Static Random Graphs

EE599: Social Network Systems

Keith M. Chugg

Fall 2014

USC University of Southern California

Overview

- Probability Review
- Static Random Models
  - Poisson random graphs (completely random)
  - The configuration model (specify the degree dist
    - *Power-law degree distribution*
  - The small world model (Watts-Strogatz)
  - Exponential and Markov models
- Growth models for random graphs
  - Preferential-attachment & power-law degree distributions

Start with purely random models and see what occurs.

Modify model to better represent real networks

Primary References

- **Static models**
  - Newman, The Structure and Function of Complex Networks, SIAM REVIEW, Vol.45, No. 2, pp.167–256, 2003. (terse and mathematical).
    - Sections IV, V, VI
  - Jackson, Chapter 4. Similar to Newman.
  - Barabasi, Chapters 3 & 4. Less mathematical, buggy.
- **Models for growth of random networks**
  - Newman, The Structure and Function of Complex Networks, SIAM REVIEW, Vol.45, No. 2, pp.167–256, 2003.
    - Sections VII
  - Jackson, Chapter 5.
  - Barabasi, Chapters 5 & 6.

Why Growth Models?

- Unlike the static random graph models we considered, growth models allow us to incorporate temporal aspects of network formation
  - May model real effects of social network formation and evolution
    - "older" nodes have more connections
    - Hubs (popular nodes) emerge
    - Age-based homophily
    - Degree correlation (positive assortativity)

Random Growth Models

- Typical growth model
  - Start with *m* nodes, typically completely connected
  - Add a node of degree *m* and connect it to m existing nodes
    - Randomly connect (Jackson 5.1)
    - Preferential attachment (Jackson 5.2, Newman VII, Barabasi Ch. 5)

General Properties of Growth Models

- Older nodes tend to have larger degree since, when each additional node is added each existing node has some chance of getting a new connection
- Node degree only grows with time
- We can use difference and/or differential equations to model the growth of the expected degrees or the evolution of the degree distribution

Properties of Growth Models

- Random growth yields an exponential distribution for the expected degree
  - For large values of "time", the fraction of nodes with expected degree above *d* is

$$e^{-\frac{d-m}{m}}$$

Preferential Attachment

- New nodes are more likely to connect with existing nodes of large degree
  - Motivated in the formation of many real world networks
    - Citation networks
    - Webpage links
    - Transportation networks
    - Wealth distribution
- "rich get richer" and the "80/20 rule"
- May be an "optimized" structure

Preferential Attachment

PREFERENTIAL ATTACHMENT: A BRIEF HISTORY

Preferential attachment has emerged repeatedly in mathematics and social sciences. Consequently today we can encounter it under different names in the scientific literature:

- It made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya (1887-1985) [2], proposed to explain the nature of certain distributions. Hence, in mathematics preferential attachment is often called a *Pólya process.*
- George Udmy Yule (1871-1951) in 1925 used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a *Yule process.*
- Rober Gibrat (1904-1980) in 1931 proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called *proportional growth* this is a form of preferential attachment.
- George Kinsley Zipf (1902-1950) in 1941 used preferential attachment to explain the fat tailed distribution of wealth in the society [5].
- Modern analytical treatments of preferential attachment use of the master equation approach pioneered by the economist Herbert Alexander Simon (1916-2001). Simon used preferential attachment in 1955 to explain the fat-tailed nature of the distributions describing city sizes, word frequencies in a text, or the number of papers published by scientists [6].
- Building on Simon's work, Derek de Solla Price (1922-1983) used preferential attachment to explain the citation statistics of scientific publications, referring to it as *cumulative advantage* [7].
- In sociology preferential attachment is often called the *Matthew effect,* named by Robert Merton (1910-2003) [8] after a passage in the Gospel of Matthew: "For everyone who has will be given more, and he will have an abundance. Whoever does not have, even what he has will be taken from him."
- The term *preferential attachment* was introduced in the 1999 paper by Barabási and Albert [1] to explain the ubiquity of power laws in networks.

Barabasi

Barabasi-Albert essentially rediscovered Price's work

Price considered directed graphs, B-A undirected (simpler)

Preferential Attachment

- New nodes are more likely to connect with existing nodes of large degree
- Consider a network with 4 nodes with

$$k1=3, k2=2, k3=1, k4=4$$

$$(1,1,1,2,3,4,4,4)$$

- Add a new node, and select target for connection by randomly selecting from the above list
  - Example: node 5 is added and connects to 2 nodes

$$(1,1,1,2,3,4,4,4,5,5,1,4)$$

Preferential Attachment

- The previous process can be captured in two ways

Expected number of new connections existing node *i* gets at time *n*:

$$m \frac{k_i(n)}{\sum_i k_i(n)} = \frac{k_i(n)}{2n} \qquad \Delta k_i(n) = \frac{k_i(n)}{2n}$$

Yields the "mean field approximation" (used by Jackson in Ch. 5)

Probability of connecting to any node with degree *k*

$$\frac{k p_n(k)}{\sum_k k p_n(k)} = \frac{k p_n(k)}{2m}$$

difference equation on $p_n(k)$

Yields the "master equation" (used by Newman VII) - following slides

Preferential Attachment

Degree distribution at time *n*: $$p_n(k)$$

Probability of attaching to node of degree *k*: $$\frac{kp_n(k)}{\sum_k kp_n(k)} = \frac{kp_n(k)}{2m}$$ (*m* edges for each node)

Expected number of nodes with degree *k* that gain an edge: $$m\frac{kp_n(k)}{2m} = \frac{1}{2}kp_n(k)$$

Expected number of nodes with degree *k* at time *n*: $$Np_n(k) = np_n(k)$$

(one node added at each time: *N*=*n*)

This yields to the so-called master equation — difference equation for degree distribution

Preferential Attachment

Master equation (k>m):

$$(n+1)p_{n+1}(k) - np_n(k) = \frac{1}{2}(k-1)p_n(k-1) - \frac{1}{2}kp_n(k)$$

Difference in expected number of degree *k* nodes from time *n* to time (*n*+1)

Expected number of degree *k-1* nodes moving to degree *k* at time *n*

Expected number of degree *k* nodes moving to degree *k+1* at time *n*

Master equation (k=m):

$$(n+1)p_{n+1}(m) - np_n(m) = 1 - \frac{1}{2}mp_n(m)$$

New node has degree *m*

Preferential Attachment

Search for a steady-state solution to the master equation:

$$(n+1)p(k) - np(k) = \frac{1}{2}(k-1)p(k-1) - \frac{1}{2}kp(k) \qquad \text{(k>m)}$$

$$(n+1)p(m) - np(m) = 1 - \frac{1}{2}mp(m) \qquad \text{(k=m)}$$

Yields: $$p(k) = \left[\frac{(k-1)}{(k-2)}\right]p(k-1)$$ & $$p(m) = \frac{2}{(m+2)}$$

Preferential Attachment

The steady-state solution to the master equation:

$$p(k) = \frac{(k-1)(k-2)\cdots m}{(k+2)(k+1)\cdots(m+3)}p(m) = \frac{2m(m+1)}{(k+2)(k+1)k} \sim k^{-3}$$

- Preferential attachment leads to power-law degree distribution!
  - recall this means "hubs" in the network, heavy tails in the degree distribution, etc.

Preferential Attachment Summary

BOX 5.6

AT A GLANCE: BARABÁSI-ALBERT MODEL

Number of nodes

$N = t$

Number of links

$N = mt$

Average Degree

$\langle k \rangle = 2m$

Degree dynamics

$k_i(t) = m\,(t/t_i)^{\beta}$

Dynamical exponent

$\beta = 1/2$

Degree distribution

$p_k \sim k^{\gamma}$

Degree exponent

$\gamma = 3$

Average distance

$\langle d \rangle \sim \log N / \log \log N$

Clustering coefficient

$\langle C \rangle \sim (\ln N)^2 / N$

Barabasi

Preferential Attachment Variations

- A hybrid model (Jackson 5.3)
  - A new node connects fraction ß of its m edges randomly to existing nodes and fraction (1-ß) of its edges to existing nodes via preferential attachment
    - Yields a power-law distribution with exponent:

$$p(k) \sim k^{-\gamma} \qquad \gamma = 1 + \frac{2}{1-\beta}$$

- An exponent of less than 3 is possible with this model

Hybrid Model Motivation

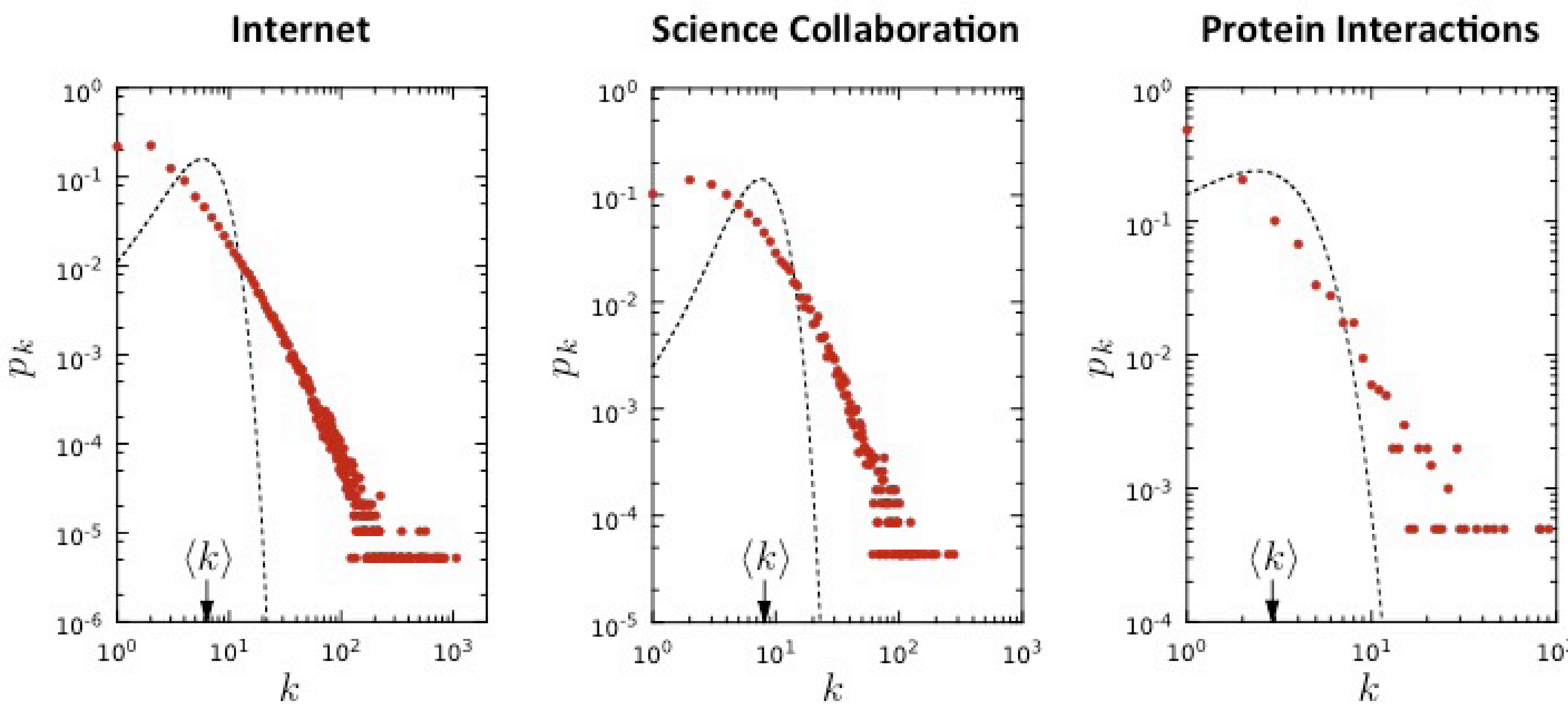


Image 3.5

Degree distribution of real networks.

The degree distribution of the Internet, science collaboration network, and the protein interaction network of yeast (Table 2.1). The dashed line corresponds to the Poisson prediction, obtained by measuring ‹k› for the real network and then plotting Eq. (8). The significant deviation between the data and the Poisson fit indicates that the random network model underestimates the size and the frequency of highly connected nodes, or hubs.

Barabasi

Preferential Attachment Variations

- Preferential attachment implies older nodes have, on average, more edges and are more likely to connect to older nodes
  - This may not always model reality as trends may fade with time
- Biaconi and Barabasi added a "fitness" term to the preferential attachment law which measures a nodes inherent "attractiveness" or fitness
  - Allows newer nodes to become hubs through high fitness scores

Preferential Attachment Variations

- Note that the probability of attaching to a degree k node is the same as the "discovered degree" distribution
  - Meetings-Based models are based on forming new connections through navigating the network (getting meetings through friends)
    - This is usually modeled for directed graphs, but yields similar results as the hybrid model
    - Adds high level of degree correlation

Scale-free Degree Distribution

- The preferential attachment model and variations yield power-law degree distributions with exponents ~ 2-3
  - Explore these degree distributions in more detail (Barabasi Ch. 4)

Power Law Degree Distribution

- Heavy-tailed degree distribution that matches many real-world network observed degree distributions

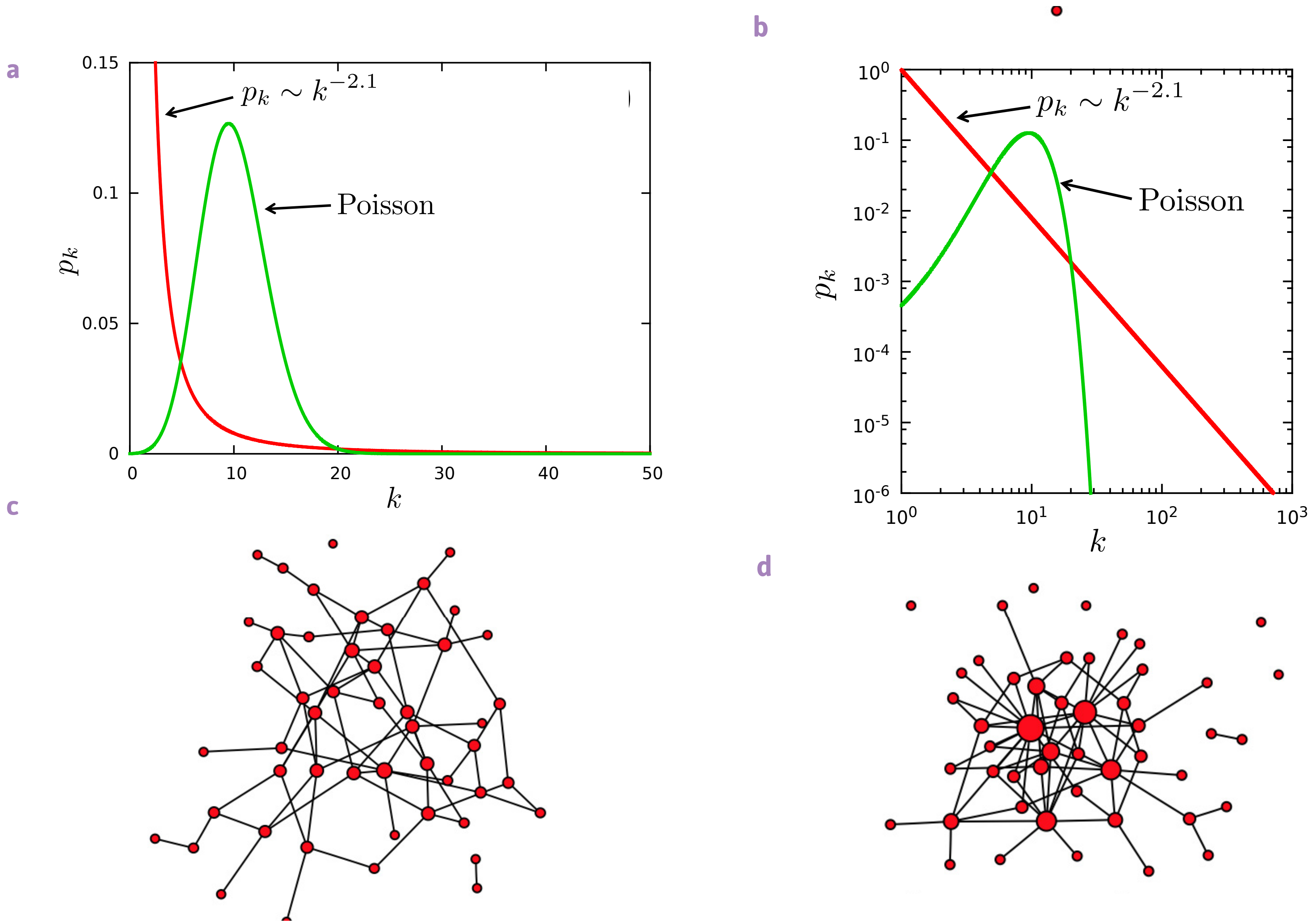


Figure 4.4
Poisson vs. power-law distributions

(a) A Poisson function and a power-law function with $\gamma = 2.1$. Both distributions have $\langle k \rangle = 10$.
(b) The curves in (a) shown on a log-log plot, offering a better view of the difference between the two functions in the high-k regime.
(c) A random network with $\langle k \rangle = 3$ and $N = 50$, illustrating that most nodes have comparable degree $k \simeq \langle k \rangle$.
(d) A scale-free network with $\langle k \rangle = 3$, illustrating that numerous small-degree nodes coexist with a few highly connected hubs.

Barabasi

Power Law Degree Distribution

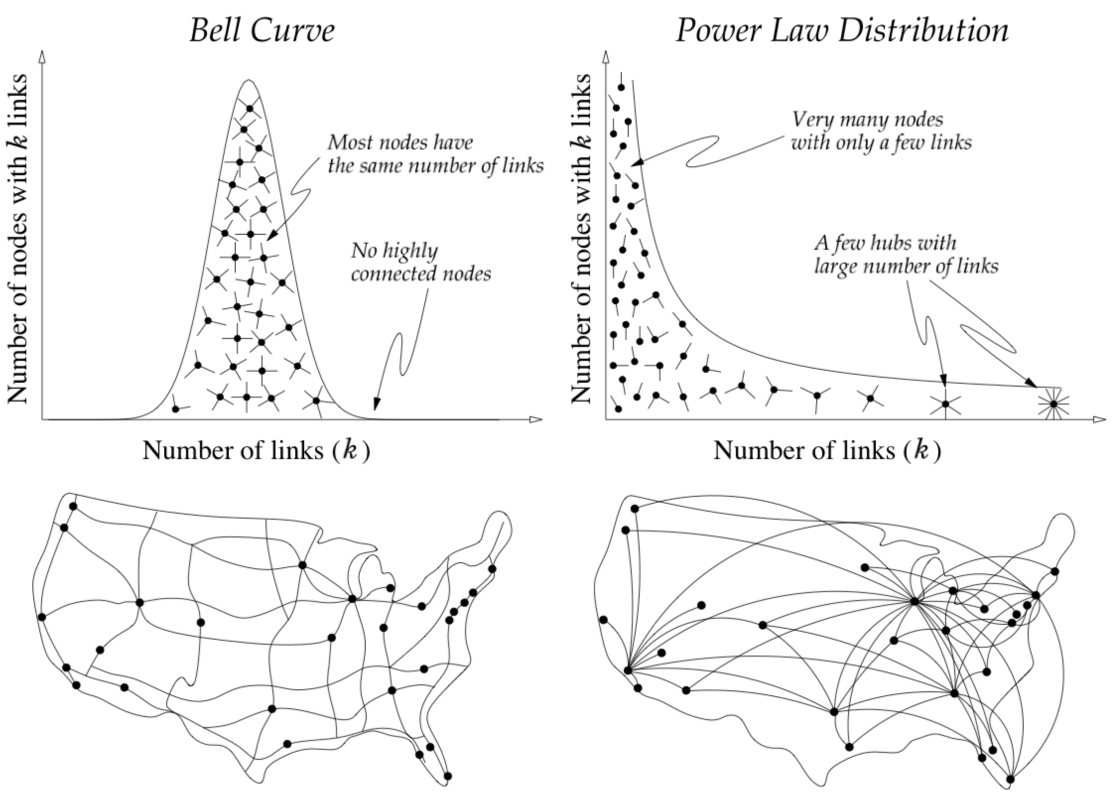


Figure 4.6
Random versus scale-free networks

Left column: the degrees of a random network follow a Poisson distribution, which is rather similar to the Bell curve shown in the figure. This indicates that most nodes have comparable degree. Hence nodes with a large number of links are absent (top panel). Consequently a random network looks a bit like a national highway network in which nodes are cities and links are the major highways connecting them (bottom panel). Indeed, there are no major cities with hundreds of highways and no city is disconnected from the highway system.

Right column: In a network with a power-law degree distribution most nodes have only a few links. These numerous small nodes are held together by a few highly connected hubs (top panel). Consequently a scale-free network looks a bit like the air-traffic network, whose nodes are airports and links are direct flights between them. Most airports are tiny, with only a few flights linking them to other airports. Yet, we can also have few very large airports, like Chicago or Atlanta, that hold hundreds of airports together, acting as major hubs (bottom panel).

Once hubs are present, they change the way we navigate the network. For example, if we travel from Boston to Los Angeles by car, we must drive through many cities (nodes). On the airplane network, however, we can reach most destinations via a single hub, like Chicago.

After [4].

Barabasi

- AKA “scale-free” networks

Power Law Degree Distribution

- Special case of configuration network
- Discussed in detail in Barabasi, chapter 4
- Networks can be grown with the Barabasi-Albert model for preferential attachment
  - Yields the power law degree distribution
  - No longer has independent degrees for different nodes
    - Model incorporates degree correlation explicitly

Cover power-law and preferential attachment later (soon)

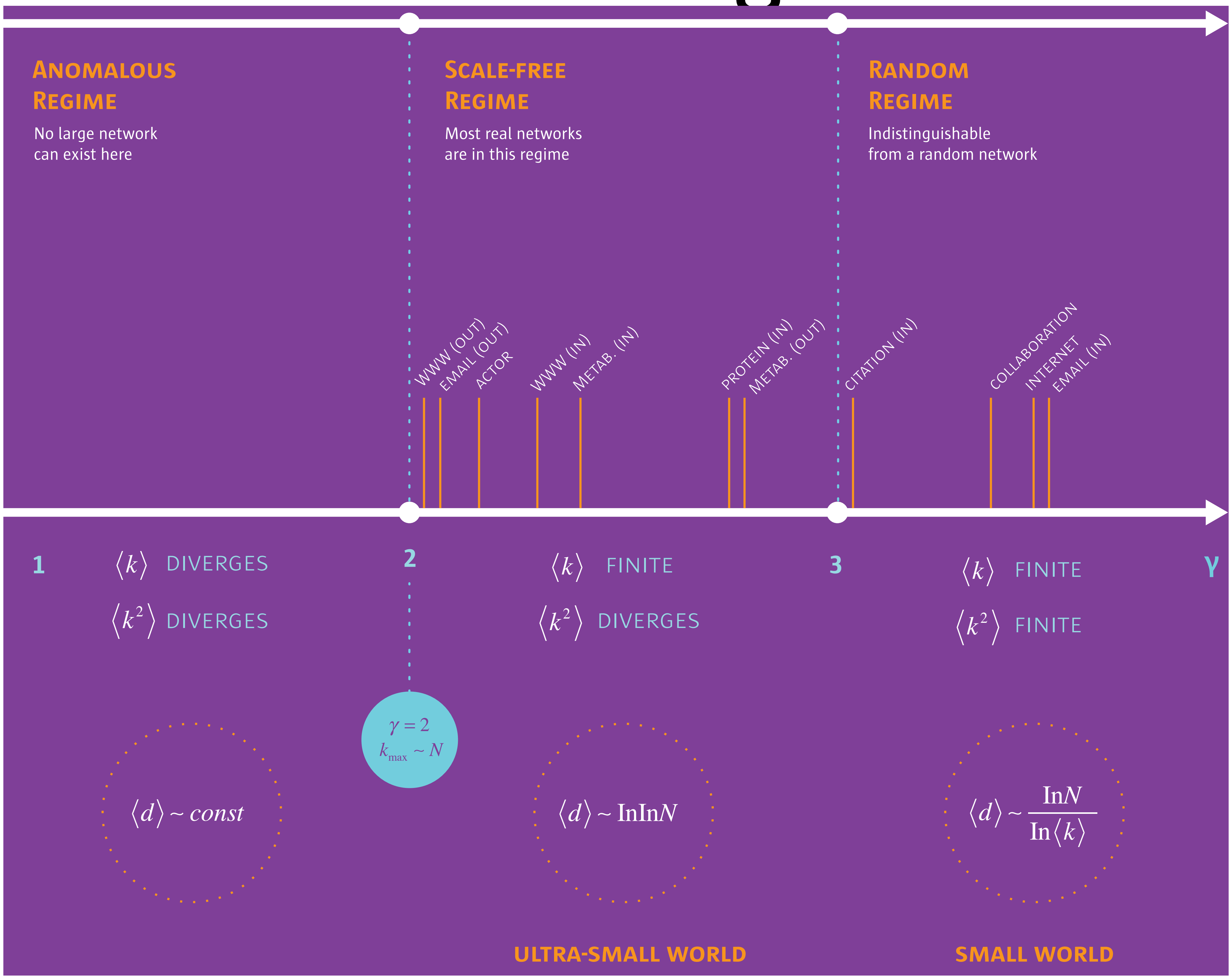
Scale-free Degree Distribution

NETWORK	*NL*		$\langle k \rangle$ $\langle k_{in} \rangle = \langle k_{out} \rangle$	σ_{in}	σ_{out}	σ	γ_{in}	γ_{out}	γ
Internet	192,244	609,066	6.34	-	-	14.14	-	-	3.42*
WWW	325,729	1,497,134	4.60	39.05	21.48	-	2.31	2.00	-
Power Grid	4,941	6,594	2.67	-	-	1.79	-	-	Exp.
Mobile Phone Calls	36,595	91,826	2.51	2.39	2.32	-	4.69*	5.01*	-
Email	57,194	103,731	1.81	9.56	34.07	-	3.43*	2.03	-
Science Collaboration	23,133	93,439	8.08	-	-	10.63	-	-	3.35
Actor Network	702,388	29,397,908	83.71	-	-	200.86	-	-	2.12
Citation Network	449,673	4,689,479	10.43	29.37	9.49	-	3.03**	4.00	-
E. Coli Metabolism	1,039	5,802	5.58	22.46	19.12	-	2.43	2.90	-
Yeast Protein Interactions	2,018	2,930	2.90	-	-	4.88	-	-	2.89*

Barabasi

Best power-law fit to real data sets

Scale-free Degree Distribution



At a glance
Scale-free networks

DEGREE DISTRIBUTION

Discrete form:

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}.$$

Continuous form:

$$p(k) = (\gamma - 1)k_{min}^{\gamma-1}\, k^{-\gamma}.$$

SIZE OF THE LARGEST HUB

$$k_{max} \sim k_{min} N^{\frac{1}{\gamma-1}}.$$

MOMENTS OF p_k

$2 < \gamma < 3$: $\langle k \rangle$ finite, $\langle k^2 \rangle$ diverges when $N \to \infty$.

$\gamma > 3$: $\langle k \rangle$ and $\langle k^2 \rangle$ finite.

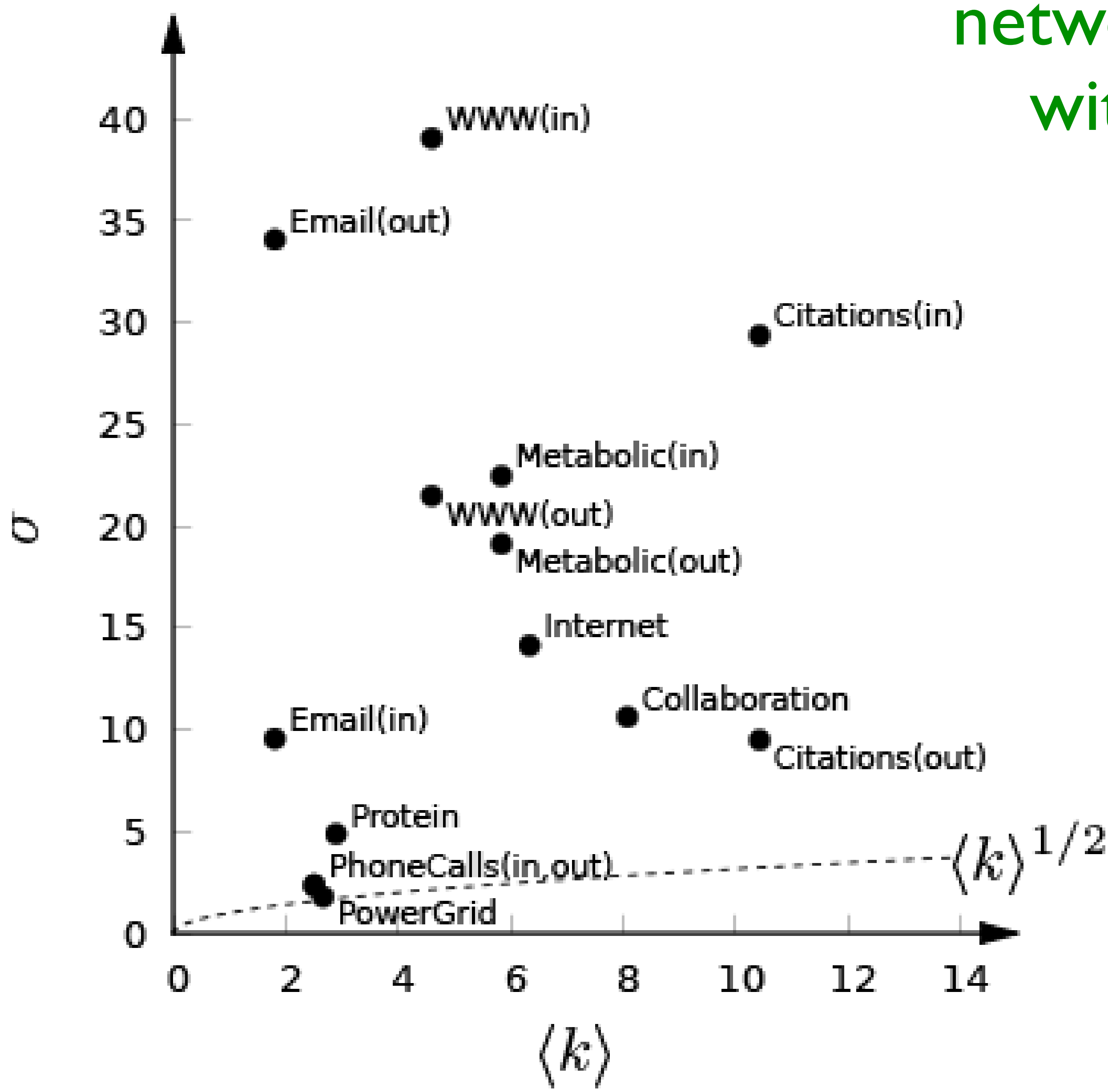
DISTANCES IN A SCALE-FREE NETWORK

$$d \sim \begin{cases} \text{const.} & \text{if } \gamma = 2, \\ \frac{\ln\ln N}{\ln(\gamma - 1)} & \text{if } 2 < \gamma < 3, \\ \frac{\ln N}{\ln\ln N} & \text{if } \gamma = 3, \\ \ln N & \text{if } \gamma > 3. \end{cases}$$

Barabasi

Scale-free Degree Distribution

another way to see that real networks have degree distributions with heavier tails than Poisson



Barabasi

Figure 4.8
Standard deviation is large in real networks

For a random network the standard deviation follows $\sigma_k = \sqrt{\langle k \rangle}$,shown as a dashed line on the figure. The symbols show σ for ten reference networks **Table 4.1**, indicating that for each σ is larger than expected for a random network with similar $\langle k \rangle$. The only exception is the power grid, which is not scale-free. While the phone call network is scale-free, it has a large γ, hence it behaves like a random network.