Network and Social Network Theory

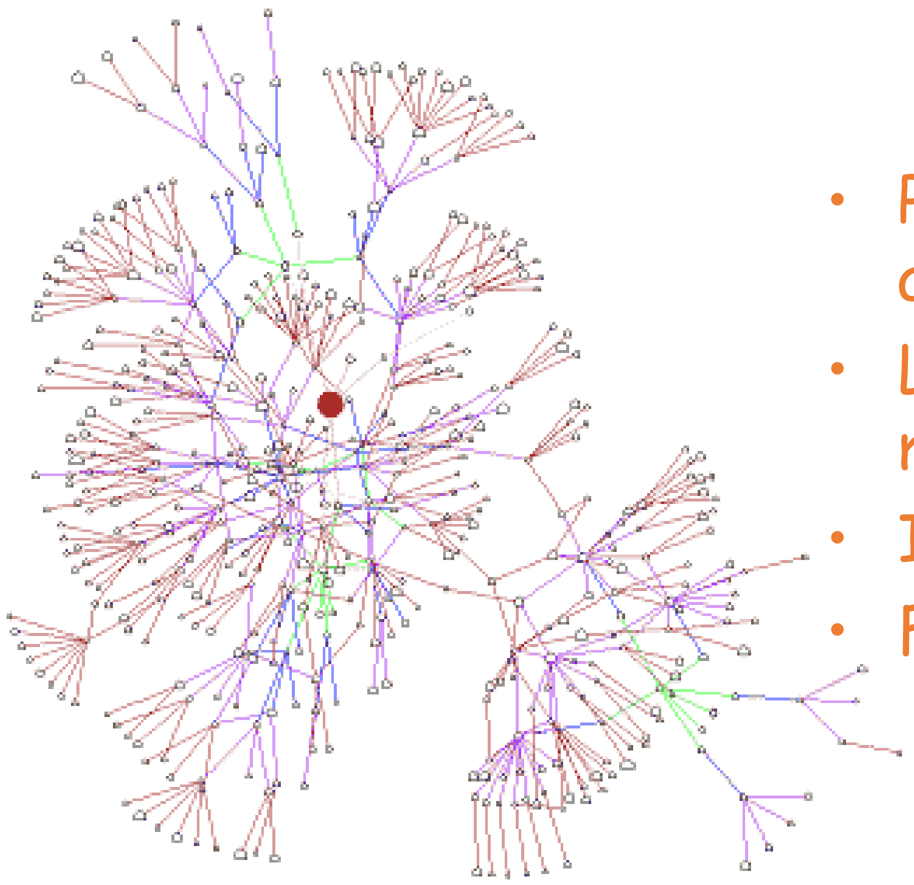
Xiaoming Fu

Acknowledgement: Prof. Lee Giles (Penn State University), Prof. Jar-Der Luo (Tsinghua University)

The Networked Nature of Society

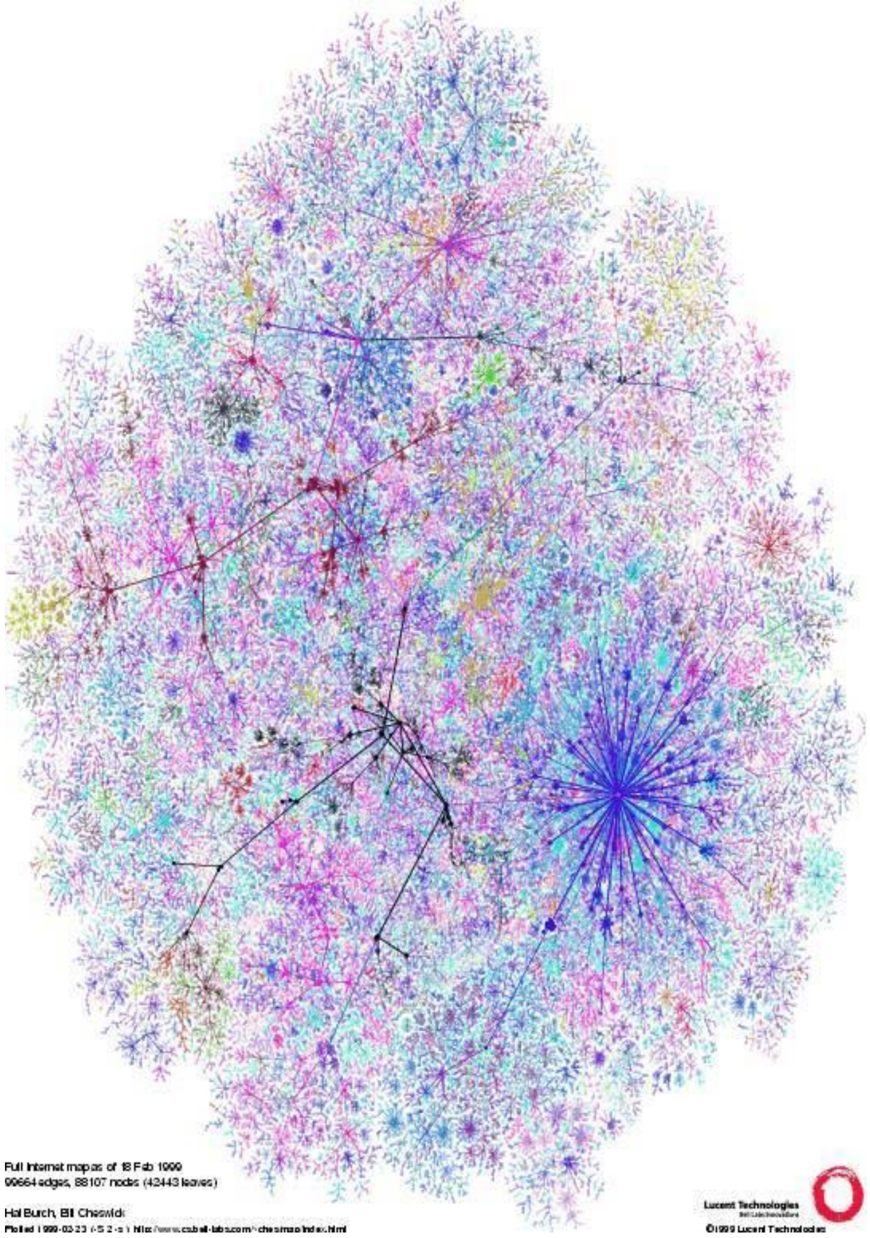
- Networks as a collection of pairwise relations
- Examples of (un)familiar and important networks
  - social networks
  - content networks
  - technological networks
  - biological networks
  - economic networks
- The distinction between structure and dynamics

A network-centric overview of modern society.



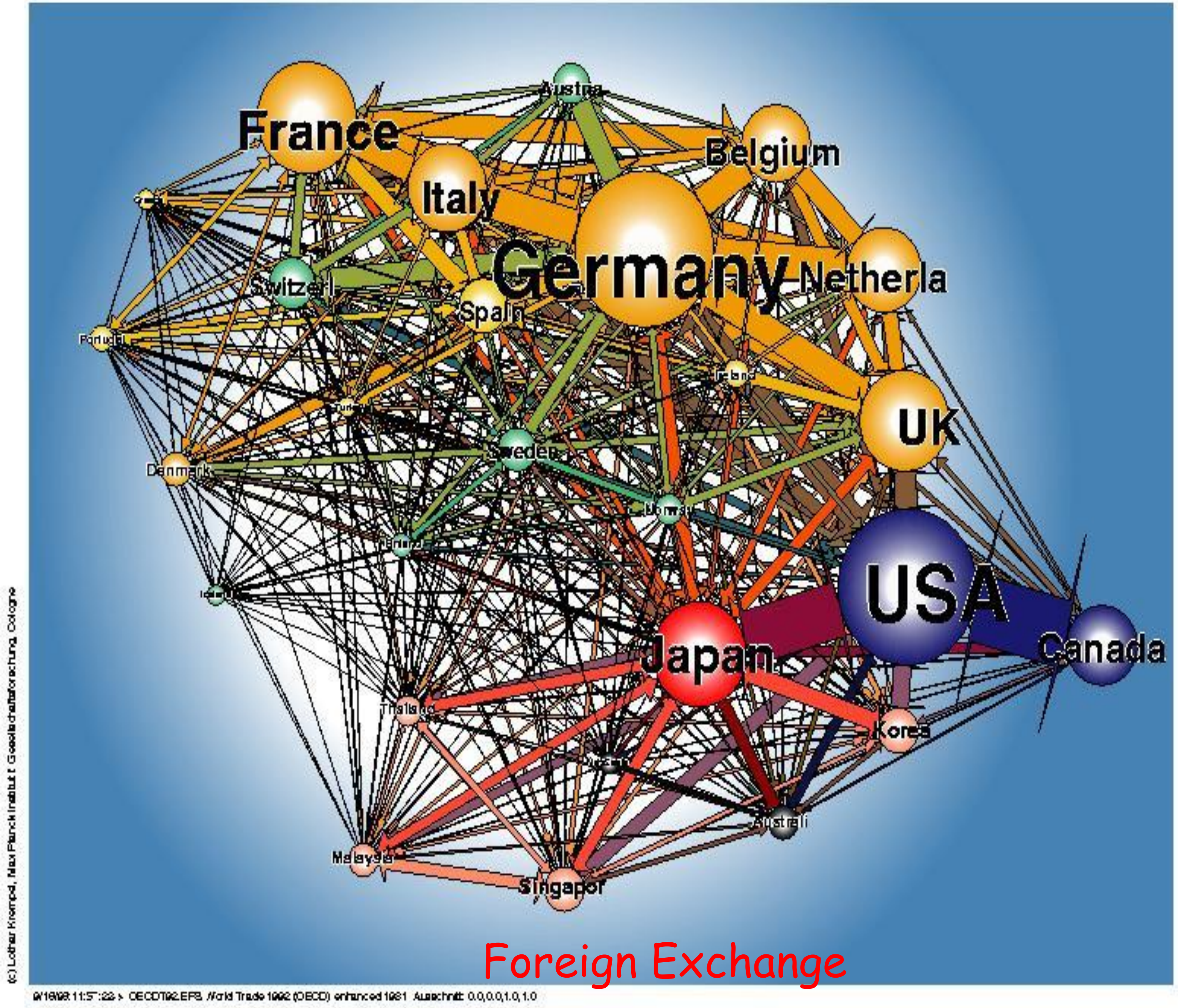
Gnutella Peers

- Points are still machines… but are associated with *people*
- Links are still physical… but may depend on *preferences*
- Interaction: content exchange
- Food for thought:
  "free riding"



Internet, Router Level

- A purely technological network?
- "Points" are physical machines
- "Links" are physical wires
- Interaction is electronic
- What more is there to say?



Foreign Exchange

- Points: sovereign nations
- Links: exchange volume
- A purely virtual network

Graph & Network Theory

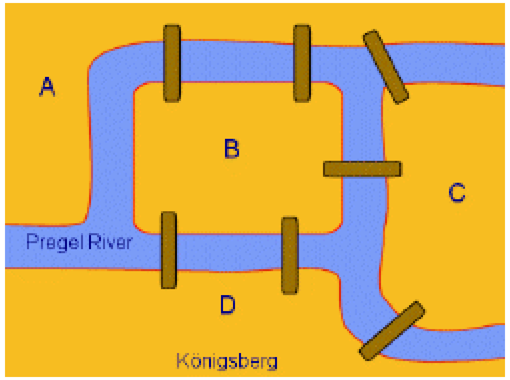
- Networks of vertices and edges
- Graph properties:
  - cliques, independent sets, connected components, cuts, spanning trees,…
  - social interpretations and significance
- Special graphs:
  - bipartite, planar, weighted, directed, regular,…
- Computational issues at a high level

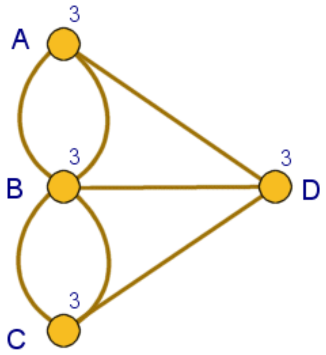
What is a network?

- Network: a collection of entities that are interconnected with links.
  - people that are friends
  - computers that are interconnected
  - web pages that point to each other
  - proteins that interact

Graphs

- In mathematics, networks are called graphs, the entities are nodes, and the links are edges
- Graph theory starts in the 18th century, with Leonhard Euler
  - The problem of Königsberg bridges
  - Since then graphs have been studied extensively.

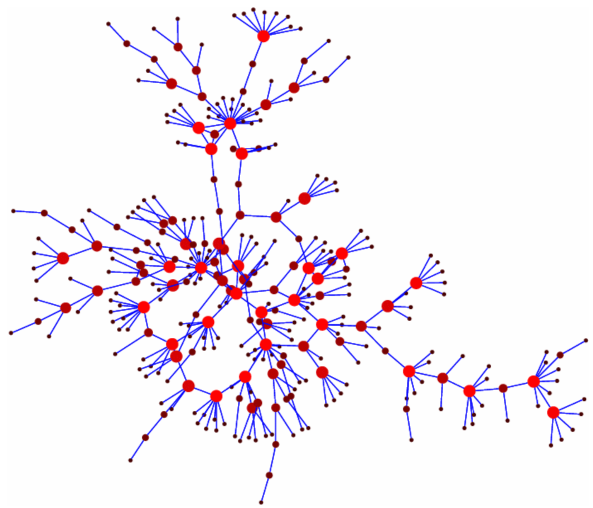




[Academic genealogy]

Networks in the past

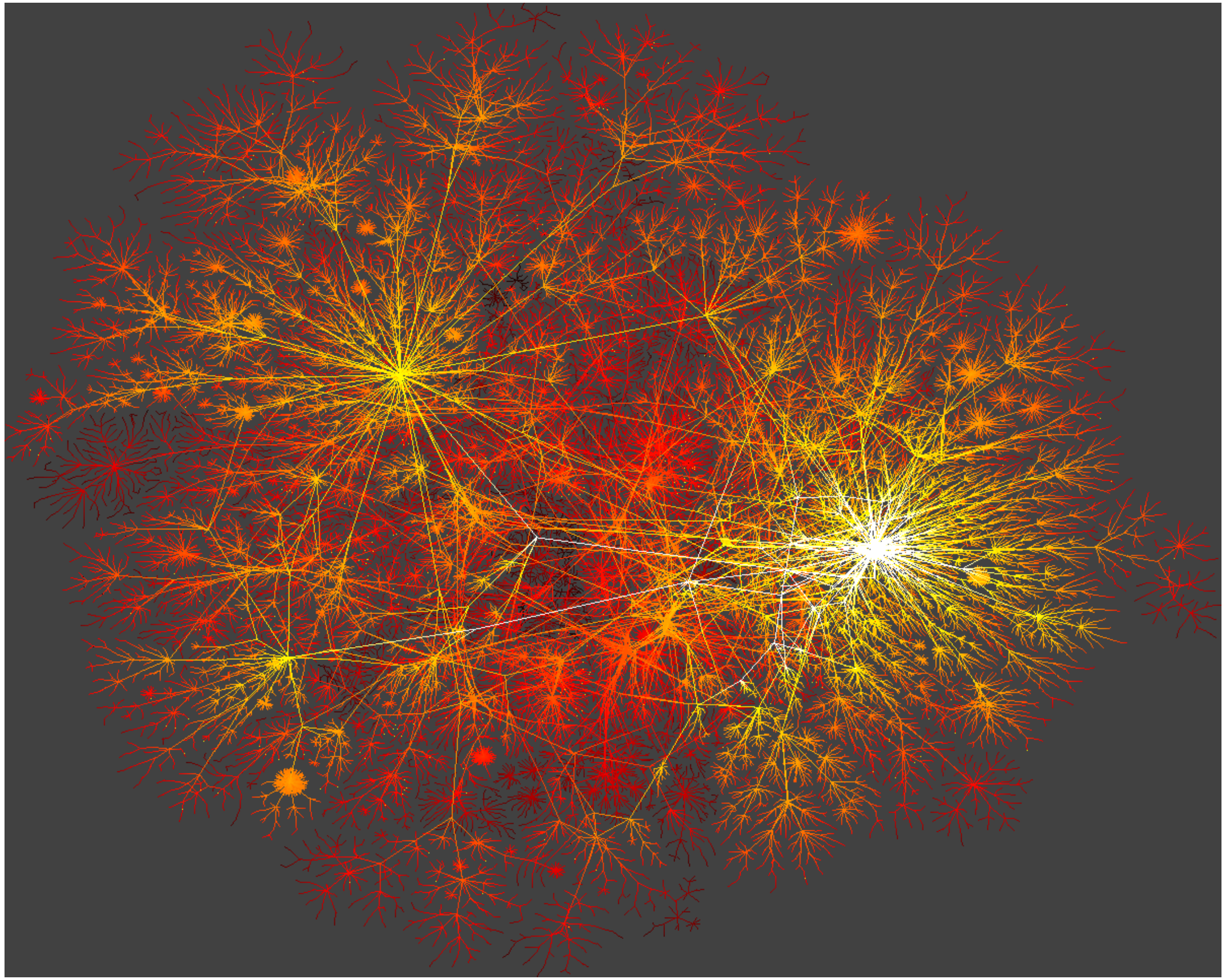
- Graphs have been used in the past to model existing networks (e.g., networks of highways, social networks)
  - usually these networks were small
  - network can be studied visual inspection can reveal a lot of information



Networks now

- More and larger networks appear
  - Products of technological advancement
    - e.g., Internet, Web
  - Result of our ability to collect more, better, and more complex data
    - e.g., gene regulatory networks
- Networks of thousands, millions, or billions of nodes
  - impossible to visualize

The internet map



Understanding large graphs

- What are the statistics of real life networks?
- Can we explain how the networks were generated?

Measuring network properties

- Around 1999
  - Watts and Strogatz, Dynamics and small-world phenomenon
  - Faloutsos, On power-law relationships of the Internet Topology
  - Kleinberg et al., The Web as a graph
  - Barabasi and Albert, The emergence of scaling in real networks

Real network properties

- Most nodes have only a small number of neighbors (degree), but there are some nodes with very high degree (power-law degree distribution)
  - scale-free networks
- If a node x is connected to y and z, then y and z are likely to be connected
  - high clustering coefficient
- Most nodes are just a few edges away on average.
  - small world networks
- Networks from very diverse areas (from internet to biological networks) have similar properties
  - Is it possible that there is a unifying underlying generative process?

Generating random graphs

- Classic graph theory model (Erdös-Renyi)
  - each edge is generated independently with probability p

- Very well studied model but:
  - most vertices have about the same degree
  - the probability of two nodes being linked is independent of whether they share a neighbor
  - the average paths are short

Modeling real networks

- Real life networks are not "random"
- Can we define a model that generates graphs with statistical properties similar to those in real life?
  - a flurry of models for random graphs

Processes on networks

- Why is it important to understand the structure of networks?

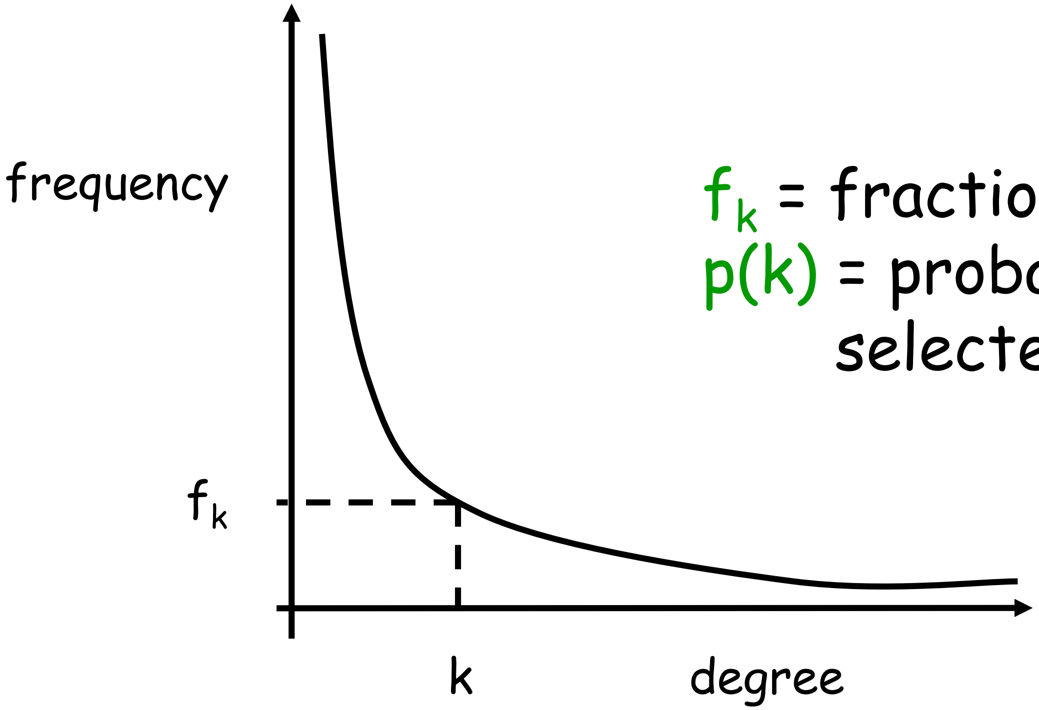
- Epidemiology: Viruses propagate much faster in scale-free networks
  - Vaccination of random nodes does not work, but targeted vaccination is very effective
  - Random sampling can be dangerous!

The basic random graph model

- The measurements on real networks are usually compared against those on "random networks"

- The basic $G_{n,p}$ (Erdös-Renyi) random graph model:
  - n : the number of vertices
  - $0 \leq p \leq 1$
  - for each pair (i,j), generate the edge (i,j) independently with probability p

Degree distributions



f_k = fraction of nodes with degree k
$p(k)$ = probability of a randomly selected node to have degree k

- Problem: find the probability distribution that best fits the observed data

Power-law distributions

- The degree distributions of most real-life networks follow a power law

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

- Right-skewed/Heavy-tail distribution
  - there is a non-negligible fraction of nodes that has very high degree (hubs)
  - scale-free: no characteristic scale, average is not informative

- In stark contrast with the random graph model!
  - Poisson degree distribution, z=np

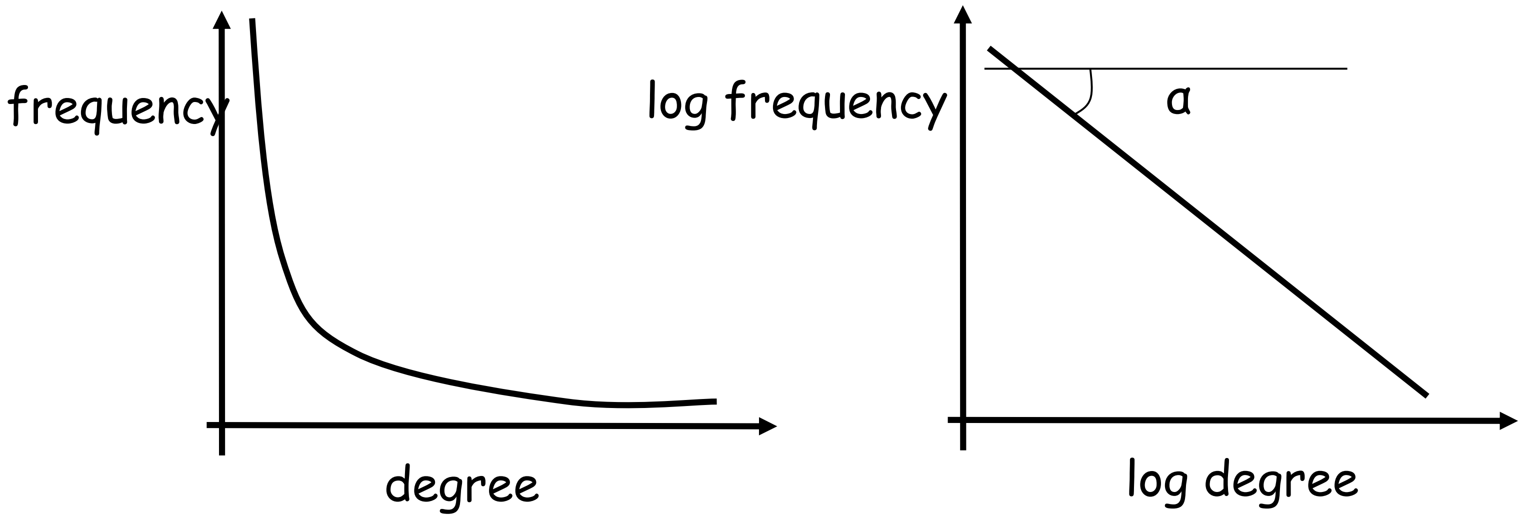
$$p(k) = P(k;z) = \frac{z^k}{k!}e^{-z}$$

  - highly concentrated around the mean
  - the probability of very high degree nodes is exponentially small

Power-law signature

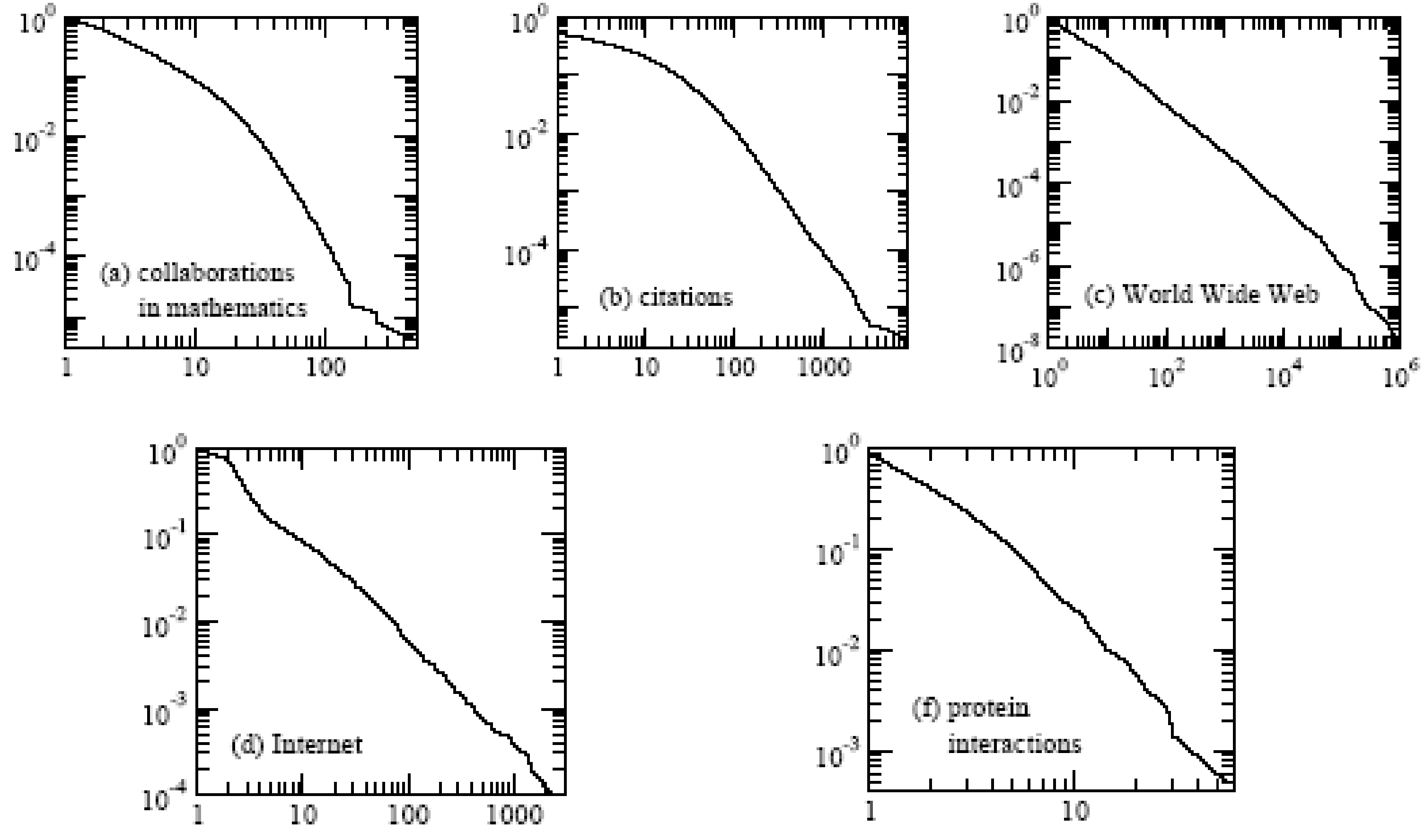
- Power-law distribution gives a line in the log-log plot

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



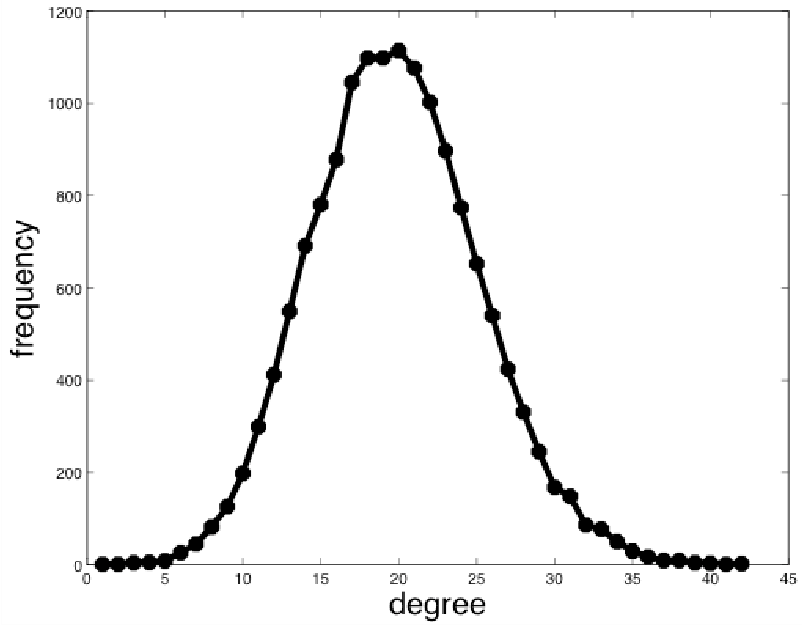
- α : power-law exponent (typically $2 \le \alpha \le 3$)

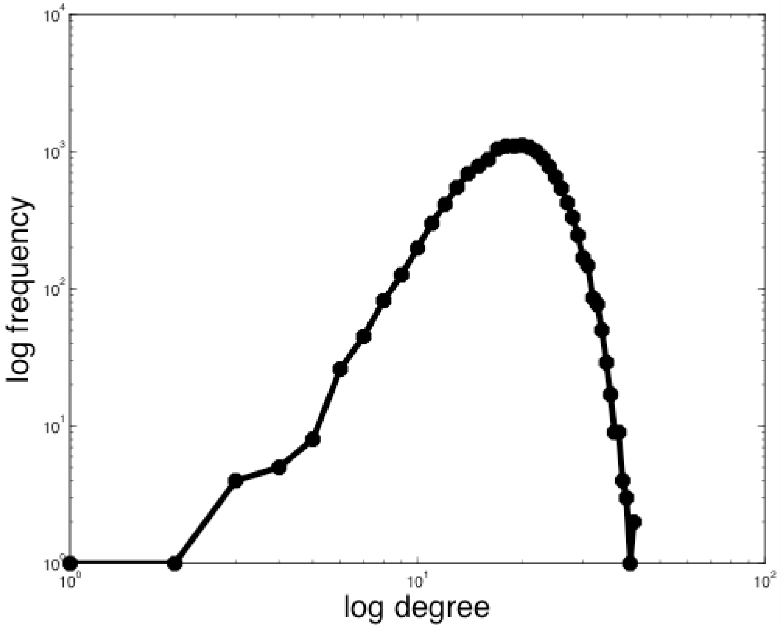
Examples of degree distribution for power laws



Taken from [Newman 2003]

A random graph example





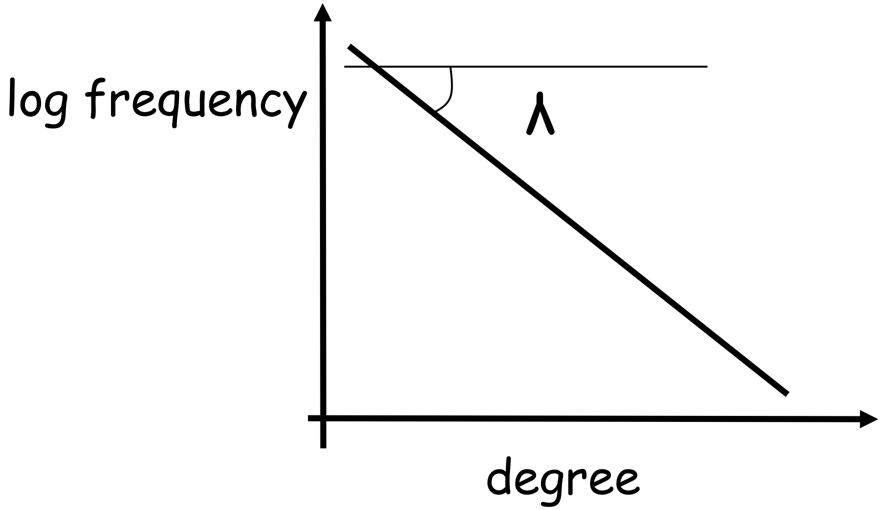
Exponential distribution

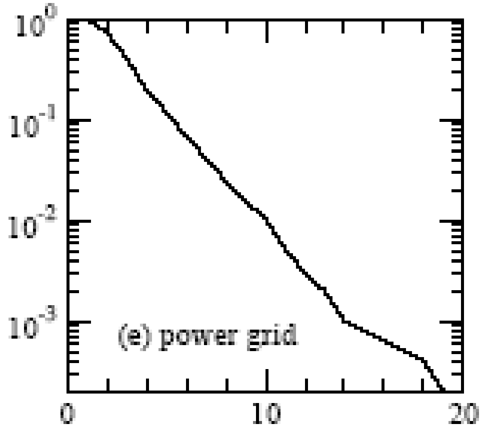
- Observed in some technological or collaboration networks

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

- Identified by a line in the log-linear plot

$$\log p(k) = -\lambda k + \log \lambda$$





Average/Expected degree

- For random graphs $z = np$

- For power-law distributed degree
  - if $\alpha \geq 2$, it is a constant
  - if $\alpha < 2$, it diverges

Maximum degree

- For random graphs, the maximum degree is highly concentrated around the average degree z
- For power law graphs

$$k_{max} \approx n^{1/(\alpha-1)}$$

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Clustering coefficient

- In graph theory, a clustering coefficient is a measure of degree to which nodes in a graph tend to cluster together.

- Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties (Holland and Leinhardt, 1971;[1] Watts and Strogatz, 1998[2]).

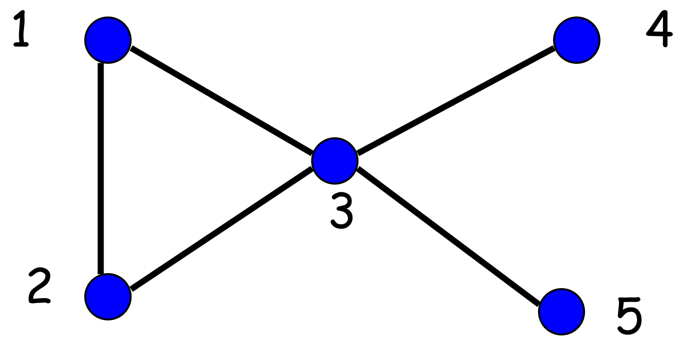
- In real-world networks, this likelihood tends to be greater than the average probability of a tie randomly established between two nodes (Holland and Leinhardt, 1971; Watts and Strogatz, 1998).

Clustering (Transitivity) coefficient

- Measures the density of triangles (local clusters) in the graph
- Two different ways to measure it, C^1 & C^2:
- The ratio of the means

$$C^{(1)} = \frac{\sum_i \text{triangles centered at node i}}{\sum_i \text{triples centered at node i}}$$

Example undirected graph



$$C^{(1)} = \frac{3}{1+1+6} = \frac{3}{8}$$

Triangles: one each centered at nodes, 1, 2, 3
Triples: none centered for nodes 4, 5

node 1 – 213
node 2 – 123
node 3 – 134, 135, 234, 235, 132, 231

Clustering (Transitivity) coefficient

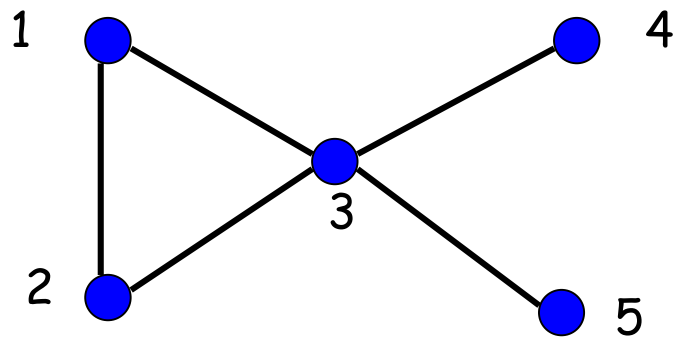
- Clustering coefficient for node i

$$C_i = \frac{\text{triangles centered at node i}}{\text{triples centered at node i}}$$

$$C^{(2)} = \frac{1}{n} C_i$$

- The mean of the ratios

Example



$$C^{(2)} = \frac{1}{5}\left(1 + 1 + 1/6\right) = \frac{13}{30}$$

$$C^{(1)} = \frac{3}{8}$$

- The two clustering coefficients give different measures
- $C^{(2)}$ increases with nodes with low degree

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Clustering coefficient for 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)

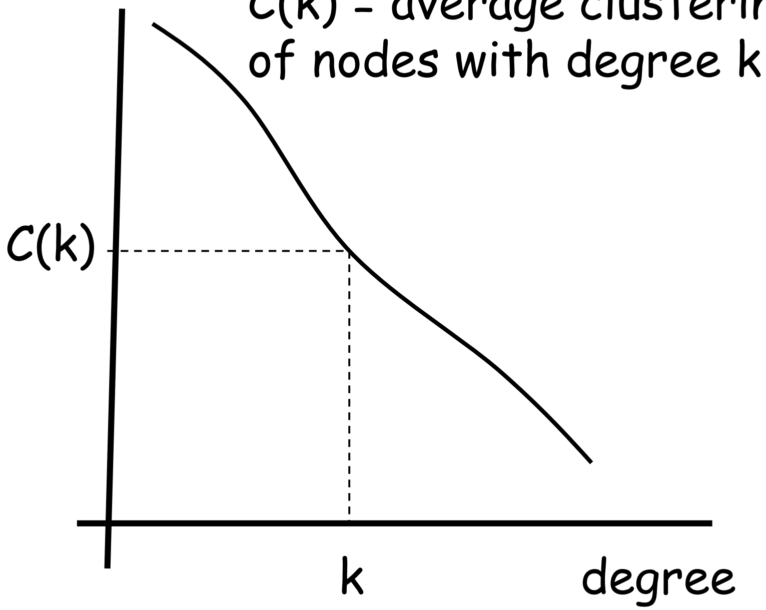
Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

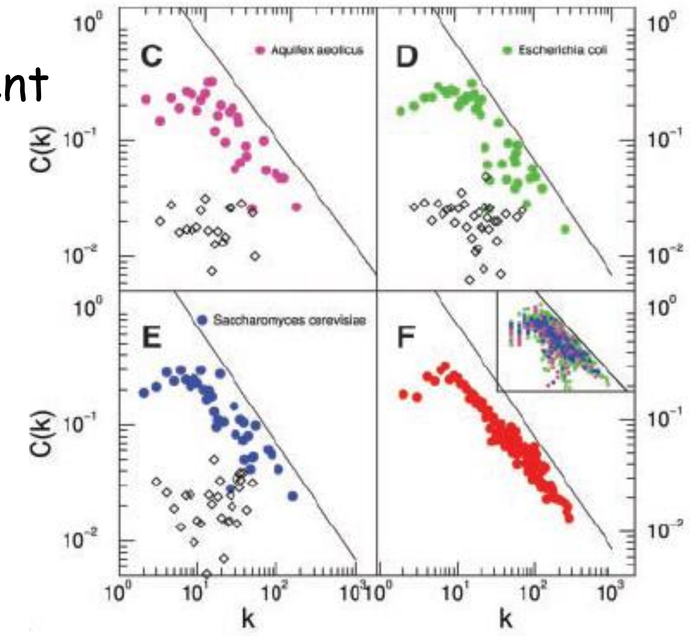
Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor 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

The C(k) distribution

- The C(k) distribution is supposed to capture the hierarchical nature of the network
  - when constant: no hierarchy
  - when power-law: hierarchy

C(k) = average clustering coefficient of nodes with degree k





Millgram’s small world experiment

- Letters were handed out to people in Nebraska to be sent to a target in Boston
- People were instructed to pass on the letters to someone they knew on first-name basis
- The letters that reached the destination followed paths of length around 6
- Six degrees of separation: (play of John Guare)

- Also:
  - The Kevin Bacon game
  - The Erdös number
- Small world project: http://smallworld.columbia.edu/index.html

Measuring the small world phenomenon

- d_{ij} = shortest path between i and j
- Diameter:

$$d = \max_{i,j} d_{ij}$$

- Characteristic path length:

$$\ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}$$

- Harmonic mean

$$\ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1}$$

- Also, distribution of all shortest paths

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Is the path length enough?

- Random graphs have diameter

$$d = \frac{\log n}{\log z}$$

- d=logn/loglogn when z=ω(log n)

- Short paths should be combined with other properties
  - ease of navigation
  - high clustering coefficient

Degree correlations

- Do high degree nodes tend to link to high degree nodes?
- Pastor Satoras et al.
  - plot the mean degree of the neighbors as a function of the degree

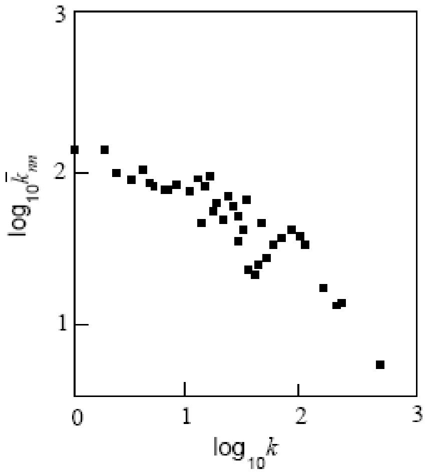


FIG. 3.13. Correlations of the degrees of nearest-neighbour vertices (autonomous systems) in the Internet at the interdomain level (after Pastor-Satorras, Vázquez, and Vespignani 2001). The empirical dependence of the average degree of the nearest neighbours of a vertex on the degree of this vertex is shown in a log–log scale. This empirical dependence was fitted by a power law with exponent approximately 0.5.

Collective Statistics (M. Newman 2003)

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW nd.edu	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

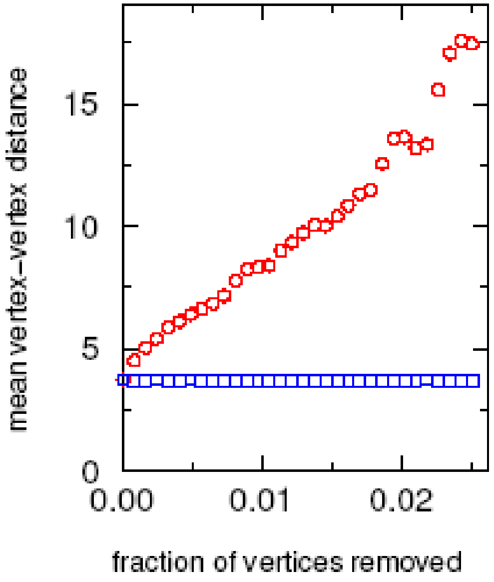
TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

Connected components

- For undirected graphs, the size and distribution of the connected components
  - is there a giant component?
- For directed graphs, the size and distribution of strongly and weakly connected components

Network Resilience

- Study how the graph properties change when performing random or targeted node deletions



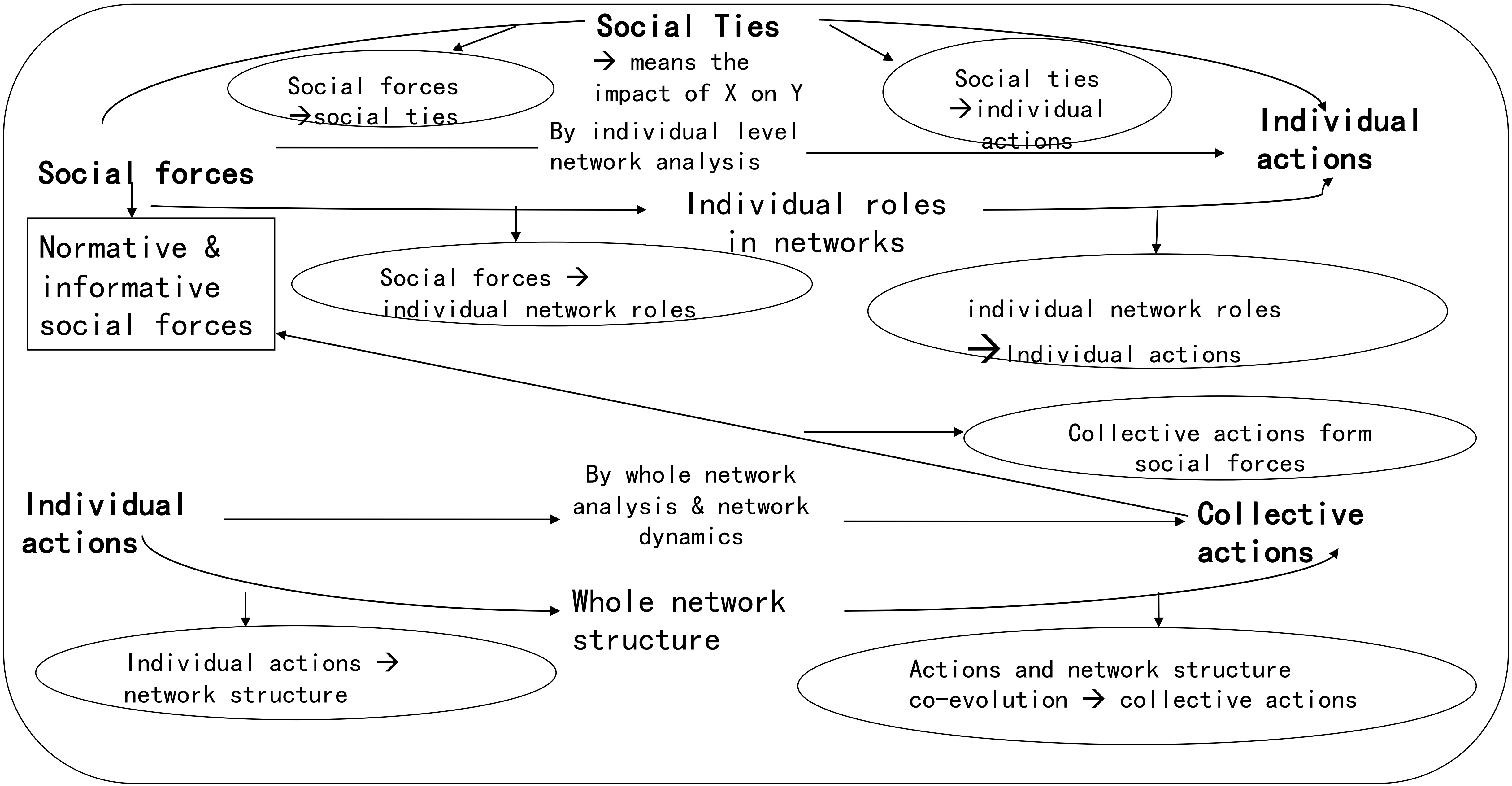
Social Networks

- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

(Source: Freeman, 2000)

Social Network Theory

- Metrics of social importance in a network:
  - degree, closeness, between-ness, clustering…
- Local and long-distance connections
- SNT "universals"
  - small diameter
  - clustering
  - heavy-tailed distributions
- Models of network formation
  - random graph models
  - preferential attachment
  - affiliation networks
- Examples from society, technology and fantasy


Social Ties
→ means the
impact of X on Y
Social forces
→social ties
Social ties
→individual
actions
Individual
actions
By individual level
network analysis
Social forces
Individual roles
in networks
Normative &
informative
social forces
Social forces →
individual network roles
individual network roles
→Individual actions
Collective actions form
social forces
By whole network
analysis & network
dynamics
Individual
actions
Collective
actions
Whole network
structure
Individual actions →
network structure
Actions and network structure
co-evolution → collective actions

Definition of Social Networks

- “A social network is a set of actors that may have relationships with one another. Networks can have few or many actors (nodes), and one or more kinds of relations (edges) between pairs of actors.” (Hannemann, 2001)

History (based on Freeman, 2000)

- 17th century: Spinoza developed first model
- 1937: J.L. Moreno introduced sociometry; he also invented the sociogram
- 1948: A. Bavelas founded the group networks laboratory at MIT; he also specified centrality

History (based on Freeman, 2000)

- 1949: A. Rapaport developed a probability based model of information flow
- 50s and 60s: Distinct research by individual researchers
- 70s: Field of social network analysis emerged.
  - New features in graph theory – more general structural models
  - Better computer power – analysis of complex relational data sets

What are **social relations/ties**?

A social relation is anything that links two actors.
Examples include:

Kinship	Co-membership
Friendship	Talking with
Love	Hate
Exchange	Trust
Coauthorship	Fighting

What properties relations are studied?

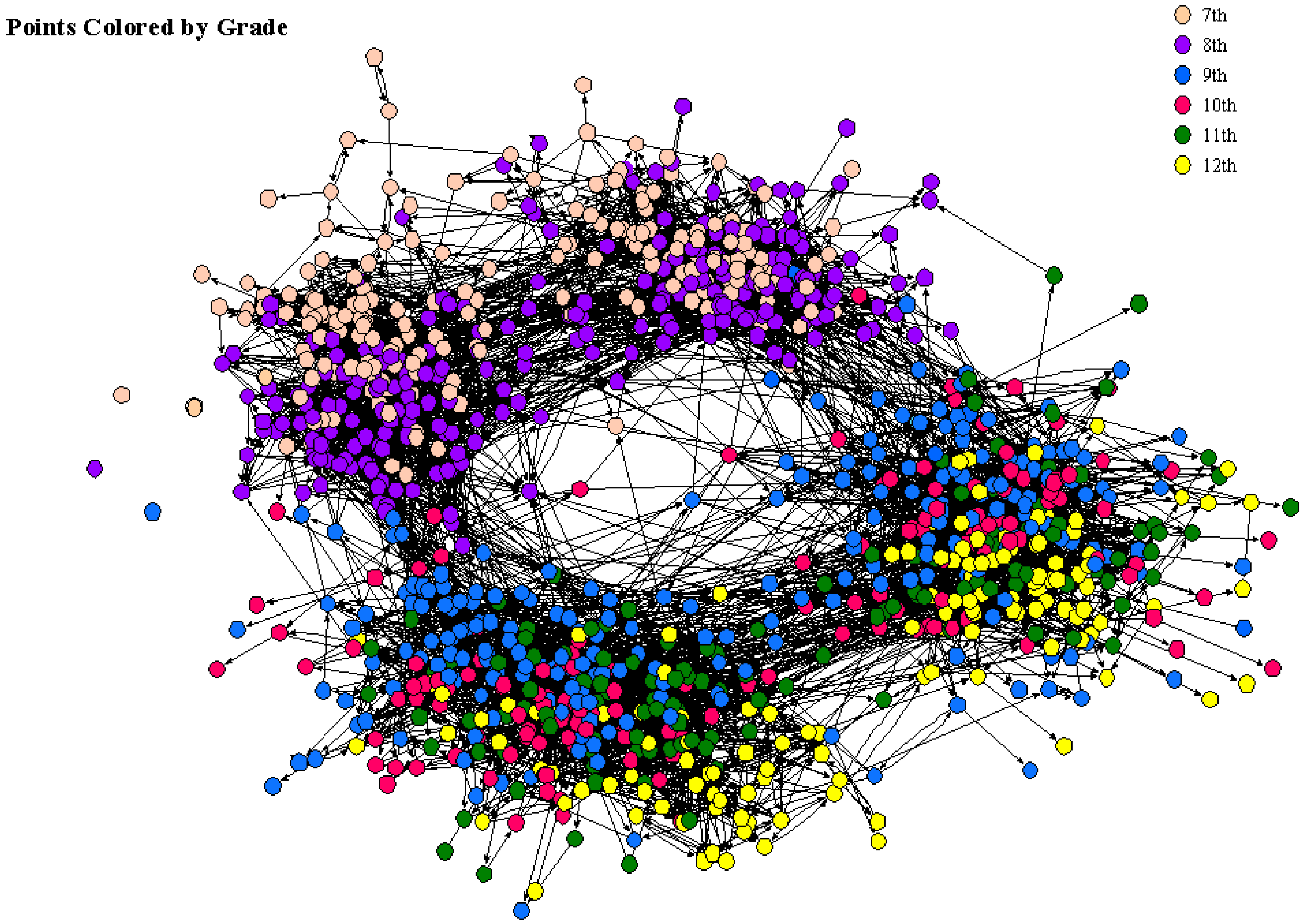
The substantive topics cross all areas of sociology. But we can identify types of questions that social network researchers ask:

1) Social network analysts often study relations as *systems*. That is, what is of interest is how the *pattern* of relations among actors affects individual behavior or system properties.

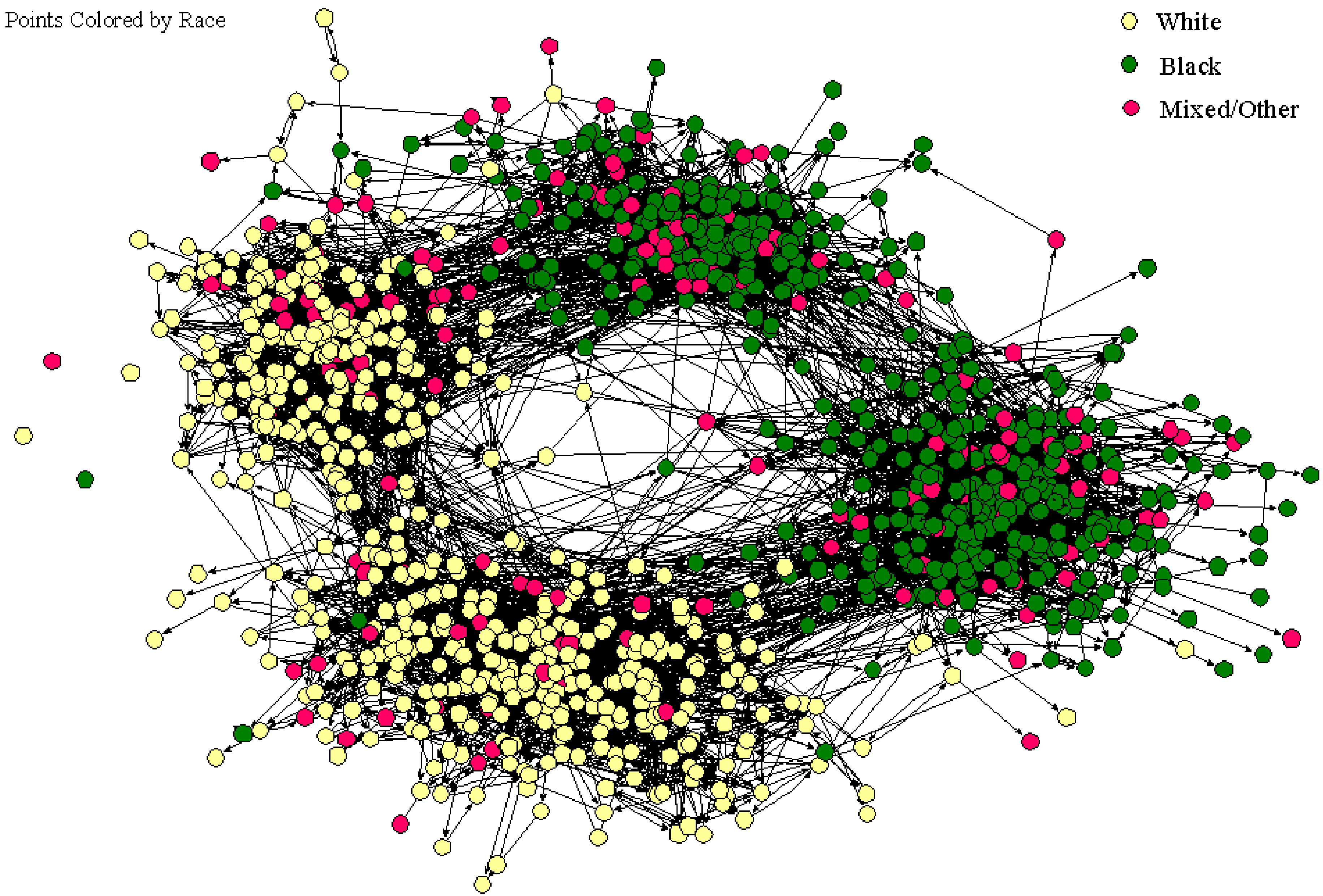
Introduction

High Schools as Networks

The Social Structure of "Countryside" School District



The Social Structure of "Countryside" School District

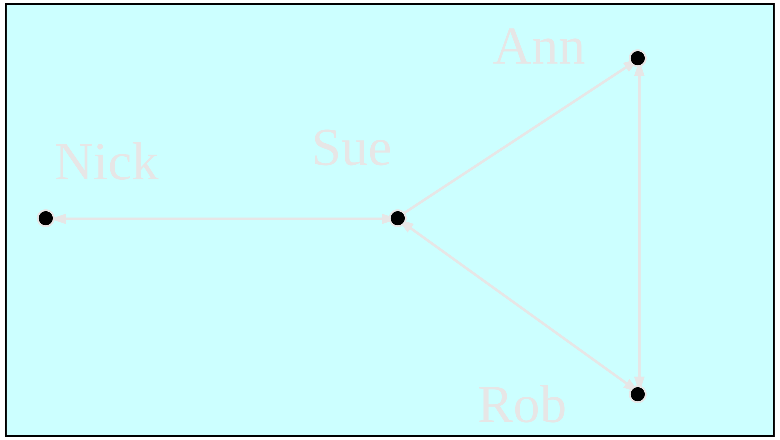


Representation of Social Networks

- Matrices

	Ann	Rob	Sue	Nick
Ann	---	1	0	0
Rob	1	---	1	0
Sue	1	1	---	1
Nick	0	0	1	---

- Graphs



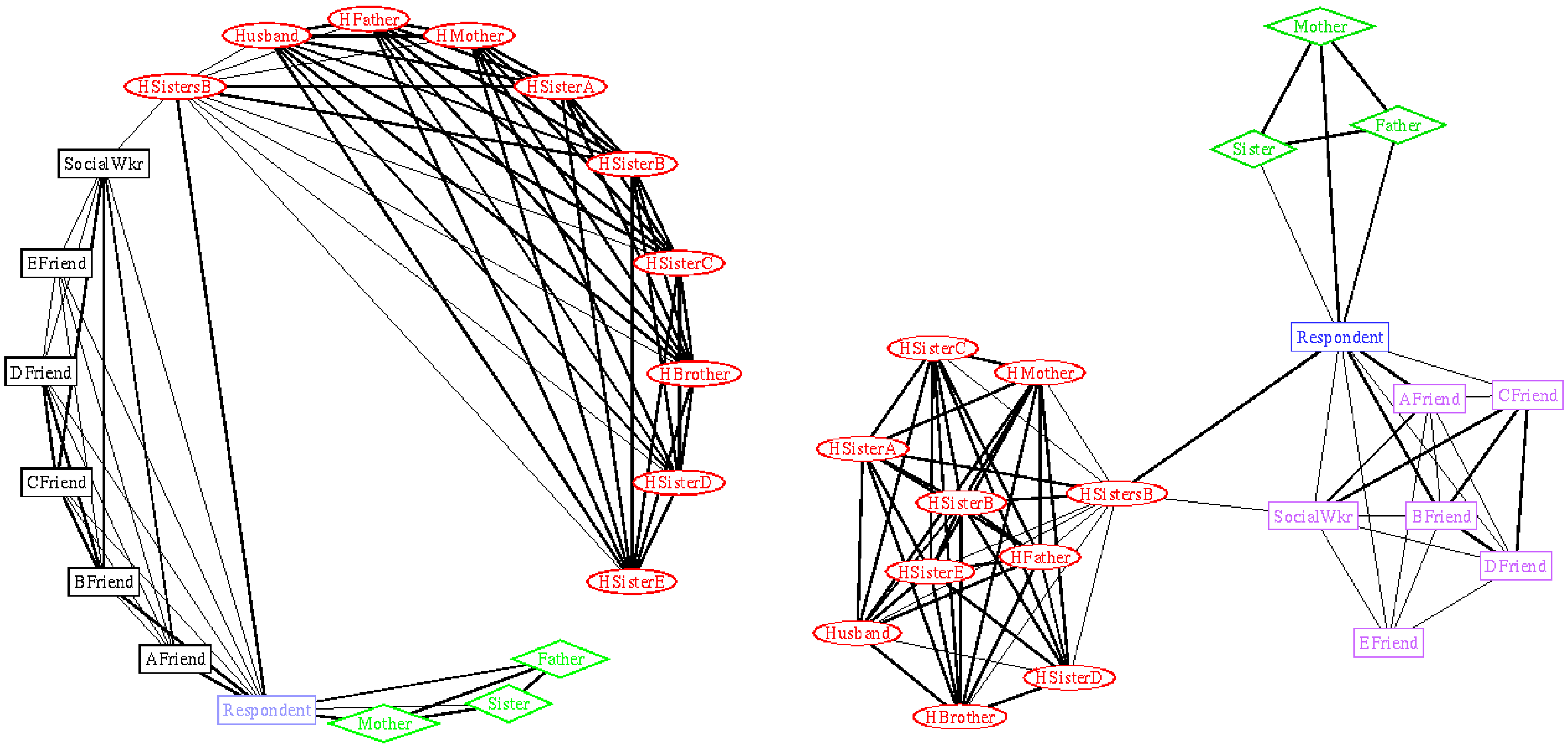
Graphs - Sociograms
(based on Hanneman, 2001)

- Labeled circles represent actors
- Line segments represent ties
- Graph may represent one or more types of relations
- Each tie can be directed or show co-occurrence
  - Arrows represent directed ties

Graphs – Sociograms
(based on Hanneman, 2001)

- Strength of ties:
  - Nominal
  - Signed
  - Ordinal
  - Valued

Visualization Software: [Krackplot]



Connections

- Size
  - Number of nodes
- Density
  - Number of ties that are present vs the amount of ties that could be present
- Out-degree
  - Sum of connections from an actor to others
- In-degree
  - Sum of connections to an actor
- Diameter
  - Maximum greatest least distance between any actor and another

Some Measures of Distance

- Walk (path)
  - A sequence of actors and relations that begins and ends with actors
- Geodesic distance (shortest path)
  - The number of actors in the shortest possible walk from one actor to another
- Maximum flow
  - The amount of different actors in the neighborhood of a source that lead to pathways to a target

Some Measures of Power

(based on Hanneman, 2001)

- Degree (indegree, outdegree)
  - Sum of connections from or to an actor
- Closeness centrality
  - Distance of one actor to all others in the network
- Betweenness centrality
  - Number that represents how frequently an actor is between other actors' geodesic paths

Cliques and Social Roles

(based on Hanneman, 2001)

- Cliques
  - Sub-set of actors
    - More closely tied to each other than to actors who are not part of the sub-set

- Social roles
  - Defined by regularities in the patterns of relations among actors

SNA applications

Many new unexpected applications plus many of the old ones

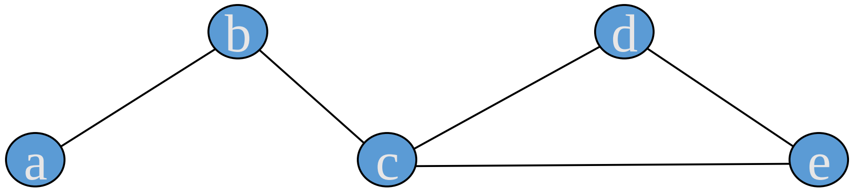
- Marketing
- Advertising
- Economic models and trends
- Political issues
    - Organization
- Services to social network actors
    - Travel; guides
    - Jobs
    - Advice
- Human capital analysis and predictions
- Medical
- Epidemiology
- Defense (terrorist networks)

The unit of interest in a network are the combined sets of actors and their relations.

We represent *actors* with points and *relations* with lines.

- Actors are referred to variously as:
  - Nodes, vertices, actors or points
- Relations are referred to variously as:
  - Edges, Arcs, Lines, Ties

Example:



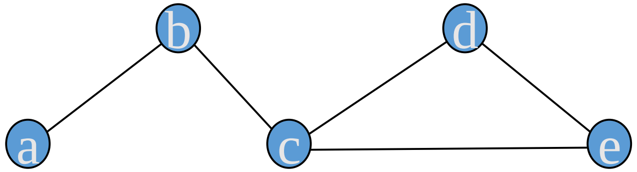
Social Network data consists of two linked classes of data:

a) **Nodes:** Information on the individuals (actors, nodes, points, vertices)
- Network nodes are most often people, but can be any other unit capable of being linked to another (schools, countries, organizations, personalities, etc.)
- The information about nodes is what we usually collect in standard social science research: demographics, attitudes, behaviors, etc.
- Often includes dynamic information about when the node is active

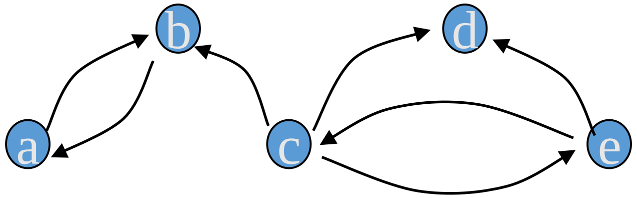
b) **Edges**: Information on the relations among individuals (lines, edges, arcs)
- Records a connection between the nodes in the network
- Can be valued, directed (arcs), binary or undirected (edges)
- One-mode (direct ties between actors) or two-mode (actors share membership in an organization)
- Includes the times when the relation is active

Graph theory notation: G(V,E)

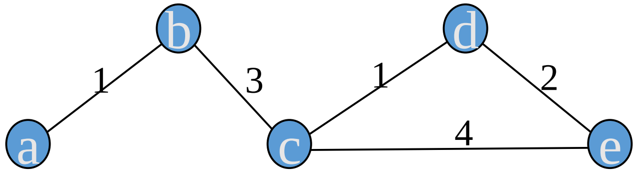
In general, a relation can be: (1) Binary or Valued (2) Directed or Undirected



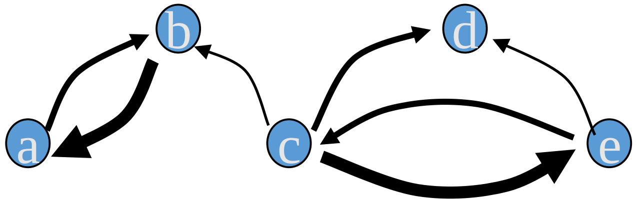
Undirected, binary



Directed, binary



Undirected, Valued

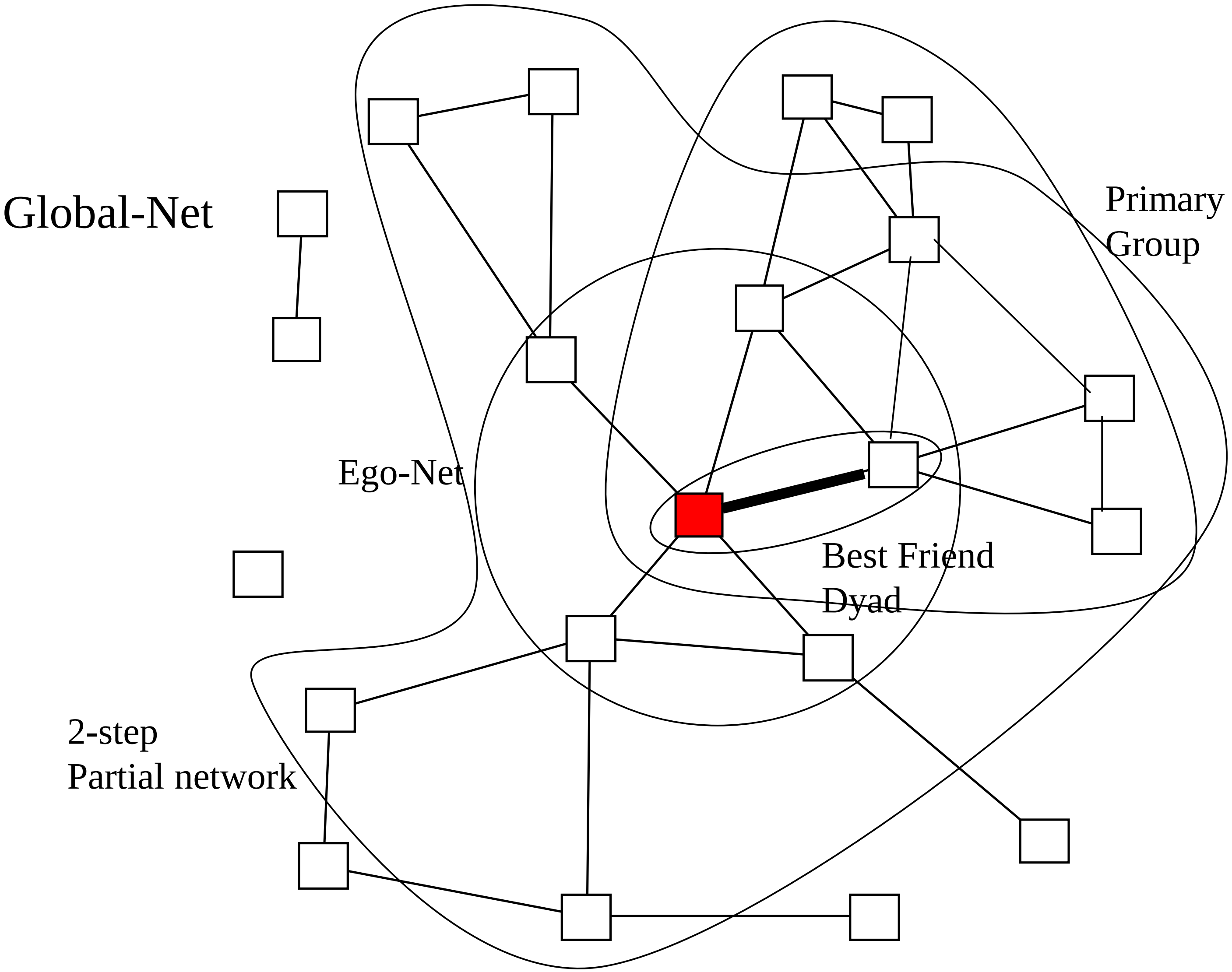


Directed, Valued

The *social process* of interest will often determine what form your data take. Almost all of the techniques and measures we describe can be generalized across data format.

Foundations

Data and social science



We can examine networks across multiple levels:

1) ***Ego-centered network***

- Have data on a respondent (ego) and the people they are connected to (alters). Example: terrorist networks

- May include the social capital embedded in this network
  "a form of economic and cultural capital in which social networks are central, transactions are marked by reciprocity, trust, and cooperation, and market agents produce goods and services not mainly for themselves, but for a common good."
- Q: how an ego-centered network influences a person's action?

2) Partial network

- Ego networks plus some amount of tracing to reach contacts of contacts

- Something less than full account of connections among all pairs of actors in the relevant population

- Example: CDC Contact tracing data

We can examine networks across multiple levels:

3) Complete or "Global" data

- Data on *all* actors within a particular (relevant) boundary

- Never exactly complete (due to missing data), but boundaries are set

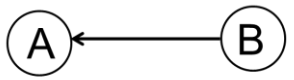
-Example: Coauthorship data among all writers in the social sciences, friendships among all students in a classroom

Network Structure Analysis

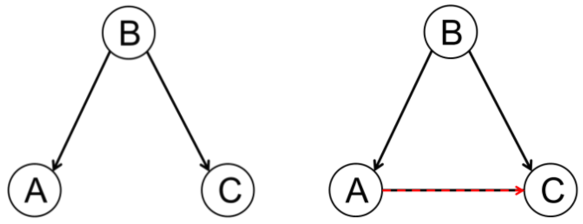
- Network structure involves research on a person’s role in a network structure depending on the dynamic process of change in his or her networking behavior.
- Some of the different roles a person can take in a network structure are: Core, Bridge (between two small groups), Outsider, and Inner circle member, etc..
- This field brings to light the dynamic processes of networking behaviors and the influence of social forces on these networking processes.
- Because of this, **social relations** and **network structure** are the two main fields in social network theory.

Different Network Structures

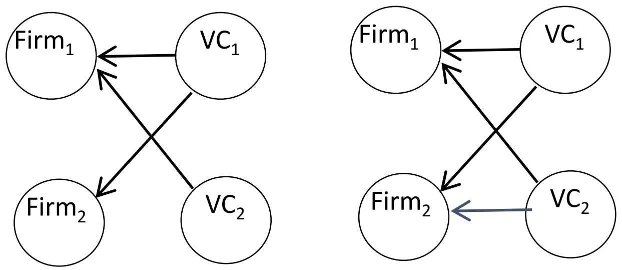
- Dyad:
  - Formed by a group of two nodes. R
  - elationship can be asymmetric or symmetric;
    Maybe based on family relation, common interests, work, trust, a joint action, …



- Triad:
  - a group of 3 nodes
    (a common way to connect new people)
  - Can be closed or open
  - Social balance theory etc

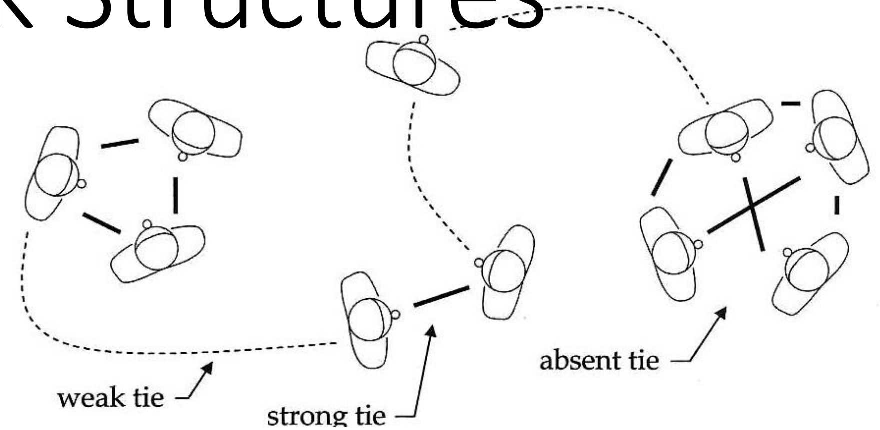


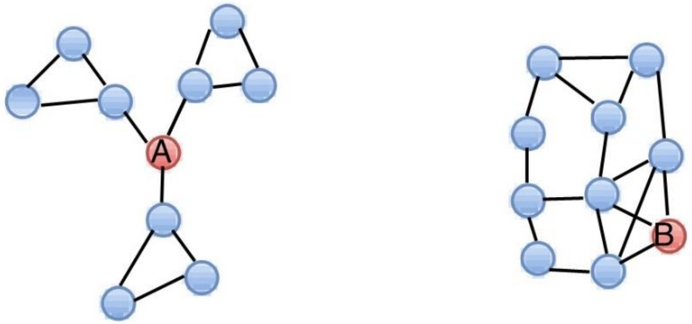
- Quadrangle:
  - A group of 4 nodes (e.g., 2-mode network)
  - Can be closed or open
  - Guanxi circle theory etc.

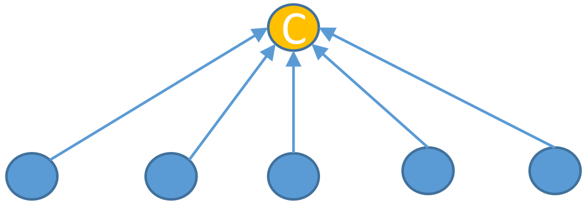


Individual Actions and Network Structures

- Weak Ties (Mark Granovetter)
  - the stronger the tie between two people is, the more likely their contacts will overlap so that they will have common ties with the same 3rd parties
- Structural Holes (Ronald Burt)
  - Structural hole spanner/broker
    - separation between non-redundant contacts
  - Related to social capital
- Opinion leaders
  - individuals who obtain more media coverage than others and are especially educated on a certain issue.
  - They seek the acceptance of others and are especially motivated to enhance their social status
  - Higher social influence







Network Dynamics

- A commonly accepted network behavior will change the structure of a whole network.
  - For example, everyone in a network attaches him-/herself to the centered person will make the whole network with high group centrality.
- Dynamics aspect includes
  - the evolution process of a network,
  - the commonly accepted network behaviors, and
  - their influence on the dynamic process of evolution

Collective Action and Network Dynamics

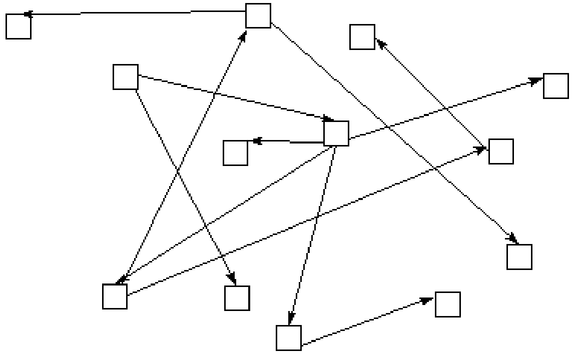
- A collective action emerges from the interlinking development of behavior and network structure.
  - It illustrates the non-linear development of a system and the emergence of collective action during network dynamics.
- Sustained collective action

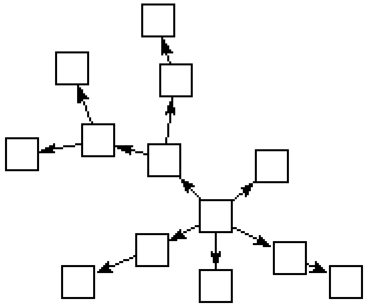
Foundations

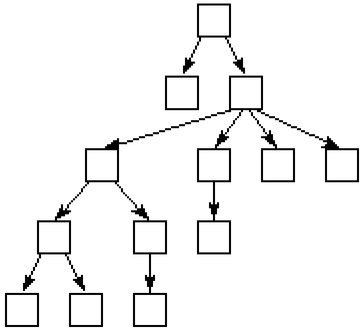
Graphs

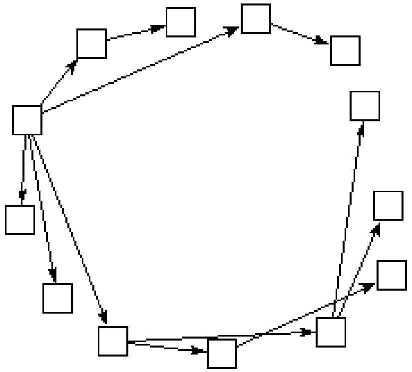
Working with pictures.

No standard way to draw a sociogram: which are equal?





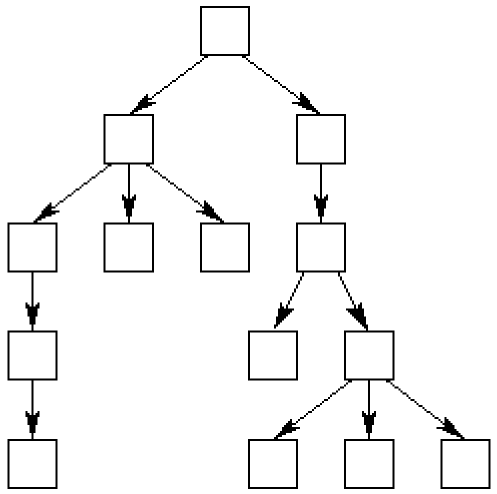




Foundations
Graphs

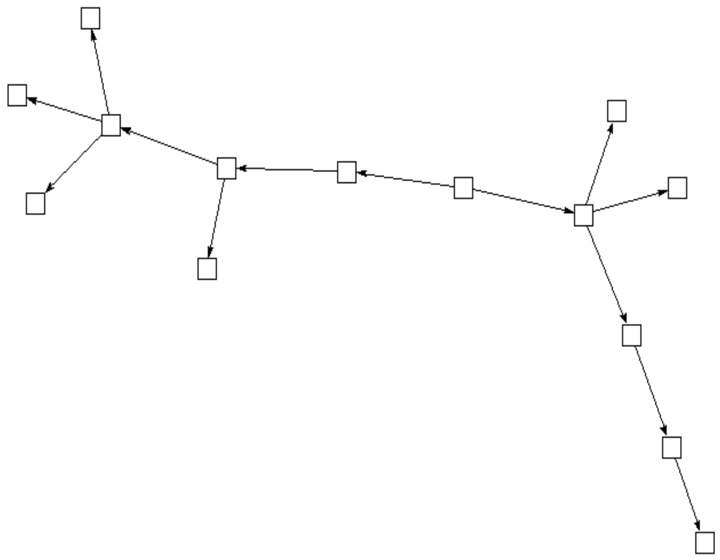
Network visualization helps build intuition, but you have to keep the drawing algorithm in mind:

Tree-Based layouts



Most effective for very sparse, regular graphs. Very useful when relations are strongly directed, such as organization charts, internet connections,

Spring-embeder layouts



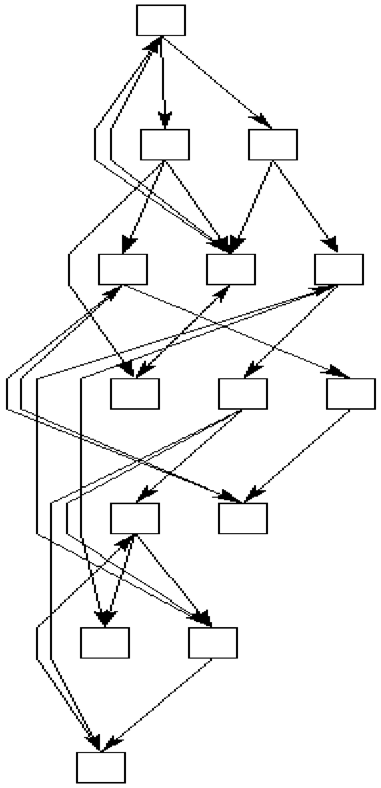
Most effective with graphs that have a strong community structure (clustering, etc). Provides a very clear correspondence between social distance and plotted distance

Two images of the same network

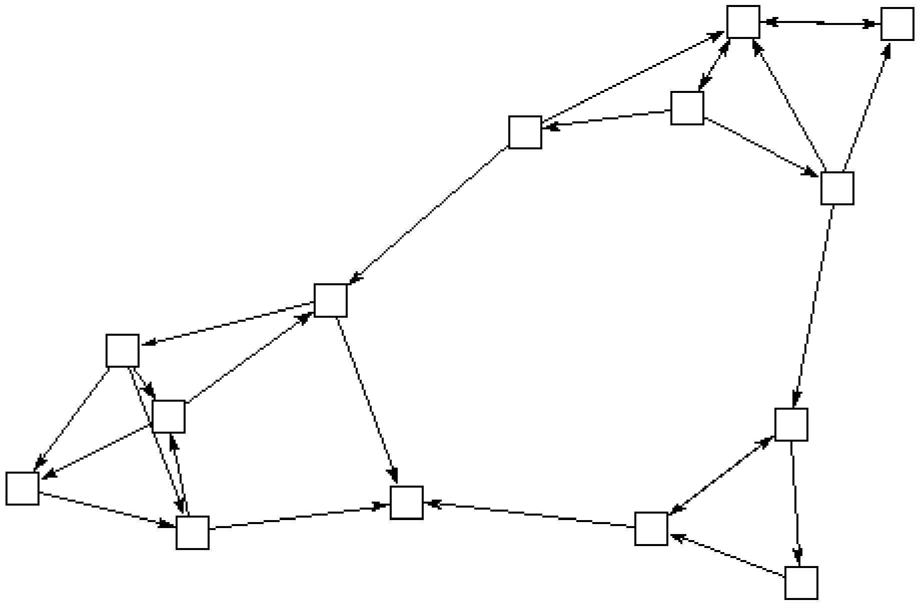
Foundations
Graphs

Network visualization helps build intuition, but you have to keep the drawing algorithm in mind:

Tree-Based layouts



Spring-embeder layouts

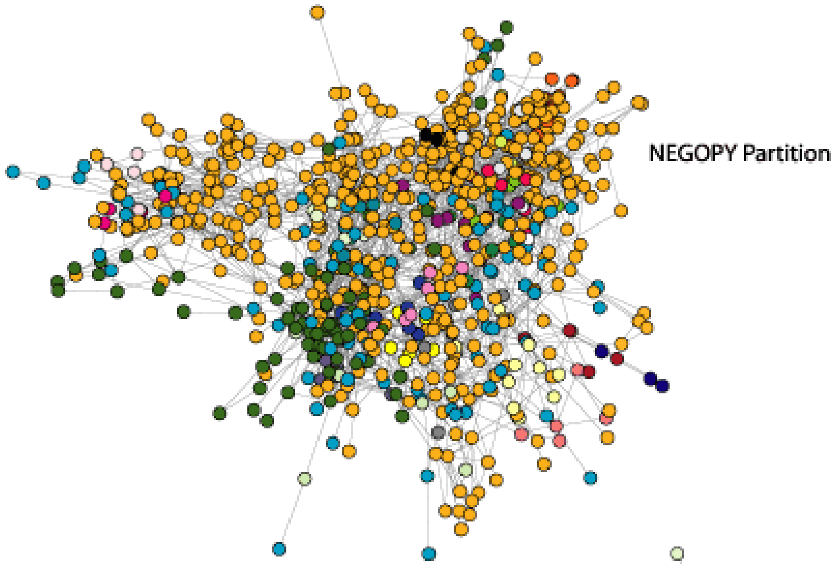


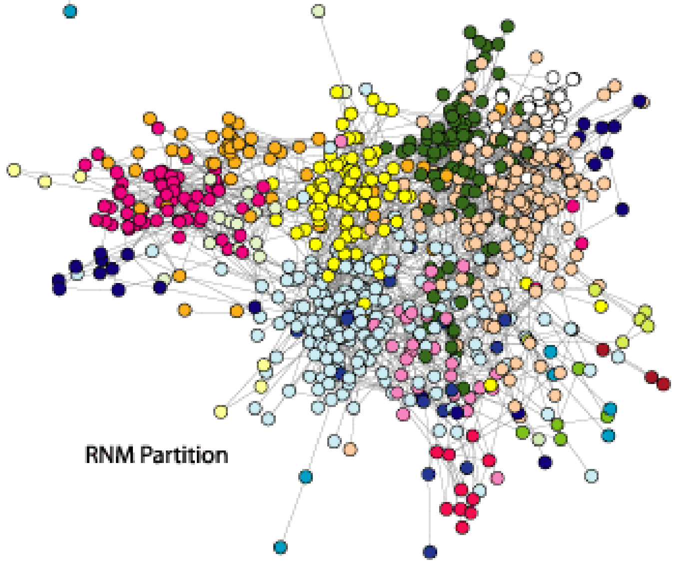
Two images of the same network

Foundations
Graphs

Using colors to code attributes makes it simpler to compare attributes to relations.

Here we can assess the effectiveness of two different clustering routines on a school friendship network.



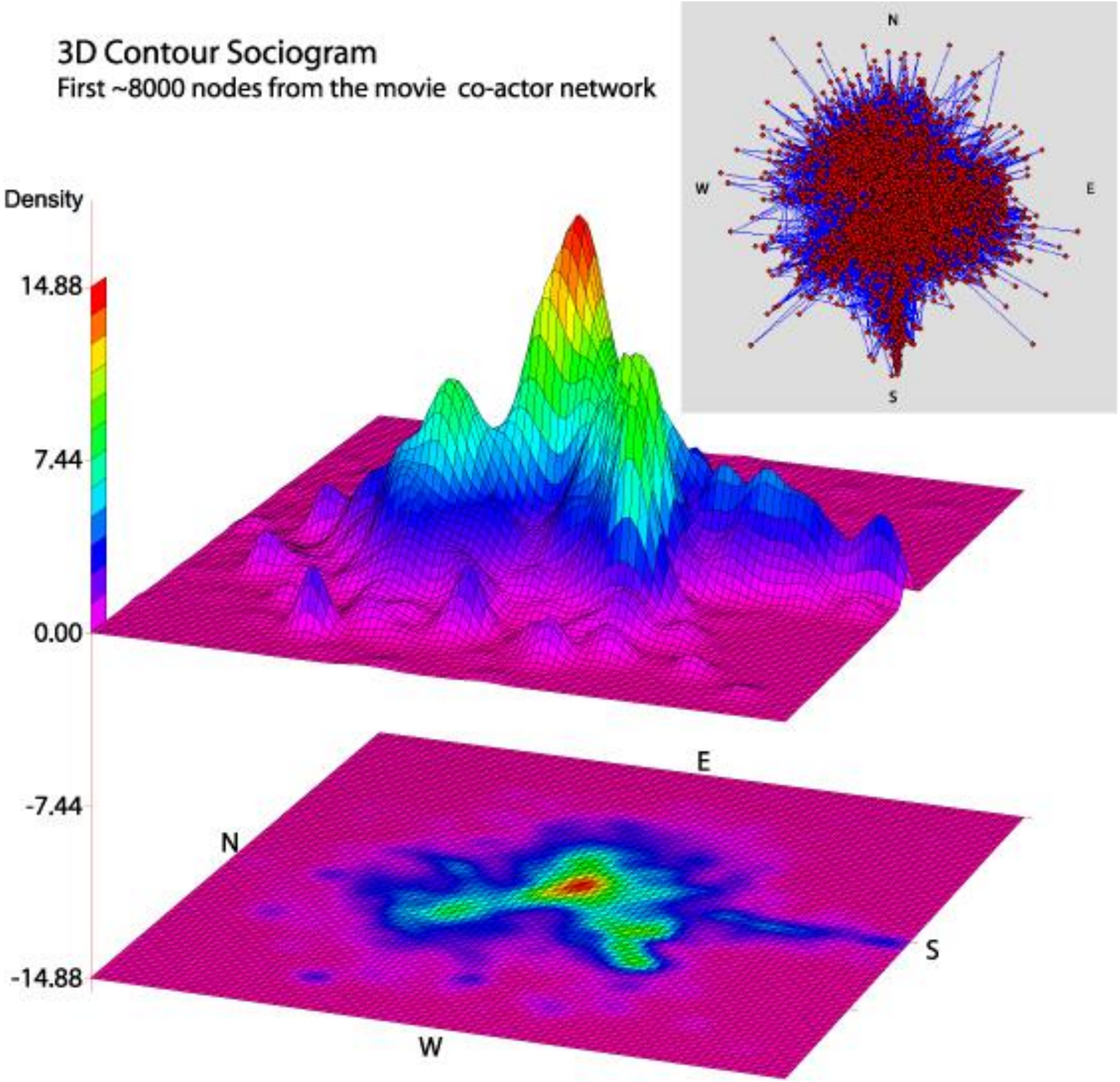


Foundations
Graphs

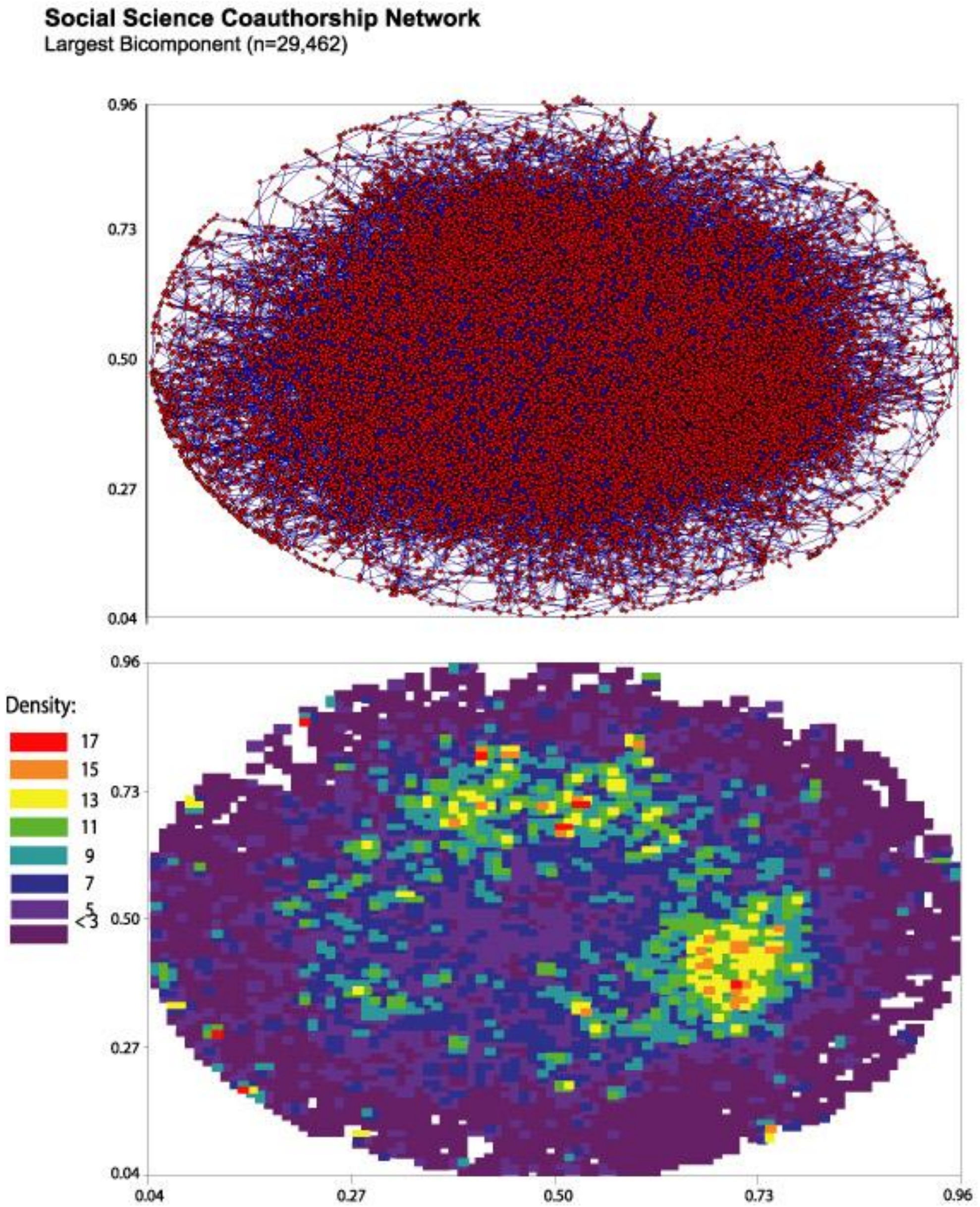
As networks increase in size, the effectiveness of a point-and-line display diminishes - run out of plotting dimensions.

Insights from the ‘overlap’ that results in from a space-based layout as information.

Here you see the clustering evident in movie co-staring for about 8000 actors.



This figure contains over 29,000 social science authors. The two dense regions reflect different topics.



Foundations

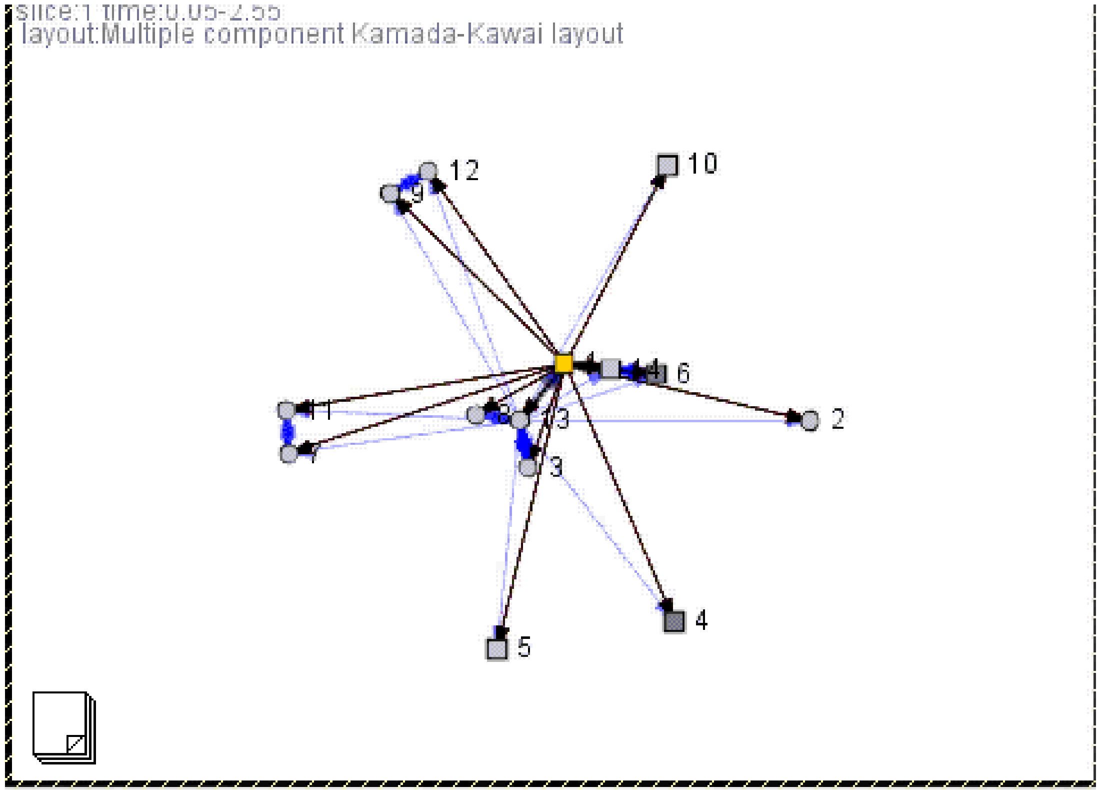
Graphs and time

Adding time to social networks is also complicated, run out of space to put time in most network figures.

One solution: *animate* the network - make a movie!

Here we see streaming interaction in a classroom, where the teacher (yellow square) has trouble maintaining order.

The SoNIA software program (McFarland and Bender-deMoll)



Foundations

Methods

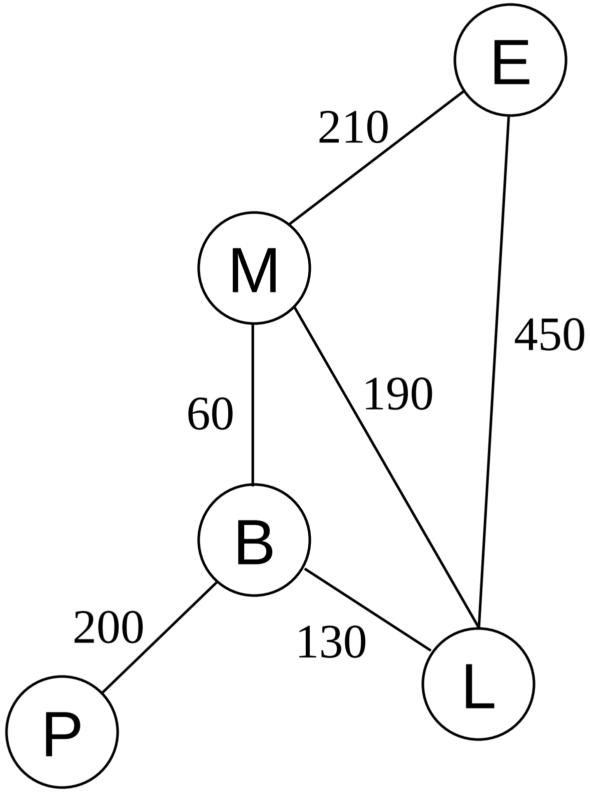
Graphs are cumbersome to work with analytically, though there is a great deal of good work to be done on using visualization to build network intuition.

Recommendation: use layouts that optimize on the feature you are most interested in.

A graph G(V,E) is a set of vertices (V) together with a set of edges (E). Some synonyms:

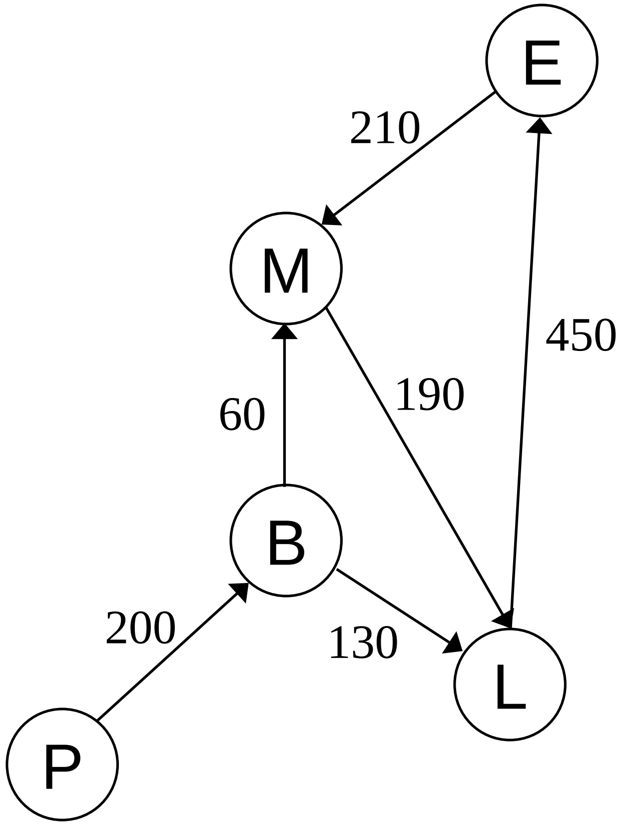
	Vertices	Edges
Mathematics	node, point	line, arc, link
Sociology	actor, agent	tie

A graph is vertices and edges



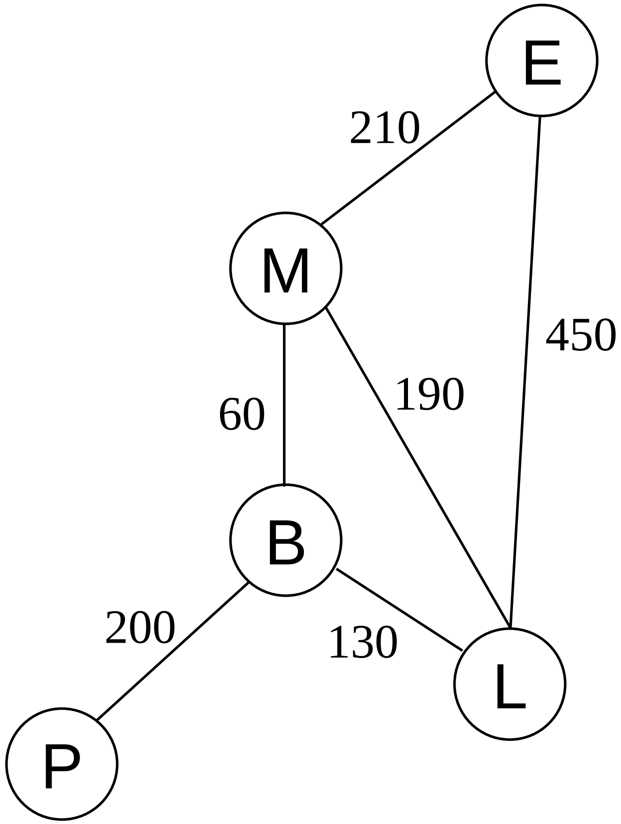
- A graph is *vertices* joined by *edges*
  - i.e. A set of vertices *V* and a set of edges *E*
- A vertex is defined by its name or label
- An edge is defined by the two vertices which it connects, plus optionally:
  - An order of the vertices (*direction*)
  - A *weight* (usually a number)
- Two vertices are *adjacent* if they are connected by an edge
- A vertex's *degree* is the no. of its edges

Directed graph (*digraph*)



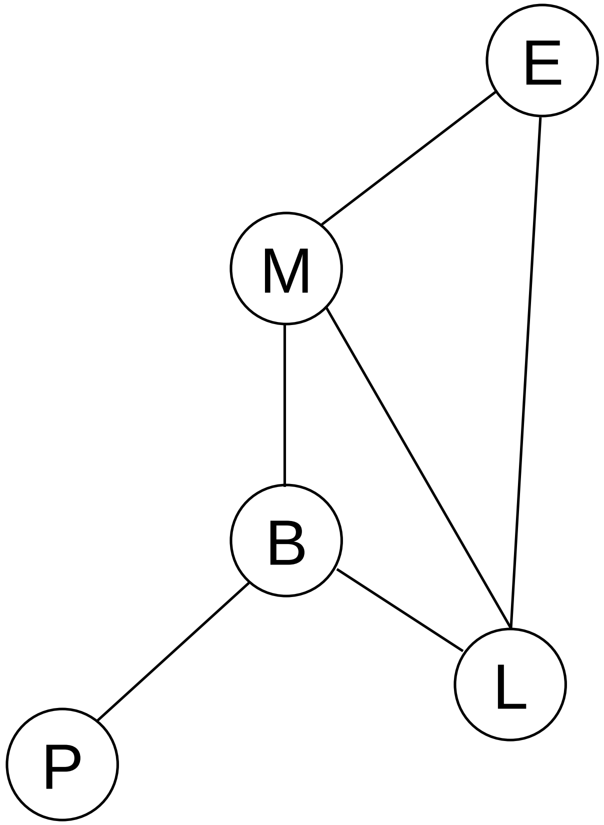
- Each edge is an *ordered* pair of vertices, to indicate direction
  - Lines become arrows
- The *indegree* of a vertex is the number of incoming edges
- The *outdegree* of a vertex is the number of outgoing edges

Traversing a graph (1)



- A *path* between two vertices exists if you can traverse along edges from one vertex to another
- A path is an ordered list of vertices
- *length*: the number of edges in the path
- *cost*: the sum of the weights on each edge in the path
- *cycle*: a path that starts and finishes at the same vertex
  - An *acyclic* graph contains no cycles

Traversing a graph (2)

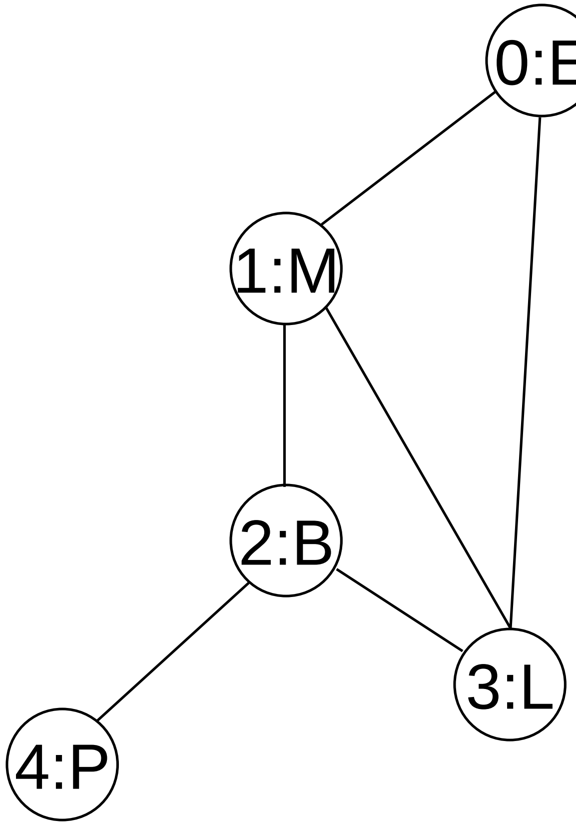


- Undirected graphs are *connected* if there is a path between any pair of vertices
- Digraphs are usually either *densely* or *sparsely* connected
  - Densely: the ratio of number of edges to number of vertices is large
  - Sparsely: the above ratio is small

Two graph representations: adjacency matrix and adjacency list

- Adjacency matrix
  - n vertices need a $n \times n$ matrix (where $n = |V|$, i.e. the number of vertices in the graph) - can store as an array
  - Each position in the matrix is 1 if the two vertices are connected, or 0 if they are not
  - For weighted graphs, the position in the matrix is the weight
- Adjacency list
  - For each vertex, store a linked list of adjacent vertices
  - For weighted graphs, include the weight in the elements of the list

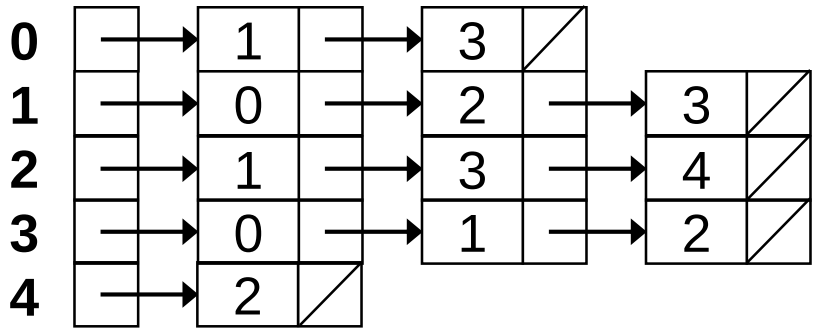
Representing an unweighted, undirected graph (example)



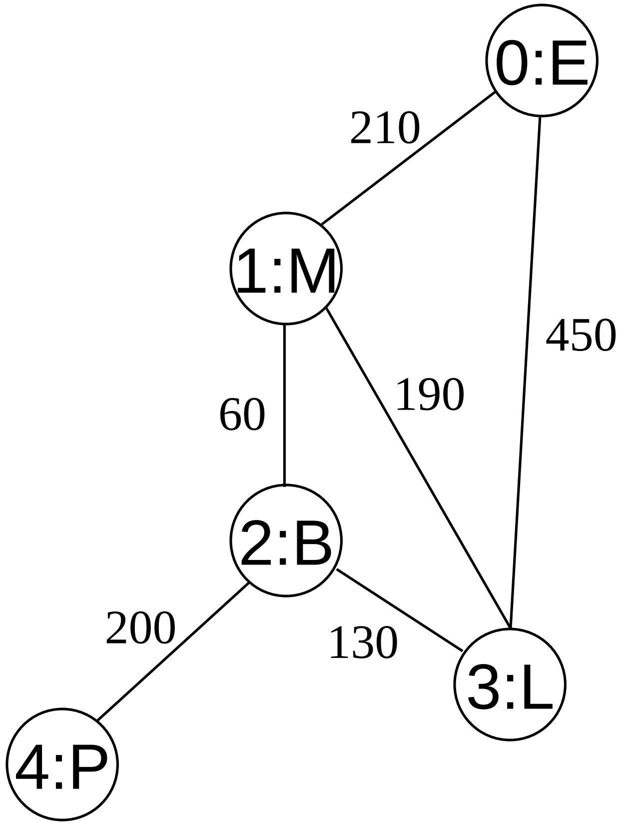
Adjacency matrix

	0	**1**	**2**	**3**	**4**
0	0	1	0	1	0
1	1	0	1	1	0
2	0	1	0	1	1
3	1	1	1	0	0
4	0	0	1	0	0

Adjacency list

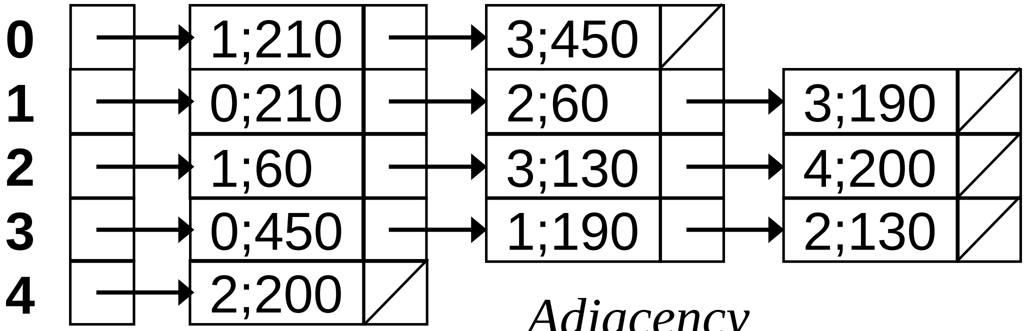


Representing a *weighted*, undirected graph (example)



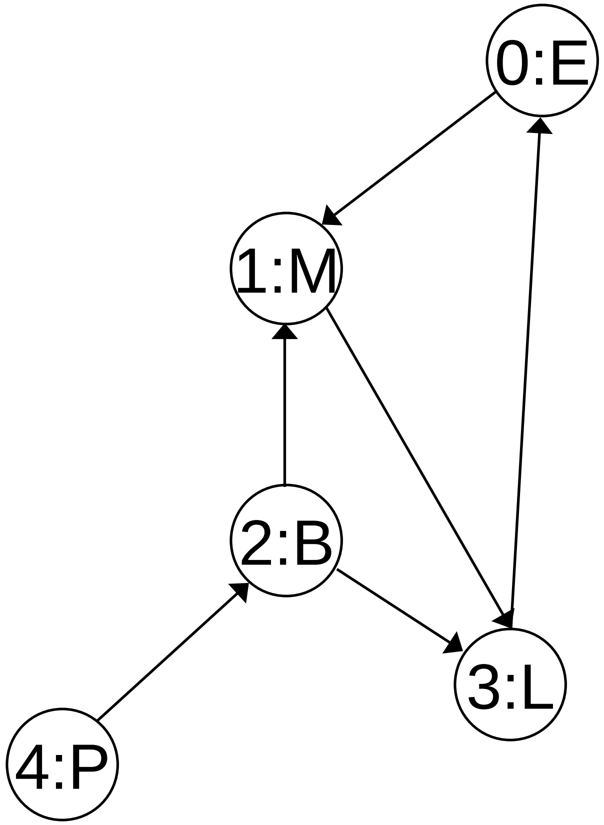
	0	1	2	3	4
0	0	210	0	450	0
1	210	0	60	190	0
2	0	60	0	130	200
3	450	190	130	0	0
4	0	0	200	0	0

Adjacency matrix



Adjacency list

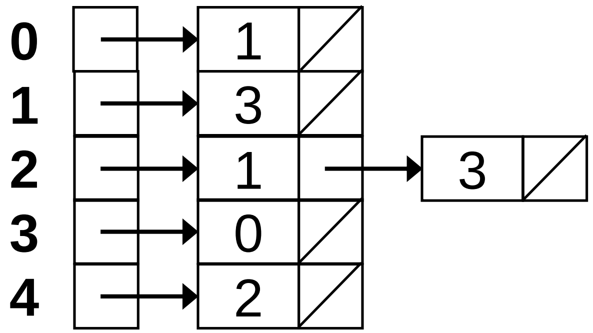
Representing an unweighted, *directed* graph (example)



Adjacency matrix

	0	1	2	3	4
0	0	1	0	0	0
1	0	0	0	1	0
2	0	1	0	1	0
3	1	0	0	0	0
4	0	0	1	0	0

Adjacency list



Comparing the two representations

- Space complexity
  - Adjacency matrix is $O(|V|^2)$
  - Adjacency list is $O(|V| + |E|)$
    - $|E|$ is the number of edges in the graph
- Static versus dynamic representation
  - An adjacency matrix is a *static* representation: the graph is built ‘in one go’, and is difficult to alter once built
  - An adjacency list is a *dynamic* representation: the graph is built incrementally, thus is more easily altered during run-time

Algorithms involving graphs

- Graph traversal
- Shortest path algorithms
  - In an unweighted graph: shortest *length* between two vertices
  - In a weighted graph: smallest *cost* between two vertices
- Minimum Spanning Trees
  - Using a tree to connect all the vertices at lowest total cost

Graph traversal algorithms

- When traversing a graph, we must be careful to avoid going round in circles!
- We do this by marking the vertices which have already been visited
- **Breadth-first search** uses a queue to keep track of which adjacent vertices might still be unprocessed
- **Depth-first search** keeps trying to move forward in the graph, until reaching a vertex with no outgoing edges to unmarked vertices

Shortest path (unweighted)

- The problem: Find the shortest path from a vertex *v* to every other vertex in a graph
- The *unweighted path* measures the number of edges, ignoring the edge’s weights (if any)

Shortest unweighted path: simple algorithm

$\boxed{\text{For a vertex } v,\ d_v \text{ is the distance between a starting vertex and } v}$

1 Mark all vertices with d_v = infinity

2 Select a starting vertex s, and set $d_s = 0$, and set *shortest* = 0

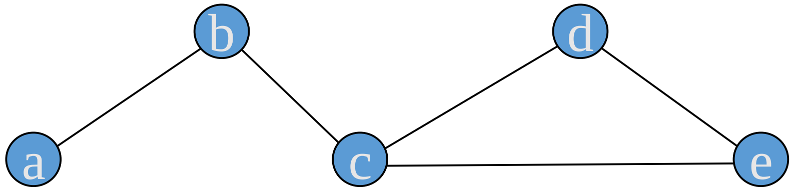
3 For all vertices v with d_v = *shortest*, scan their adjacency lists for vertices w where d_w is infinity

- For each such vertex w, set d_w to *shortest*+1

4 Increment *shortest* and repeat step **3**, until there are no vertices w

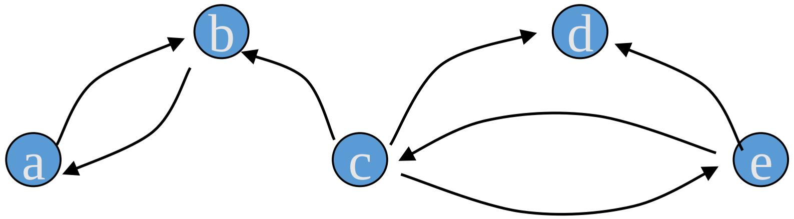
Foundations
Build a socio-matrix

From pictures to matrices



Undirected, binary

	a	b	c	d	e
a		1			
b	1		1		
c		1		1	1
d			1		1
e			1	1	



Directed, binary

	a	b	c	d	e
a		1			
b	1				
c		1		1	1
d					
e			1	1	

From matrices to lists

	a	b	c	d	e
a		1			
b	1		1		
c		1		1	1
d			1		1
e			1	1	

Adjacency List

a	b
b	a c
c	b d e
d	c e
e	c d

Arc List

a b
b a
b c
c b
c d
c e
d c
d e
e c
e d

Basic Measures

For greater detail, see:

http://www.analytictech.com/networks/graphtheory.htm

Volume

The first measure of interest is the simple volume of relations in the system, known as density, which is the average relational value over all dyads. Under most circumstances, it is calculated as:

$$\Delta = \frac{\sum X}{N(N-1)} \qquad 1 \geq \Delta > 0$$

Volume

At the individual level, volume is the number of relations, sent or received, equal to the row and column sums of the adjacency matrix.

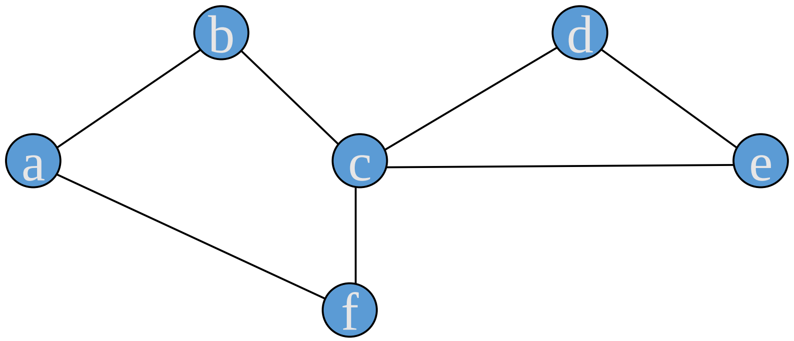
	a	b	c	d	e
a		1			
b	1				
c		1		1	1
d					
e			1	1	

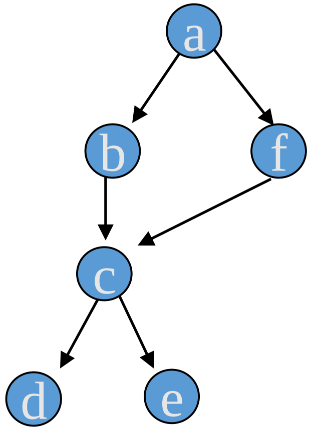
Node	In-Degree	Out-Degree
a	1	1
b	2	1
c	1	3
d	2	0
e	1	2
Mean:	7/5	7/5

Basic Measures

Reachability

Indirect connections are what make networks systems. One actor can *reach* another if there is a *path* in the graph connecting them.





SNA disciplines

More diverse than expected!

- Sociology
- Political Science
- Business
- Economics
- Sciences
- Computer science
- Information science
- Others?

SNA and the Web 2.0

- Wikis
- Blogs
- Folksonomies
- Collaboratories
- What next?

Computational SNA Models

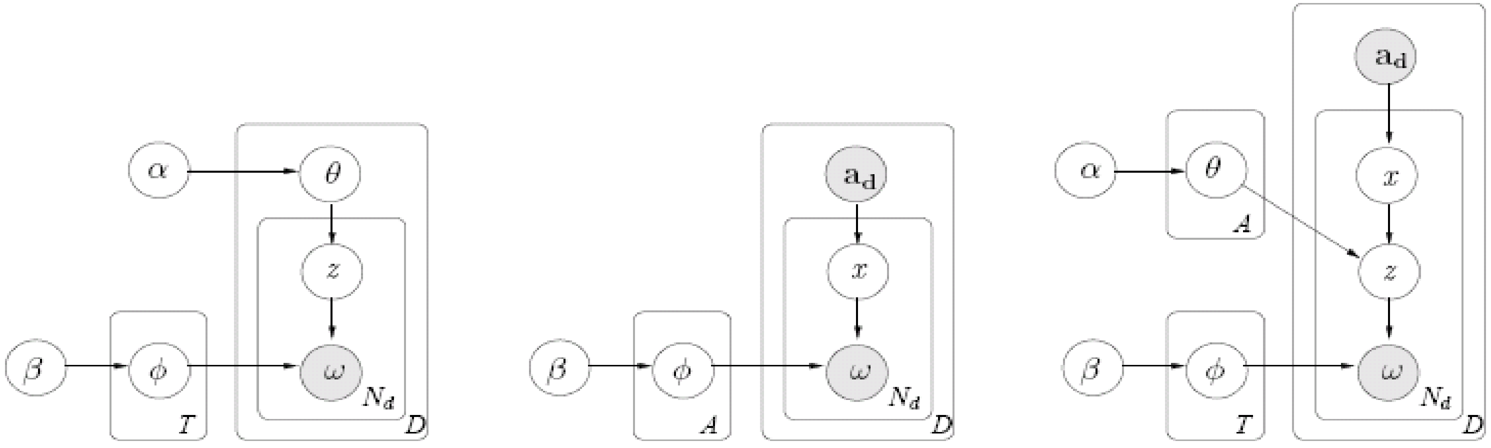
New models are emerging

Very large network analysis is possible!

- Deterministic - algebraic
  - Early models still useful
- Statistical
  - Descriptive using many features
    - Diameter, betweeness,
- Probabilistic graphs
  - Generative
    - Creates SNA based on agency, documents, geography, etc.
    - Community discovery and prediction

Graphical models

- Modeling the document generation



(a) Topic-Word (LDA) (b) Author-Word (c) Author-Topic

Existing three generative models.
Three variables in the generation of documents are considered:
(1) authors; (2) words; and (3) topics (latent variable)

Theories used in SNA

- Graph/network
  - Heterogeneous graphs
  - Hypergraphs
  - Probabilistic graphs
- Economics/game theory
- Optimization
- Visualization/HCI
- Actor/Network
- Many more

Future of social networks?

Top End User Predictions for 2010 - Gartner

- By 2012, Facebook will become the hub for social networks integration and Web socialization.
- Internet marketing will be regulated by 2015, controlling more than $250 billion in Internet marketing spending worldwide.
- By 2014, more than three billion of the world's adult population will be able to transact electronically via mobile and Internet technology.
- By 2015, context will be as influential to mobile consumer services and relationships as search engines are to the Web.
- By 2013, mobile phones will overtake PCs as the most common Web access device worldwide.

Open questions

- Scalability
- Data acquisition and data rights
- Search (socialnetworkrank?)
    - CollabSeer
- Trust
- Heterogeneous network analysis
- Business models!

Social networks vs social networking

- **Social networks** are links of actors and their relationships usually represented as a graph or network

- **Social networking** is the actual implementation of social networks in the digital world or media
  - A social network service focuses on building and reflecting of social networks or social relations among people, e.g., who share interests and/or activities. A social network service essentially consists of a representation of each user (often a profile), his/her social links, and a variety of additional services. Most social network services are web based and provide means for users to interact over the internet, such as e-mail and instant messaging.

Facebook vs Google

Great Wall of Facebook: The Social Network's Plan to Dominate the Internet — and Keep Google Out

By Fred Vogelstein 06.22.09

Instead of working together to reach the promised land of online brand advertising, Facebook and Google are racing to see who can get there first.
Illustration: Brent Humphreys

Web 2.0

- A perceived second generation of web development and design, that aims to facilitate communication, secure information sharing, interoperability, and collaboration on the World Wide Web.
- Web 2.0 concepts have led to the development and evolution of web-based communities, hosted services, and applications such as social-networking sites, video-sharing sites, wikis, blogs, and folksonomies.

Social Media

- Information content created by people using highly accessible and scalable publishing technologies that is intended to facilitate communications, influence and interaction with peers and with public audiences, typically via the Internet and mobile communications networks.
- The term most often refers to activities that integrate technology, telecommunications and social interaction, and the construction of words, pictures, videos and audio.
- Businesses also refer to social media as user-generated content (UGC) or consumer-generated media (CGM).

Social Media on Web 2.0

- **Multimedia**
  - Photo-sharing: Flickr
  - Video-sharing: YouTube
  - Audio-sharing: imeem
- **Entertainment**
  - Virtual Worlds: Second Life
  - Online Gaming: World of Warcraft
- **News/Opinion**
  - Social news: Digg, Reddit
  - Reviews: Yelp, epinions
- **Communication**
  - Microblogs: Twitter, Pownce
  - Events: Evite
- ***Social Networking Services:***
  - ***Facebook, LinkedIn, MySpace***

Social Search

- Social search engines are an important web development which utilise the popularity of social networking services.
- There are various kinds of social search engine, but sites like Wink and Spokeo generate results by searching across the public profiles of multiple social networking sites, allowing the creation of web-based dossiers on individuals.
- This type of people search cuts across the traditional boundaries of social networking site membership, although any data retrieved should already be in the public domain.

Things you can do in a Social Network

- Communicating with existing networks, making and developing friendships/contacts
- Represent themselves online, create and develop an online presence
- Viewing content/finding information
- Creating and customizing profiles
- Authoring and uploading content
- Adding and sharing content
- Posting messages – public & private
- Collaborating with other people