Complex Networks
the large-scale structure of networks

2018.12.24(Mon)

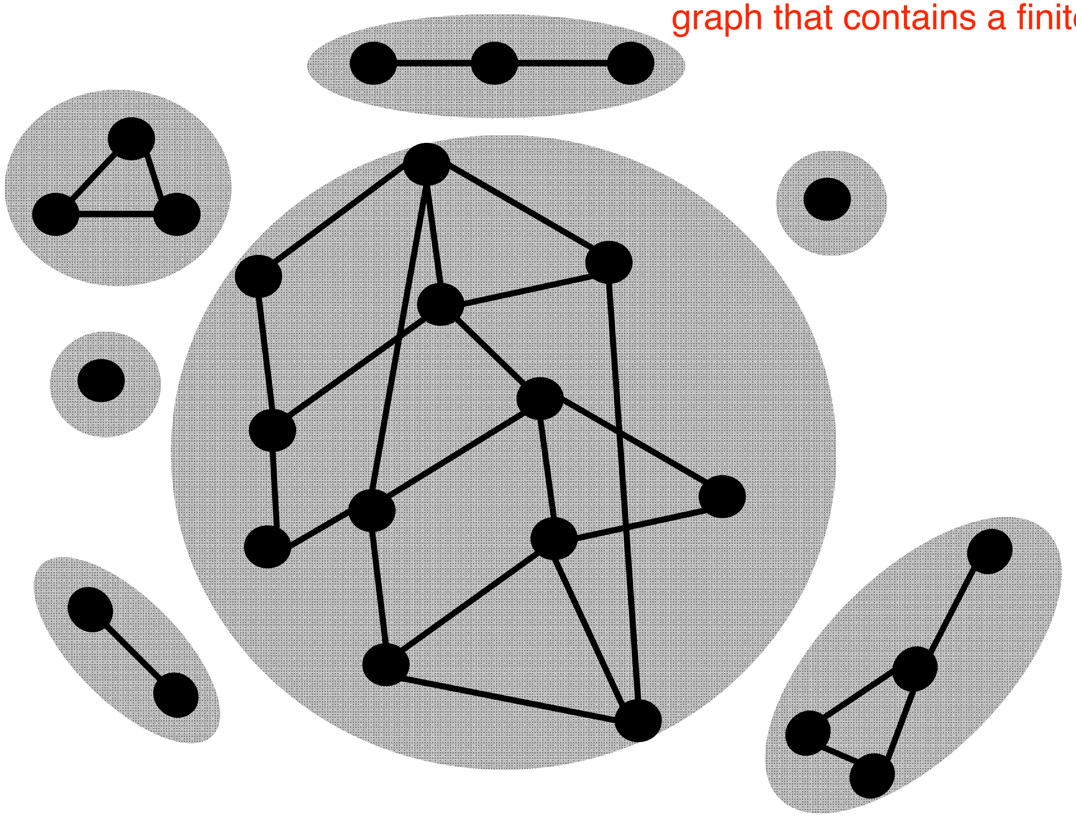
contents of this chapter

- component sizes
- path lengths and small-world effect
- degree distributions and power law
- clustering coefficient

components

- large component: usually more than half and not infrequently over 90%

a giant component is a connected component of a given random graph that contains a finite fraction of the entire graph's vertices.



metrics in real networks

S: the size of the largest component as a fraction of total network size

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).	S
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416	0.980
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323	0.876
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182	0.822
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313	0.838
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313	0.918
	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	0.952
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321	0.590
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45	0.503
	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	1.000
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74	0.914
	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	0.977
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157	1.000
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148	1.000
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416	1.000
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366	1.000
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318	0.998
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395	1.000
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155	1.000
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354	0.805
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214	0.996
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212	0.689
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204	1.000
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272	1.000
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421	0.967

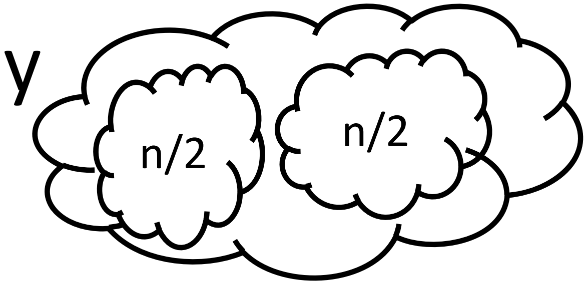
TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (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.

M. Newman "The structure and function of complex networks"
http://arxiv.org/abs/cond-mat/0303516

two large component
no large component

- two large components -> $(n/2)^2$ pairs

-> no connection is highly unlikely



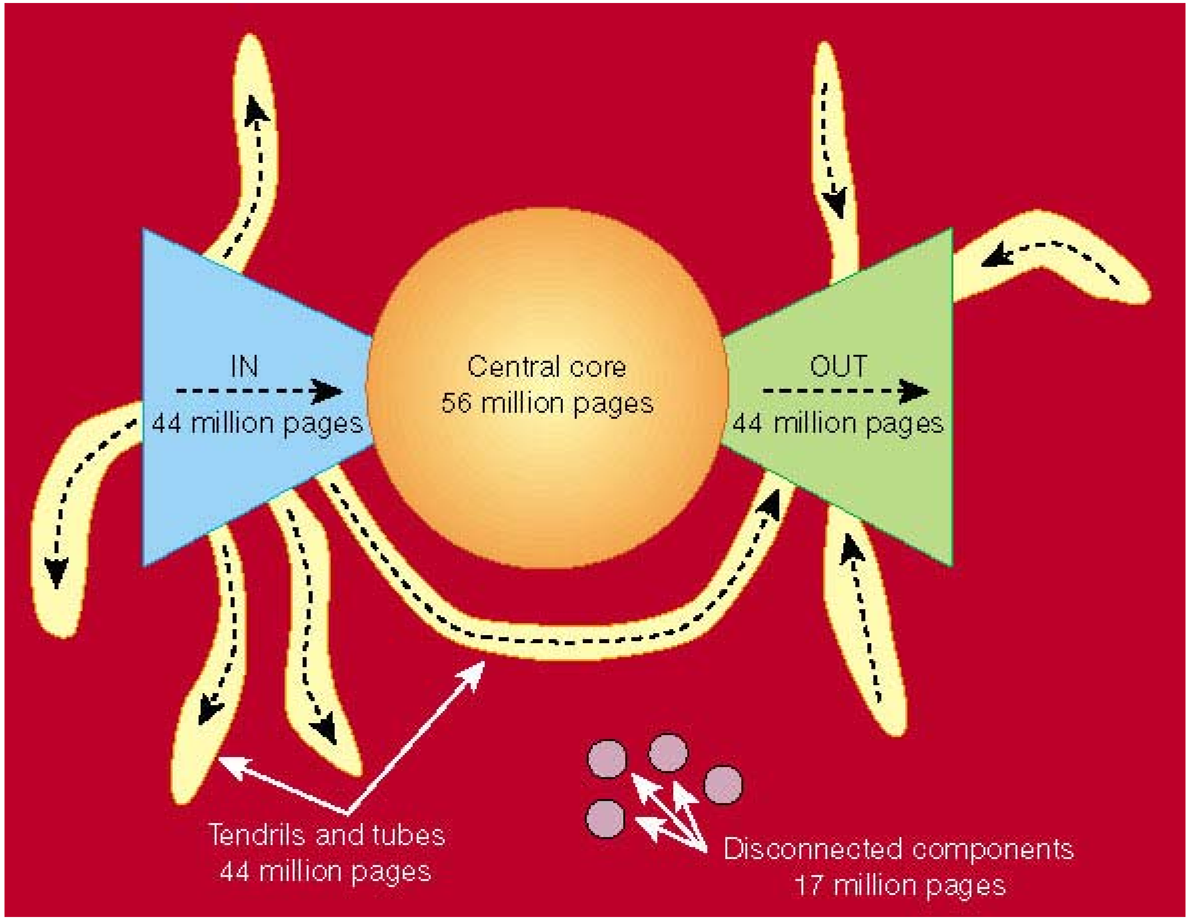
- no large component

-> people don't usually represent such situations by networks at all.

reason why we use network for analysis interruption
because it is highly likely there exsists large component

components in directed networks

- SCC, in-component, and out-component



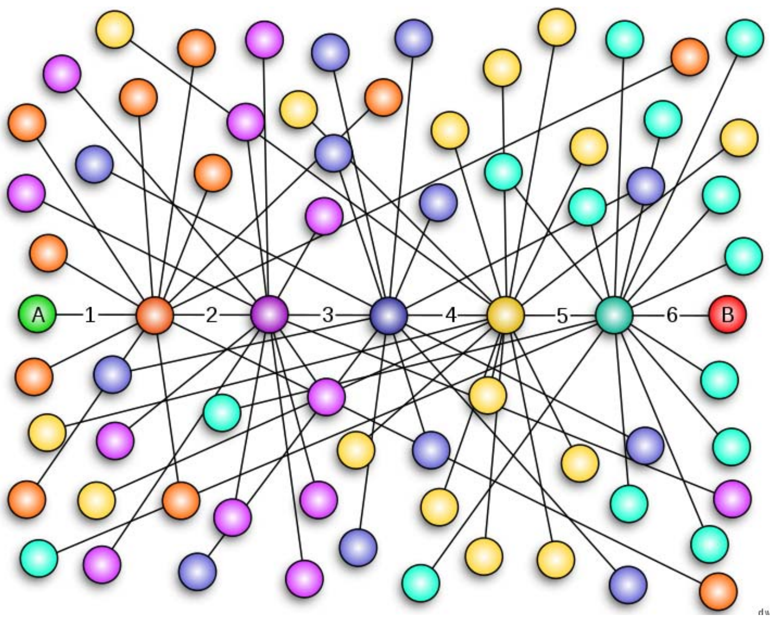
“The web is a bow tie”
http://www.nature.com/nature/journal/v405/n6783/full/405113a0.html

small-world effect

- Stanley Milgram's letter-passing experiment
  - people were asked to send a letter to a distant target person by passing it from acquaintance to acquaintance
  - # of hops between two arbitrary persons is around six on average
    - remarkably small (although the network have millions of vertices)

that's why the number of hub is quite small

- path length scale as log n with the number n of network vertices



On Facebook, it's now 4.74 degrees of separation

- http://edition.cnn.com/2011/11/22/tech/social-media/facebook-six-degrees/index.html

(CNN) -- In the Facebook age -- when digital "friends" are just a click away -- the distance between people seems to be shrinking, according to data the social network released on Monday night.

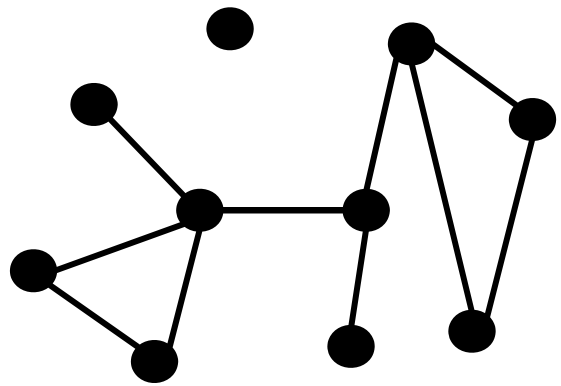
The adage maintains there are "six degrees of separation" between any two people on Earth, meaning that any two people would know each other through no more than six intermediary contacts.

On Facebook, however, the average user is only 4.74 degrees away from any other Facebooker.

degree distributions

- p_k : fraction of the vertices that have degree k

$$p_0 = \frac{1}{10}, p_1 = \frac{2}{10}, p_2 = \frac{4}{10}, p_3 = \frac{2}{10}, p_4 = \frac{1}{10}$$



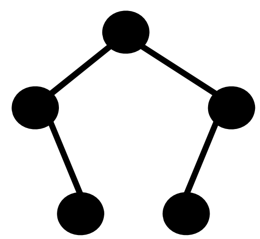
- degree sequence:$\{k_1, k_2, \ldots, k_n\}$

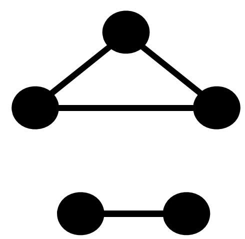
$\{0,1,1,2,2,2,2,3,3,4\}$

graph -> degree sequence: easy!

degree sequence -> graph: more than one possibilities!

- there is more than one network with the same degree distribution

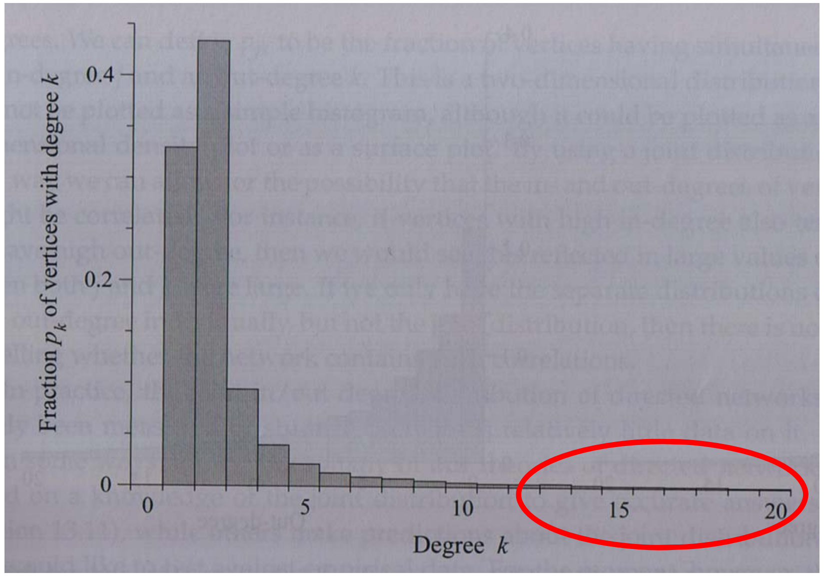




$\{1,1,2,2,2\}$ $\{1,1,2,2,2\}$

the degree distribution of the Internet

- x axis: degree (k)
- y axis: fraction (p_k) of vertices with degree k
- most of the vertices have low degree
- significant "tail" -- hubs
- the degree distribution is right-skewed

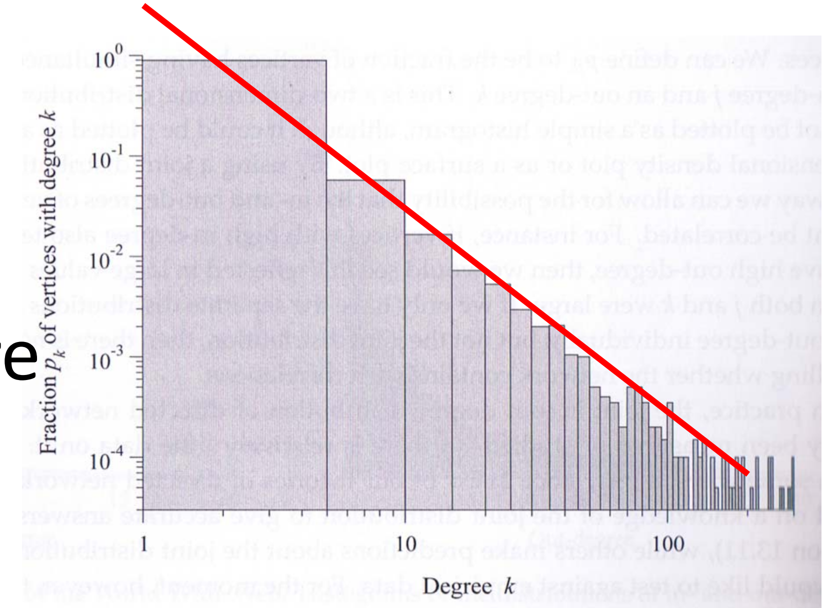


power laws and scale-free networks

- both axes are logarithmic

$$\ln p_k = -\alpha \ln k + c \qquad p_k = Ck^{-\alpha}$$

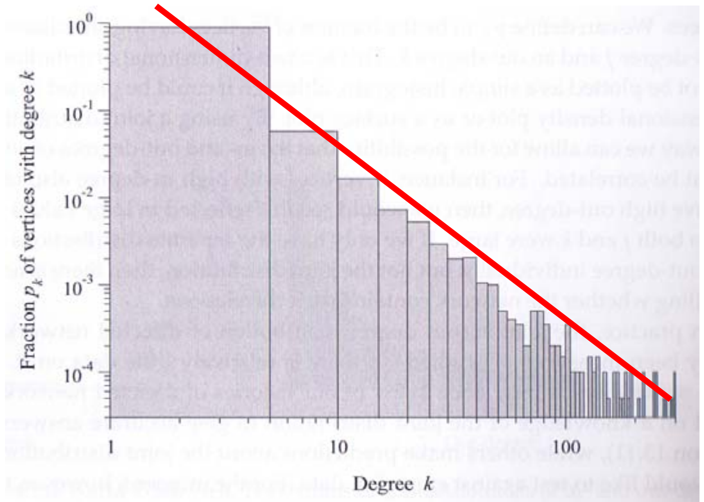
- distributions of this form are called as "power laws"
- values in the range $2 \leq \alpha \leq 3$ are typical
- in many cases, the power law is obeyed only in the tail of the distribution
- "scale-free networks"



detecting and visualizing power laws

- a simple histogram presents some problems
  - poor statistics in the tail of the distribution
  - noisy signal will make it difficult to detect power laws
- solutions
  - use a histogram with larger bins
  - using bins of different sizes
    - wide bins in the tail
    - narrow ones at the left-hand end

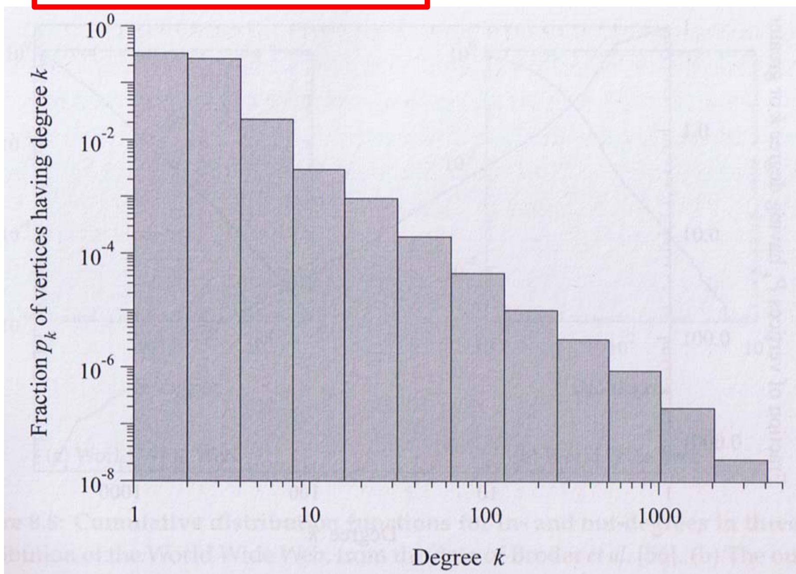
    are desirable



logarithmic binning

- each bin is made wider than its predecessor by a constant factor a (a=2 is common)
  - 1st bin: $1 \le k < 2$
  - nth bin: $a^{n-1} \le k < a^n$
  - width : $a^n - a^{n-1} = (a-1)a^{n-1}$
- the histogram is much less noisy
- the bins have equal width on a log-scale histogram

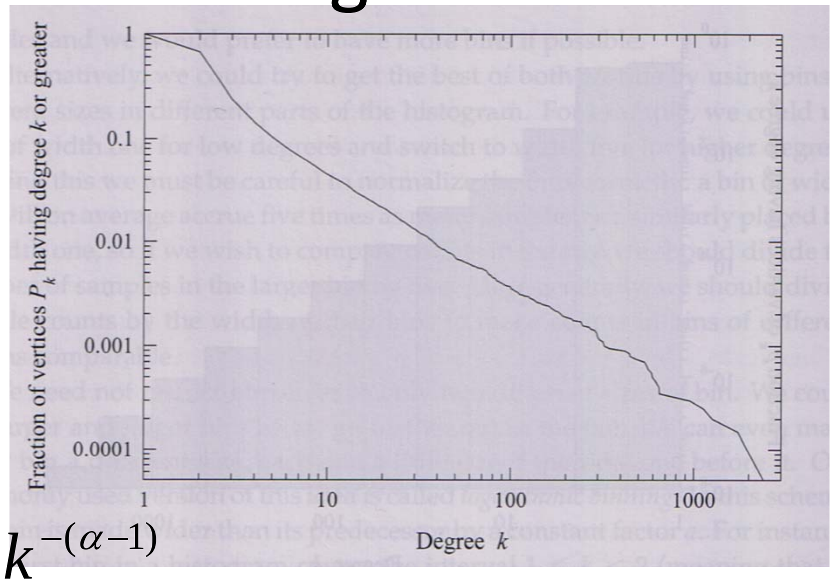
degree 1
degree 2 ~ 3
degree 4 ~ 7
...
degree 2^n ~ 2^{n+1} - 1



cumulative distribution function

- P_k: fraction of vertices that have degree k or greater $P_k = \sum_{k'=k}^{\infty} p_{k'}$
- p_k follows a power law
  - $p_k = Ck^{-\alpha}$ for $k \geq k_{\min}$
- for $k \geq k_{\min}$

$$P_k = C\sum_{k'=k}^{\infty} k'^{-\alpha} \cong C\int_k^{\infty} k'^{-\alpha}\, dk' = \frac{C}{\alpha - 1} k^{-(\alpha-1)}$$



- if the distribution p_k follows a power law, so does the cumulative distribution function P_k

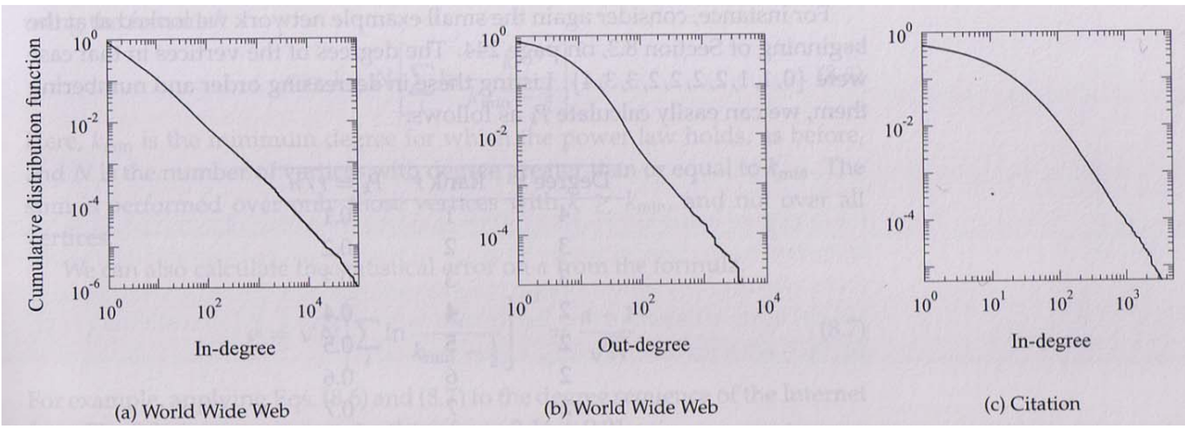
advantages of cumulative distribution functions

- P_k does not require binning
- easy to calculate: sort the degrees of vertices in descending order and number them from 1 to n in that order

x axis

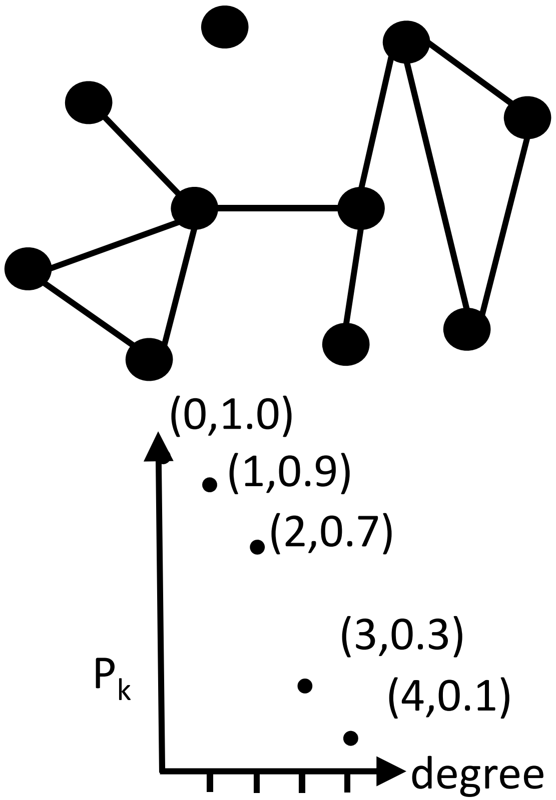
degree :	(highest)				k_i		(lowest)	
rank:	1	2	3	....	r_i	...	n	
P_k:	1/n	2/n	3/n		r_i/n		n/n	

y axis



(a) World Wide Web (b) World Wide Web (c) Citation

example



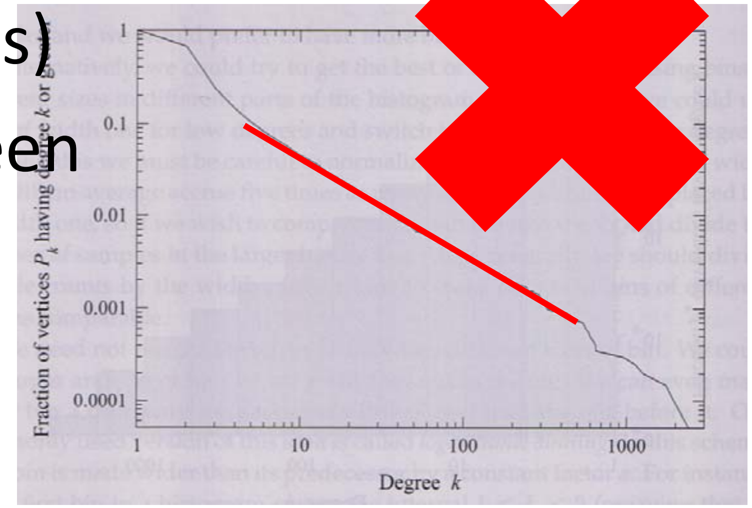
x axis

y axis

degree k	rank r	$P_k = r/n$
4	1	0.1
3	2	0.2
3	3	0.3
2	4	0.4
2	5	0.5
2	6	0.6
2	7	0.7
1	8	0.8
1	9	0.9
0	10	1.0

disadvantages of cumulative distribution functions

- less easy to interpret
- successive points on a plot are correlated
  - not valid to extract the exponent of a power law distribution by fitting the slope on the straight-line portion of a plot and equating the result with α-1
  - fitting (such as least squares) assume independence between the data points



calculating α directly from the data

- not good to evaluate exponent (α) from cumulative distribution functions or ordinary histograms
- calculating α directly

$$\alpha = 1 + N\left[\sum_i \ln \frac{k_i}{k_{\min} - \frac{1}{2}}\right]^{-1}$$

k_{min}: the minimum degree for which the power law holds

N: # of vertices with degree ≥ k_{min}

- statistical error on α

$$\sigma = \sqrt{N}\left[\sum_i \ln \frac{k_i}{k_{\min} - \frac{1}{2}}\right]^{-1} = \frac{\alpha - 1}{\sqrt{N}}$$

example : Internet (Fig. 8.3)

$$\alpha = 2.11 \pm 0.01$$

properties of power-law distributions

- power-laws appear in a wide varieties of places
  - the size of city populations, earthquakes, moon creators, solar flares, computer files, wars
  - the frequency of use of words in human languages
  - the frequency of occurrence of personal names
  - the number of papers scientists write
  - the number of hits on Web pages
  - the sales of books, music recordings, and almost every other branded commodity

normalization

- pure power-law distribution(k starts from 1)

$$\sum_{k=0}^{\infty} p_k = 1 \qquad p_k = Ck^{-\alpha}$$

$$C\sum_k k^{-\alpha} = 1 \qquad C = \frac{1}{\sum_{k=1}^{\infty} k^{-\alpha}} = \frac{1}{\zeta(\alpha)}$$

Riemann zeta function

$$p_k = \frac{k^{-\alpha}}{\zeta(\alpha)}$$ k>0, p_0=0

- deviation from power-law for small k

$$p_k = \frac{k^{-\alpha}}{\sum_{k=k_{\min}}^{\infty} k^{-\alpha}} = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})}$$

generalized zeta function

- or the tail is approximated by an integral

$$C \cong \frac{1}{\int_{k_{\min}}^{\infty} k^{-\alpha} dk} = (\alpha - 1)k_{\min}^{\alpha-1} \qquad p_k \cong \frac{\alpha - 1}{k_{\min}}\left(\frac{k}{k_{\min}}\right)^{-\alpha}$$

moments (1)

- moments of degree distribution

average

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k \qquad \langle k^2 \rangle = \sum_{k=0}^{\infty} k^2 p_k \qquad \langle k^m \rangle = \sum_{k=0}^{\infty} k^m p_k$$

- if p_k has a power-law tail for $k \geq k_{min}$:

$$\langle k^m \rangle = \sum_{k=0}^{k_{\min}-1} k^m p_k + C \sum_{k=k_{\min}}^{\infty} k^{m-\alpha}$$

because $p_k = Ck^{-\alpha}$

$$\langle k^m \rangle \cong \sum_{k=0}^{k_{\min}-1} k^m p_k + C \int_{k_{\min}}^{\infty} k^{m-\alpha} dk$$

approximate by an integral

$$= \sum_{k=0}^{k_{\min}-1} k^m p_k + \frac{C}{m-\alpha+1} \left[k^{m-\alpha+1} \right]_{k_{\min}}^{\infty}$$

depends on m and α

well-defined if and only if α > m+1

moments (2)

- second moment $\langle k^2 \rangle$ arises in many calculations
  - mean degree of neighbors
  - robustness calculations
  - epidemiological processes
- for large network, it is finite if and only if α > 3
  - but 2 ≤ α ≤ 3 for most real-world networks
- for any finite networks, it is finite

$$\langle k^m \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i^m$$

if we were to calculate for an arbitrarily large network with the same power-law degree distribution,

top-heavy distribution

- the fraction W of end of edges attached to a fraction P of the highest-degree vertices in a network

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

- Lorenz curves
- example: WWW

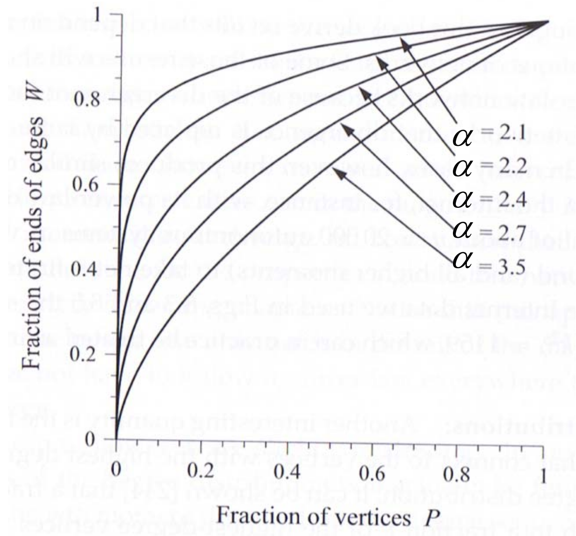
  α = 2.2

  P=0.5 ⇨ W=0.89

  – 89% of all hyperlinks link to pages in the top half of the degree distribution

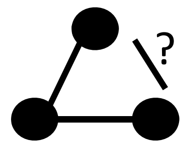
  W=0.5⇨ P=0.015

  – 50% of links go to 1.5% richest vertices



clustering coefficient (C)

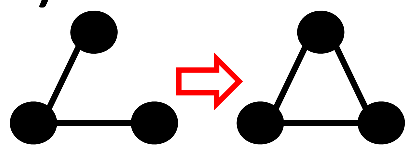
- average probability that two neighbors of a vertex are themselves neighbors



- density of triangles in a network

$$C = \frac{1}{n}\frac{\left[\langle k^2\rangle - \langle k\rangle\right]^2}{\langle k\rangle^3}$$

- random network: C is small
- social network : C is large (10% - 60%)
  – because of “triadic closure”



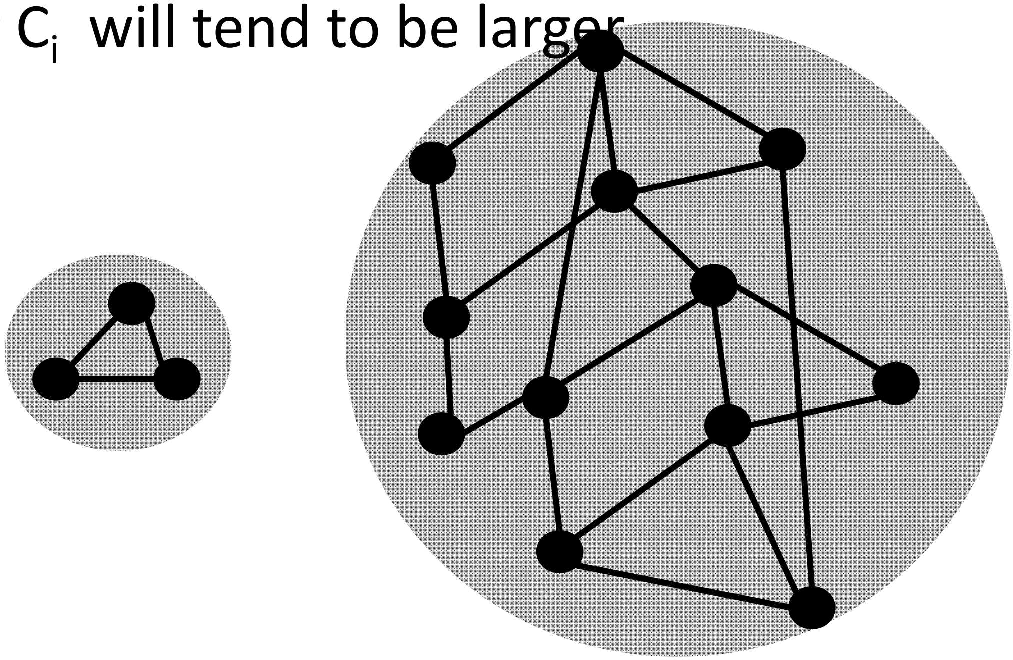
- Internet: observed C is smaller than expected
  – C =0.012, but expected value=0.84

local clustering coefficient

- $C_i = \frac{\text{(\# of pairs of neighbors of i that are connected)}}{\text{(\# of paths of neighbors of i)}}$
- C_i decrease with k $\quad C_i \approx k^{-0.75}$

vertices of higher degree tend to have lower local clustering coefficient

  - because of community structure
    - vertices in a small community are constrained to have low degree, and their C_i will tend to be larger



assortative mix, homophily

- high-degree vertices tend to connect high-degree ones
- correlation coefficient $r = \dfrac{\sum_{ij}(A_{ij} - k_i k_j / 2m) k_i k_j}{\sum_{ij}(k_i \delta_{ij} - k_i k_j / 2m) k_i k_j}$
- faster computation of r

$$r = \frac{S_1 S_e - S_2^2}{S_1 S_3 - S_2^2} \qquad S_e = \sum_{ij} A_{ij} k_i k_j = 2 \sum_{edges(i,j)} k_i k_j$$

$$S_1 = \sum_i k_i, S_2 = \sum_i k_i^2, S_3 = \sum_i k_i^3$$

2m

- social networks: positive r
- other networks: negative r

degree distribution with R+igraph

GML file is available at Mark Newman's Website (http://www-personal.umich.edu/~mejn/netdata/).

```
> ig<-read.graph("as-22july06.gml",format="gml")
> summary(ig)
```

summary of the network

```
Vertices: 22963
Edges: 48436
Directed: FALSE
No graph attributes.
Vertex attributes: id, label.
No edge attributes.
> vcount(ig)
```

the number of vertices

```
[1] 22963
> ecount(ig)
```

the number of edges

```
[1] 48436
> no.clusters(ig)
```

the number of clusters

```
[1] 1
> average.path.length(ig)
```

average path length

```
[1] 3.842426
> transitivity(g)
```

clustering coefficient

```
[1] 0
> mean(degree(ig))
```

average degree

```
[1] 4.218613
> max(degree(ig))
```

the maximum degree

```
[1] 2390
> min(degree(ig))
```

the minimum degree

```
[1] 1
> tkplot(ig)
```

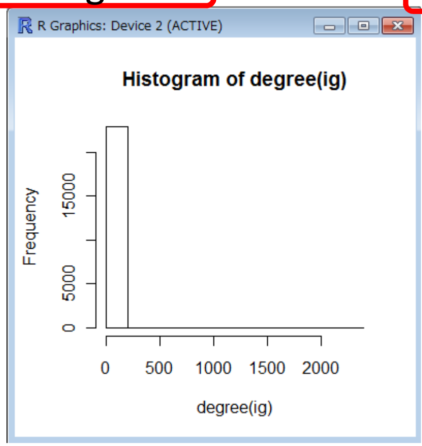
too large to visualize

```
…
> power.law.fit(degree(ig))
```

fits a power-law distribution

```
Call:
mle(minuslogl = mlogl, start = list(alpha = start))
Coefficients:
  alpha
1.874345
> hist(degree(ig))
> plot(degree.distribution(ig),log="xy")
```

histogram



power-law distribution