Foundations of Data Science

Lecture 7: Graph and Patterns

MING GAO

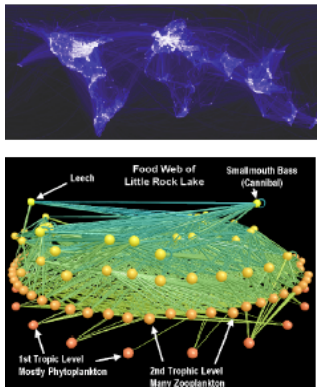
DaSE@ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

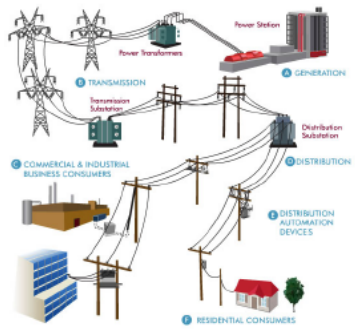
Oct. 28, 2016

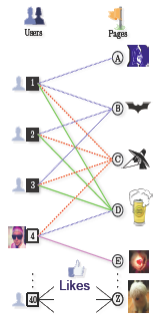
Outline

Outline

Graphs - why should we care?







Networks in real world

- "YahooWeb graph": 1B vertices(Web sites), 6B edges (http links)
- Facebook, Twitter, etc: more than 1B users
- Food Web: all biologies, food chain
- Power-grid: vertices (plants or consumers), edges (power lines)
- Airline route: vertices (airports), edges (flights)
- Adoption: users purchase products, adopt services, etc.

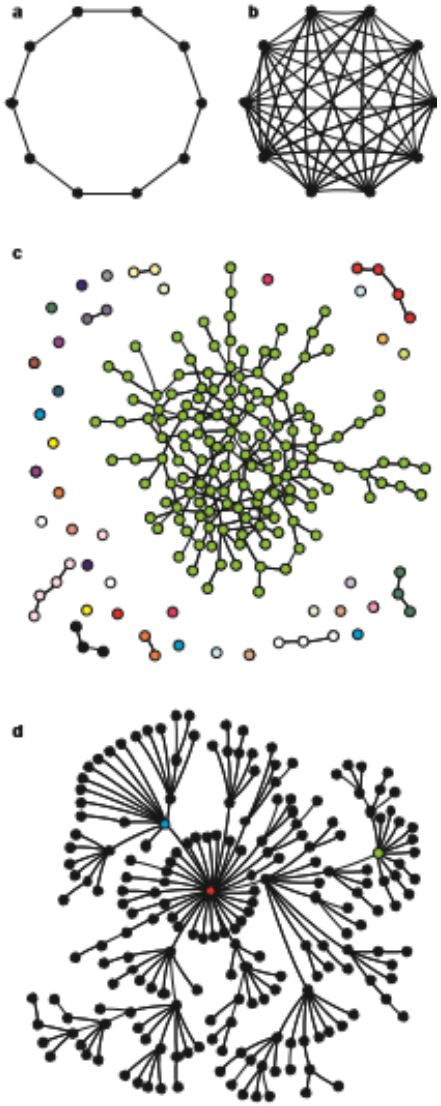
Motivation questions

Questions

- What do real graphs look like?
  - What properties of vertices, edges are important to model?
  - What local and global properties are important to measure?
- Are graphs helpful to understand the real world?
  - Social influence
  - Recommendation
  - Information propagation
  - Human behaviors
- Is a sub-graph "normal" (Water army, fraud detection, spam filtering, etc)?
- How to generate realistic graphs?
- How to get a "good" sample of a network?
- How to design an efficient algorithm to handle large-scale graphs?

Models for complex networks

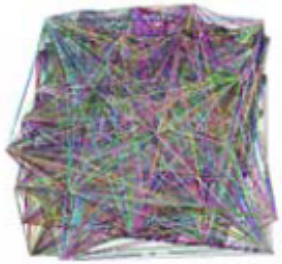
Steven H. S. proposes the model for complex networks in Nature 2001.



Model

- Regular network: each node has exactly the same number of edges.
- Random network: it is obtained by starting with a set of n isolated vertices and adding successive edges between them at random.
- Scale-free network: it grows via attaching new nodes to previously existing nodes randomly, while the probability is proportional to the degree of the target node, i.e., richly connected nodes tend to get richer, leading to the formation of hubs and a skewed degree distribution with a heavy tail.(Matthew Effect or Pareto's Law)

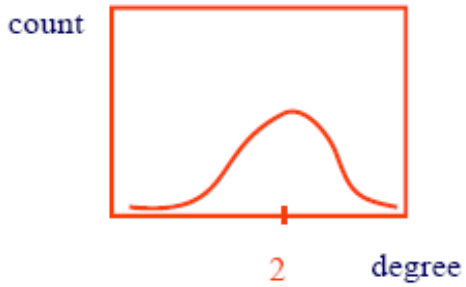
Are real graphs random?

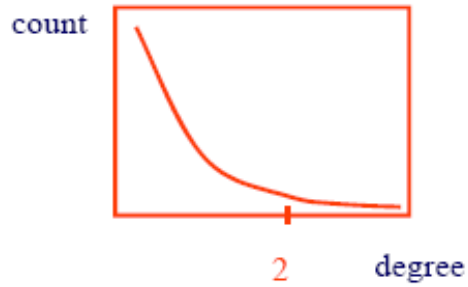


Looks random - right?

How does the Internet look like? Any rules?

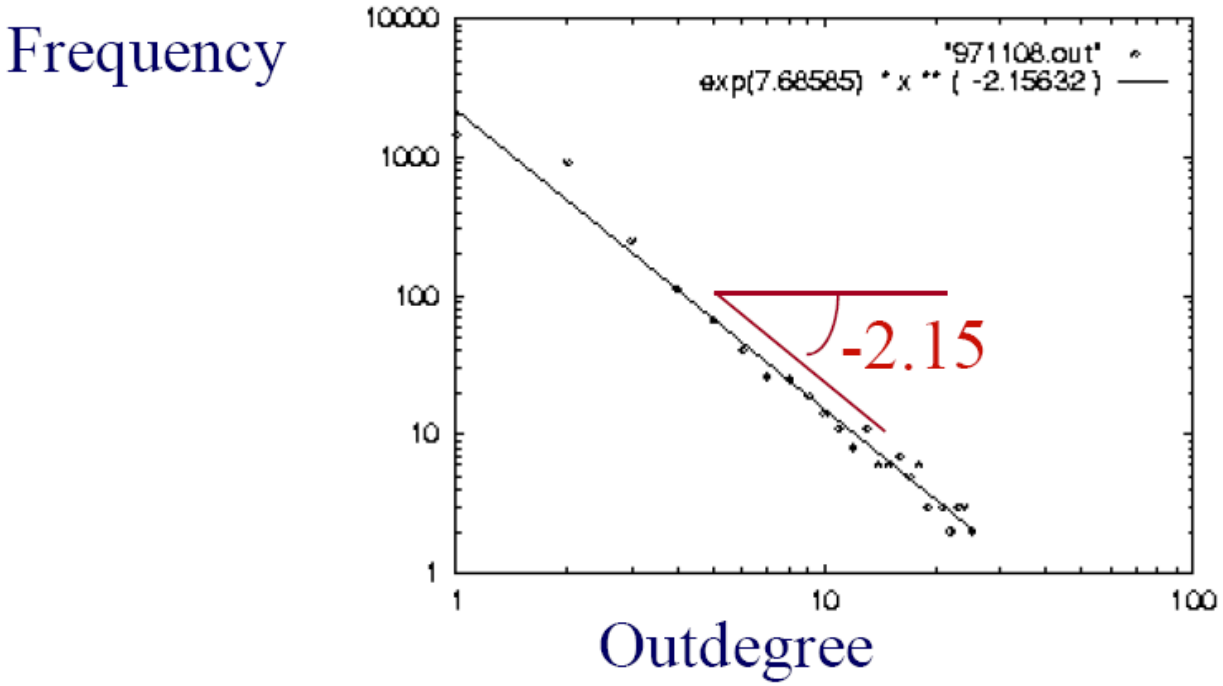
- Diameter: would you like to guess?
- In- and out- degree distributions: if average degree is 2, what is the most probable degree?
- Other (surprising) patterns?





Outline

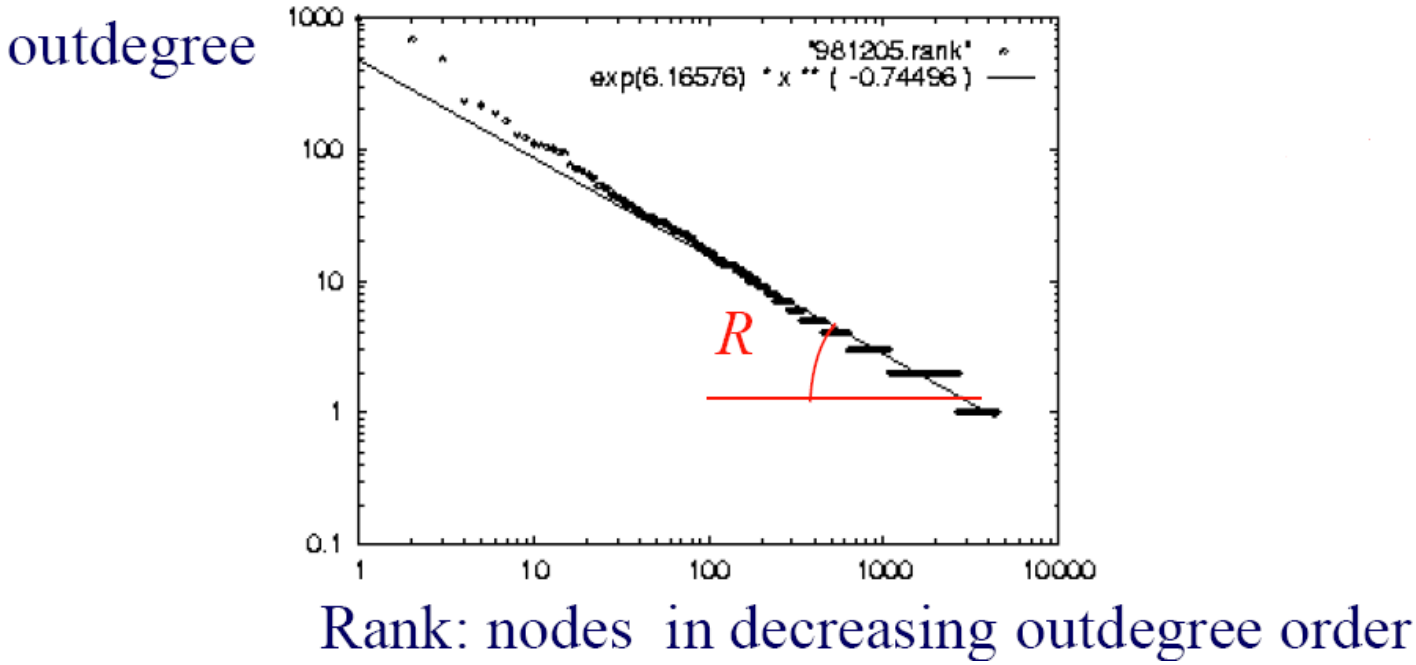
Power-law I



Internet topology [SIGCOMM 99]

- Out-degree distribution is plotted in log-log scale.
- It forms a line with a slope ~ -2.15
- $freq. = deg.^{-2.15}$

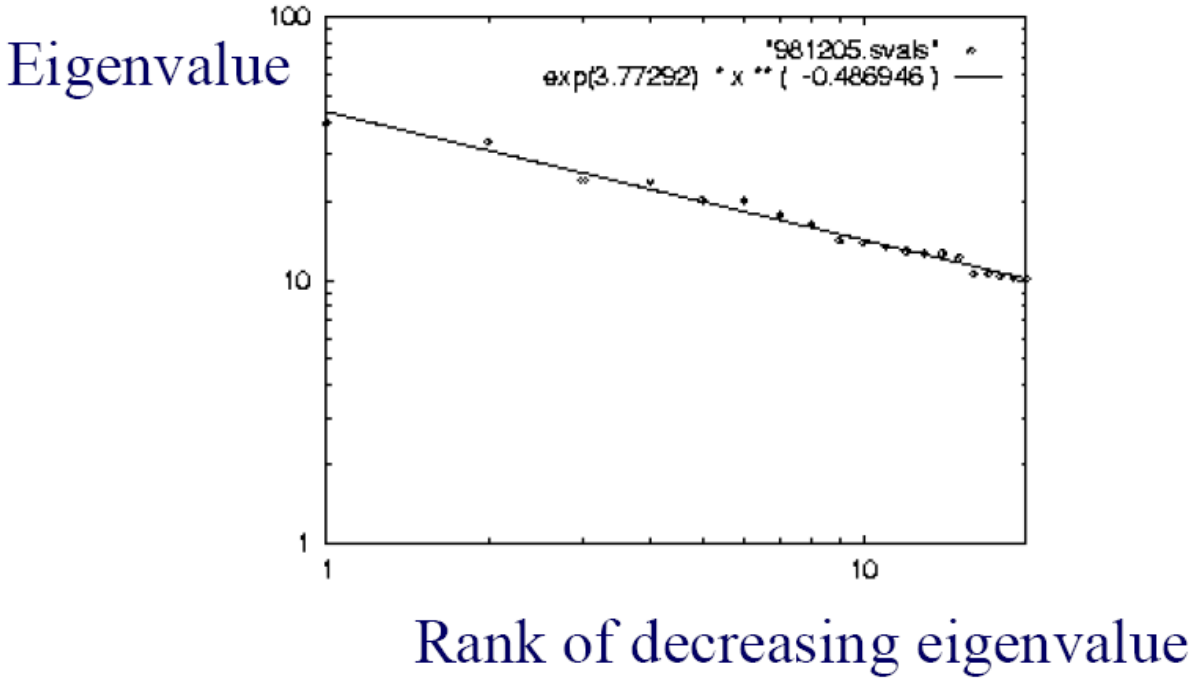
Power-law II



Rank of out-degrees [ICDE 09]

- Vertices are ranked in decreasing out-degree order, and plotted in log-log scale.
- It forms a line with a slope ~ -0.74
- $deg. = rank^{-0.74}$

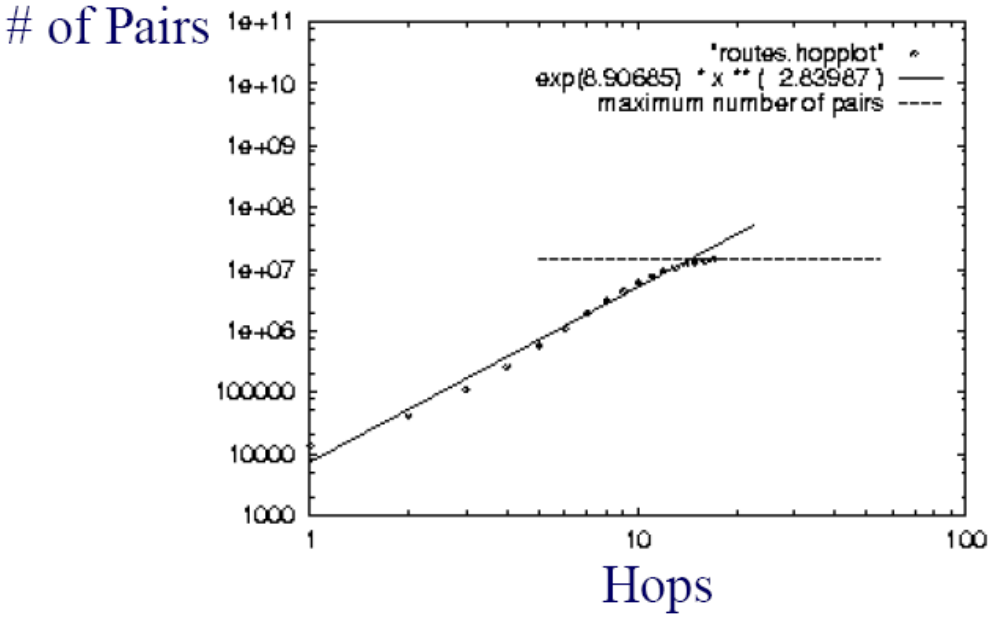
Power-law III

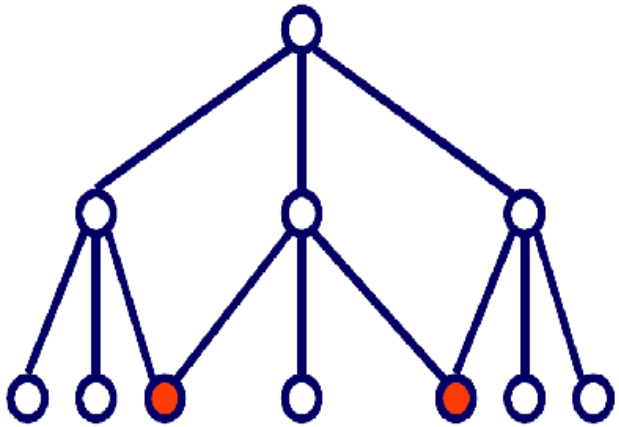


Rank of eigenvalues [ICDE 09]

- Eigenvalues of adjacency matrix (top 20) are ranked in decreasing order, and plotted in log-log scale.
- It forms a line with a slope ~ -0.48
- $eigen. = rank^{-0.48}$

Power-law IV

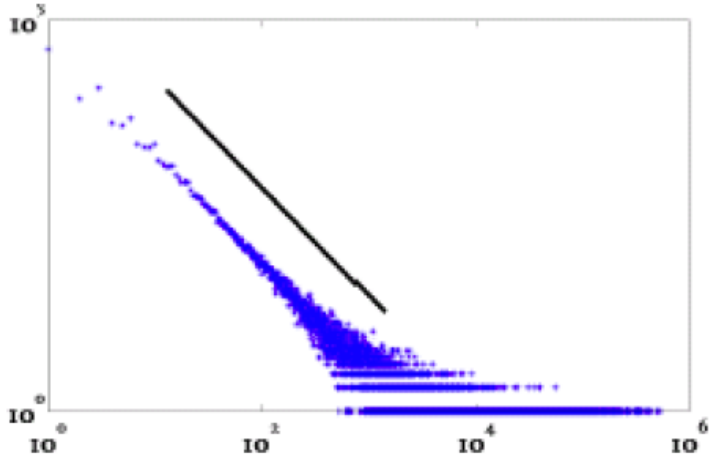




Hop plot [ICDE 09]

- How many neighbors within $1, 2, \cdots, h$ hops? $(\sum_{i=1}^{h} avg.^i)$
- Pairs of vertices are plotted in log-log scale. It forms a line with a slope ~ 2.83
- $pairs. = hop^{2.83}$

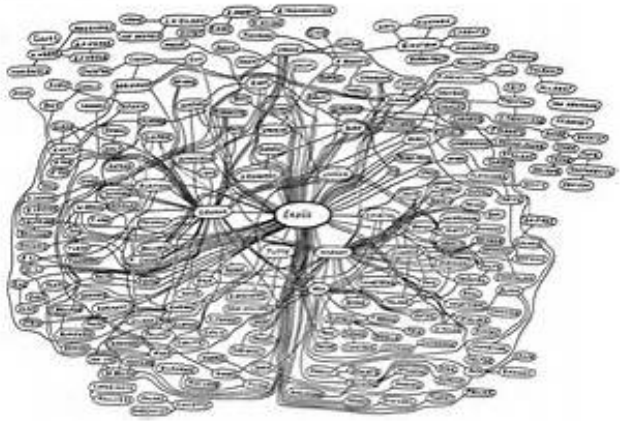
Power-law V

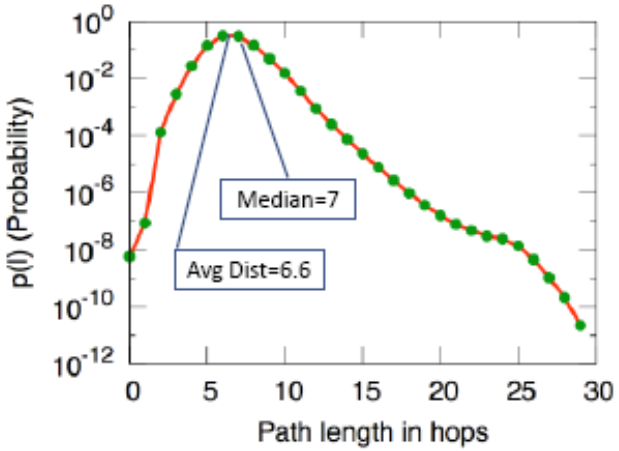


Counting of triangle [ICDM 08]

- X-axis: # of triangles a vertex participates in
- Y-axis: count of such vertices
- In log-log scale, the plot is almost linear.

Erdös number



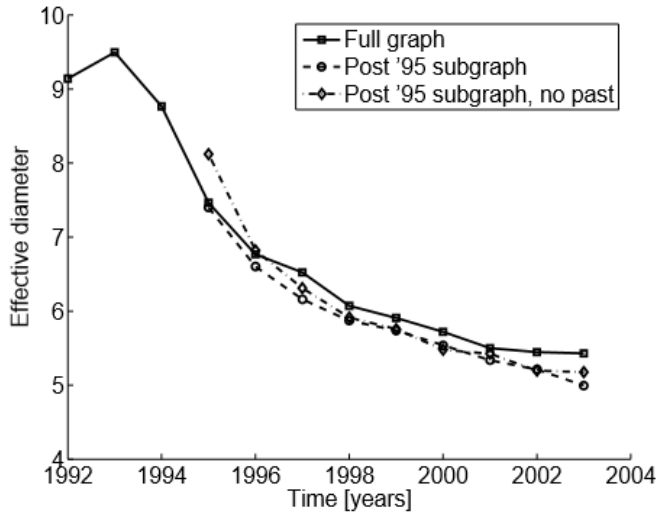


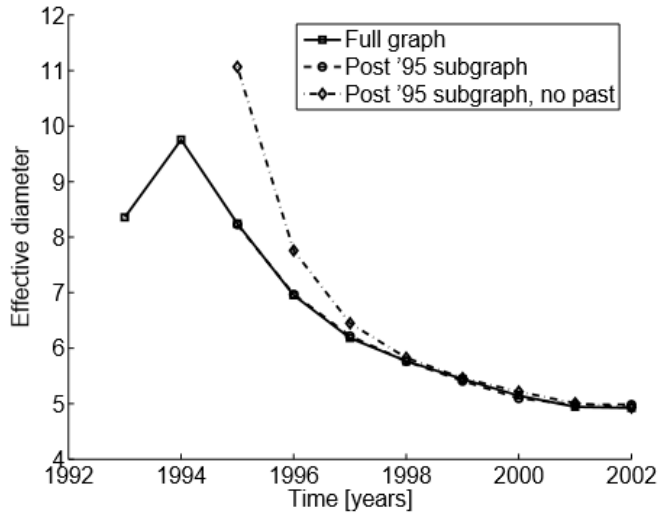
Small world - six degrees of separation

The world looks "small" when you think of how short a path of friends it takes to get from you to almost anyone else

- Stanley Milgram and his colleagues in the 1960s did an experiment.
- 296 randomly chosen starters asked to forward a letter to a "target" person, a stockbroker in Bostons suburb.
- They found the six degrees of separation, and the same observation found by Jure Leskovec on Miscrosoft Instant Message[WWW 2008].

Shrinking diameter





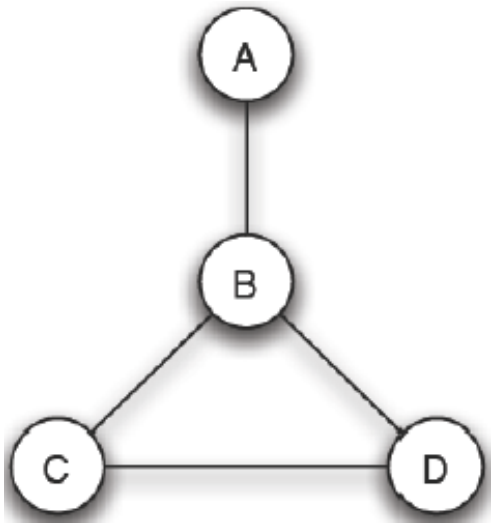
Citation or patents networks [KDD 05]

For citation network, they collected citations among Physics papers.

- 11 years data
  - 29,555 papers
  - 352,807 citations
- For each month, create a graph of all citations up to the month.
- The diameters are plotted in the figures.

Outline

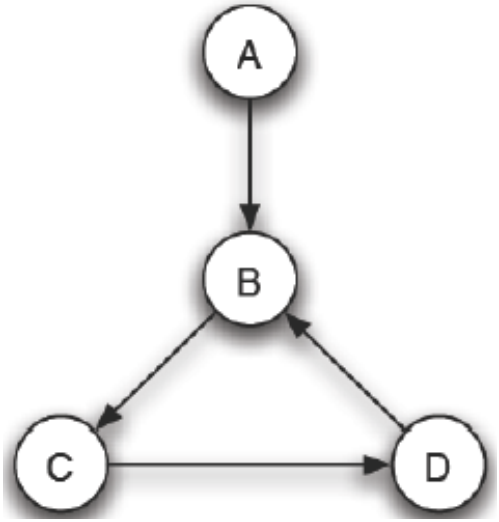
Graph types



Undirected graph

A undirected graph on 4 vertices

- Degree: # edges connected to the vertex
- Degree 0 vertex: isolated vertex

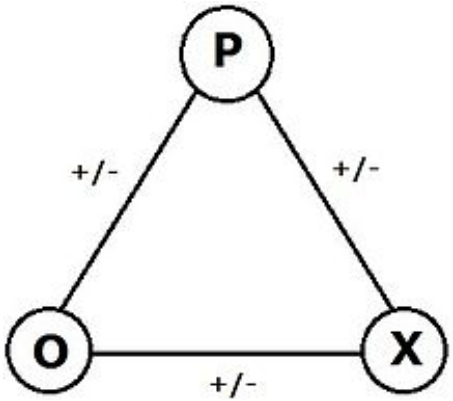


Directed graph

A directed graph on 4 vertices

- In-degree: # incoming edges to the vertex
- Out-degree: # outgoing edges to the vertex
- Degree: in-degree + outdegree

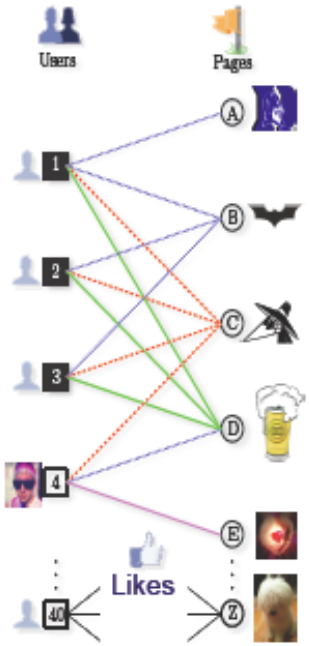
Graph types cont.



Signed graph

A signed graph on 3 vertices

- Positive-degree: # edges associated with positive labels
- Negative-degree: # edges associated with negative labels



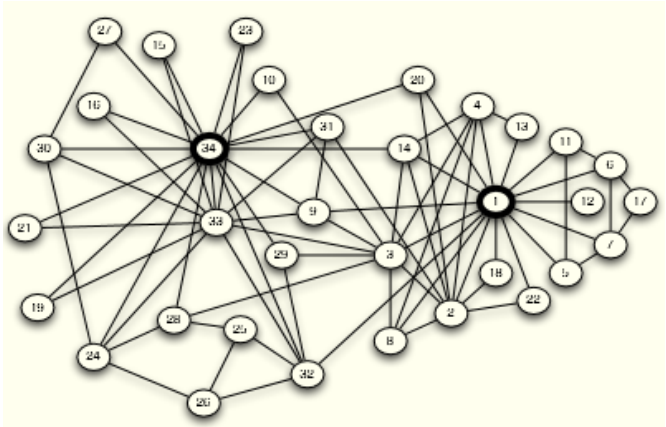
Bipartite graph

Users interact on social platforms

- Reply network
- Retweet network
- Adoption network

Outline

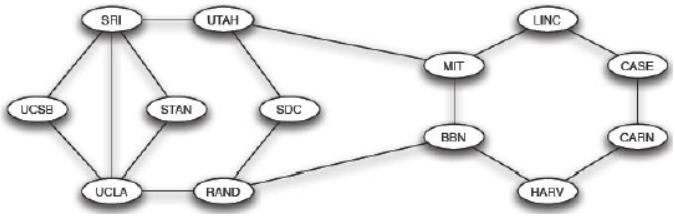
Paths



Path

Path is a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge

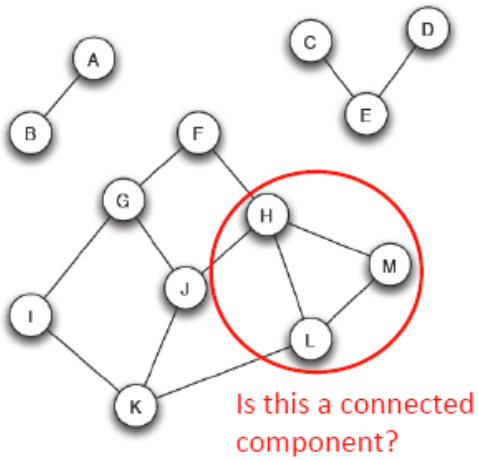
- Simple path does not repeat nodes.
- The length of path is the number of nodes in the path



Cycle

Cycle is a path with at least three edges, in which the first and last nodes are the same. Every edge in the 1970 Arpanet belongs to a cycle, and this was by design. Why?

Connectivity



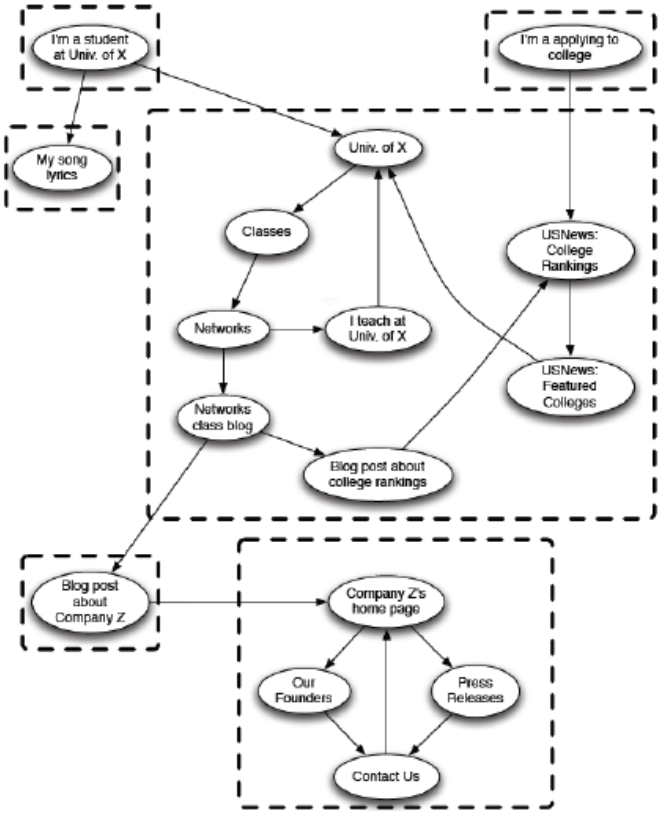
Connected component

A connected component is a subset of nodes s.t.:

- Every node in the subset has a path to every other; and
- The subset is not part of some larger set with the property that every node can reach every other.

A graph is connected if for every pair of nodes, there is a path between them, i.e., the whole graph is a connected component.

Strongly connected component

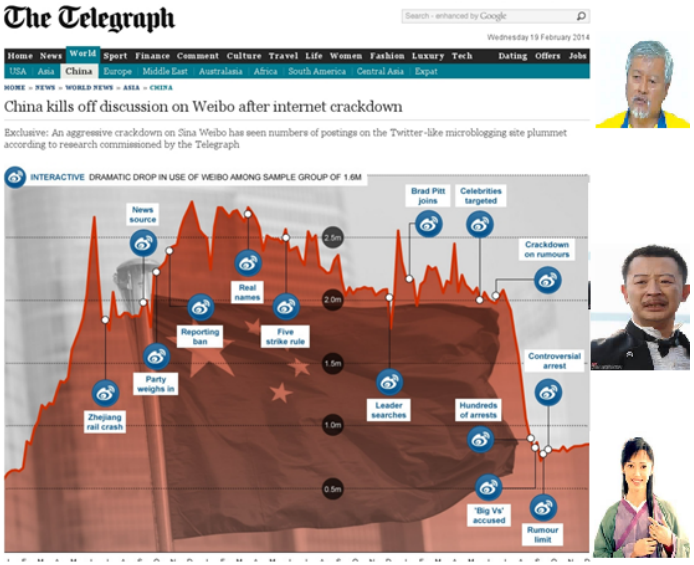


Strongly connected component

A *directed graph* is strongly connected if there is a path from every node to every other node.

- Edges of the path must follow the forward direction.
- A undirected graph can be treated as a bidirectional graph. Thus connected component in a directed graph is also a SCC.
- In a strongly connected component, there are followers and followees for each node.
- SCCs can be treated as super-nodes.

Giant component

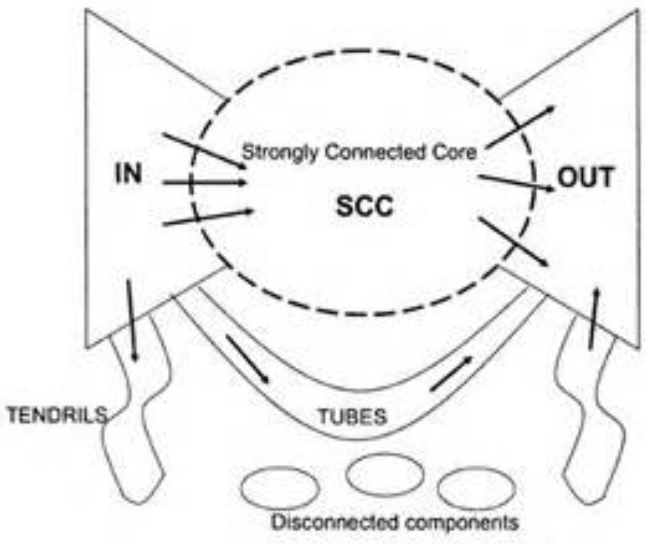


Giant connected component

A connected component that contains a significant fraction of all the nodes.

- When a network (e.g., friendship network) contains a giant component, it almost always contains only one.
- The other connected components are very small by comparison.
- The largest connected component would break apart into three distinct components if this node were removed [related to robustness of network].

Web giant component



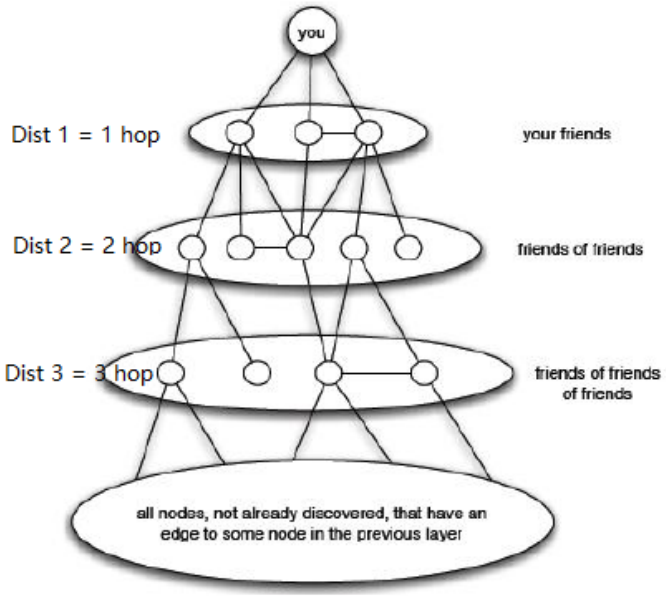
200 M pages, 1.5 B hyperlinks

Web graph

Web contains a giant strongly connected component (containing home pages of many of the major commercial, governmental, and nonprofit organizations)

- IN: nodes that can reach the giant SCC but cannot be reached from it, i.e., nodes that are "upstream" of it.
- OUT: nodes that can be reached from the giant SCC but cannot reach it, i.e., nodes are "downstream" of it.

Distance and diameter



Distance or Geodesic distance

The distance between two vertices in a graph is the number of edges in a shortest path.

- Diameter is the length of the "longest shortest path" between any two vertices of a graph.
- Erdös number is bounded by diameter of a graph.
- Research community is a small world [Duncan Watts and Steven Strogatz 1998].

Mean Geodesic distance of undirected networks

Definition [SIAM review 45 2003]

$$L = \frac{1}{\frac{1}{2}n(n+1)\sum_{i \geq j} d_{ij}},$$

where n denotes # of nodes, and d_{ij} is the shortest distance between nodes i and j.

- Mean Geodesic distance includes distance to itself.
- Can be computed in $O(mn)$ using breadth first search, where m denotes # of edges.
- What happens if the network has multiple connected components?
- Harmonic mean (can have multiple connected components):

$$L^{-1} = \frac{1}{\frac{1}{2}n(n+1)\sum_{i \geq j} d_{ij}^{-1}}$$

Summarization

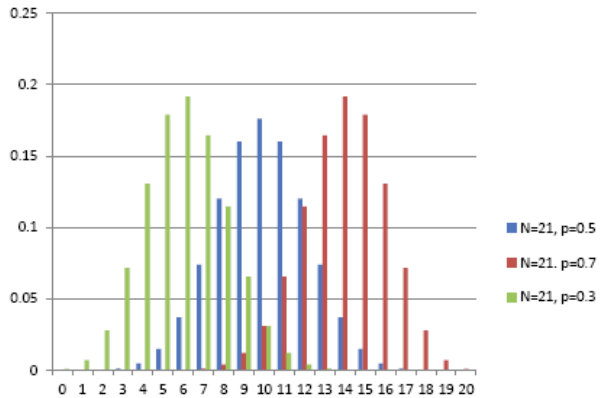
	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 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

Network generation

Erdös-Renyi model

Erdös-Renyi model is known as the random graph model, which generates undirected random graphs.

- Parameters: N (# vertices) and p (probability of forming an edge)
- For each possible node pair, the approach generates an edge with probability p. Thus, # edges $= \frac{pN(N-1)}{2}$.
- Degree distribution:
  - $P(\text{node has degree k}) = \binom{N-1}{k} p^k (1-p)^{N-1-k}$
  - Follows binomial distribution with mean $(N-1)p$ and variance $(N-1)p(1-p)$ (not power-law distribution).



Network generation cont.

Preferential attachment model

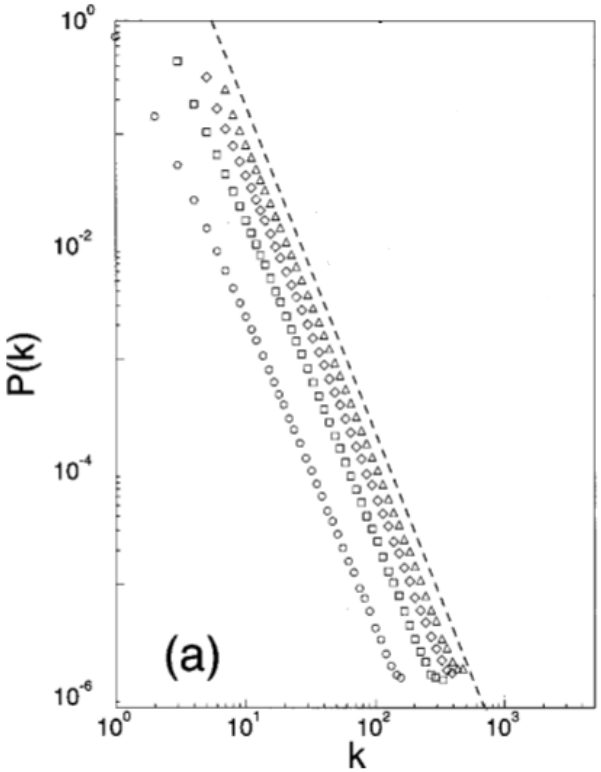
The more connected a node is, the more likely it is to receive new links (namely, Rich gets Richer, Matthew Effect or Paretos Law, etc.).

- Price Model
- Barabasi Albert Model

Price's preferential attachment model for citation networks

- Each new paper is generated with m citations (mean).
- New papers cite previous papers with probability proportional to their indegree (citations).
  - Each new paper is generated with m citations (mean).
  - New papers cite previous papers with probability proportional to their indegree (citations).
  - Power law with exponent $\alpha = 2 + \frac{1}{m}$ [Science 1965]

Network generation cont.



Barabasi Albert Model

- Start with an initial network of m_0 (≥ 2) nodes, and the degree of each node ≥ 1, otherwise it will always remain isolated.
- For each new node, connect it to m existing nodes i with a probability p_i, where $p_i = \frac{k_i}{\sum_j k_j}$, where k_i is degree of node i.
- Results in a single connected component with power-law degree distribution with $\alpha = 3$ [Reviews of Modern Physics 2003].

Take-home messages

- Graph
  - Motivations
  - Patterns
- Graph aspects
  - Graph types
  - Properties
- Network generation