Machine Learning with Graphs (MLG)

Network Properties

Unfold Common Properties across Real Graphs

Cheng-Te Li (李政德)

Institute of Data Science

National Cheng Kung University

chengte@mail.ncku.edu.tw

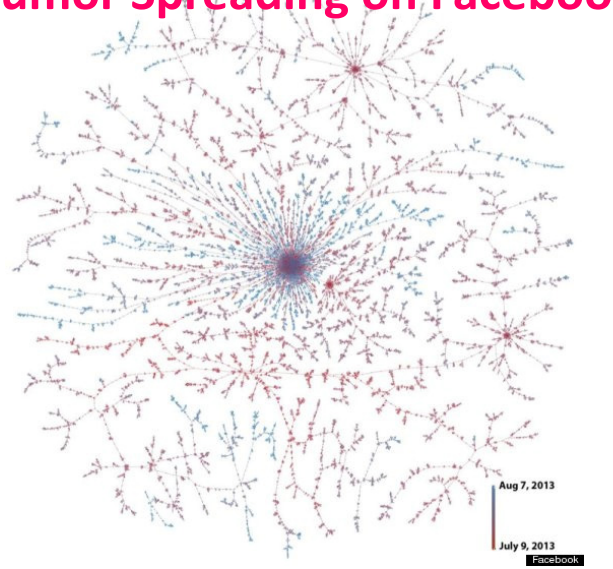
3 Essential Network Properties

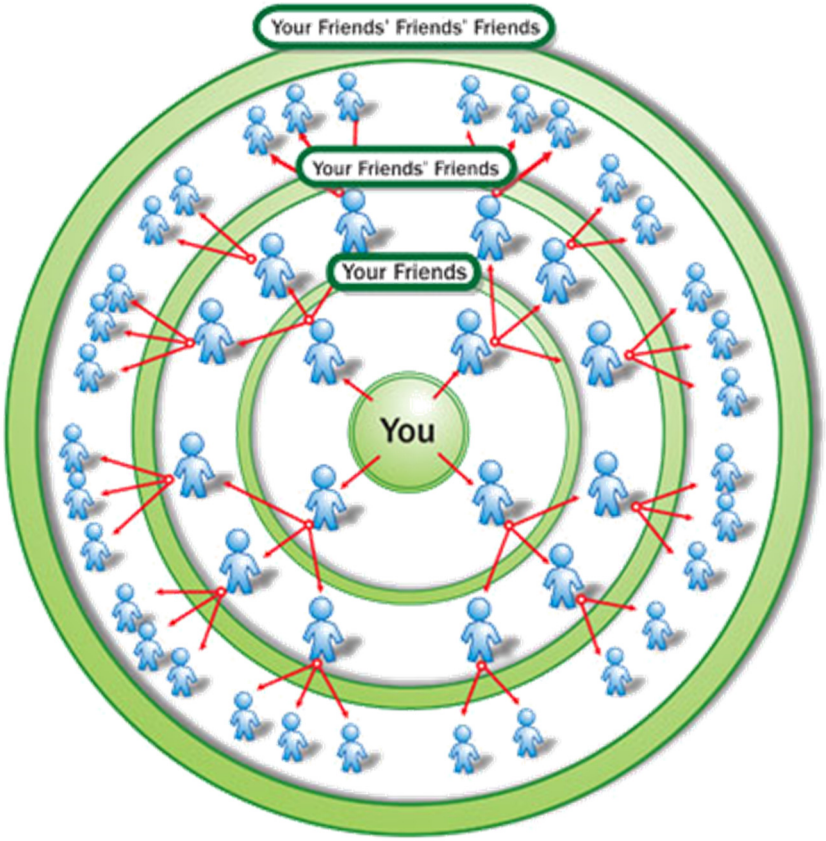
- **Short** Average Path Length
  - Small-world Effect (小世界現象)
- **High** Clustering Coefficient
  - Friends of Friends are Friends (群聚現象)
- **Power-law** Degree Distribution
  - Long-tail Effect (長尾效應)

How Small is the World?

- A rumor is spreading over a social network
  - Assume all users pass it immediately to all of their friends
- Questions
  1) How long does it take to reach almost all of the nodes in the network?
  2) What is the maximum time?
  3) What is the average time?

Rumor Spreading on Facebook



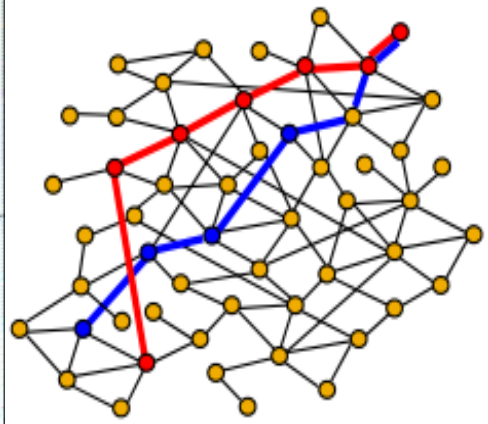


Milgram's Small-world Experiment

- Letter Delivery (1967)
  - 296 Random people in Omaha
  - Ask each to send a letter to an unknown stockbroker in Boston
  - Letters can only be passed to acquaintances
  - Choose one believed to successfully find the stockbroker
  - Tell the receiver the rule

Stanley Milgram
(1933-1984)

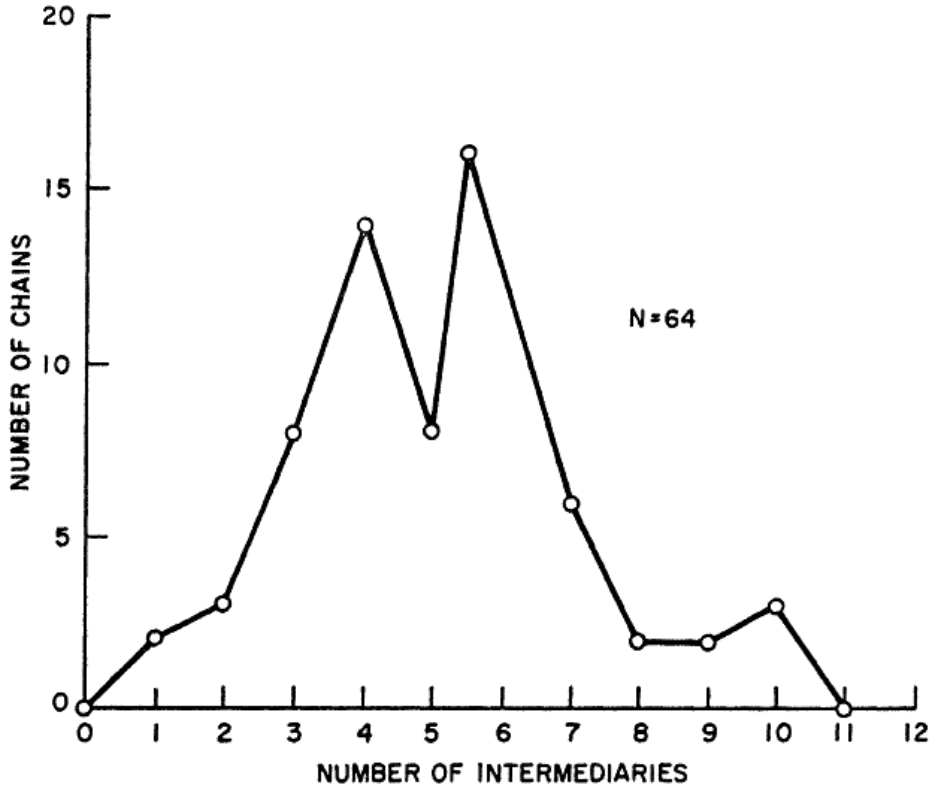
股票經紀人

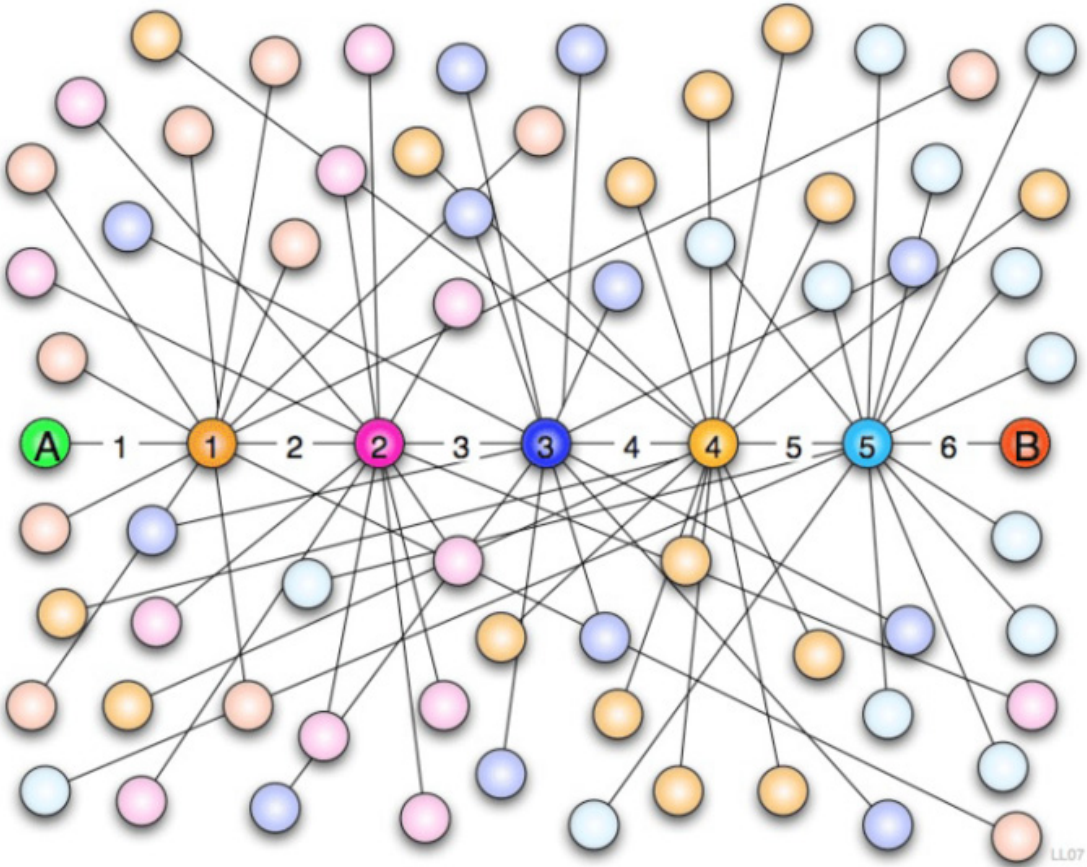


Six Degree of Separation (六度分離)

- 64 out of 296 letters reached the goal
- The average steps is 6 to reach the destination
- Small World
  - Short connections between people in the world

Travers, Jeffrey, and Stanley Milgram. "An experimental study of the small world problem." *Sociometry* (1969): 425-443.



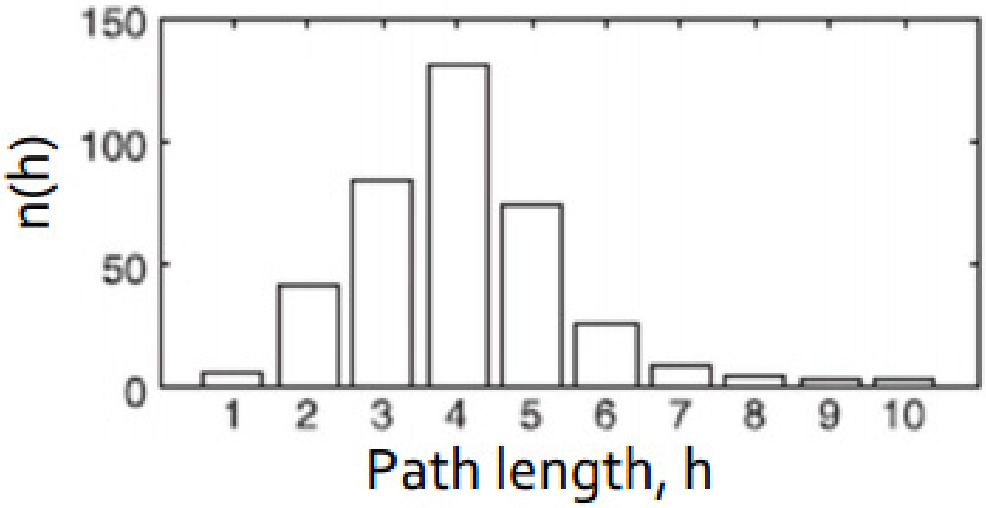


Small World on Emails

Outlook

- 18 targets
- 13 countries
- 60,000 participants
- 384 success (~1.5%) with

average path length = **4.01**



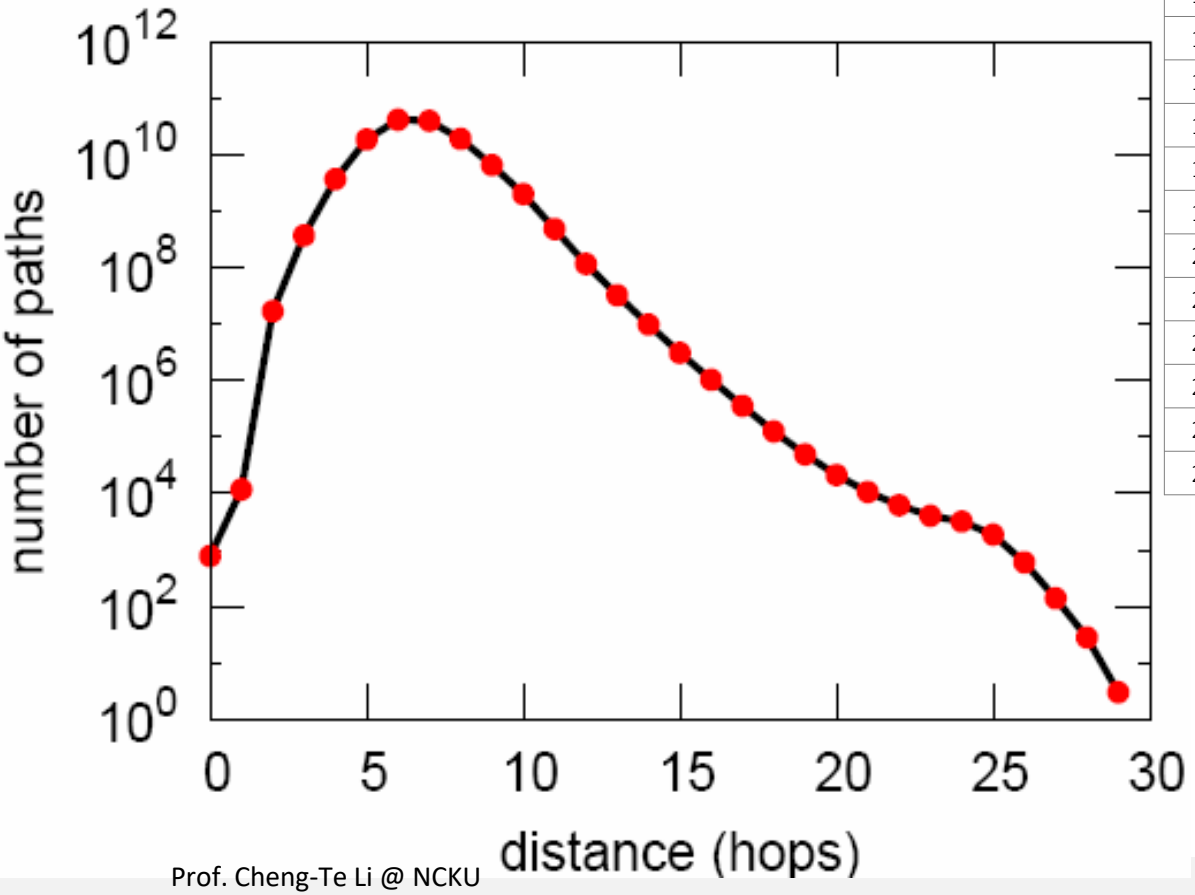
Small World on MSN

- 180 million users
- 1.3 billion messages sending within one month
- Average path length = **6.6**
  - 90% of nodes is reachable < 8 steps

Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

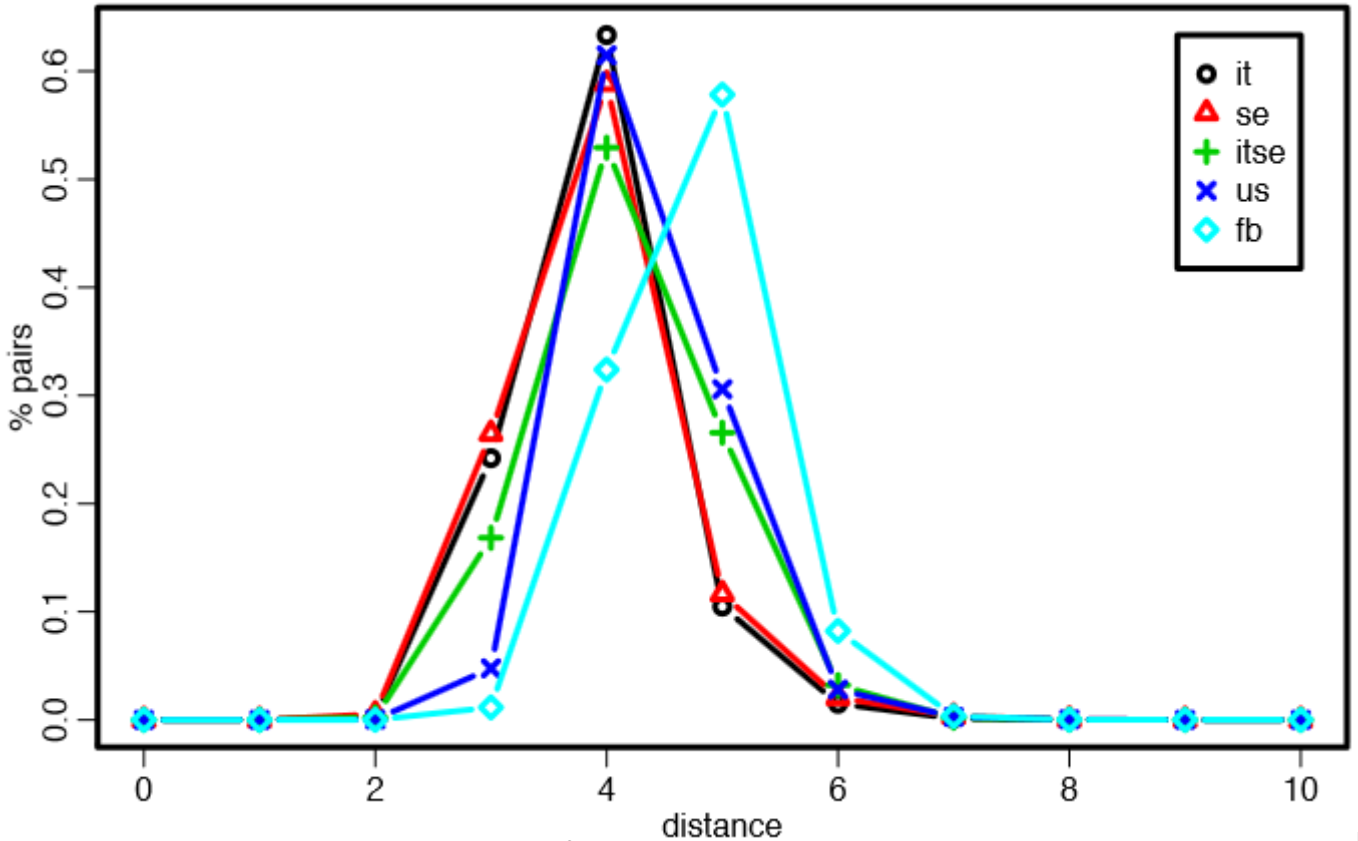
Implications

- Information (viruses) spread quickly
- Shortest paths exists



Small World on Facebook

- 720 million active users
- 69 billion friendship
- Average path length = **4.74**
  - 92%: 4-degree separation, 99%: five-degree separation



Measuring Small World Effects

- Let d_{ij} be the shortest path length between node i and j
- Diameter: $D = max_{i,j}\{d_{ij}\}$
- Average path length: $l = \frac{1}{n(n-1)/2}\sum_{i>j} d_{ij}$

Facebook

May 2011:

✓ Average path length was **4.7**

✓ **4.3** for US users

Four degrees of separation!

Web	Facebook	Flickr	LiveJournal	Orkut	YouTube
16.12	4.7	5.67	5.88	4.25	5.10

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

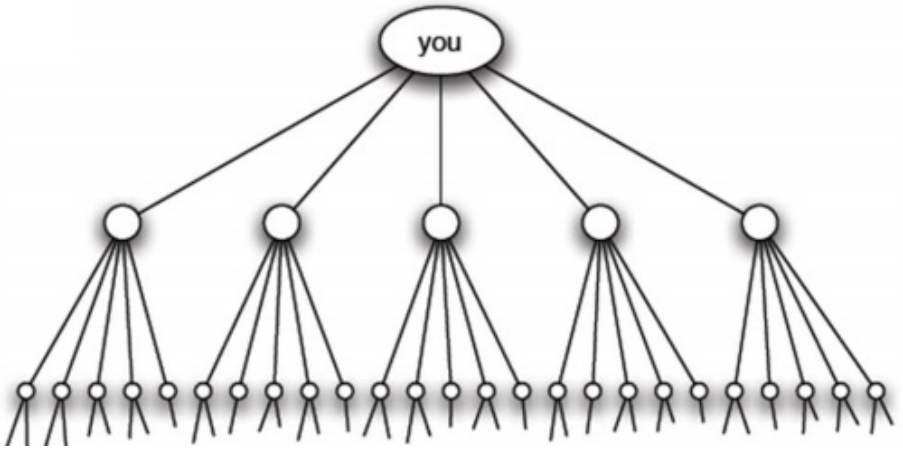
TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law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

Small-world Effect

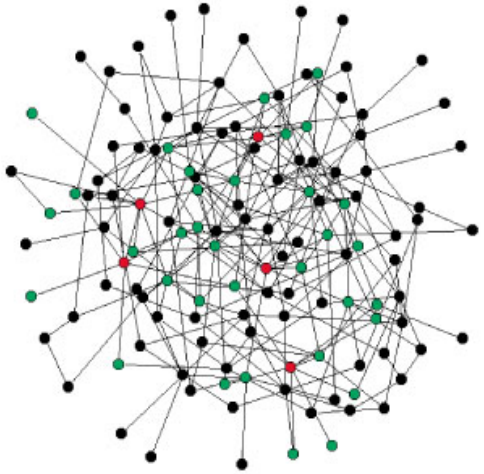
- Social Networks are highly likely to be compact, almost independent of network size
  - Consider only the giant connected part of the graph
  - **The average length between all pairs of nodes** in the network is small (around **6**)
- Open Question:
  with the increase of graph size, the average path length increases, still around 6, or decreases?

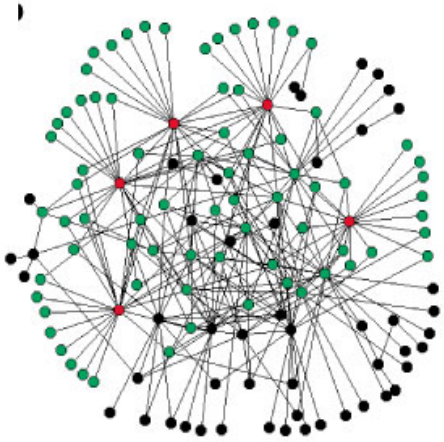
Is it truly surprising to have small world?

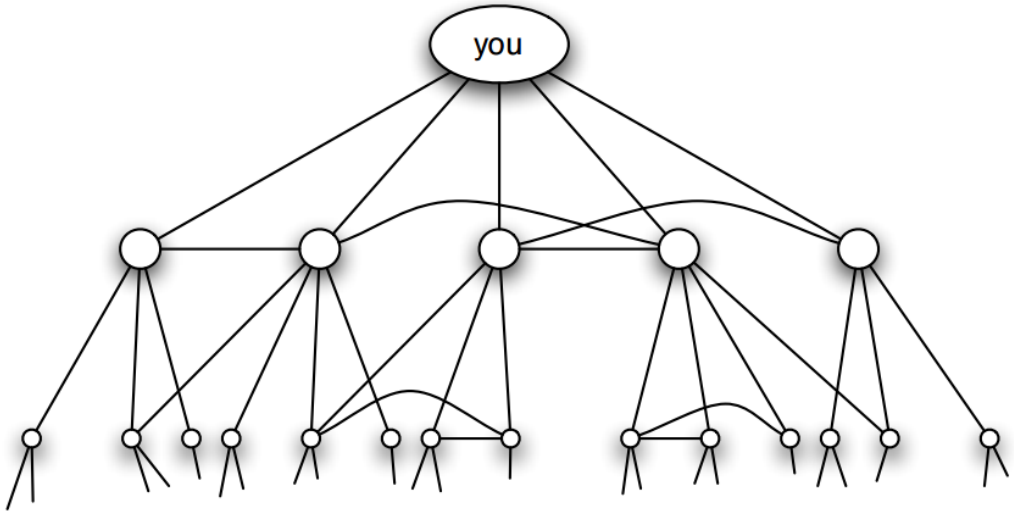
- Consider each person has 100 friends
  - 1st level: 100
  - 2nd level: 100*100 = 10,000
  - 3rd level: 100*100*100 = 100萬
  - 4th level: 100*100*100*100 = 1億
  - 5th level = **100億**



- In addition to small world, any other things we should notice in network?







3 Essential Network Properties

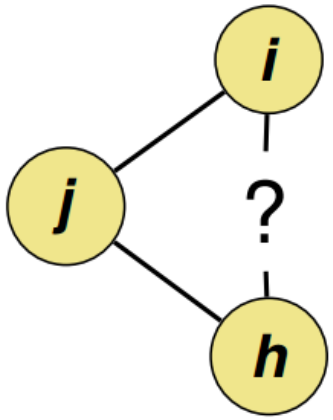
- **Short** Average Path Length
  - Small-world Effect (小世界現象)
- **High** Clustering Coefficient
  - Friends of Friends are Friends (群聚現象)
- **Power-law** Degree Distribution
  - Long-tail Effect (長尾效應)

Transitivity

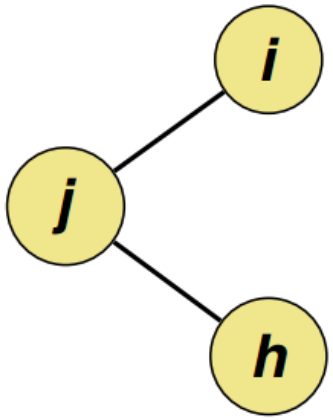
- Transitivity of a relation means that when there is a tie from node i to j, and also from j to h, then there is also a tie from i to h:

friends of my friends are my friends

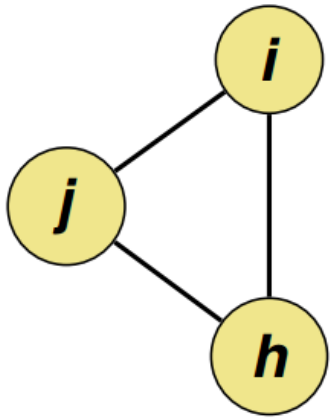
- Transitivity depends on triads, subgraphs formed by 3 nodes



Potentially transitive



Intransitive



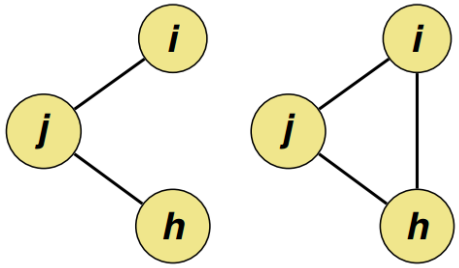
Transitive

Global Clustering Coefficient

- **Clustering coefficient** (CC) measures the transitivity in undirected graphs

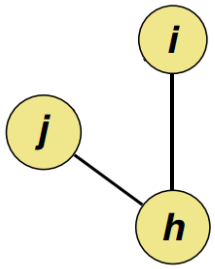
$$C = \frac{(\text{Number of Triangles}) \times 3}{\text{Number of Connected Triples of Nodes}}$$

- **Triple**: an **ordered set** of three nodes,
  - connected by two (open triple) edges or
  - three edges (closed triple)



Triple Triangle

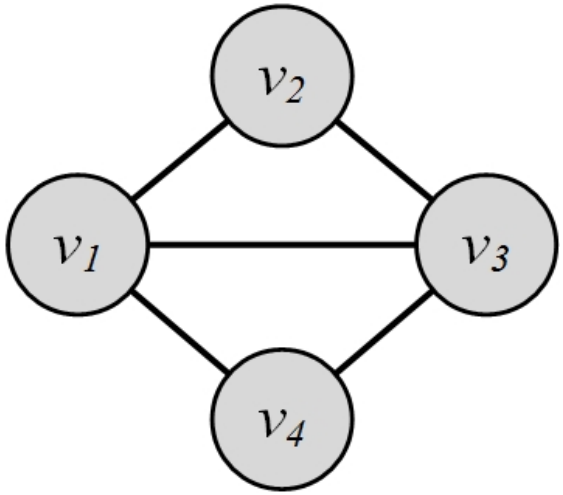
- A triangle can miss any of its three edges
  - A triangle has **3 Triples**



$v_i v_j v_h$ and $v_j v_h v_i$ are different triples

- The same members
- First missing edge $e(v_k, v_i)$ and second missing $e(v_i, v_j)$

Global Clustering Coefficient: Example



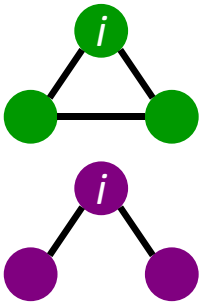
$$CC = \frac{(\text{Number of Triangles}) \times 3}{\text{Number of Connected Triples of Nodes}}$$

$$= \frac{2 \times 3}{2 \times 3 + \underbrace{2}_{v_2v_1v_4, v_2v_3v_4}} = 0.75.$$

Local Clustering Coefficient

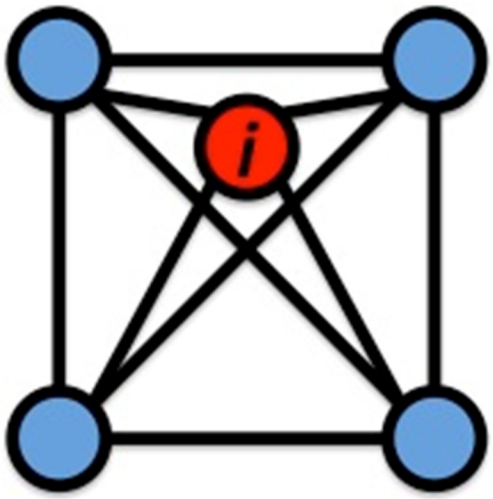
- Local clustering coefficient measures transitivity at the node level in undirected graphs
  - Imply how strongly neighbors of a node v (nodes adjacent to v) are themselves connected
  - The tendency that a friend of your friend being your friend

$$C(v_i) = \frac{\text{Number of Pairs of Neighbors of } v_i \text{ That Are Connected}}{\text{Number of Pairs of Neighbors of } v_i}.$$

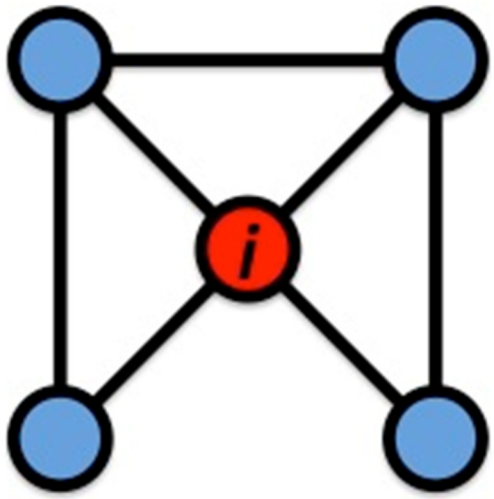


In an undirected graph, the denominator can be rewritten as: $\binom{d_i}{2} = d_i(d_i - 1)/2$

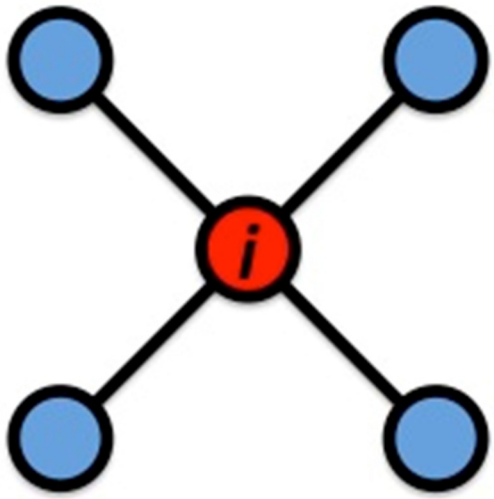
Local Clustering Coefficient: Example



$C_i = 1$



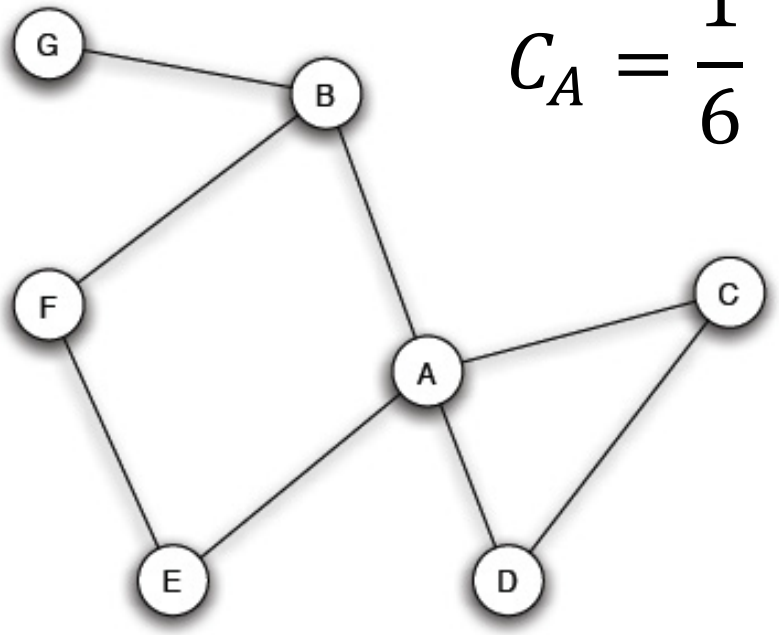
$C_i = 1/2$



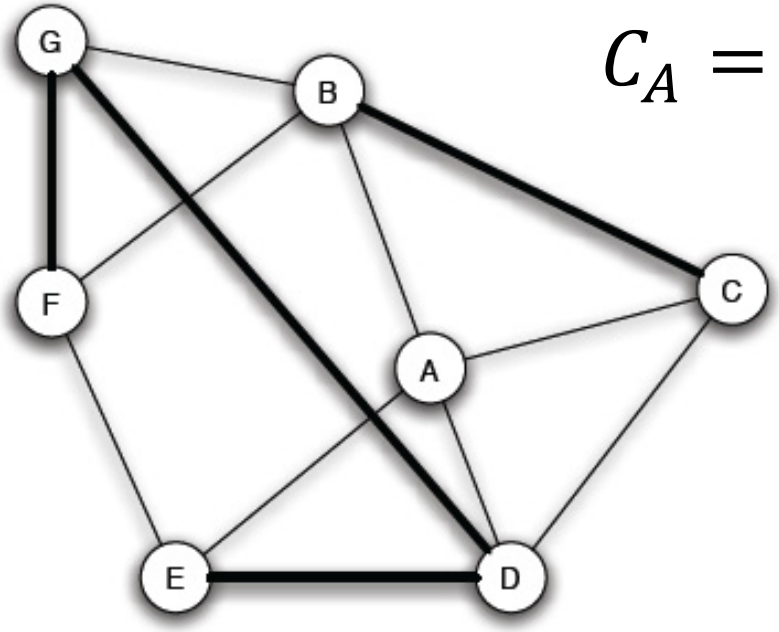
$C_i = 0$

Local Clustering Coefficient

- The clustering coefficient of a node A is also described as the probability that two randomly selected friends of *A* are friends with each other



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



$$C_A = \frac{1}{2}$$

Collective Statistics (M. Newman 2003)

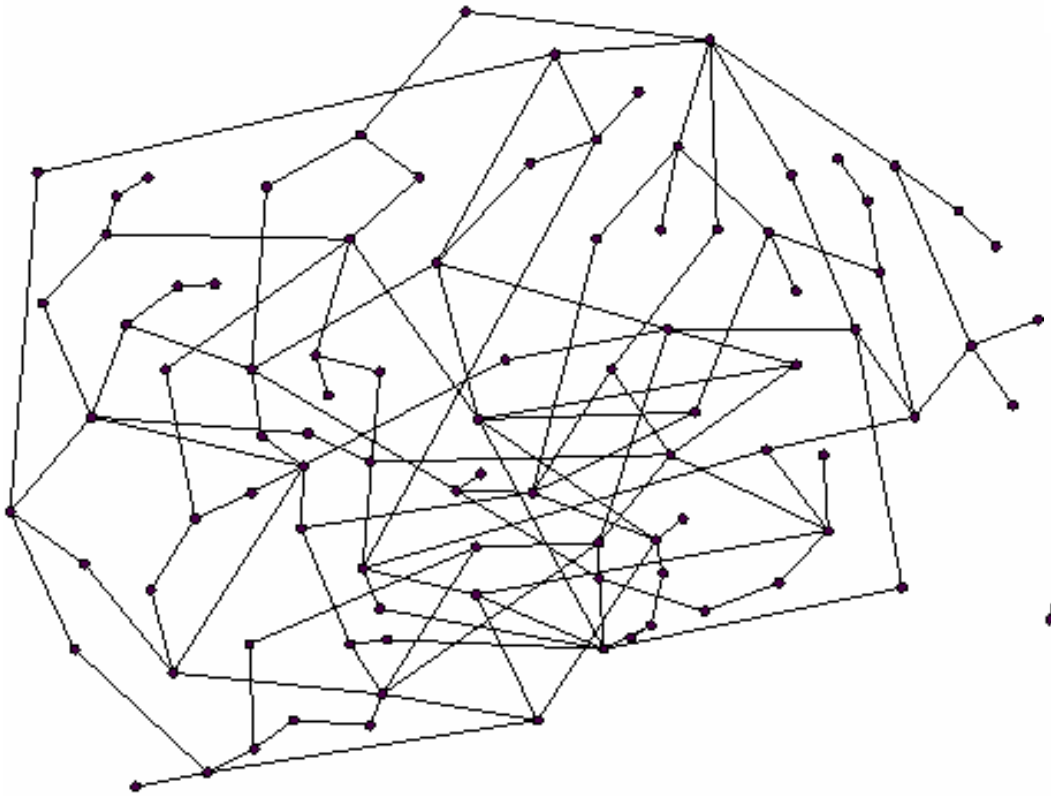
	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107,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

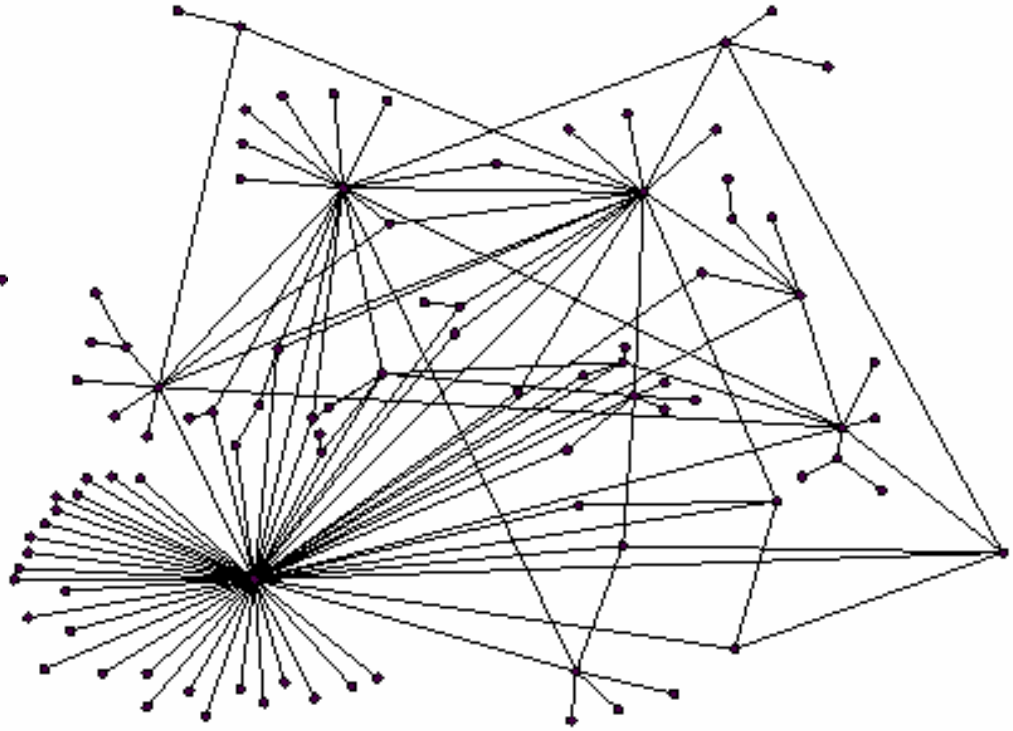
TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law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

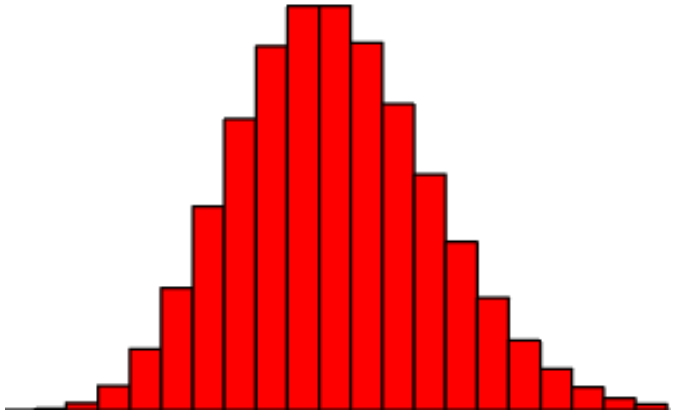
3 Essential Network Properties

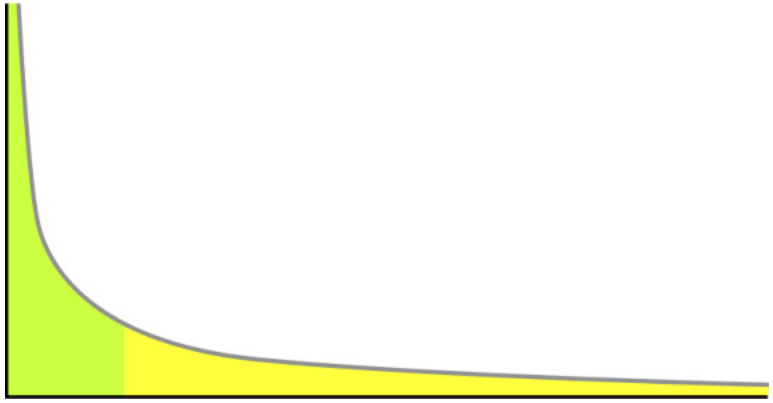
- **Short** Average Path Length
  - Small-world Effect (小世界現象)
- **High** Clustering Coefficient
  - Friends of Friends are Friends (群聚現象)
- **Power-law** Degree Distribution
  - Long-tail Effect (長尾效應)

Which is the Real-world Case?



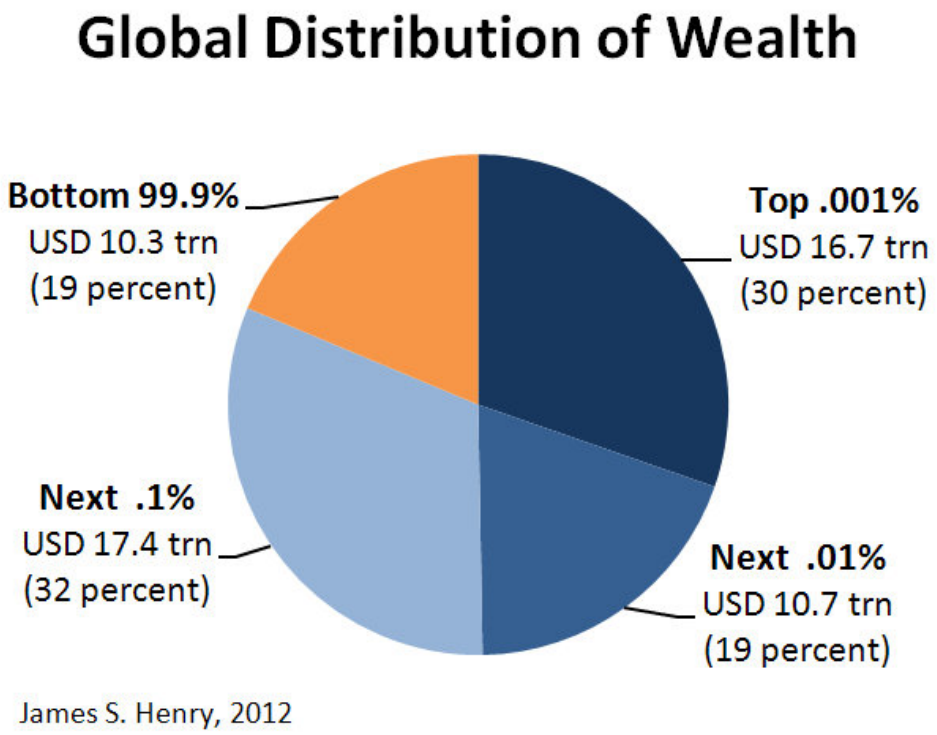






Wealth Distribution

- Most individuals have average capitals
- Few are considered wealthy
- Exponentially more individuals with average capital than the wealthier ones



James S. Henry, 2012

Herbert A Simon,
On a Class of Skew Distribution Functions, 1955

The **Pareto principle**
(80–20 rule): 80% of the effects come from 20% of the causes

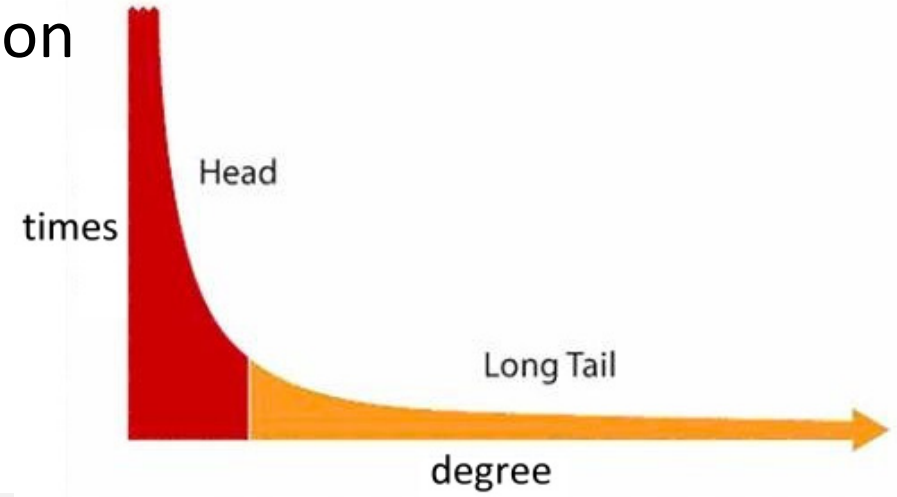
Power-law Degree Distribution

- When the frequency of an event changes as a power of an attribute → the frequency follows a **power-law**

$$P_k = ck^{-\alpha}$$

k : node degree
P_k : fraction of nodes with degree k
α : the power-law exponent and its value is typically in the range of [2, 3]
c : constant

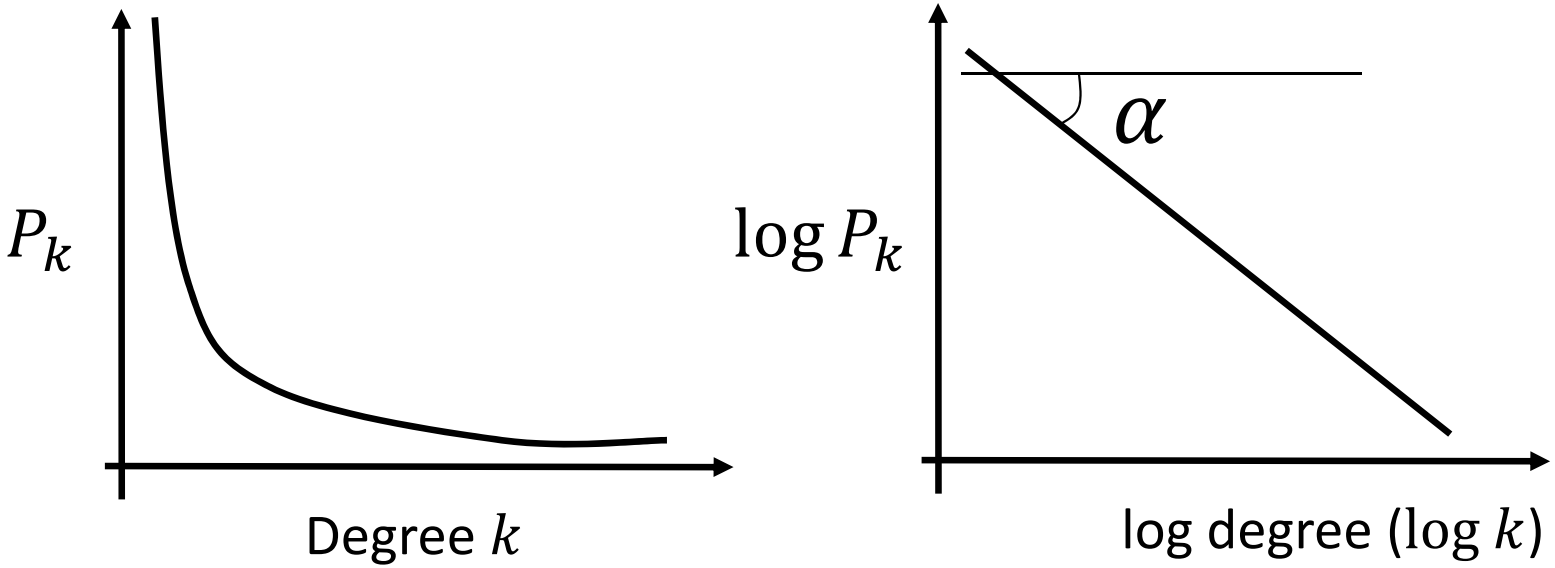
- A power-law (heavy-tailed) distribution
  - **Small occurrences**: common
  - **Large instances**: extremely rare
  - A non-negligible fraction of nodes that has very high degree



Power-law Degree Distribution

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

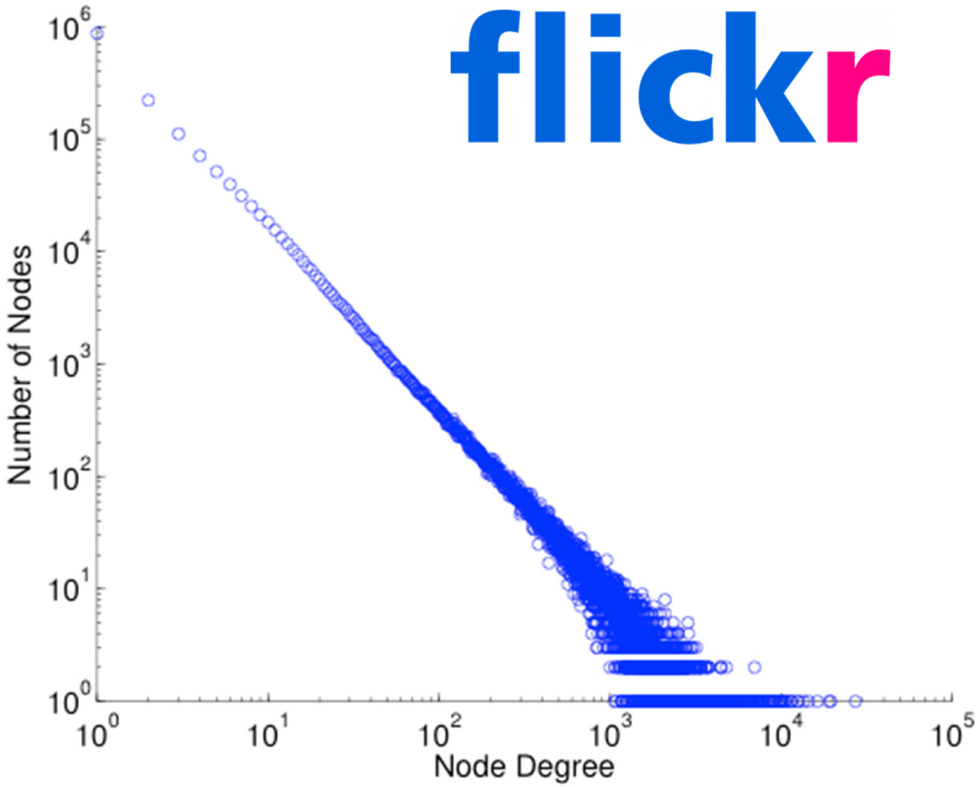
$$P_k = ck^{-\alpha} \rightarrow \log P_k = -\alpha \log k + \log C$$



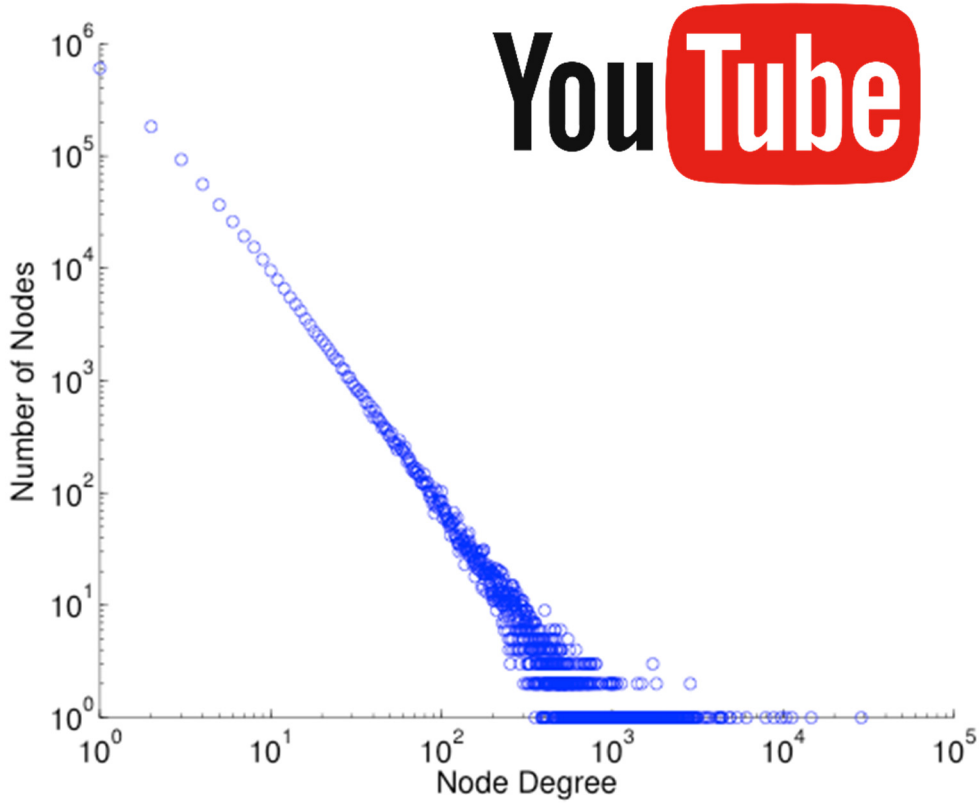
- α : power-law exponent (typically $2 \leq \alpha \leq 3$)

Power-law Degree Distribution

- Many real-world networks exhibit a power-law degree distribution



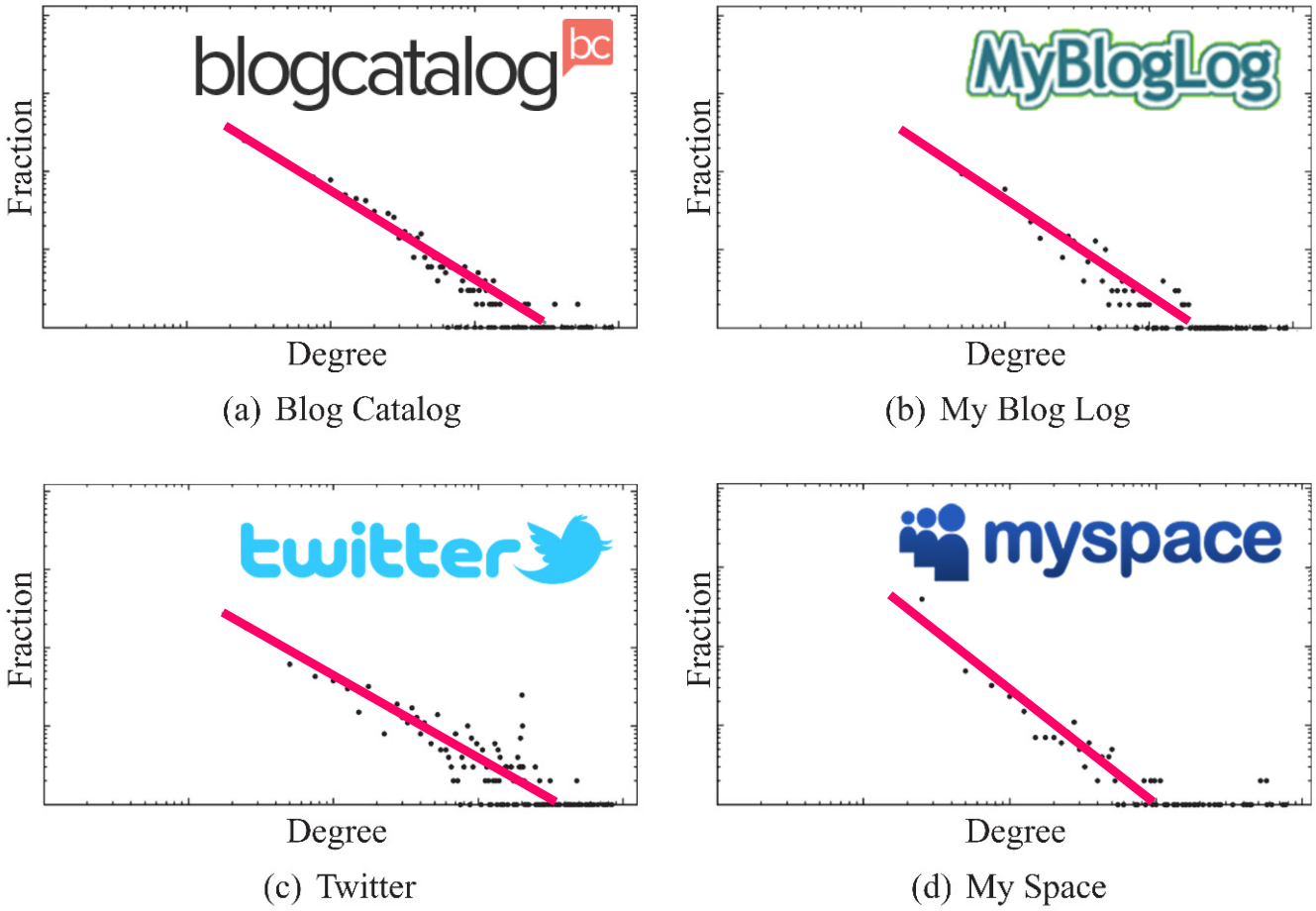
Friendship Network in Flickr



Friendship Network in YouTube

Power-law Degree Distribution

- Networks with a power-law degree distribution are called Scale-Free networks



Normal vs. Power-Law

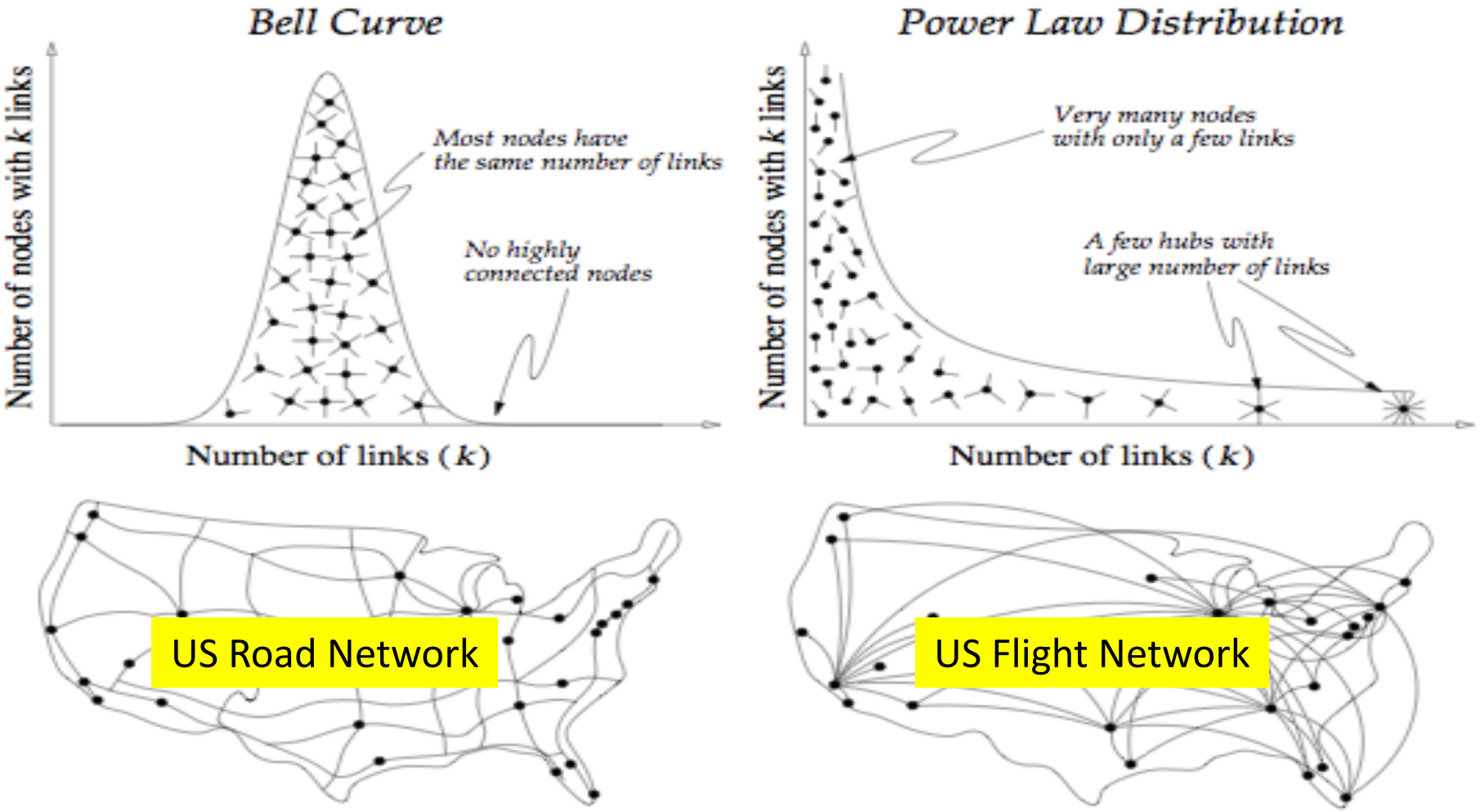


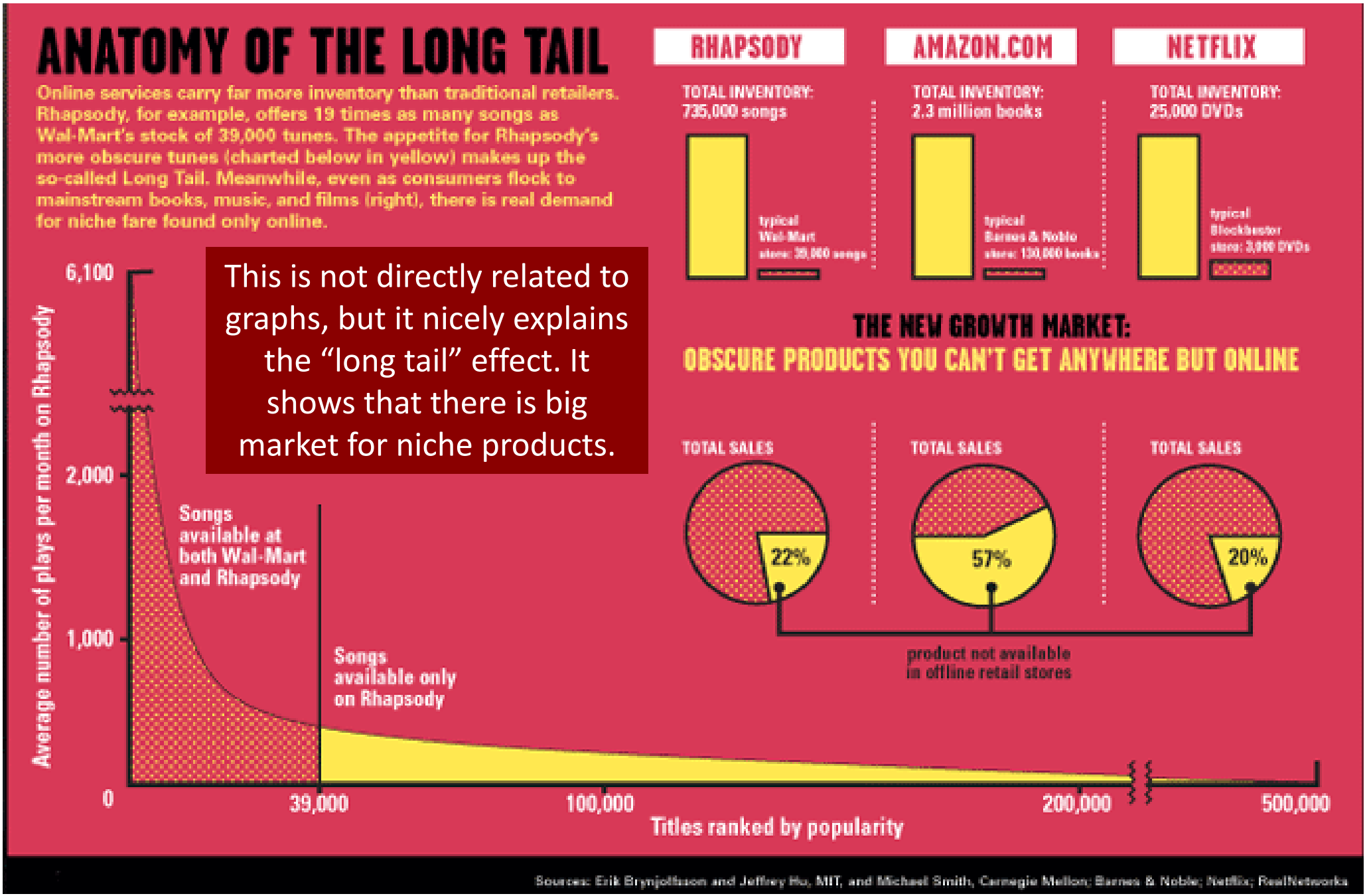
FIGURE (6.1)

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

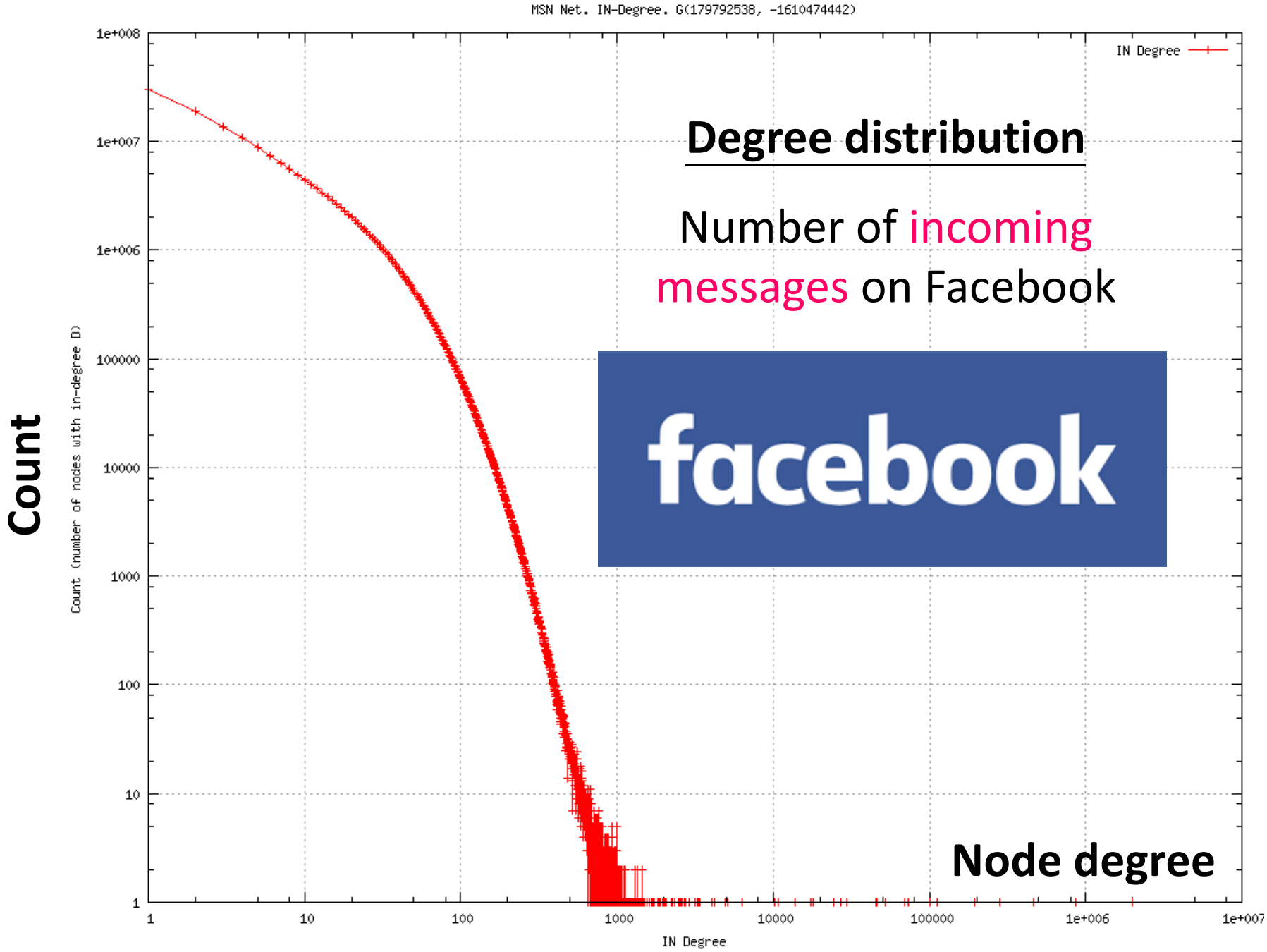
TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law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

Detour: How long is the long tail?

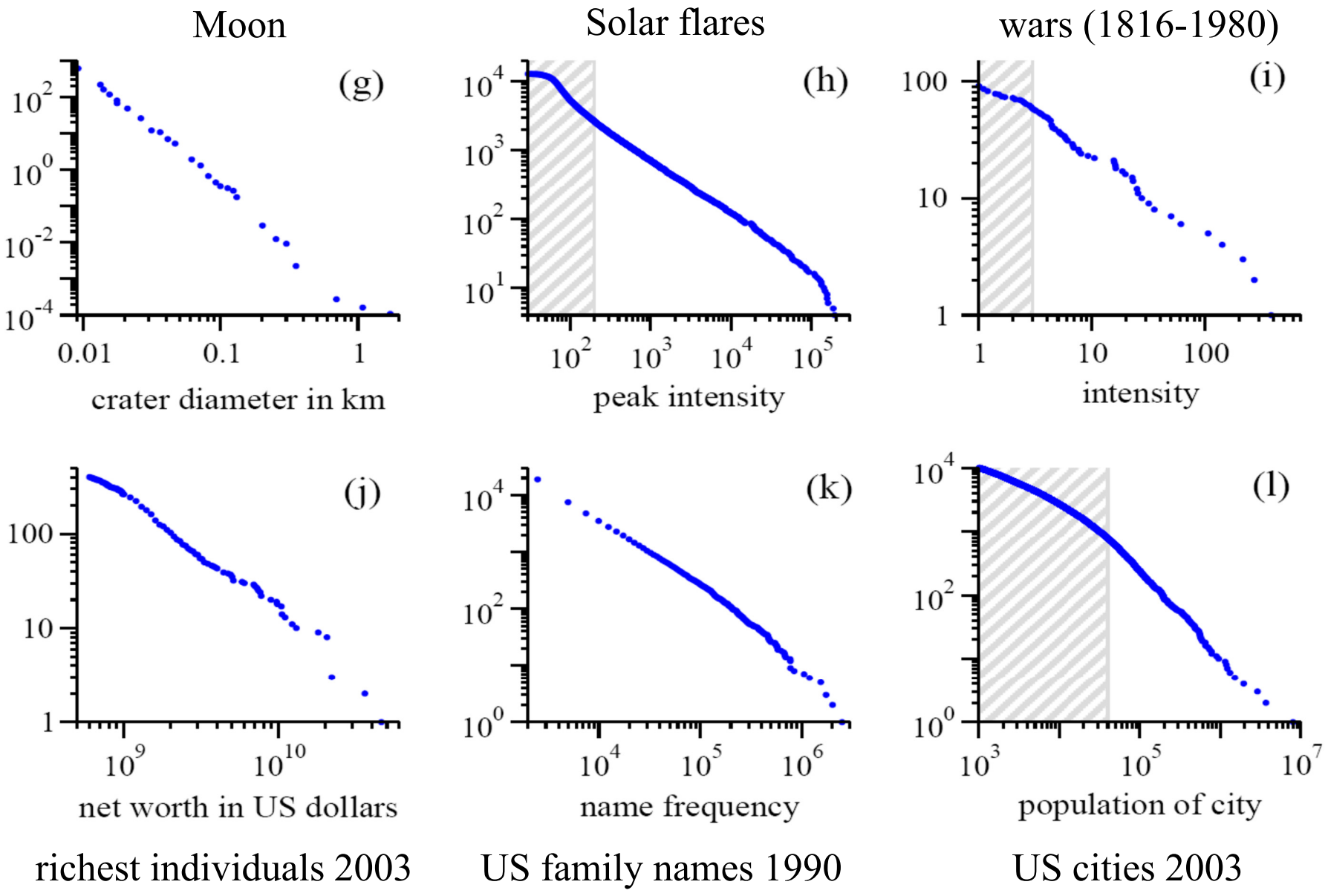


This is not directly related to graphs, but it nicely explains the "long tail" effect. It shows that there is big market for niche products.

Degree Distribution (cont.)

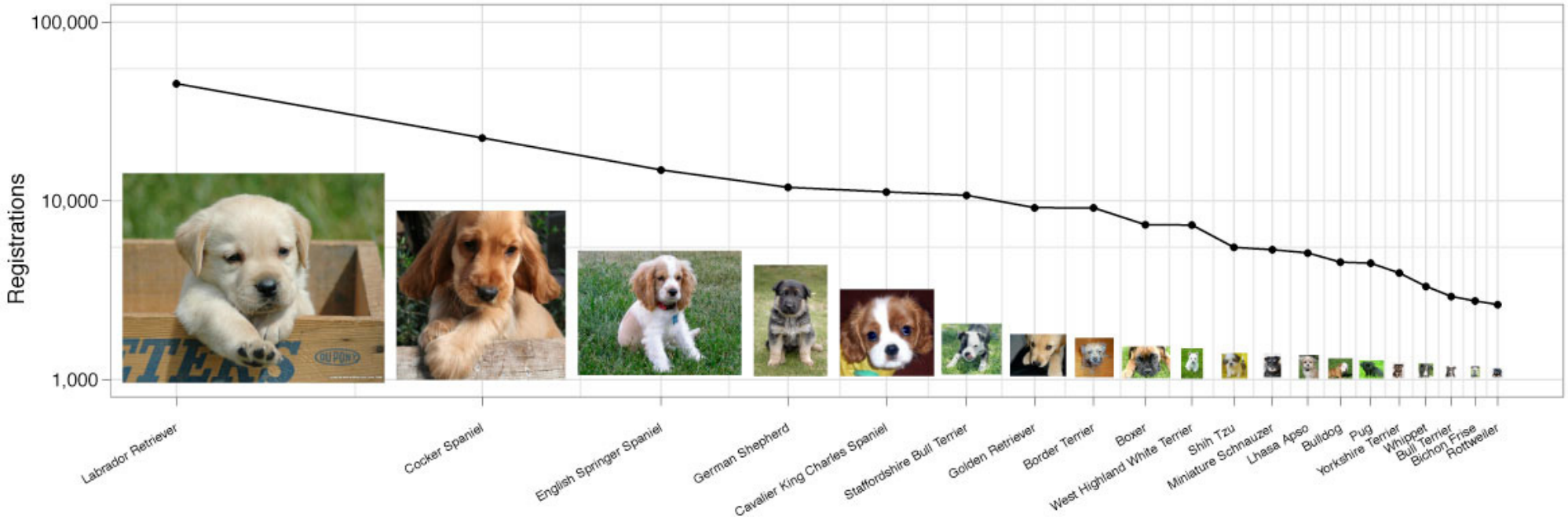


Power Law is Ubiquitous!



Power Law is Ubiquitous!

Puppy Power Law



拉布拉多犬　美國獵犬　英國獵犬　德國牧羊犬

3 Essential Network Properties

- **Short** Average Path Length
  - Small-world Effect (小世界現象)
- **High** Clustering Coefficient
  - Friends of Friends are Friends (群聚現象)
- **Power-law** Degree Distribution
  - Long-tail Effect (長尾效應)
- Question
  - Why do real-world networks exhibit these properties?
  - Can we model the generative process of networks such that properties are exhibited?

More Network Properties

Static	• **High** Clustering Coefficient [Watts'98] • **Low** Average Path Length [Watts'98] • **Power-Law** Degree Distribution [Barabasi'99] • Emergence of Giant Component [Erdos'60] • Triangle Power Law [Tsourakakis'08] • Eigenvalue Power Law [Siganos'03] • Community Structure [Newman'02]
Dynamic	• **Densification** Power Law [Leskovec'05] • **Shrinking Diameter** [Leskovec'05] • Next-Large Connected Component [McGlohon'08] • Principal Eigenvalue Power Law [Akoglu'08] • Busty/self-similar edge/weight additions [McGlohon'08] • Weight Power Law [McGlohon'08]