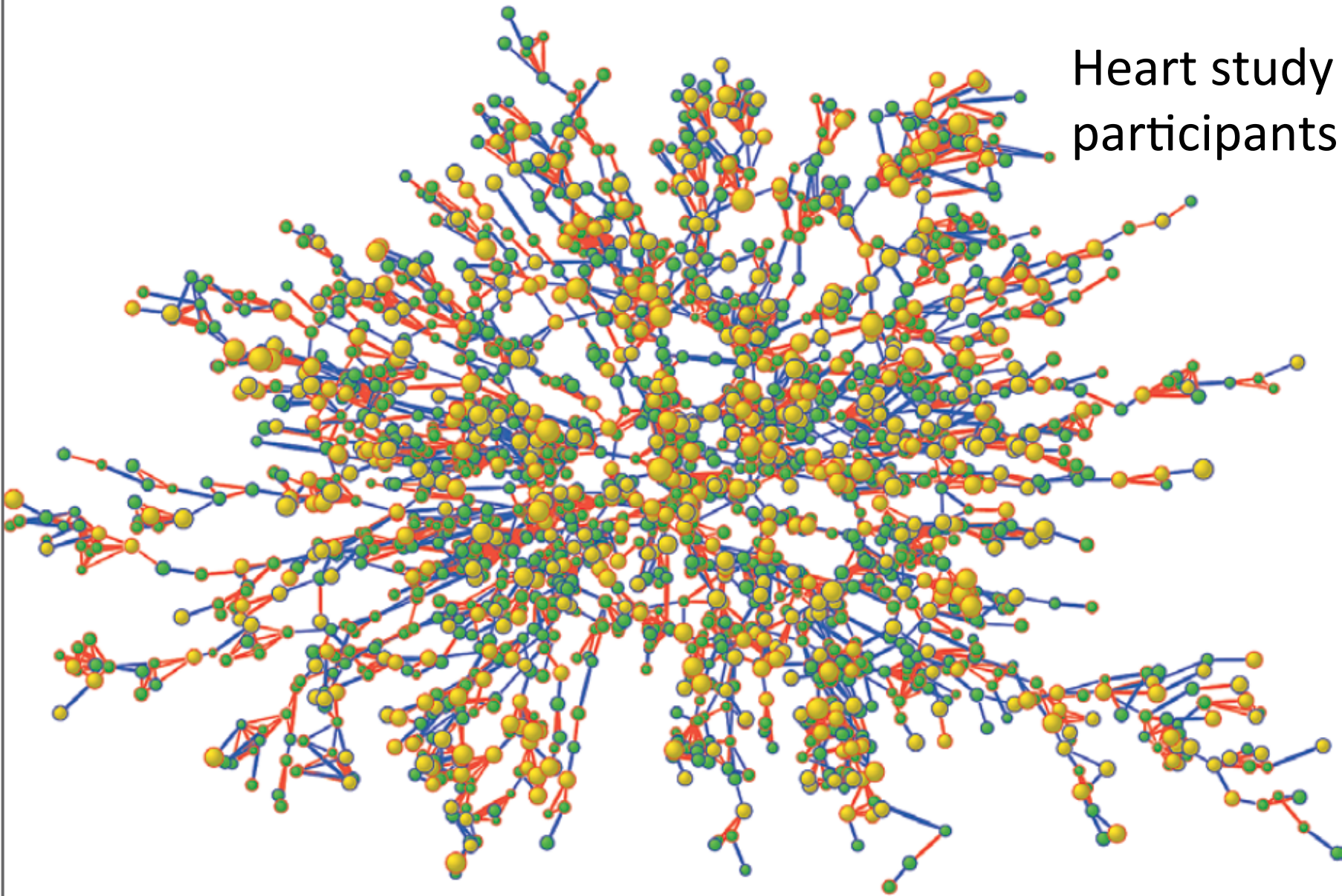
Karate club



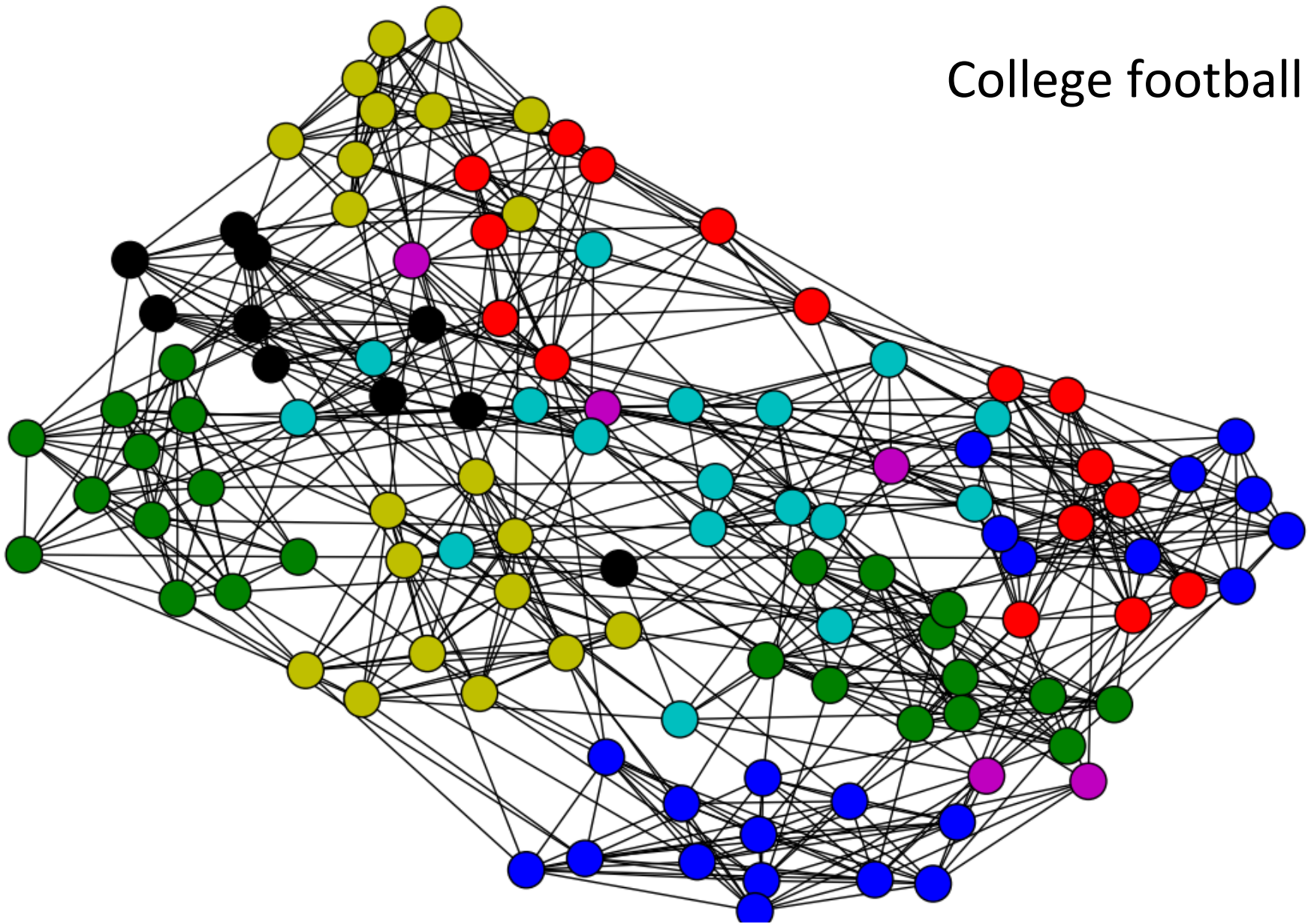
Schoolkids



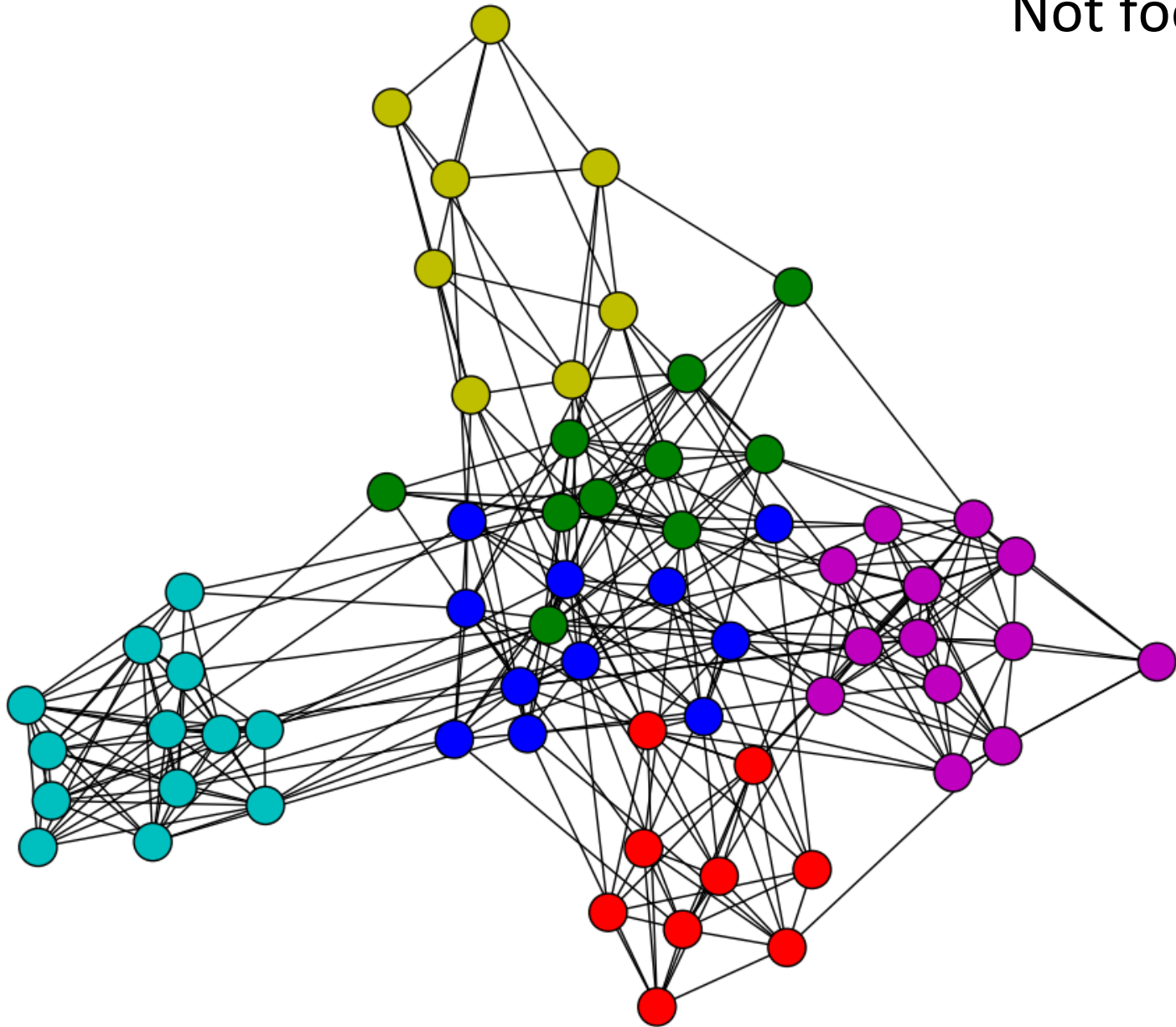
Heart study
participants



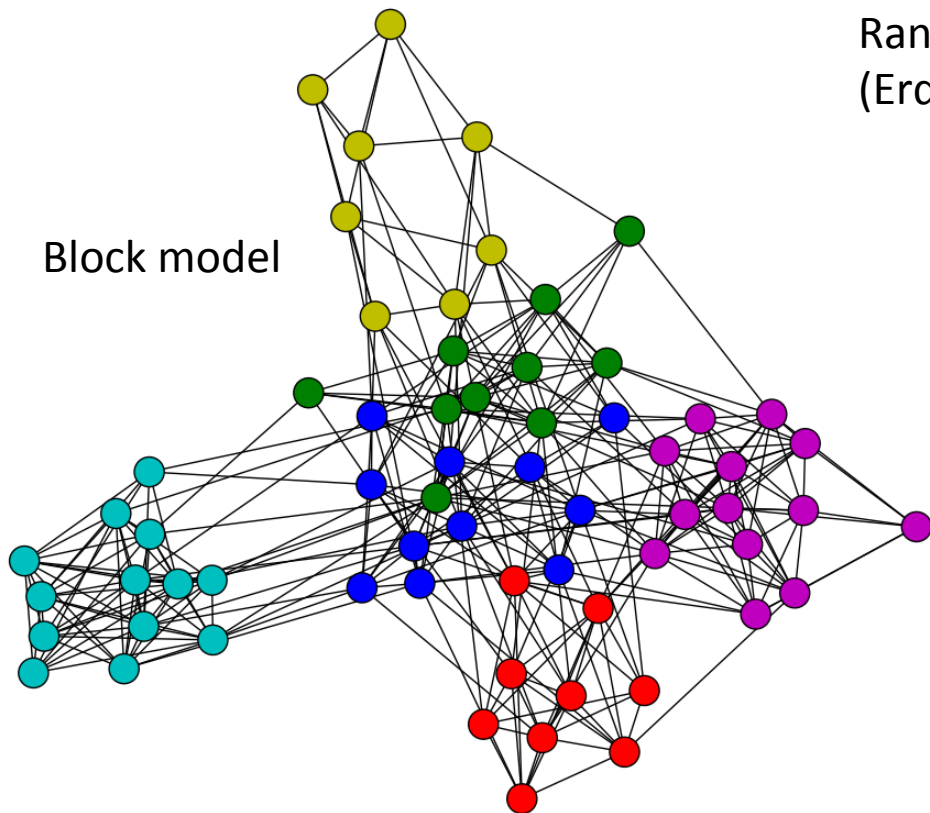
College football



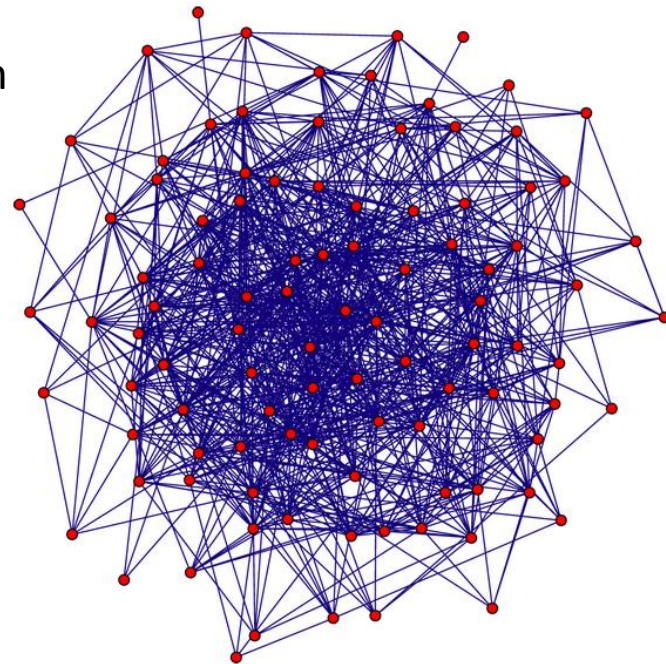
Not football



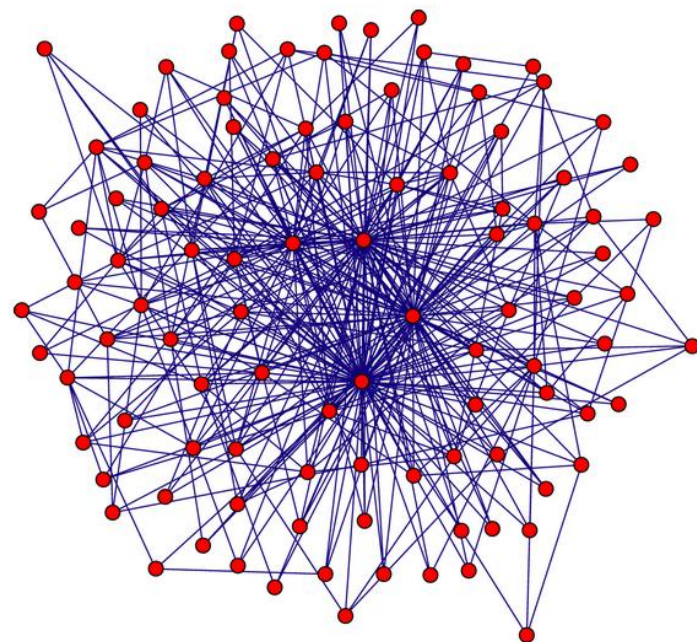
Block model



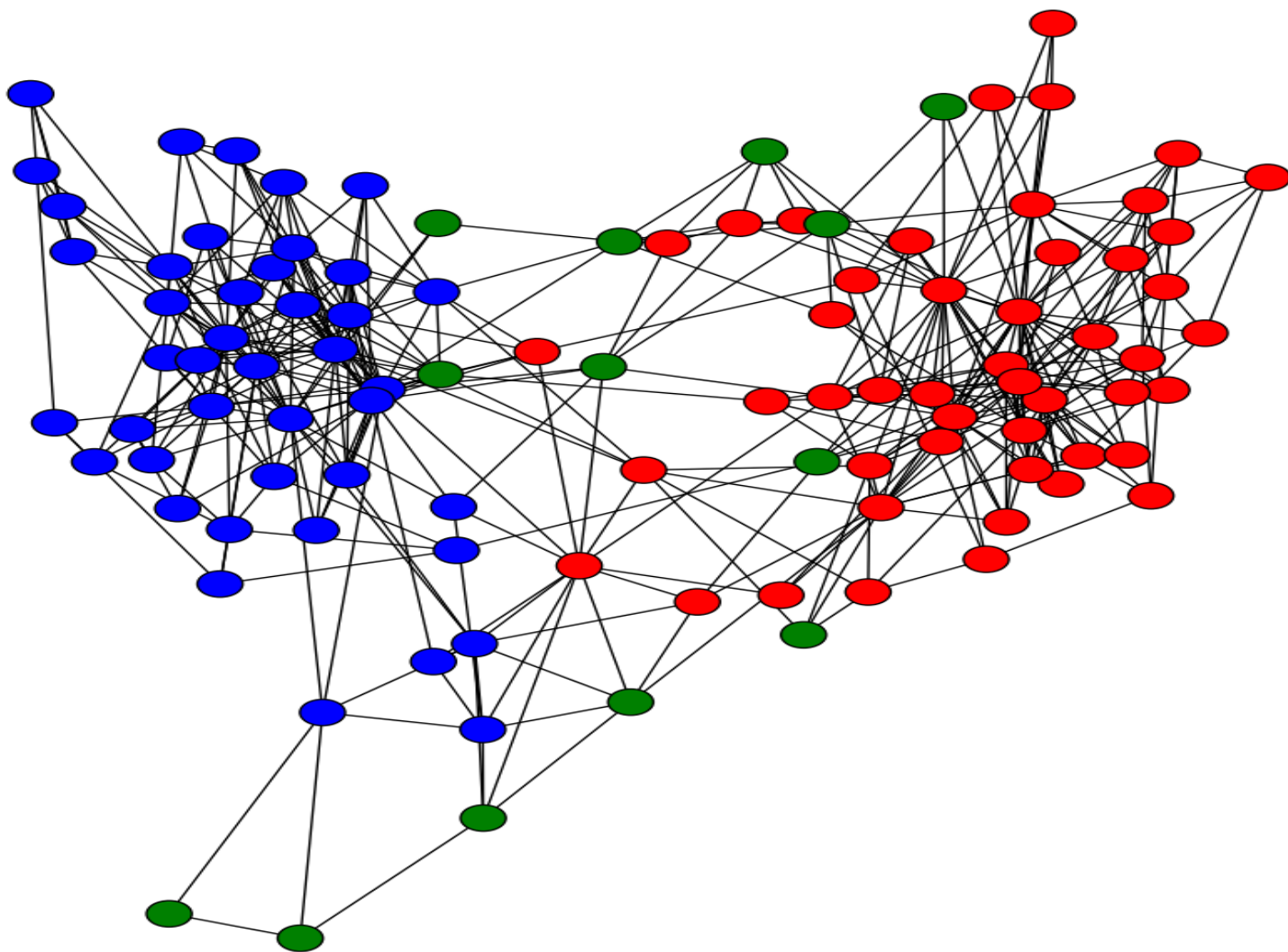
Random graph
(Erdos Renyi)

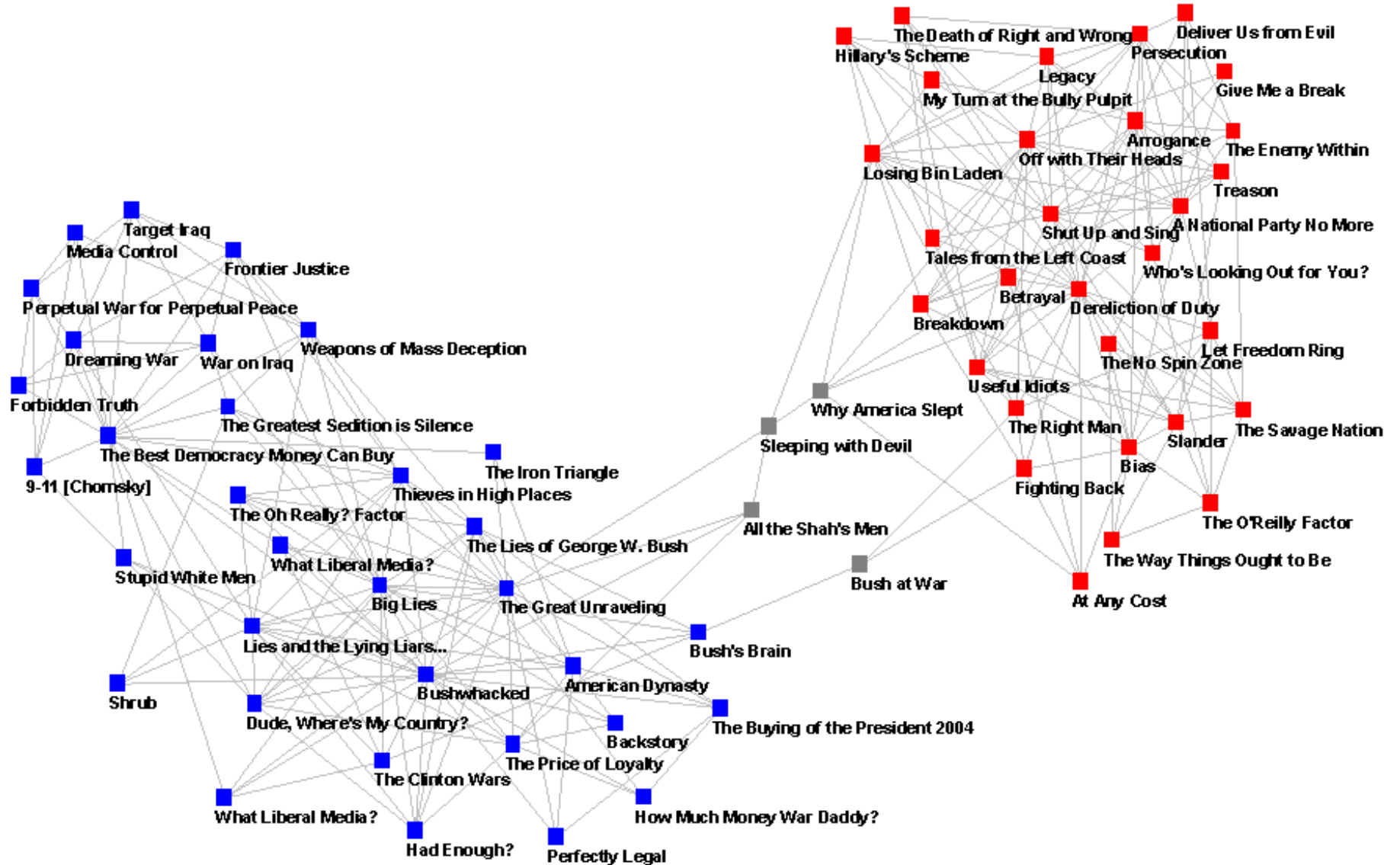


Preferential attachment
(Barabasi-Albert)

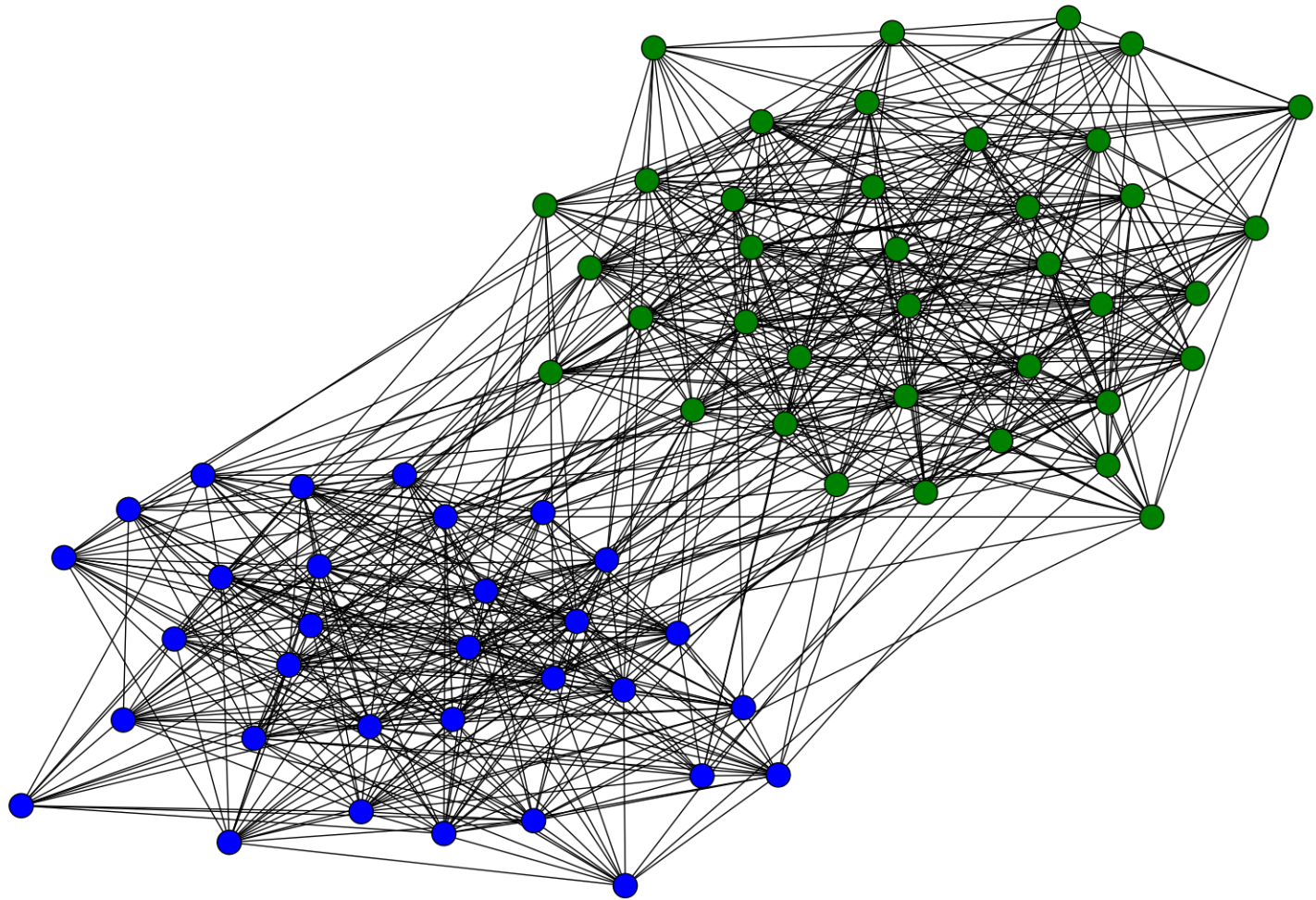


books

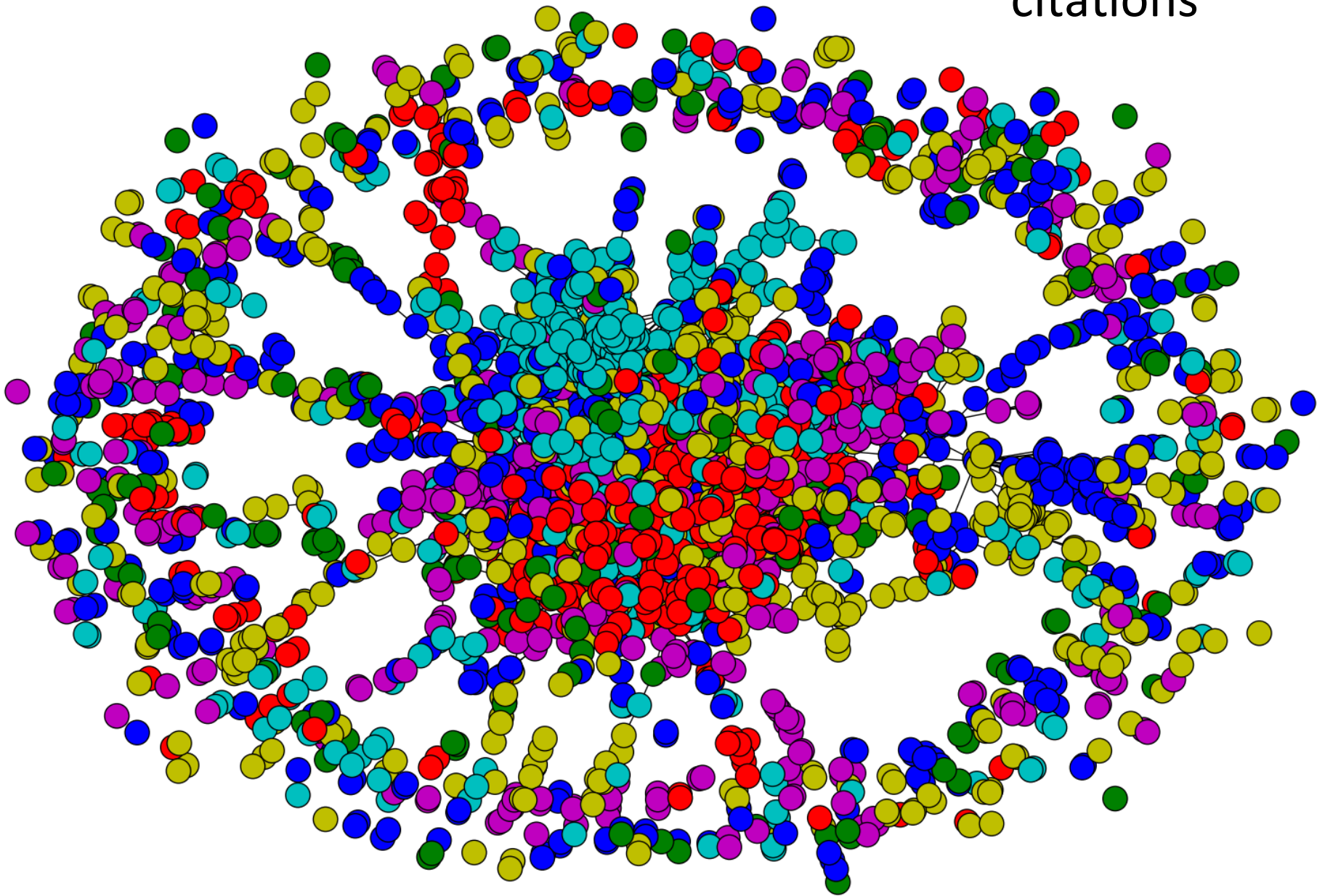


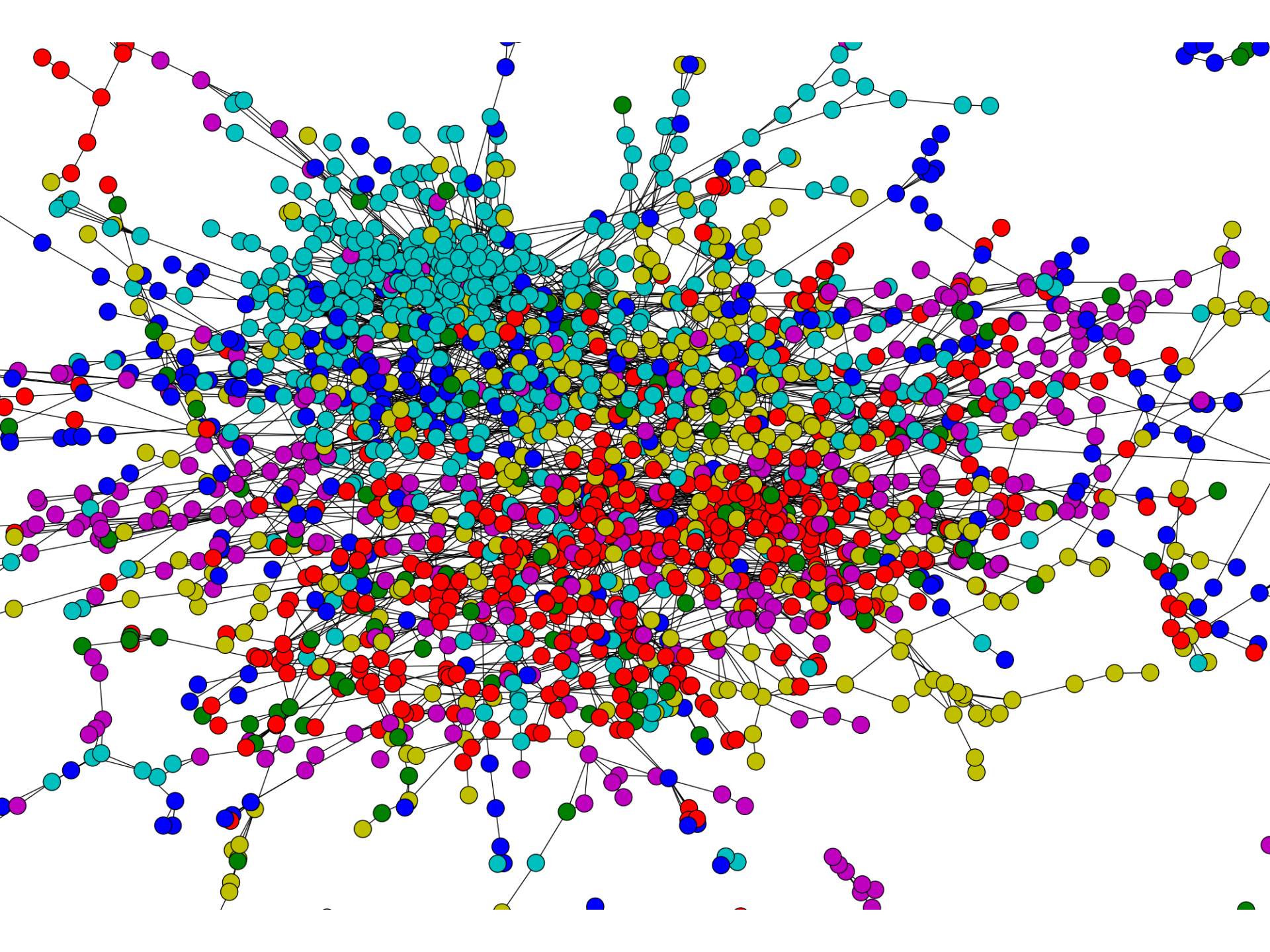


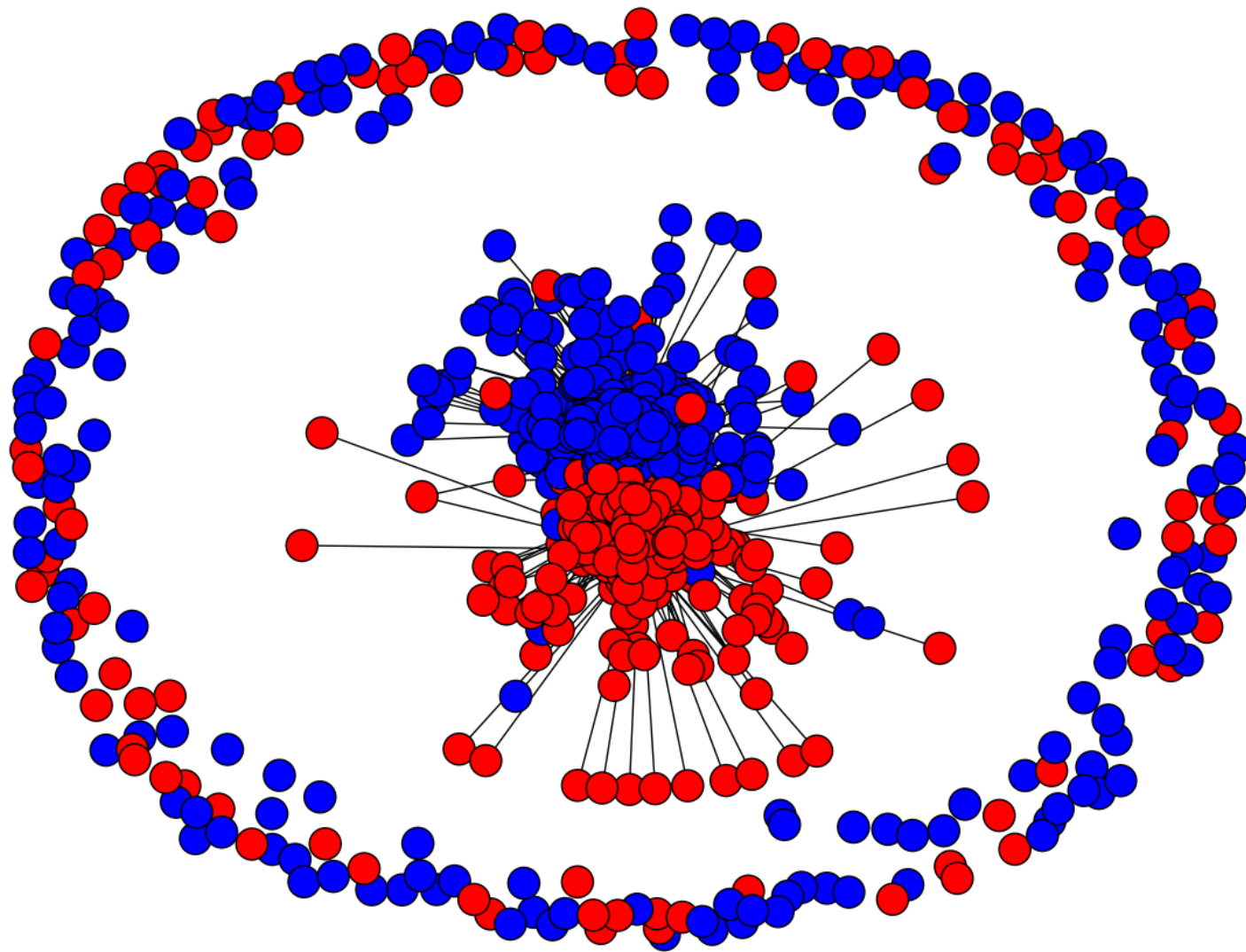
Not books

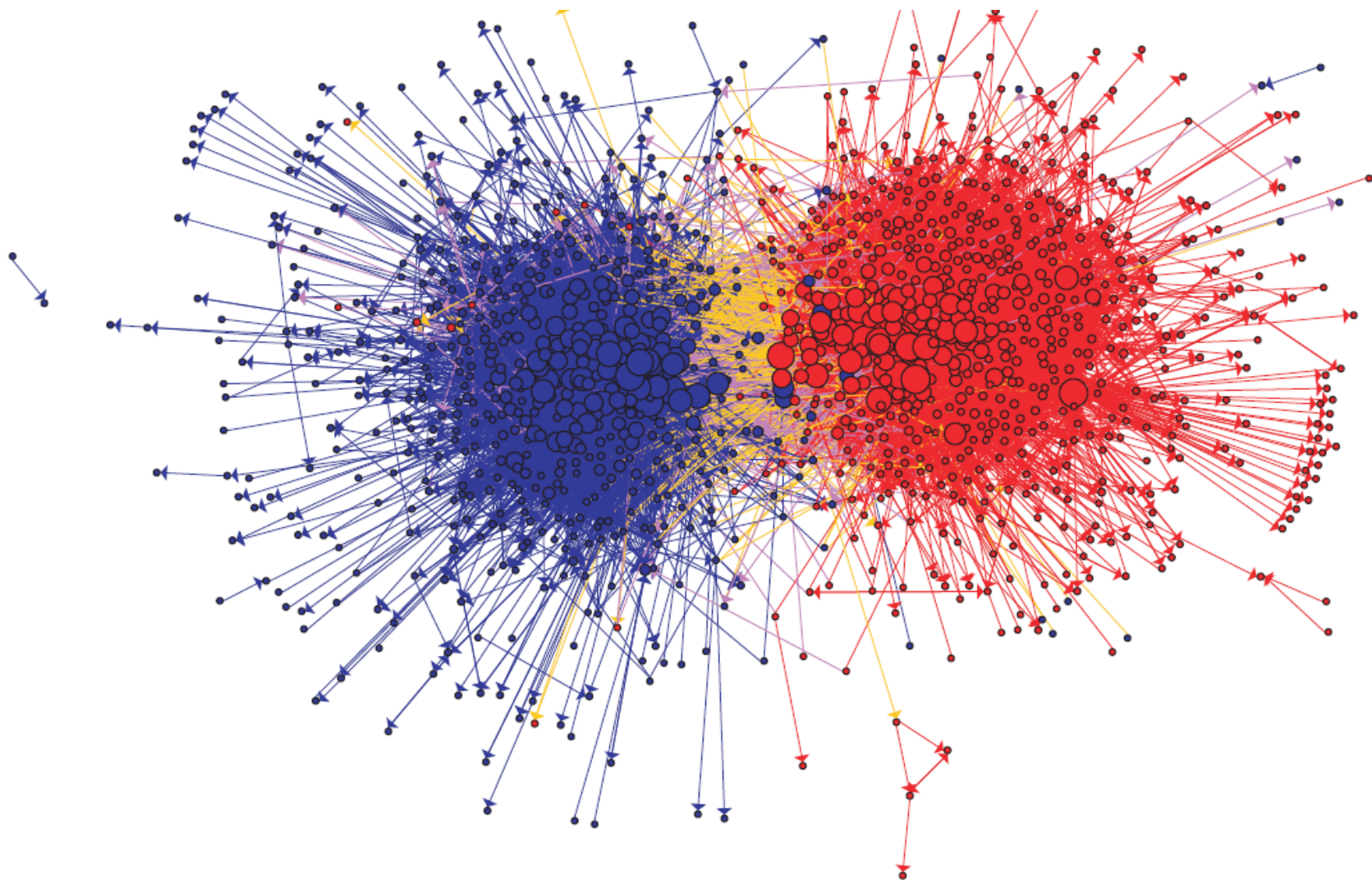


citations

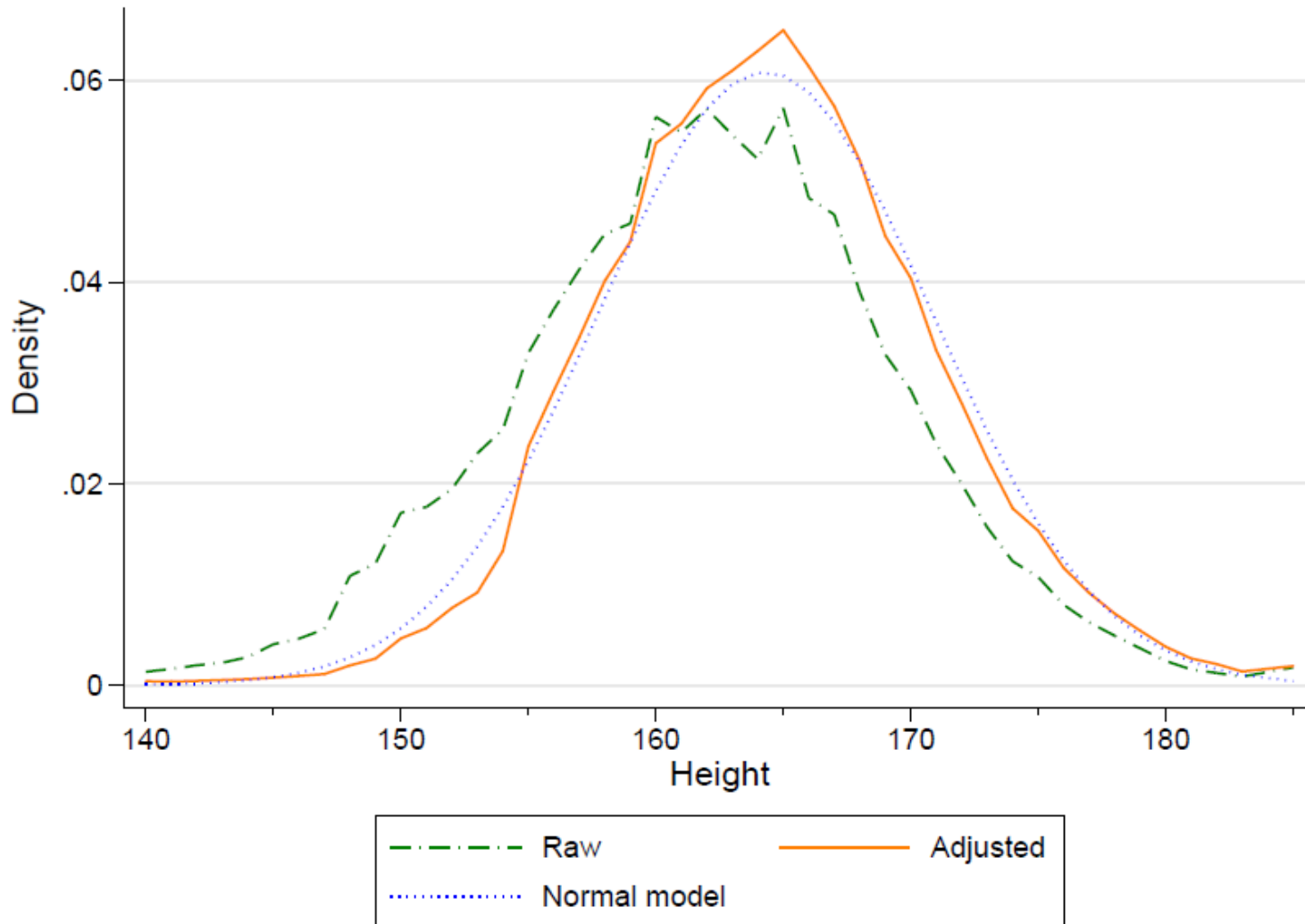








How do you model graphs?



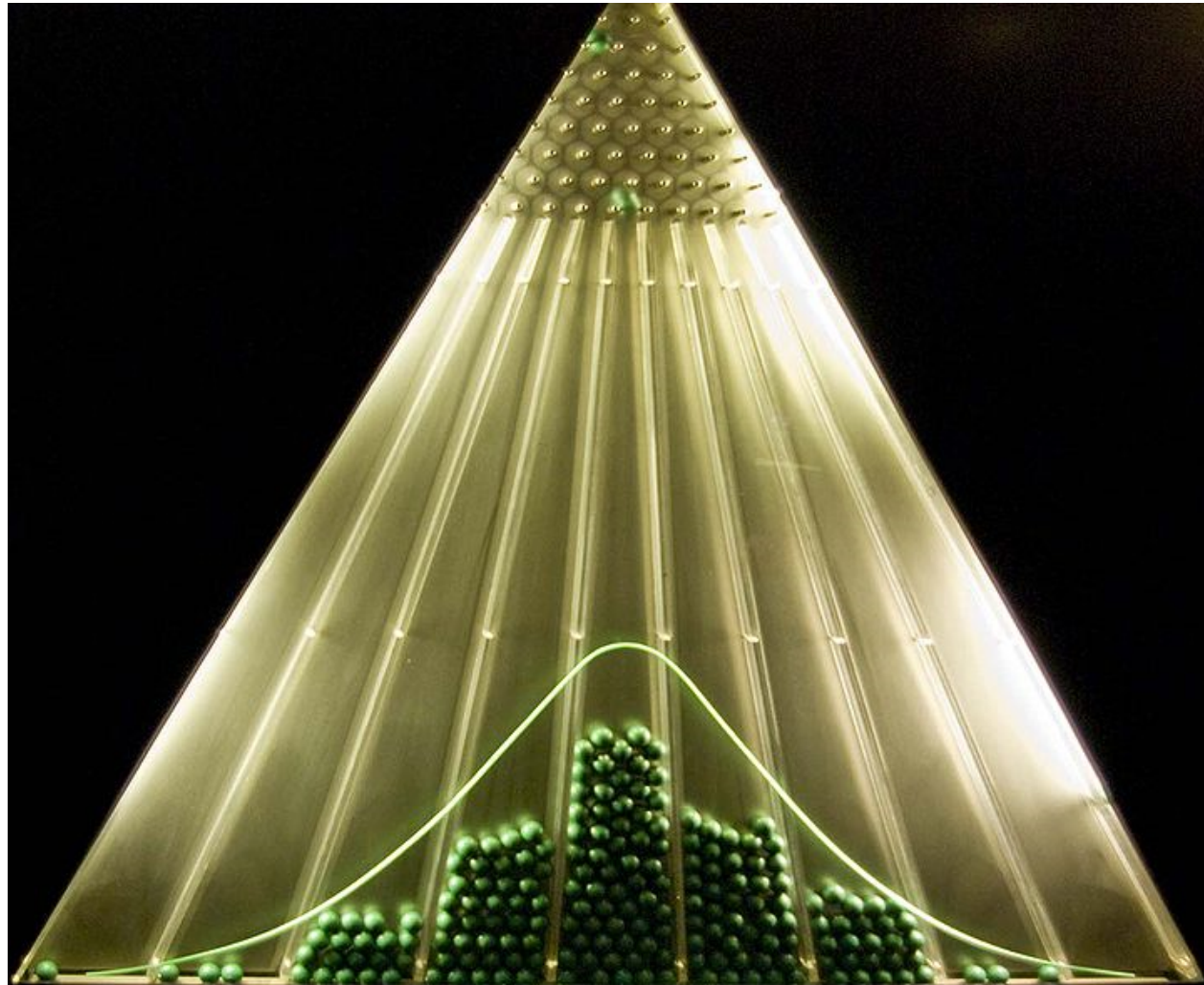
What are we trying to do?

Normal curve:

- Data to fit
- Underlying process that “explains” the data
- Closed-form parametric

form $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

- Only models a small part of the data



Graphs

- Set V of vertices/nodes v_1, \dots , set E of edges $(u, v), \dots$
 - Edges might be directed and/or weighted and/or labeled
- *Degree of v* is #edges touching v
 - *Indegree* and *Outdegree* for directed graphs
- *Path* is sequence of edges $(v_0, v_1), (v_1, v_2), \dots$
- *Geodesic path between u and v* is shortest path connecting them
 - *Diameter* is $\max_{u, v \in V} \{\text{length of geodesic between } u, v\}$
 - *Effective diameter* is 90th percentile
 - *Mean diameter* is over connected pairs
- *(Connected) component* is subset of nodes that are all pairwise connected via paths
- *Clique* is subset of nodes that are all pairwise connected via *edges*
- *Triangle* is a clique of size three

Graphs

- Some common properties of graphs:
 - Distribution of node degrees
 - Distribution of cliques (e.g., triangles)
 - Distribution of paths
 - Diameter (max shortest-path)
 - Effective diameter (90th percentile)
 - Connected components
 - ...
- Some types of graphs to consider:
 - Real graphs (social & otherwise)
 - Generated graphs:
 - Erdos-Renyi “Bernoulli” or “Poisson”
 - Watts-Strogatz “small world” graphs
 - Barabasi-Albert “preferential attachment”
 - ...

Erdos-Renyi graphs

- Take n nodes, and connect each pair with probability p
 - Mean degree is $z=p(n-1)$

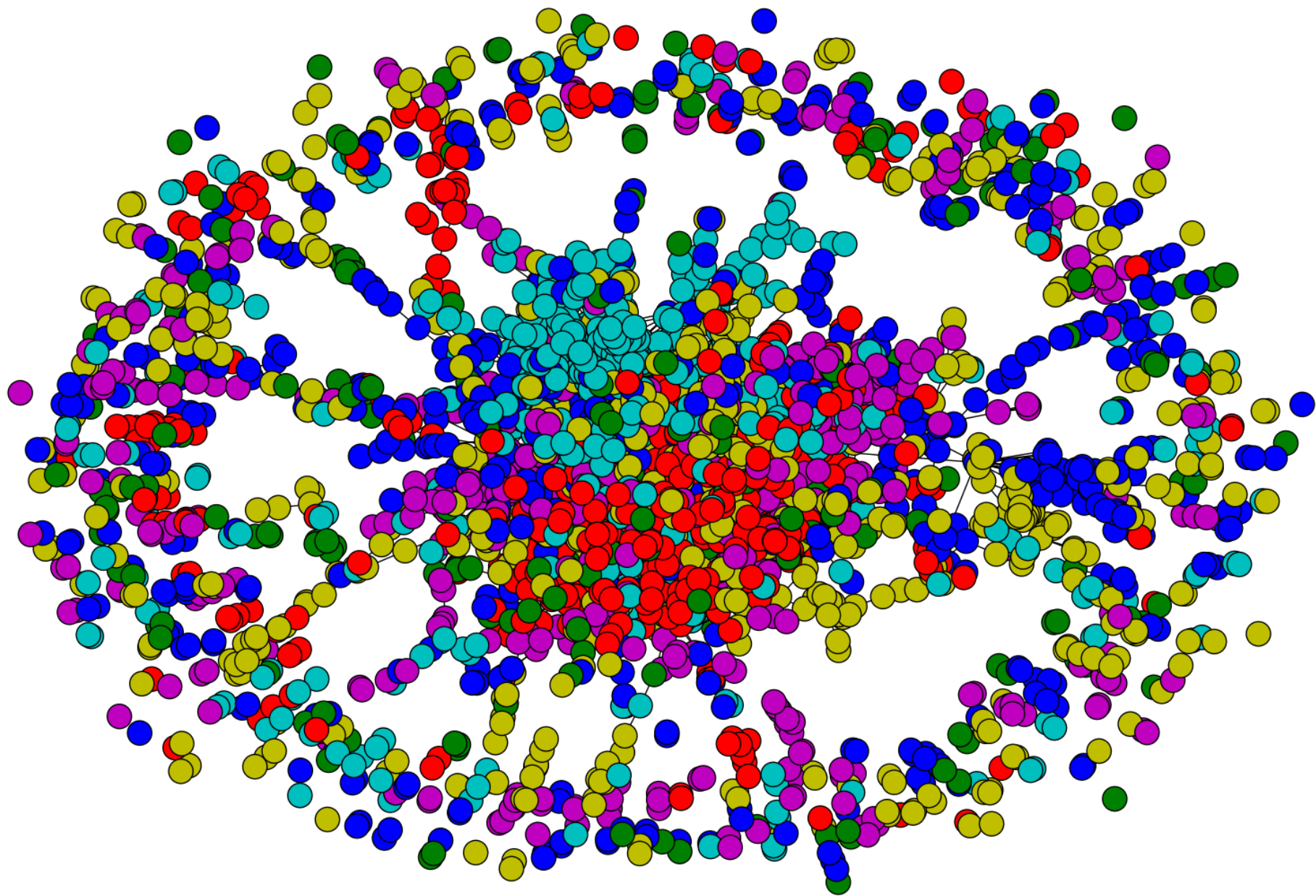
$$\Pr[\text{degree}(v) = k] = p_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{z^k e^{-z}}{k!} \quad \text{for fixed } z, \text{ large } n$$

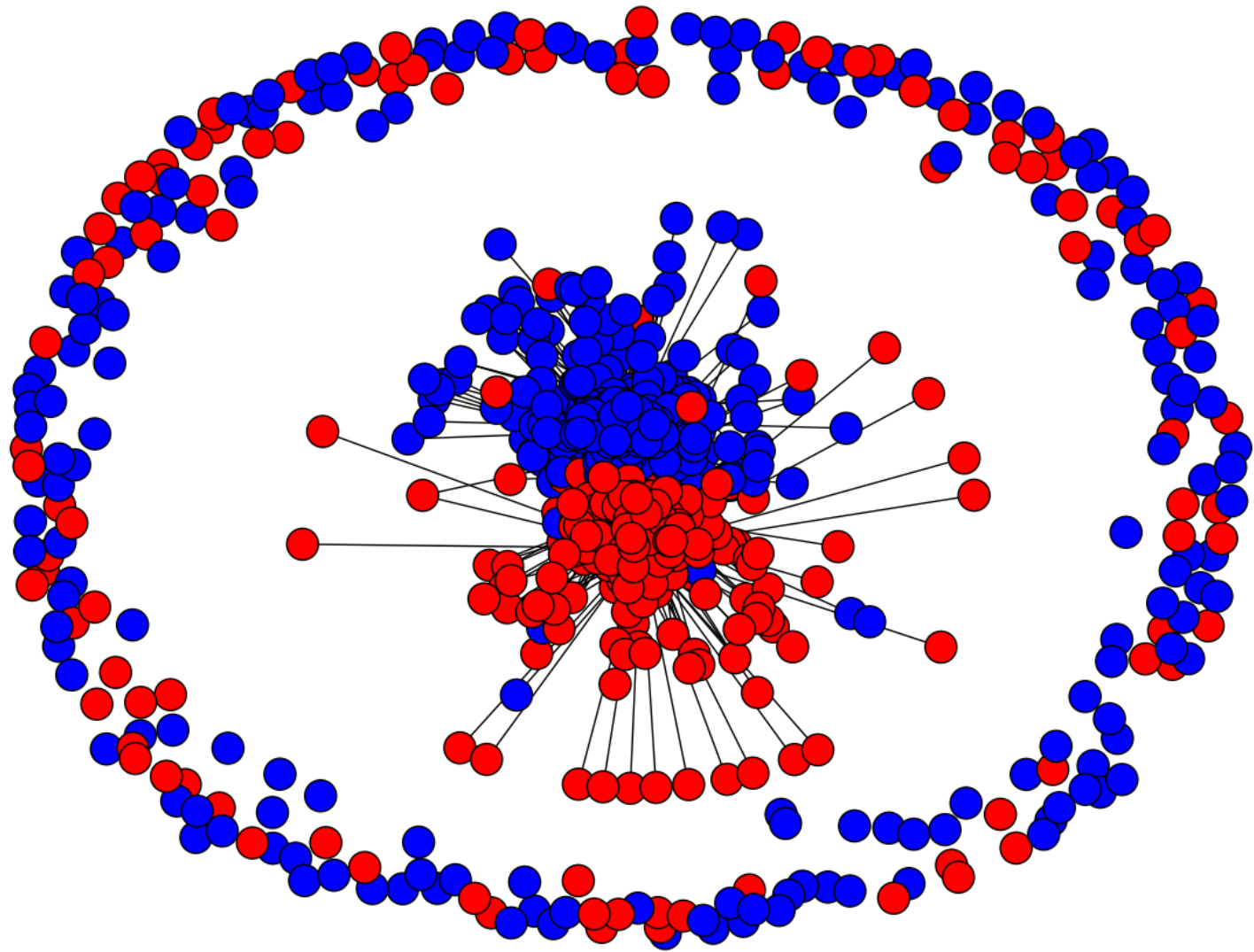
$$e^{-z}$$

$$\binom{n}{k} p^k \approx \frac{n^k p^k}{k!} = \frac{(np)^k}{k!} \approx \frac{z^k}{k!}$$

Erdos-Renyi graphs

- Take n nodes, and connect each pair with probability p
 - Mean degree is $z=p(n-1)$
 - Mean number of neighbors distance d from v is z^d
 - How large does d need to be so that $z^d \geq n$?
 - If $z > 1$, $d = \log(n)/\log(z)$
 - If $z < 1$, you can't do it
 - So:
 - *There tend to be either many small components ($z < 1$) or one large one ($z > 1$) giant connected component*
 - Another intuition:
 - If there are a two large connected components, then with high probability a few random edges will link them up.





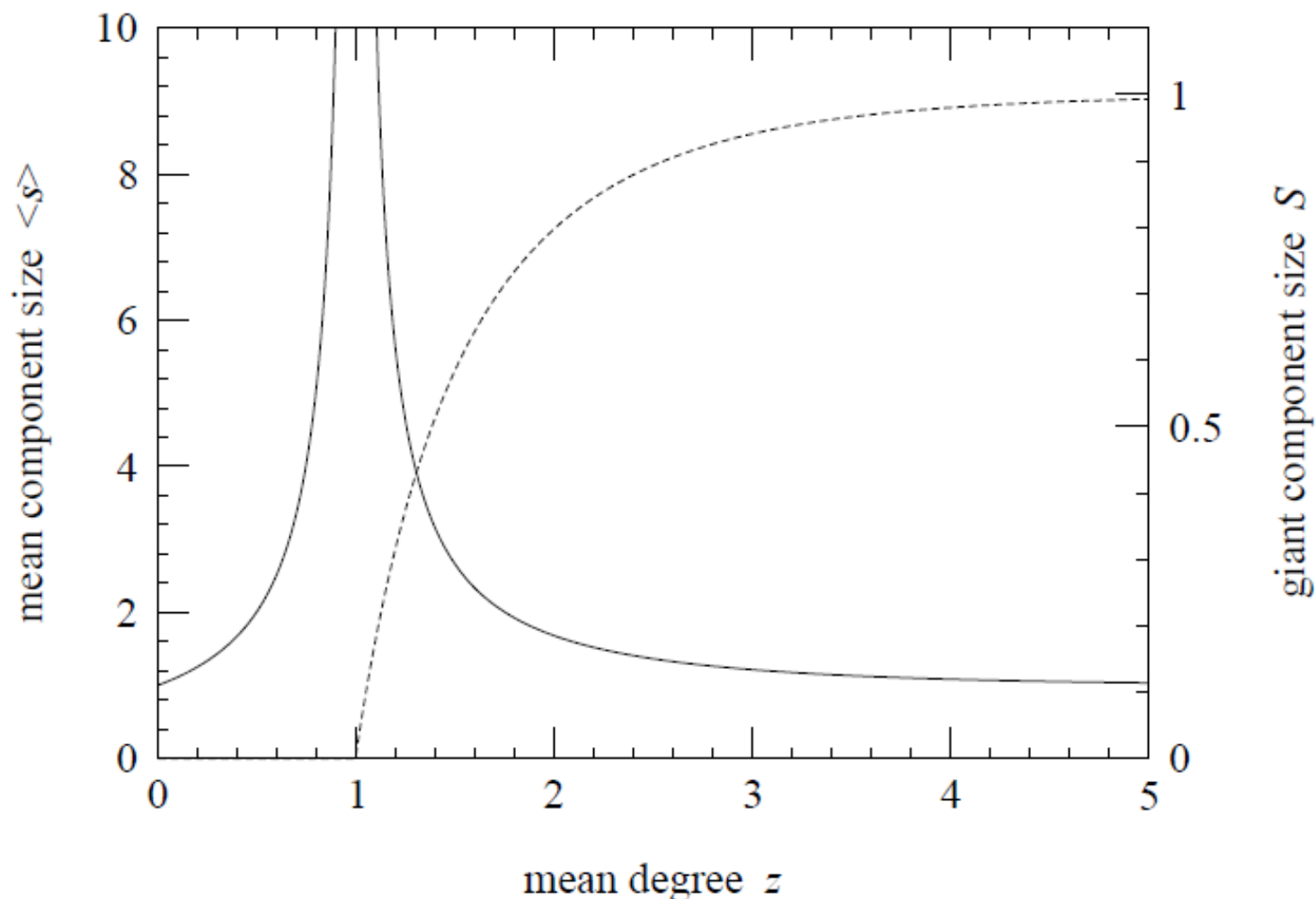


FIG. 10 The mean component size (solid line), excluding the giant component if there is one, and the giant component size (dotted line), for the Poisson random graph, Eqs. (20) and (21).

Erdos-Renyi graphs

- Take n nodes, and connect each pair with probability p
 - Mean degree is $z=p(n-1)$
 - Mean number of neighbors distance d from v is z^d
 - How large does d need to be so that $z^d \geq n$?
 - If $z > 1$, $d = \log(n)/\log(z)$
 - If $z < 1$, you can't do it
 - So:
 - *If $z > 1$, diameters tend to be small (relative to n)*

Sociometry, Vol. 32, No. 4. (Dec., 1969), pp. 425-443.

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals ($N=296$) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target

64 of 296
chains
succeed,
avg chain
length is
6.2

Illustrations of the Small World

- Millgram's experiment
- Erdős numbers
 - <http://www.ams.org/mathscinet/searchauthors.html>
- Bacon numbers
 - <http://oracleofbacon.org/>
- LinkedIn
 - <http://www.linkedin.com/>
 - Privacy issues: the whole network is *not* visible to all

THE DEAD WHAT?

WALK THE
EARTH!

I HAVE
TO GO.

$\sum_{i=1}^n i \frac{A_i}{i} \log(n)$
 $\sqrt{163}$ SCRIBBLE
SCRIBBLE

THE DEAD RETURN!
EVERYONE, QUICK,
GET YOUR NAMES
ON HERE!

MATH
DEPT

AT LAST!

I HOPE
THERE'S
TIME!

HURRHHH

CEMETERY

PAUL ERDŐS?

WE NEED YOU
TO SIGN THIS.

YES?

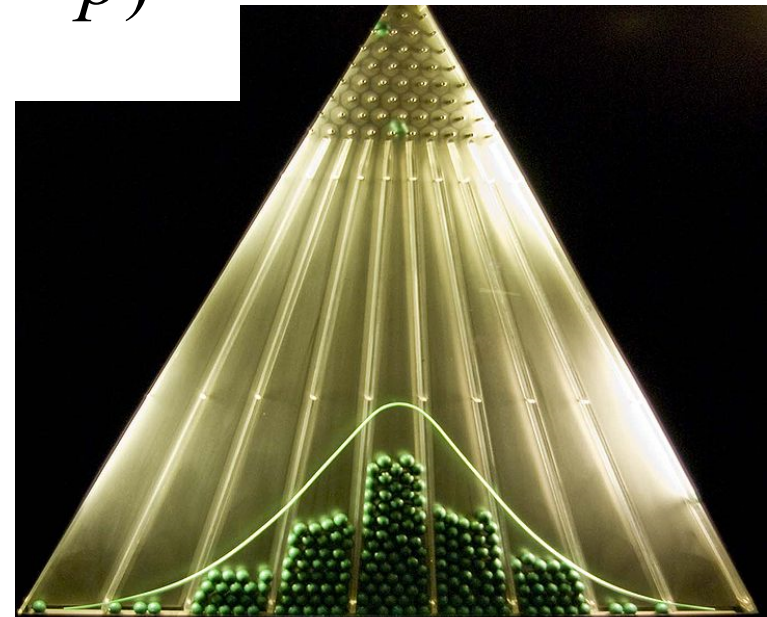
	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1		
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001
	sexual contacts	undirected	2 810				3.2		
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7		
	citation network	directed	783 339	6 716 198	8.57		3.0/–		
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080
	train routes	undirected	587	19 603	66.79	2.16	–		0.69
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28

Erdos-Renyi graphs

- Take n nodes, and connect each pair with probability p
 - Mean degree is $z=p(n-1)$

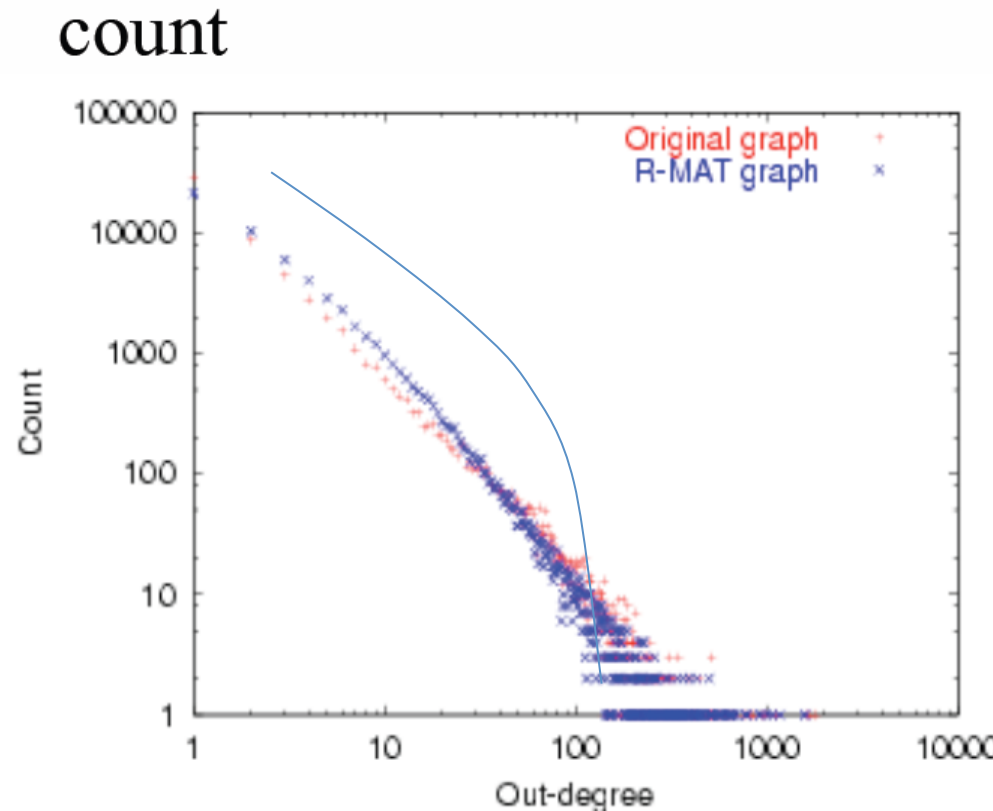
$$\Pr[\text{degree}(v) = k] = p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

This is usually *not* a good model of degree distribution in natural networks



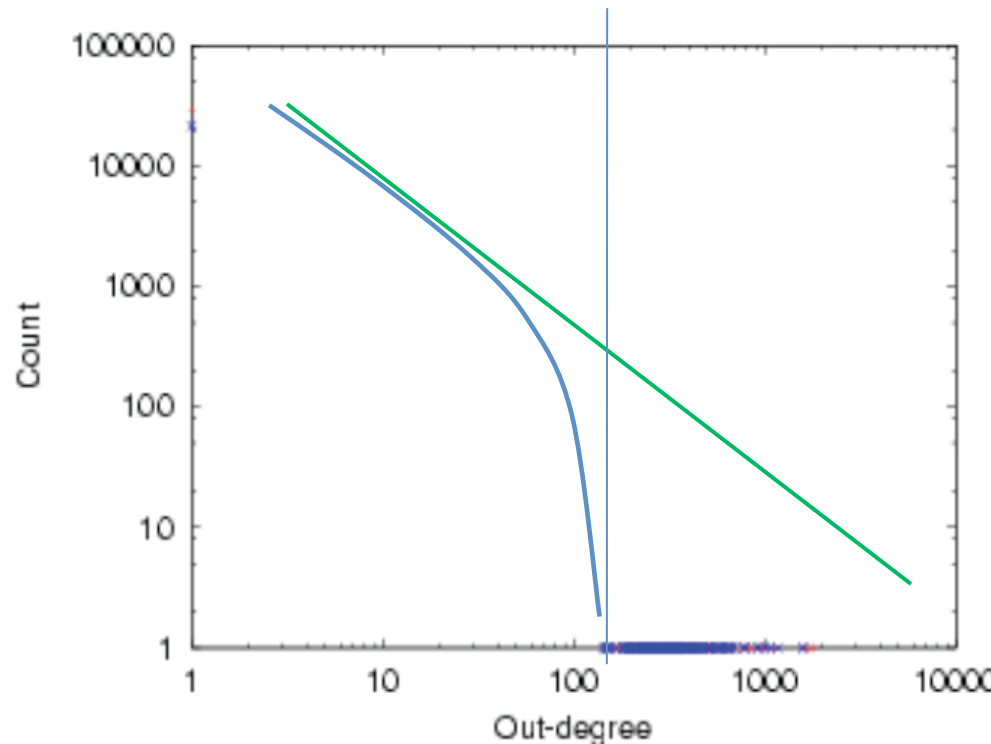
Degree distribution

- Plot cumulative degree
 - X axis is degree
 - Y axis is #nodes that have degree at least k
- Typically use a log-log scale
 - Straight lines are a power law; normal curve dives to zero at some point
 - Left: trust network in epinions web site from Richardson & Domingos



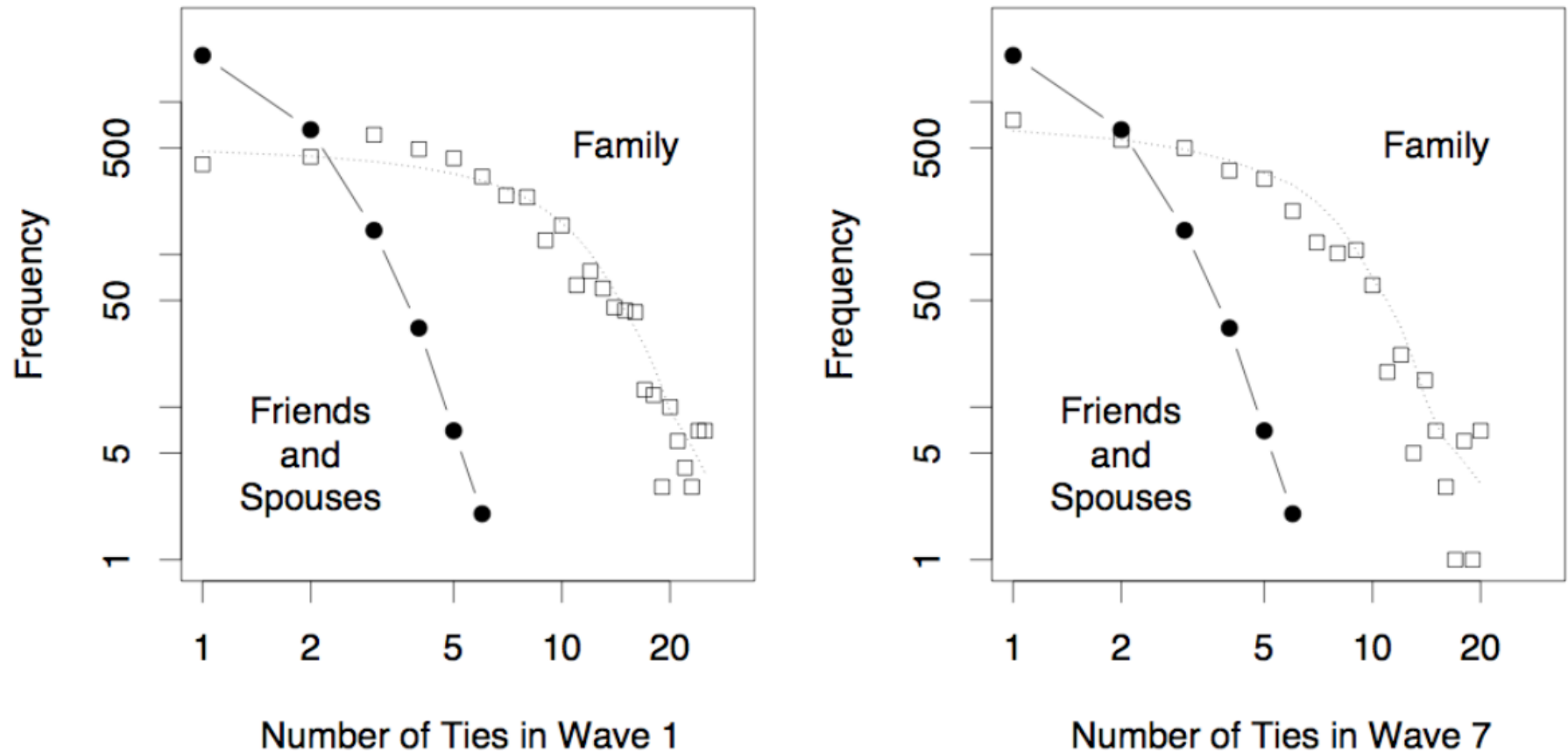
Degree distribution

- Plot cumulative degree
 - X axis is degree
 - Y axis is #nodes that have degree at least k
- Typically use a log-log scale
 - Straight lines are a power law; normal curve dives to zero at some point
 - This defines a “scale” for the network
 - Left: trust network in epinions web site from Richardson & Domingos



$$p_k \propto k^{-\alpha}$$

Figure S1: Degree Distribution of the FHS-Net



Friendship network in Framington Heart Study

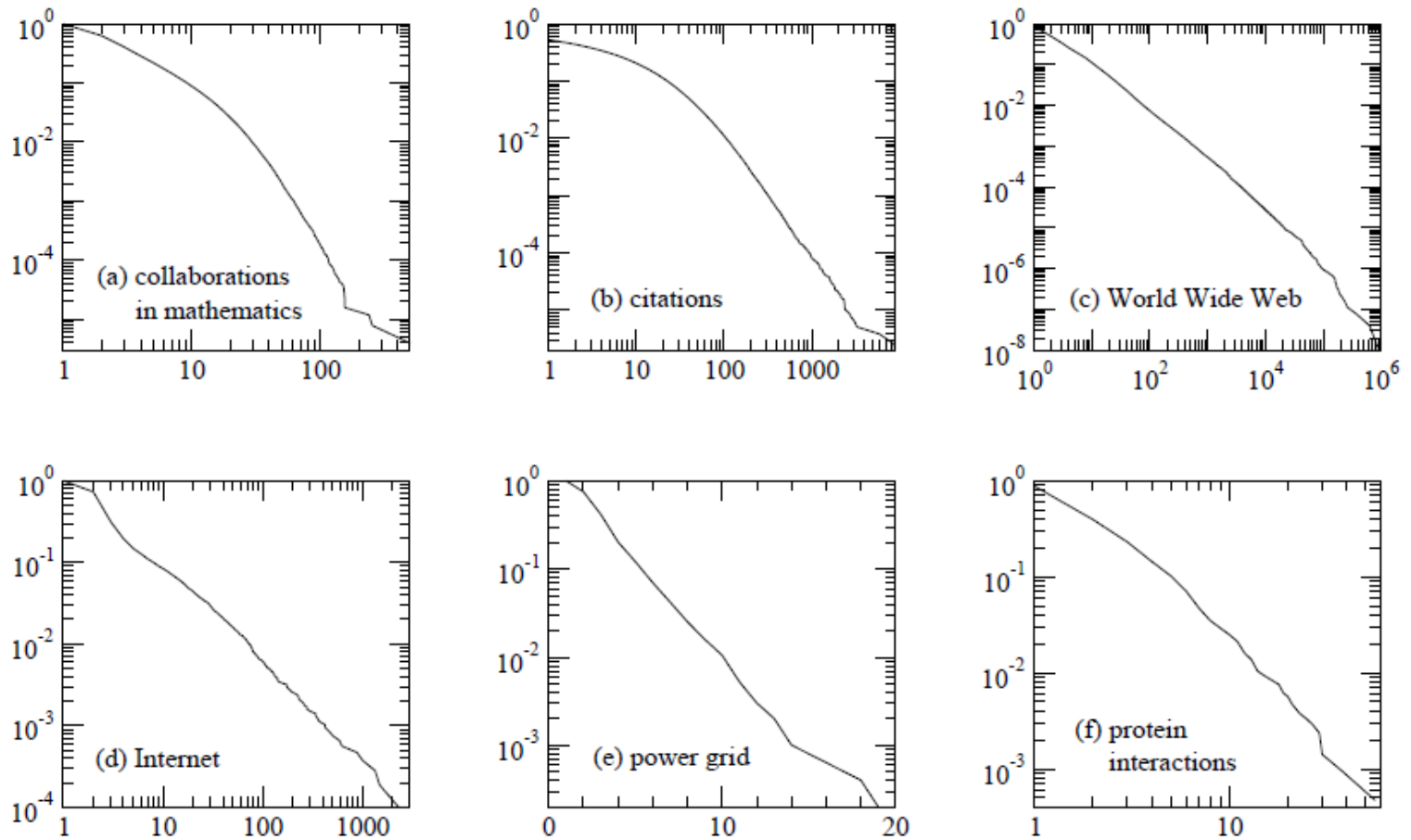


FIG. 6 Cumulative degree distributions for six different networks. The horizontal axis for each panel is vertex degree k (or in-degree for the citation and Web networks, which are directed) and the vertical axis is the cumulative probability distribution of degrees, i.e., the fraction of vertices that have degree greater than or equal to k . The networks shown are: (a) the collaboration network of mathematicians [182]; (b) citations between 1981 and 1997 to all papers cataloged by the Institute for Scientific Information [351]; (c) a 300 million vertex subset of the World Wide Web, *circa* 1999 [74]; (d) the Internet at the level of autonomous systems, April 1999 [86]; (e) the power grid of the western United States [416]; (f) the interaction network of proteins in the metabolism of the yeast *S. Cerevisiae* [212]. Of these networks, three of them, (c), (d) and (f), appear to have power-law degree distributions, as indicated by their approximately straight-line forms on the doubly logarithmic scales, and one (b) has a power-law tail but deviates markedly from power-law behavior for small degree. Network (e) has an exponential degree distribution (note the log-linear scales used in this panel) and network (a) appears to have a truncated power-law degree distribution of some type, or possibly two separate power-law regimes with different exponents.

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1		
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001
	sexual contacts	undirected	2 810				3.2		
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7		
	citation network	directed	783 339	6 716 198	8.57		3.0/–		
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080
	train routes	undirected	587	19 603	66.79	2.16	–		0.69
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28

Graphs

- Some common properties of graphs:
 - **Distribution of node degrees**
 - Distribution of cliques (e.g., triangles)
 - **Distribution of paths**
 - **Diameter** (max shortest-path)
 - Effective diameter (90th percentile)
 - **Connected components**
 - ...
- Some types of graphs to consider:
 - Real graphs (social & otherwise)
 - Generated graphs:
 - **Erdos-Renyi** “Bernoulli” or “Poisson”
 - Watts-Strogatz “small world” graphs
 - Barabasi-Albert “preferential attachment”
 - ...

Graphs

- Some common properties of graphs:

- **Distribution of node degrees: often scale-free**

- Distribution of cliques (e.g., triangles)

- **Distribution of paths**

- **Diameter** (max shortest-path)
- Effective **diameter** (90th percentile) **often small**
- **Connected components usually one giant CC**

- ...

- Some types of graphs to consider:

- Real graphs (social & otherwise)

- Generated graphs:

- **Erdos-Renyi**
“Bernoulli” or “Poisson”

- Watts-Strogatz “small world” graphs

- Barabosi-Albert
“preferential attachment”
generates scale-free graphs

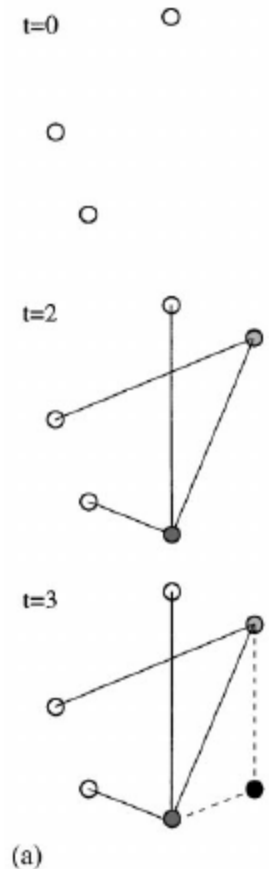
- ...

Barabasi-Albert Networks

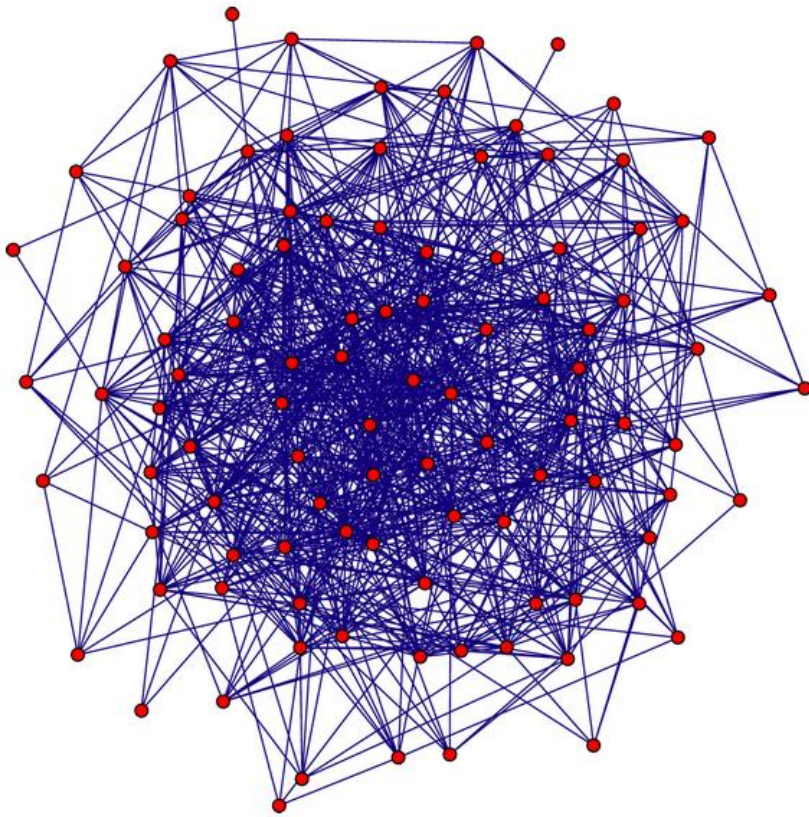
- *Science* **286** (1999)
- Start from a small number of node, add a new node with m links
- **Preferential Attachment**
 - Probability of these links to connect to existing nodes is proportional to the node's degree

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

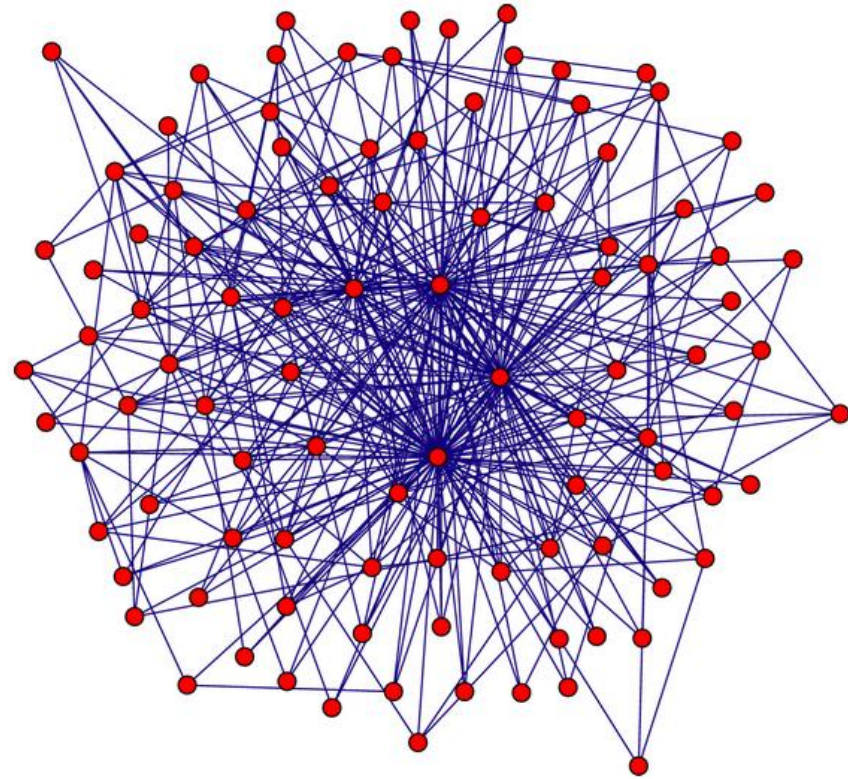
- 'Rich gets richer'
- This creates 'hubs': few nodes with very large degrees



Random graph
(Erdos Renyi)

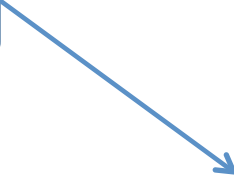


Preferential attachment
(Barabasi-Albert)



Stopped here 9/20

Graphs

- Some common properties of graphs:
 - **Distribution of node degrees: often scale-free**
 - Distribution of cliques (e.g., triangles)
 - **Distribution of paths**
 - **Diameter** (max shortest-path)
 - Effective **diameter** (90th percentile) **often small**
 - **Connected components usually one giant CC**
 - ...
 - Some types of graphs to consider:
 - Real graphs (social & otherwise)
 - Generated graphs:
 - **Erdos-Renyi “Bernoulli” or “Poisson”**
 - Watts-Strogatz “small world” graphs
 - Barabasi-Albert “preferential attachment” **generates scale-free graphs**
 - ...
- 

Homophily

- One definition: excess edges between similar nodes
 - E.g., assume nodes are male and female and $\Pr(\text{male})=p$, $\Pr(\text{female})=q$.
 - Is $\Pr(\text{gender}(u) \neq \text{gender}(v) \mid \text{edge}(u,v)) \geq 2pq$?
- Another def'n: excess edges between common neighbors of v

$$CC(v) = \frac{\# \text{triangles connected to } v}{\# \text{pairs connected to } v}$$

$$CC(V, E) = \frac{1}{|V|} \sum_v CC(v)$$

Homophily

- Another def'n: excess edges between common neighbors of v

$$CC(v) = \frac{\# \text{triangles connected to } v}{\# \text{pairs connected to } v}$$

$$CC(V, E) = \frac{1}{|V|} \sum_v CC(v)$$

$$CC'(V, E) = \frac{\# \text{triangles in graph}}{\# \text{length 3 paths in graph}}$$

Homophily

- In a random Erdos-Renyi graph:

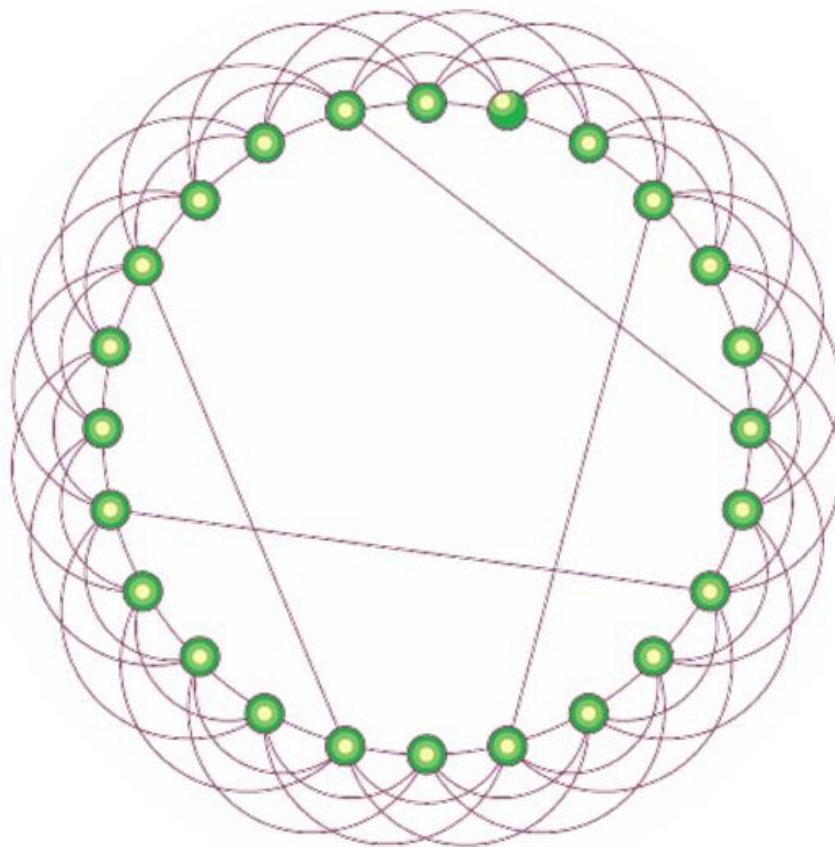
$$CC'(V, E) = \frac{\text{\# triangles in graph}}{\text{\# length 3 paths in graph}} \approx \frac{1}{n} \text{ for large } n$$

In natural graphs two of your mutual friends might well be friends:

- Like you they are both in the same class (club, field of CS, ...)
- You introduced them

Watts-Strogatz model

- Start with a ring
- Connect each node to k nearest neighbors
 - \rightarrow homophily
- Add some random shortcuts from one point to another
 - \rightarrow small diameter
- Degree distribution *not* scale-free
- Generalizes to d dimensions



	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1		
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001
	sexual contacts	undirected	2 810				3.2		
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7		
	citation network	directed	783 339	6 716 198	8.57		3.0/–		
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080
	train routes	undirected	587	19 603	66.79	2.16	–		0.69
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28

A big question

- Which is cause and which is effect?
 - Do birds of a feather flock together?
 - Do you change your behavior based on the behavior of your peers?

In natural graphs two of your mutual friends might well be friends:

- Like you they are both in the same class (club, field of CS, ...)
- You introduced them

A big question

- Which is cause and which is effect?
 - Do birds of a feather flock together?
 - Do you change your behavior based on the behavior of your peers?
- How can you tell?
 - Look at when links are added and see what patterns emerge:
 $\text{Pr}(\text{new link btwn } u \text{ and } v \mid \# \text{common friends})$

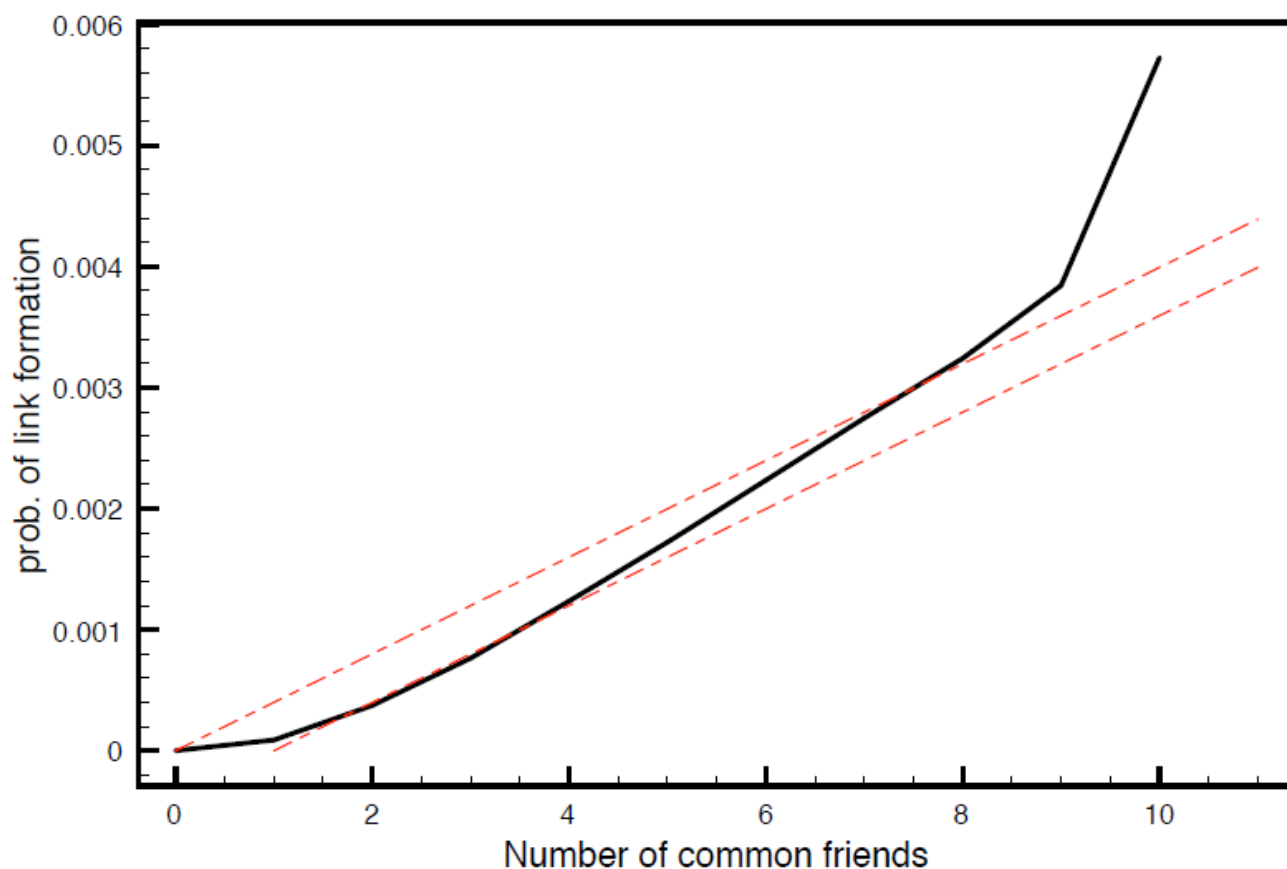


Figure 4.9: Quantifying the effects of triadic closure in an e-mail dataset [256]. The curve determined from the data is shown in the solid black line; the dotted curves show a comparison to probabilities computed according to two simple baseline models in which common friends provide independent probabilities of link formation.

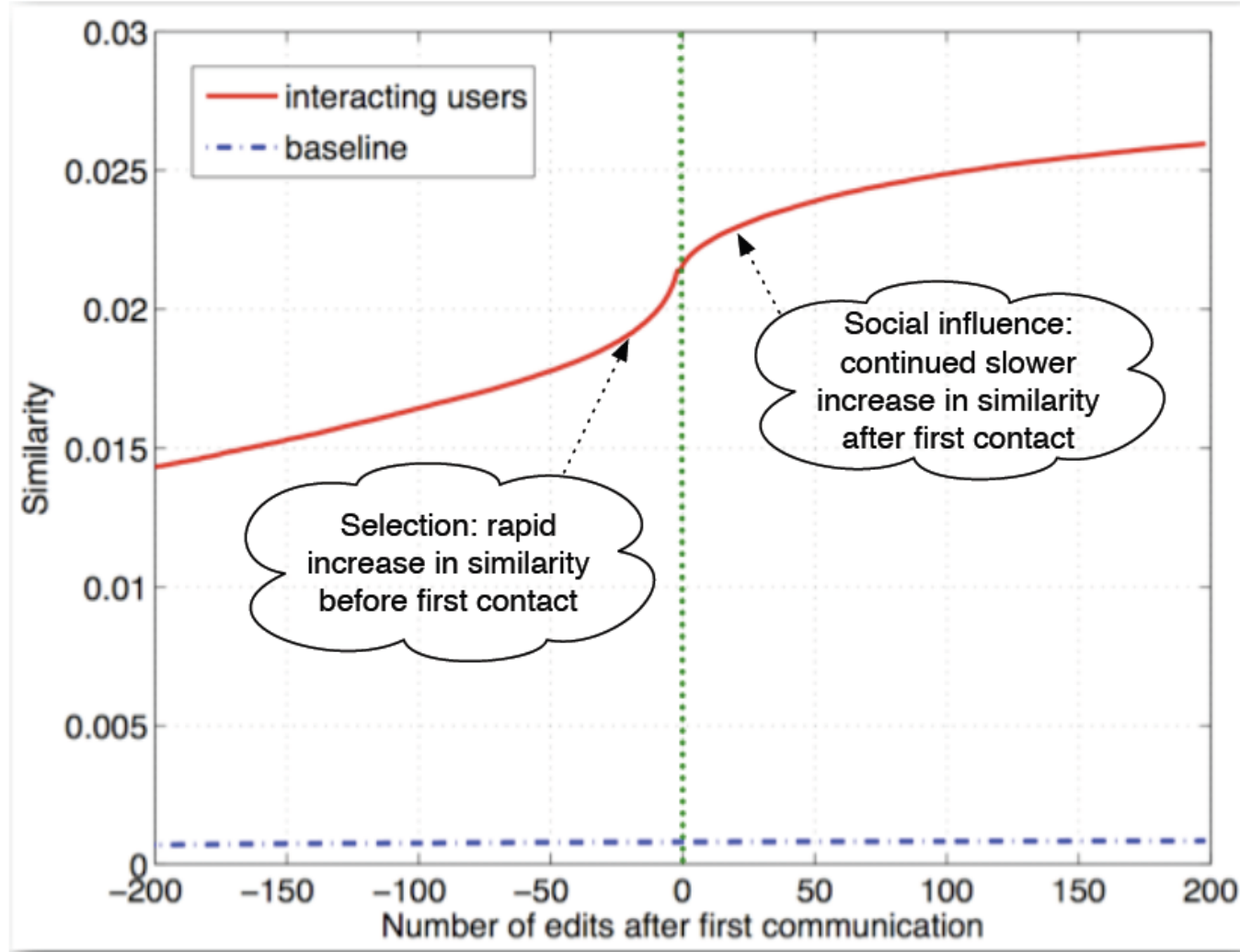
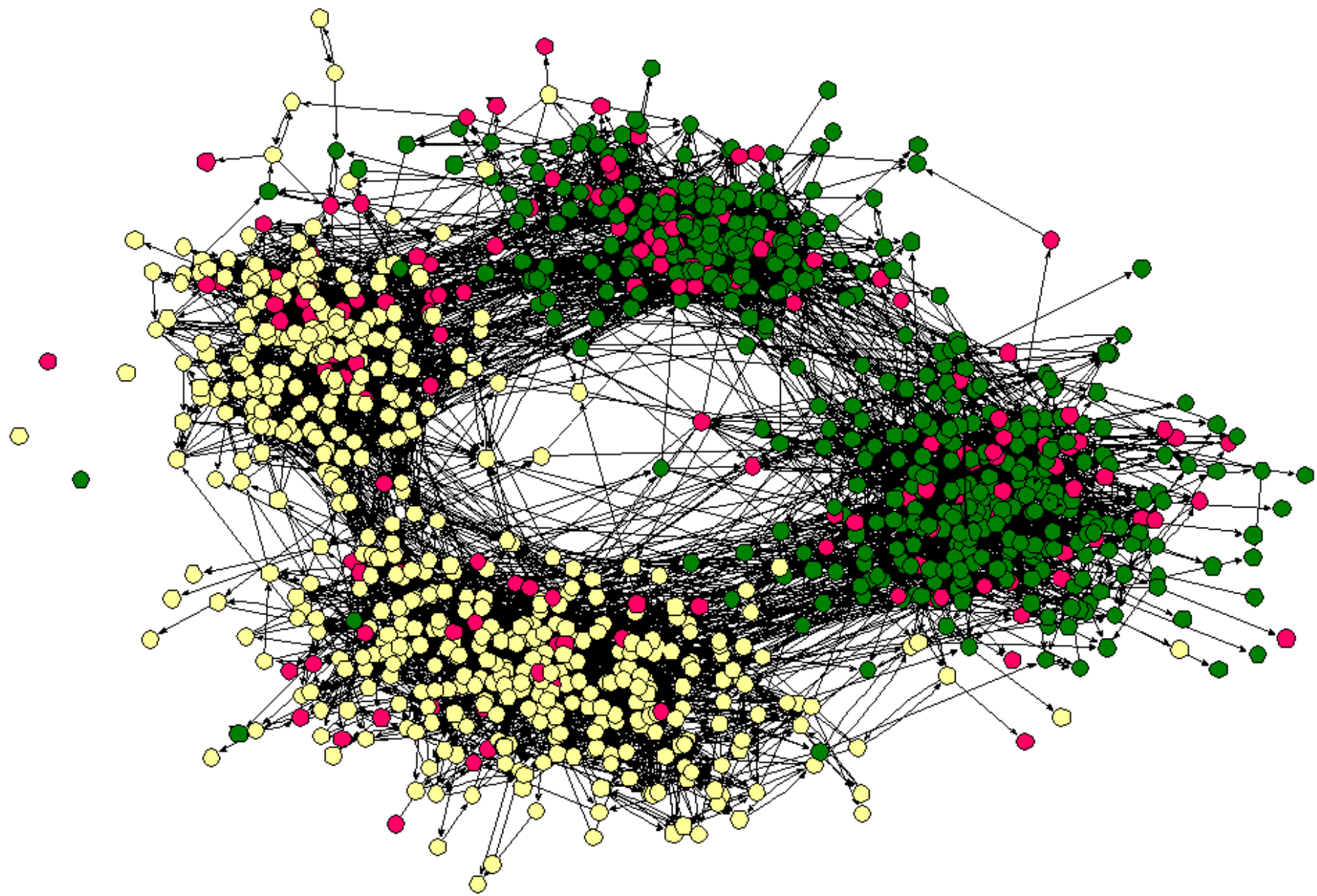
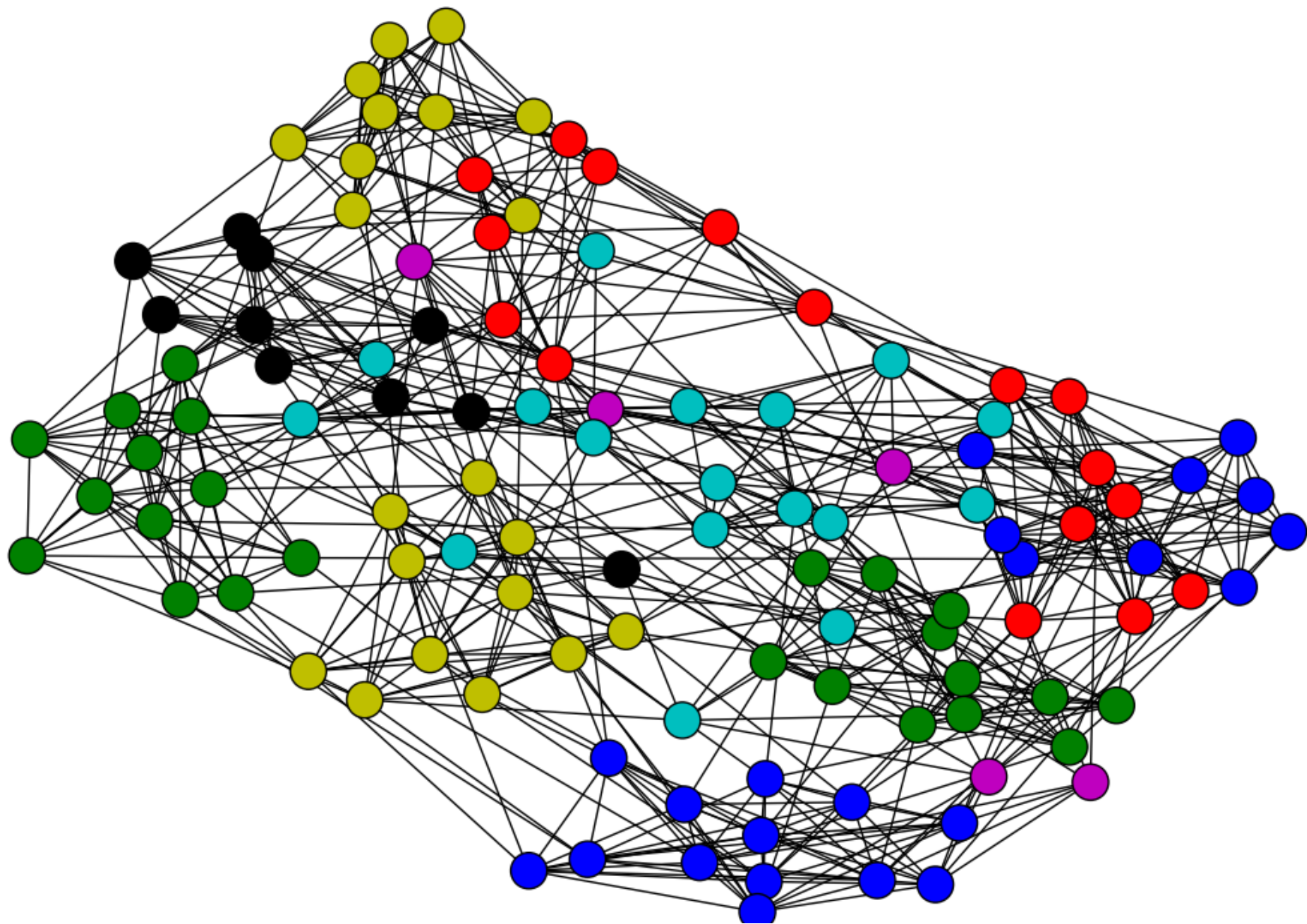


Figure 4.12: The average similarity of two editors on Wikipedia, relative to the time (0) at which they first communicated [121]. Time, on the x -axis, is measured in discrete units, where each unit corresponds to a single Wikipedia action taken by either of the two editors. The curve increases both before and after the first contact at time 0, indicating that both selection and social influence play a role; the increase in similarity is steepest just before time 0.





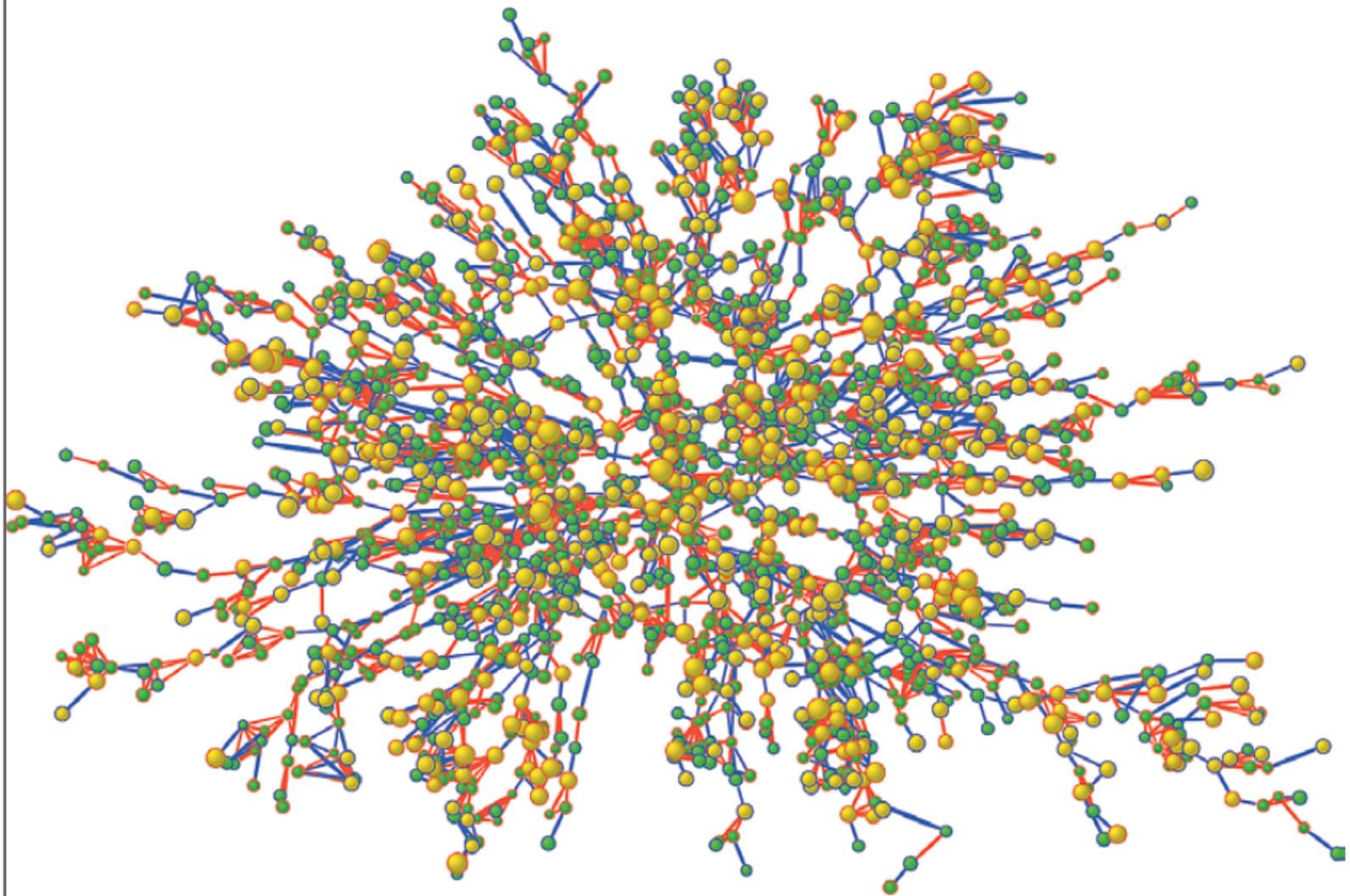


Figure 1. Largest Connected Subcomponent of the Social Network in the Framingham Heart Study in the Year 2000.

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.

Final example: spatial segregation

- <http://names.mappinglondon.co.uk/>
- How picky do people have to be about their neighbors for homophily to arise?
- Imagine a grid world where
 - Agents are red or blue
 - Agents move to a random location if they are unhappy
 - Agents are happy unless $< k$ neighbors are the same color they are ($k =$
 - i.e., they prefer not to be in a small minority
 - What's the result over time?
- <http://ccl.northwestern.edu/netlogo/models/Segregation>