INTRODUCTION

Hubs represent the most striking difference between a random and a scale-free network. On the World Wide Web, they are websites with an exceptional number of links, like google.com or facebook.com; in the metabolic network they are molecules like ATP or ADP, energy carriers involved in an exceptional number of chemical reactions. The very existence of these hubs and the related scale-free topology raises two fundamental questions:

• Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

• Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in real networks?

The first question is particularly puzzling given the fundamental differences in the nature, origin, and scope of the systems that display the scale-free property:

• The *nodes* of the cellular network are metabolites or proteins, while the nodes of the WWW are documents, representing information without a physical manifestation.

• The *links* within the cell are chemical reactions and binding interactions, while the links of the WWW are URLs, or small segments of computer code.

• The *history* of these two systems could not be more different: The cellular network is shaped by 4 billion years of evolution, while the WWW is less than three decades old.

• The *purpose* of the metabolic network is to produce the chemical components the cell needs to stay alive, while the purpose of the WWW is information access and delivery.

Online Resource 5.1
Scale-free Sonata

Listen to a recording of Michael Edward Edgerton's *1 sonata for piano*, music inspired by scale-free networks.

To understand why so *different* systems converge to a *similar* architecture we need to first understand the mechanism responsible for the emergence of the scale-free property. This is the main topic of this chapter. Given the diversity of the systems that display the scale-free property, the explanation must be simple and fundamental. The answers will change the way we model networks, forcing us to move from describing a network's topology to modeling the evolution of a complex system.

GROWTH AND PREFERENTIAL ATTACHMENT

We start our journey by asking: Why are hubs and power laws absent in random networks? The answer emerged in 1999, highlighting two hidden assumptions of the Erdős-Rényi model, that are violated in real networks [1]. Next we discuss these assumptions separately.

Networks Expand Through the Addition of New Nodes

The random network model assumes that we have a *fixed* number of nodes, *N*. Yet, *in real networks the number of nodes continually grows thanks to the addition of new nodes*.

Consider a few examples:

• In 1991 the WWW had a single node, the first webpage build by Tim Berners-Lee, the creator of the Web. Today the Web has over a trillion (10^{12}) documents, an extraordinary number that was reached through the continuous addition of new documents by millions of individuals and institutions (Figure 5.1a).

• The collaboration and the citation network continually expands through the publication of new research papers (Figure 5.1b).

• The actor network continues to expand through the release of new movies (Figure 5.1c).

• The protein interaction network may appear to be static, as we inherit our genes (and hence our proteins) from our parents. Yet, it is not: The number of genes grew from a few to the over 20,000 genes present in a human cell over four billion years.

Consequently, if we wish to model these networks, we cannot resort to a static model. Our modeling approach must instead acknowledge that networks are the product of a steady growth process.

Nodes Prefer to Link to the More Connected Nodes

The random network model assumes that we randomly choose the interaction partners of a node. Yet, *most real networks new nodes prefer to link to the more connected nodes*, a process called *preferential attachment* (Figure 5.2).

Consider a few examples:

• We are familiar with only a tiny fraction of the trillion or more documents available on the WWW. The nodes we know are not entirely random: We all heard about Google and Facebook, but we rarely encounter the billions of less-prominent nodes that populate the Web. As our knowledge is biased towards the more popular Web documents, we are more likely to link to a high-degree node than to a node with only few links.

• No scientist can attempt to read the more than a million scientific papers published each year. Yet, the more cited is a paper, the more likely that we hear about it and eventually read it. As we cite what we read, our citations are biased towards the more cited publications, representing the high-degree nodes of the citation network.

• The more movies an actor has played in, the more familiar is a casting director with her skills. Hence, the higher the degree of an actor in the actor network, the higher are the chances that she will be considered for a new role.

In summary, the random network model differs from real networks in two important characteristics:

(A) Growth

Real networks are the result of a growth process that continuously increases N. In contrast the random network model assumes that the number of nodes, N, is fixed.

(B) Preferential Attachment

In real networks new nodes tend to link to the more connected nodes. In contrast nodes in random networks randomly choose their interaction partners.

There are many other differences between real and random networks, some of which will be discussed in the coming chapters. Yet, as we show next, these two, *growth* and *preferential attachment*, play a particularly important role in shaping a network's degree distribution.
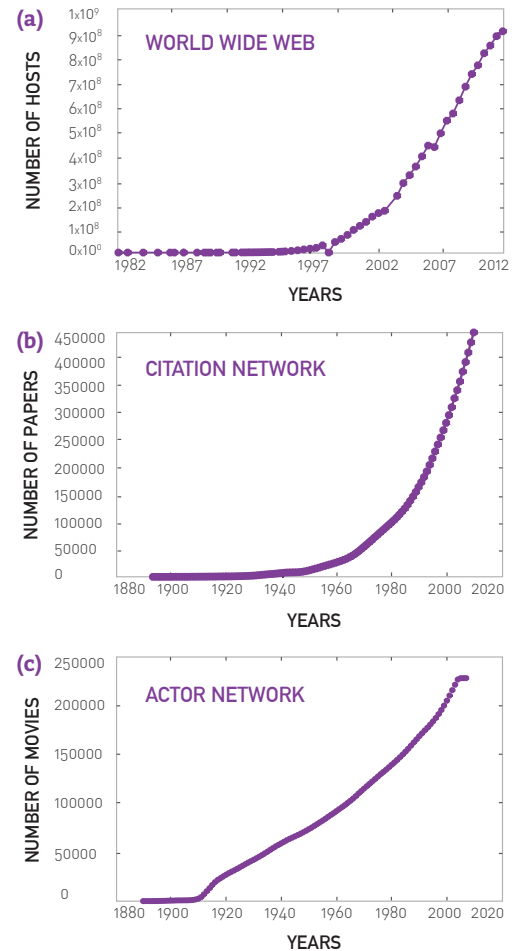
(a) WORLD WIDE WEB

(b) CITATION NETWORK

(c) ACTOR NETWORK

Figure 5.1
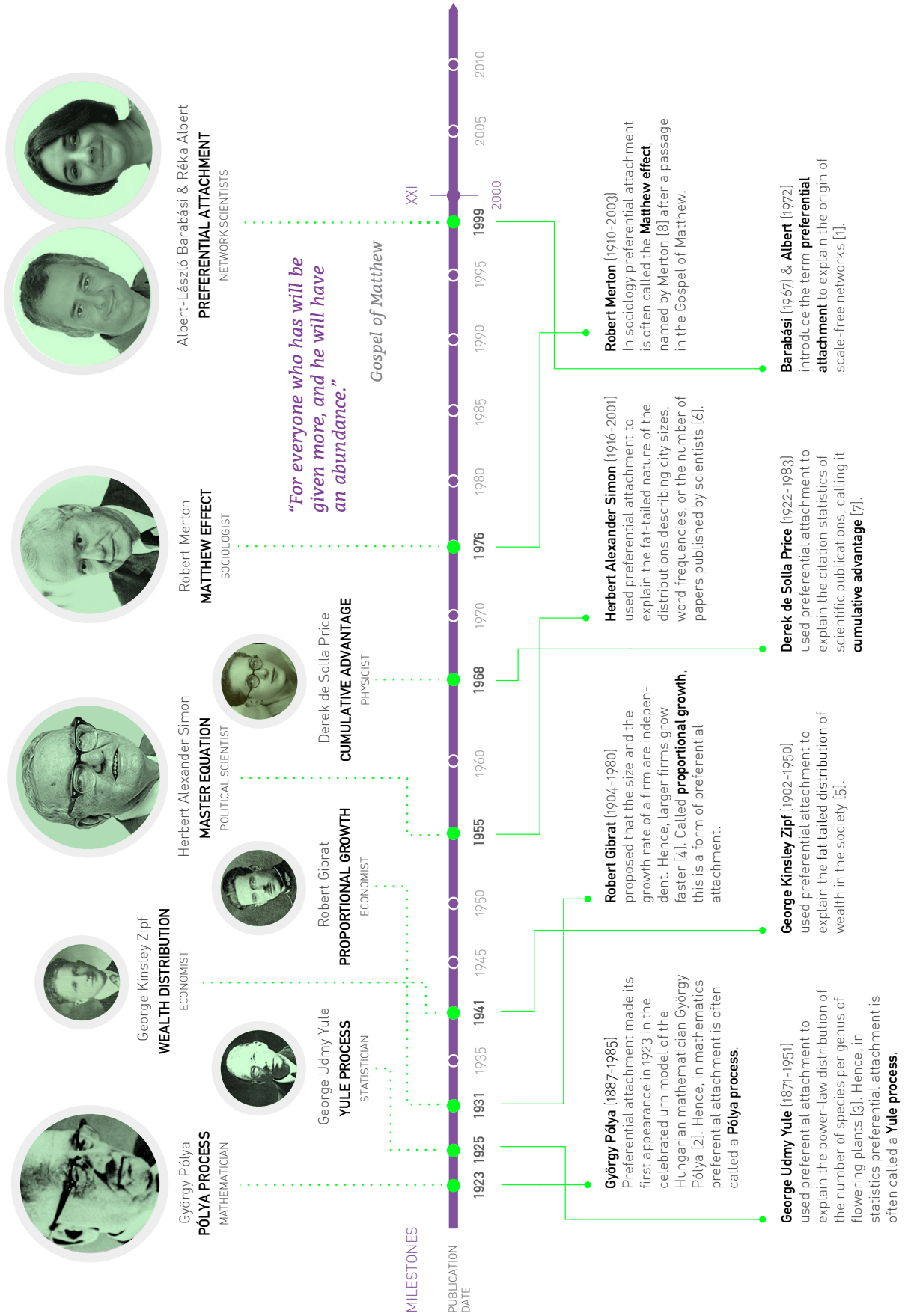
The Growth of Networks

Networks are not static, but grow via the addition of new nodes:

(a) The evolution of the number of WWW hosts, documenting the Web's rapid growth. After *http://www.isc.org/solutions/survey/history*.

(b) The number of scientific papers published in *Physical Review* since the journal's founding. The increasing number of papers drives the growth of both the science collaboration network as well as of the citation network shown in the figure.

(c) Number of movies listed in IMDB.com, driving the growth of the actor network.

FIG 5.2

PREFERENTIAL ATTACHMENT: A BRIEF HISTORY

Preferential attachment has emerged independently in many disciplines, helping explain the presence of power laws characterising various systems. In the context of *networks* preferential attachment was introduced in 1999 to explain the scale-free property.

György Pólya
PÓLYA PROCESS
MATHEMATICIAN

George Kinsley Zipf
WEALTH DISTRIBUTION
ECONOMIST

George Udmy Yule
YULE PROCESS
STATISTICIAN

Herbert Alexander Simon
MASTER EQUATION
POLITICAL SCIENTIST

Robert Gibrat
PROPORTIONAL GROWTH
ECONOMIST

Derek de Solla Price
CUMULATIVE ADVANTAGE
PHYSICIST

Robert Merton
MATTHEW EFFECT
SOCIOLOGIST

Albert-László Barabási & Réka Albert
PREFERENTIAL ATTACHMENT
NETWORK SCIENTISTS

"For everyone who has will be given more, and he will have an abundance."

Gospel of Matthew

MILESTONES

PUBLICATION DATE

XXI

1923   1925   1931   1935   1941   1945   1950   1955   1960   1968   1970   1976   1980   1985   1990   1995   1999   2000   2005   2010

György Pólya (1887–1985)
Preferential attachment made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya [2]. Hence, in mathematics preferential attachment is often called a **Pólya process.**

George Udmy Yule (1871–1951)
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.**

Robert Gibrat (1904–1980)
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)
used preferential attachment to explain the fat tailed distribution of wealth in the society [5].

Herbert Alexander Simon (1916–2001)
used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].

Derek de Solla Price (1922–1983)
used preferential attachment to explain the citation statistics of scientific publications, calling it **cumulative advantage** [7].

Robert Merton (1910–2003)
In sociology preferential attachment is often called the **Matthew effect,** named by Merton [8] after a passage in the Gospel of Matthew.

Barabási (1967) **& Albert** (1972)
introduce the term **preferential attachment** to explain the origin of scale-free networks [1].

THE BARABÁSI-ALBERT MODEL

The recognition that growth and preferential attachment coexist in real networks has inspired a minimal model called the *Barabási-Albert* model, which can generate scale-free networks [1]. Also known as the *BA model* or the *scale-free model*, it is defined as follows:

We start with m_0 nodes, the links between which are chosen arbitrarily, as long as each node has at least one link. The network develops following two steps (Figure 5.3):

(A) Growth

At each timestep we add a new node with m ($\leq m_0$) links that connect the new node to m nodes already in the network.

(B) Preferential attachment

The probability $\Pi(k)$ that a link of the new node connects to node i depends on the degree k_i as

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}.$$

(5.1)

Preferential attachment is a probabilistic mechanism: A new node is free to connect to *any* node in the network, whether it is a hub or has a single link. Equation (5.1) implies, however, that if a new node has a choice between a degree-two and a degree-four node, it is twice as likely that it connects to the degree-four node.

After t timesteps the Barabási-Albert model generates a network with $N = t + m_0$ nodes and $m_0 + mt$ links. As Figure 5.4 shows, the obtained network has a power-law degree distribution with degree exponent $\gamma=3$. A mathematically self-consistent definition of the model is provided in BOX 5.1.

As Figure 5.3 and Online Resource 5.2 indicate, while most nodes in the network have only a few links, a few gradually turn into hubs. These hubs are the result of a *rich-gets-richer phenomenon*: Due to preferential attach-
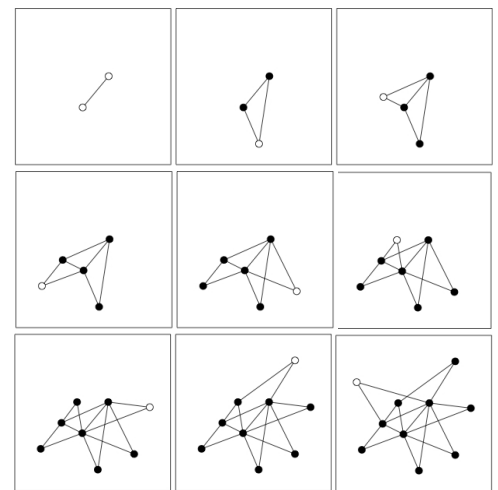
Figure 5.3
Evolution of the Barabási-Albert Model

The sequence of images shows nine subsequent steps of the Barabási-Albert model. Empty circles mark the newly added node to the network, which decides where to connect its two links (*m*=2) using preferential attachment (5.1). After [9].
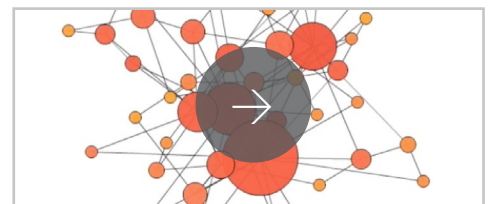
Online Resource 5.2
Emergence of a Scale-free Network

Watch a video that shows the growth of a scale-free network and the emergence of the hubs in the Barabási-Albert model. Courtesy of Dashun Wang.

ment new nodes are more likely to connect to the more connected nodes than to the smaller nodes. Hence, the larger nodes will acquire links at the expense of the smaller nodes, eventually becoming hubs.

In summary, the Barabási-Albert model indicates that two simple mechanisms, *growth* and *preferential attachment*, are responsible for the emergence of scale-free networks. The origin of the power law and the associated hubs is a *rich-gets-richer phenomenon* induced by the coexistence of these two ingredients. To understand the model's behavior and to quantify the emergence of the scale-free property, we need to become familiar with the model's mathematical properties, which is the subject of the next section.
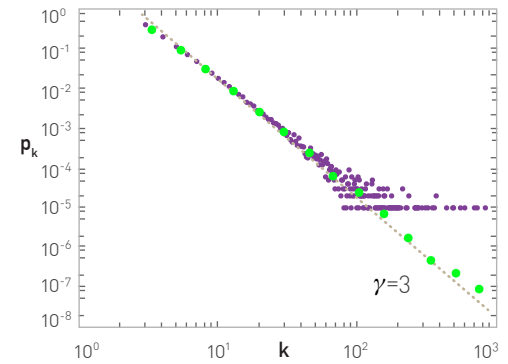
Figure 5.4

The Degree Distribution

The degree distribution of a network generated by the Barabási-Albert model. The figure shows p_k for a single network of size N=100,000 and m=3. It shows both the linearly-binned (purple) and the log-binned version (green) of p_k. The straight line is added to guide the eye and has slope γ=3, corresponding to the network's predicted degree exponent.

BOX 5.1

The definition of the Barabási-Albert model leaves many mathematical details open:

• It does not specify the precise initial configuration of the first m_0 nodes.

• It does not specify whether the m links assigned to a new node are added one by one, or simultaneously. This leads to potential mathematical conflicts: If the links are truly independent, they could connect to the same node i, resulting in multi-links.

Bollobás and collaborators [10] proposed the *Linearized Chord Diagram* (LCD) to resolve these problems, making the model more amenable to mathematical approaches.

According to the LCD, for $m=1$ we build a graph $G_1^{(t)}$ as follows (Figure 5.5):

(1) Start with $G_1^{(0)}$, corresponding to an empty graph with no nodes.

(2) Given $G_1^{(t-1)}$ generate $G_1^{(t)}$ by adding the node v_t and a single link between v_t and v_i, where v_i is chosen with probability

$$p = \begin{cases} \dfrac{k_i}{2t-1} & if\ 1 \le i \le t-1 \\ \dfrac{1}{2t-1}, & if\ i = t \end{cases} \qquad (5.2)$$

That is, we place a link from the new node v_t to node v_i with probability $k_i/(2t-1)$, where the new link already contributes to the degree of v_t. Consequently node v_t can also link to itself with probability $1/(2t-1)$, the second term in (5.2). Note also that the model permits self-loops and multi-links. Yet, their number becomes negligible in the $t \to \infty$ limit.

For $m > 1$ we build $G_m^{(t)}$ by adding m links from the new node v_t one by one, in each step allowing the outward half of the newly added link to contribute to the degrees.
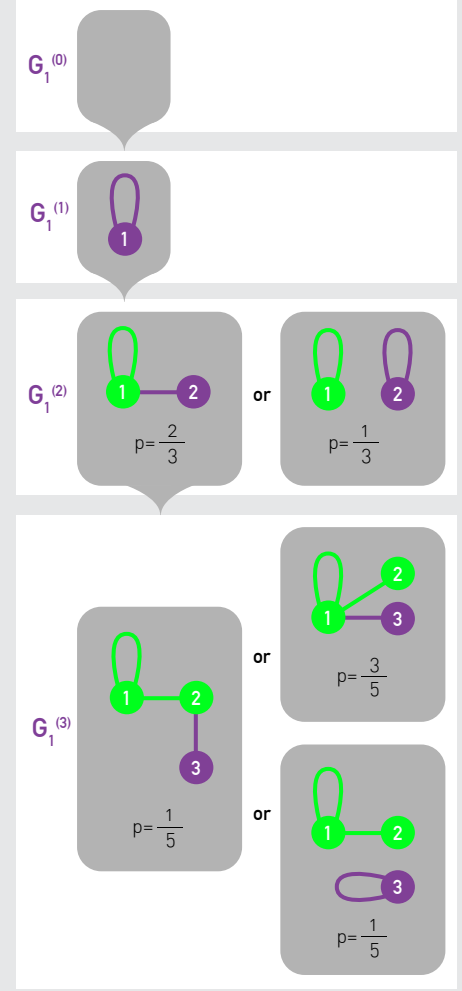
Figure 5.5
The Linearized Chord Diagram (LCD)

The construction of the LCD, the version of the Barabási-Albert model amenable to exact mathematical calculations [10]. The figure shows the first four steps of the network's evolution for $m=1$:

$G_1^{(0)}$: We start with an empty network.
$G_1^{(1)}$: The first node can only link to itself, forming a self-loop. Self-loops are allowed, and so are multi-links for $m>1$.
$G_1^{(2)}$: Node 2 can either connect to node 1 with probability 2/3, or to itself with probability 1/3. According to (5.2), half of the links that the new node 2 brings along is already counted as present. Consequently node 1 has degree $k_1=2$ at node 2 has degree $k_2=1$, the normalization constant being 3.
$G_1^{(3)}$: Let us assume that the first of the two $G_1^{(t)}$ network possibilities have materialized. When node 3 comes along, it again has three choices: It can connect to node 2 with probability 1/5, to node 1 with probability 3/5 and to itself with probability 1/5.

DEGREE DYNAMICS

To understand the emergence of the scale-free property, we need to focus on the time evolution of the Barabási-Albert model. We begin by exploring the time-dependent degree of a single node [11].

In the model an existing node can increase its degree each time a *new* node enters the network. This new node will link to m of the $N(t)$ nodes already present in the system. The probability that one of these links connects to node i is given by (5.1).

Let us approximate the degree k_i with a continuous real variable, representing its expectation value over many realizations of the growth process. The rate at which an existing node i acquires links as a result of new nodes connecting to it is

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_{j=1}^{N-1}k_j}. \tag{5.3}$$

The coefficient m describes that each new node arrives with m links. Hence, node i has m chances to be chosen. The sum in the denominator of (5.3) goes over all nodes in the network except the newly added node, thus

$$\sum_{j=1}^{N-1}k_j = 2mt - m. \tag{5.4}$$

Therefore (5.4) becomes

$$\frac{dk_i}{dt} = \frac{k_i}{2t-1}. \tag{5.5}$$

For large t the (-1) term can be neglected in the denominator, obtaining

$$\frac{dk_i}{k_i} = \frac{1}{2}\frac{dt}{t}. \tag{5.6}$$

By integrating (5.6) and using the fact that $k_i(t_i)=m$, meaning that node i joins the network at time t_i with m links, we obtain

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta}.$$ (5.7)

We call β the *dynamical exponent* and has the value

$$\beta = \frac{1}{2}.$$

Equation (5.7) offers a number of predictions:

- The degree of each node increases following a power-law with the same dynamical exponent $\beta = 1/2$ (Figure 5.6a). Hence all nodes follow the same dynamical law.

- The growth in the degrees is sublinear (i.e. $\beta < 1$). This is a consequence of the growing nature of the Barabási-Albert model: Each new node has more nodes to link to than the previous node. Hence, with time the existing nodes compete for links with an increasing pool of other nodes.

- The earlier node i was added, the higher is its degree $k_i(t)$. Hence, hubs are large because they arrived earlier, a phenomenon called *first-mover advantage* in marketing and business.

- The rate at which the node i acquires new links is given by the derivative of (5.7)

$$\frac{dk_i(t)}{dt} = \frac{m}{2}\frac{1}{\sqrt{t_i t}},$$ (5.8)

indicating that in each time step older nodes acquire more links (as they have smaller t_i). Furthermore the rate at which a node acquires links decreases with time as $t^{-1/2}$. Hence, fewer and fewer links go to a node.

In summary, the Barabási-Albert model captures the fact that in real networks nodes arrive one after the other, offering a dynamical description of a network's evolution. This generates a competition for links during which the older nodes have an advantage over the younger ones, eventually turning into hubs.
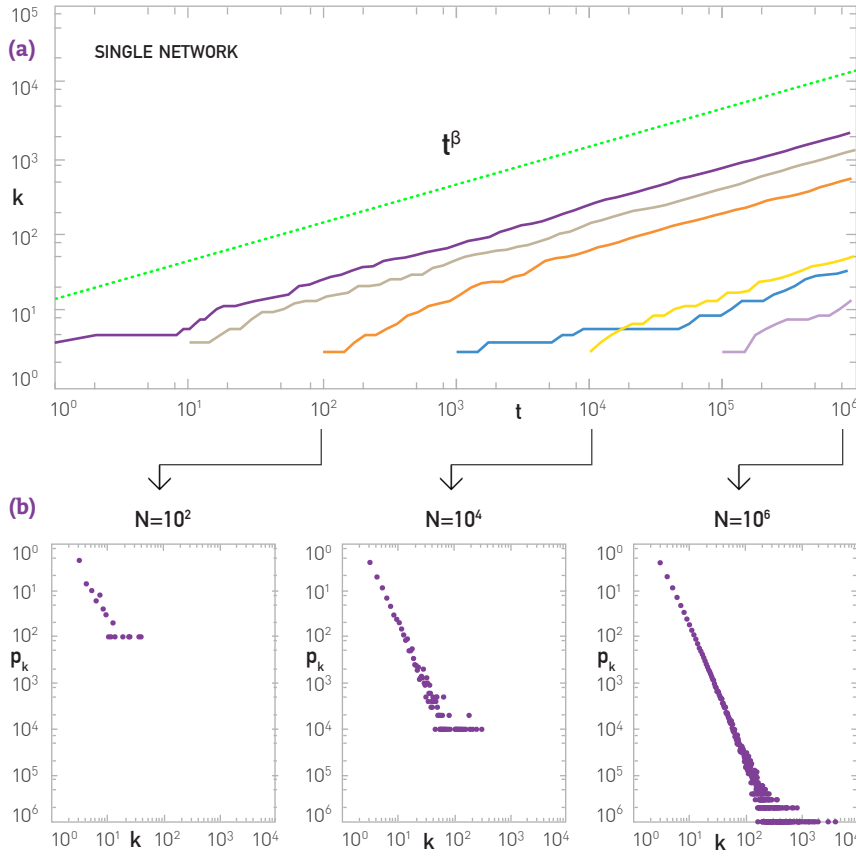
Figure 5.6

Degree Dynamics

(a) The growth of the degrees of nodes added at time $t = 1, 10, 10^2, 10^3, 10^4, 10^5$ (continuous lines from left to right) in the Barabási-Albert model. Each node increases its degree following (5.7). Consequently at any moment the older nodes have higher degrees. The dotted line corresponds to the analytical prediction (5.7) with $\beta = 1/2$.

(b) Degree distribution of the network after adding $N = 10^2, 10^4$, and 10^6 nodes, i.e. at time $t = 10^2, 10^4$, and 10^6 (illustrated by arrows in (a)). The larger the network, the more obvious is the power-law nature of the degree distribution. Note that we used linear binning for p_k to better observe the gradual emergence of the scale-free state.

DEGREE DISTRIBUTION

The distinguishing feature of the networks generated by the Barabá-si-Albert model is their power-law degree distribution (Figure 5.4). In this section we calculate the functional form of p_k, helping us understand its origin.

A number of analytical tools are available to calculate the degree distribution of the Barabási-Albert network. The simplest is the *continuum theory* that we started developing in the previous section [1, 11]. It predicts the degree distribution (BOX 5.3),

$$p(k) \approx 2m^{1/\beta} k^{-\gamma}$$ (5.9)

with

$$\gamma = \frac{1}{\beta} + 1 = 3.$$ (5.10)

Therefore the degree distribution follows a power law with degree exponent γ=3, in agreement with the numerical results (Figures 5.4 and 5.7). Moreover (5.10) links the degree exponent, γ, a quantity characterizing the network topology, to the dynamical exponent β that characterizes a node's temporal evolution, revealing a deep relationship between the network's topology and dynamics.

While the continuum theory predicts the correct degree exponent, it fails to accurately predict the pre-factors of (5.9). The correct pre-factors can be obtained using a master [12] or rate equation [13] approach or calculated exactly using the LCD model [10] (BOX 5.2). Consequently the *exact degree distribution* of the Barabási-Albert model is (ADVANCED TOPICS 5.A)

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \;.$$ (5.11)

Equation (5.11) has several implications:

• For large k (5.11) reduces to $p_k \sim k^{-3}$, or $\gamma = 3$, in line with (5.9) and (5.10).

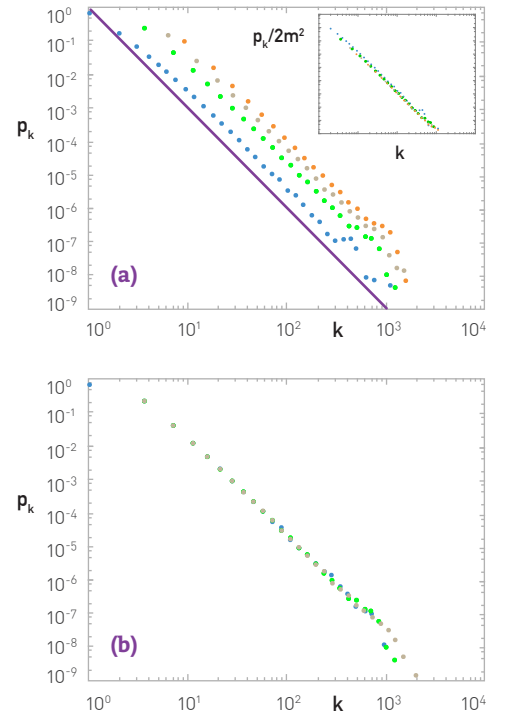• The degree exponent γ is independent of m, a prediction that agrees

Figure 5.7
Probing the Analytical Predictions

(a) We generated networks with N=100,000 and m_0=m=1 (blue), 3 (green), 5 (grey), and 7 (orange). The fact that the curves are parallel to each other indicates that γ is independent of m and m_0. The slope of the purple line is -3, corresponding to the predicted degree exponent γ=3. Inset: (5.11) predicts $p_k \sim 2m^2$, hence $p_k/2m^2$ should be independent of m. Indeed, by plotting $p_k/2m^2$ vs. k, the data points shown in the main plot collapse into a single curve.

(b) The Barabási-Albert model predicts that p_k is independent of N. To test this we plot p_k for N = 50,000 (blue), 100,000 (green), and 200,000 (grey), with m_0=m=3. The obtained p_k are practically indistinguishable, indicating that the degree distribution is *stationary*, i.e. independent of time and system size.

with the numerical results (Figure 5.7a).

- The power-law degree distribution observed in real networks describes systems of rather different age and size. Hence, an appropriate model should lead to a time-independent degree distribution. Indeed, according to (5.11) the degree distribution of the Barabási-Albert model is independent of both t and N. Hence the model predicts the emergence of a *stationary scale-free state*. Numerical simulations support this prediction, indicating that p_k observed for different t (or N) fully overlap (Figure 5.7b).

- Equation (5.11) predicts that the coefficient of the power-law distribution is proportional to $m(m+1)$ (or m^2 for large m), again confirmed by numerical simulations (Figure 5.7a, inset).

In summary, the analytical calculations predict that the Barabási-Albert model generates a scale-free network with degree exponent $\gamma=3$. The degree exponent is independent of the m and m_0 parameters. Furthermore, the degree distribution is stationary (i.e. time invariant), explaining why networks with different history, size and age develop a similar degree distribution.

BOX 5.3

CONTINUUM THEORY

To calculate the degree distribution of the Barabási-Albert model in the continuum approximation we first calculate the number of nodes with degree smaller than k, i.e. $k_i(t) < k$. Using (5.7), we write

$$t_i < t\left(\frac{m}{k}\right)^{1/\beta}. \qquad (5.12)$$

In the model we add a node at equal time step (BOX 5.2). Therefore the number of nodes with degree smaller than k is

$$t\left(\frac{m}{k}\right)^{1/\beta}. \qquad (5.13)$$

Altogether there are $N=m_0+t$ nodes, which becomes $N \approx t$ in the large t limit. Therefore the probability that a randomly chosen node has degree k or smaller, which is the cumulative degree distribution, follows

$$P(k) = 1 - \left(\frac{m}{k}\right)^{1/\beta}. \qquad (5.14)$$

By taking the derivative of (5.14) we obtain the degree distribution

$$p_k = \frac{\partial P(k)}{\partial k} = \frac{1}{\beta}\frac{m^{1/\beta}}{k^{1/\beta+1}} = 2m^2 k^{-3},, \qquad (5.15)$$

which is (5.9).

THE ABSENCE OF GROWTH OR PREFERENTIAL ATTACHMENT

The coexistence of growth and preferential attachment in the Barabási-Albert model raises an important question: Are they both necessary for the emergence of the scale-free property? In other words, could we generate a scale-free network with only one of the two ingredients? To address these questions, next we discuss two limiting cases of the model, each containing only one of the two ingredients [1, 11].

MODEL A

To test the role of preferential attachment we keep the growing character of the network (ingredient A) and eliminate preferential attachment (ingredient B). Hence, *Model A* starts with m_0 nodes and evolves following these steps:

(A) Growth

At each time step we add a new node with $m(\leq m_0)$ links that connect to m nodes added earlier.

(B) Preferential Attachment

The probability that a new node links to a node with degree k_i is

$$\Pi(k_i) = \frac{1}{(m_0 + t - 1)}.$$
(5.16)

That is, $\Pi(k_i)$ is independent of k_i, indicating that new nodes choose randomly the nodes they link to.

The continuum theory predicts that for Model A $k_i(t)$ increases logarithmically with time

$$k_i(t) = m \ln\left(e \frac{m_0 + t - 1}{m_0 + t_i - 1} \right),$$
(5.17)

a much slower growth than the power law increase (5.7). Consequently the degree distribution follows an exponential (Figure 5.8a)

$$p(k) = \frac{e}{m} \exp\left(-\frac{k}{m}\right). \tag{5.18}$$

An exponential function decays much faster than a power law, hence it does not support hubs. Therefore the lack of preferential attachment eliminates the network's scale-free character and the hubs. Indeed, as all nodes acquire links with equal probabilty, we lack a rich-get-richer process and no clear winner can emerge.

MODEL B

To test the role of growth next we keep preferential attachment (ingredient B) and eliminate growth (ingredient A). Hence, *Model B* starts with *N* nodes and evolves following this step:

(B) Preferential Attachment

At each time step a node is selected randomly and connected to node *i* with degree k_i already present in the network, where *i* is chosen with probability $\Pi(k)$. As $\Pi(0)=0$ nodes with $k=0$ are assumed to have $k=1$, otherwise they can not acquire links.

In Model B the number of nodes remains constant during the network's evolution, while the number of links increases linearly with time. As a result for large *t* the degree of each node also increases linearly with time (Figure 5.7b, inset)

$$k_i(t) \approx \frac{2}{N} t. \tag{5.19}$$

Indeed, in each time step we add a new link, without changing the number of nodes.

At early times, when there are only a few links in the network (i.e. *L* ≪ *N*), each new link connects previously unconnected nodes. In this stage the model's evolution is indistinguishable from the Barabási-Albert model with *m*=1. Numerical simulations show that in this regime the model develops a degree distribution with a power-law tail (Figure 5.8b).

Yet, p_k is not stationary. Indeed, after a transient period the node degrees converge to the average degree (5.19) and the degree develops a peak (Figure 5.8b). For $t \to N(N\text{-}1)/2$ the network becomes a complete graph in which all nodes have degree $k_{max}=N\text{-}1$, hence $p_k = \delta(N\text{-}1)$.

In summary, the absence of preferential attachment leads to a growing network with a stationary but exponential degree distribution. In contrast the absence of growth leads to the loss of stationarity, forcing the network to converge to a complete graph. This failure of Models A and B to reproduce the empirically observed scale-free distribution indicates that growth and preferential attachment are simultaneously needed for the emergence of the scale-free property.
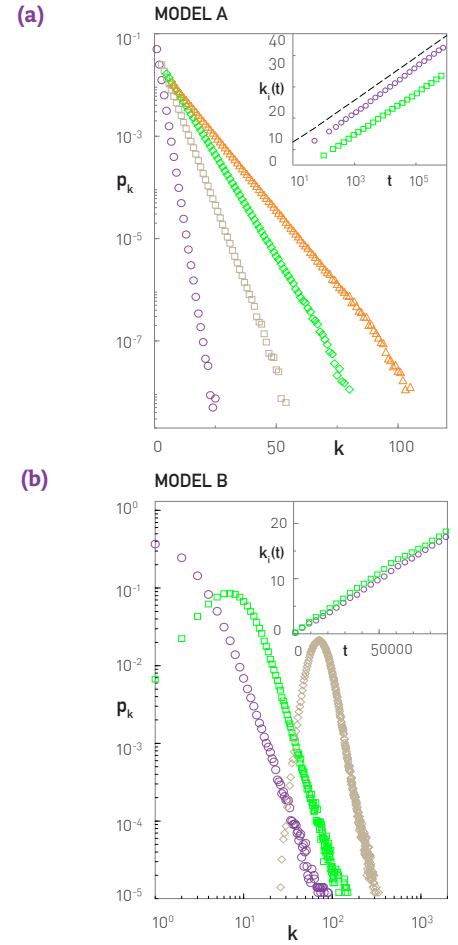

(a) MODEL A

(b) MODEL B

Figure 5.8
Model A and Model B

Numerical simulations probing the role of growth and preferential attachment.

(a) Model A
Degree distribution for Model A, that incorporates growth but lacks preferential attachment. The symbols correspond to $m_0=m=1$ (circles), 3 (squares), 5 diamonds), 7 (triangles) and *N*=800,000. The linear-log plot indicates that the resulting network has an exponential p_k, as predicted by (5.18).

Inset: Time evolution of the degree of two nodes added at $t_1=7$ and $t_2=97$ for $m_0=m=3$. The dashed line follows (5.17).

(b) Model B
Degree distribution for Model B, that lacks growth but incorporates preferential attachment, shown for *N*=10,000 and *t*=*N* (circles), *t*=5*N* (squares), and *t*=40*N* (diamonds). The changing shape of p_k indicates that the degree distribution is not stationary.

Inset: Time dependent degrees of two nodes (*N*=10,000), indicating that $k_i(t)$ grows linearly, as predicted by (5.19). After [11].

MEASURING PREFERENTIAL ATTACHMENT

In the previous section we showed that growth and preferential attachment are jointly responsible for the scale-free property. The presence of growth in real systems is obvious: All large networks have reached their current size by adding new nodes. But to convince ourselves that preferential attachment is also present in real networks, we need to detect it experimentally. In this section we show how to detect preferential attachment by measuring the $\Pi(k)$ function in real networks.

Preferential attachment relies on two distinct hypotheses:

Hypothesis 1
The likelihood to connect to a node depends on that node's degree k. This is in contrast with the random network model, for which $\Pi(k)$ is independent of k.

Hypothesis 2
The functional form of $\Pi(k)$ is linear in k.

Both hypotheses can be tested by measuring $\Pi(k)$. We can determine $\Pi(k)$ for systems for which we know the time at which each node joined the network, or we have at least two network maps collected at not too distant moments in time [14, 15].

Consider a network for which we have two different maps, the first taken at time t and the second at time $t + \Delta t$ (Figure 5.9a). For nodes that changed their degree during the Δt time frame we measure $\Delta k_i = k_i(t+\Delta t)-k_i(t)$. According to (5.1), the relative change $\Delta k_i/\Delta t$ should follow

$$\frac{\Delta k_i}{\Delta t} \sim \Pi(k_i),\qquad (5.20)$$

providing the functional form of preferential attachment. For (5.20) to be valid we must keep Δt small, so that the changes in Δk are modest. But Δt must not be too small so that there are still detectable differences between the two networks.

(a)

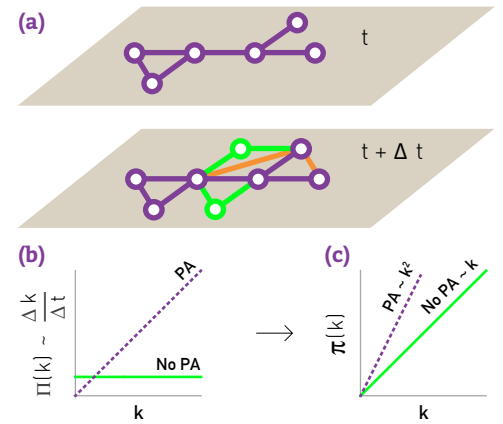**(b)** $\Pi(k) \sim \frac{\Delta k}{\Delta t}$

(c) $\pi(k)$

Figure 5.9
Detecting Preferential Attachment

(a) If we have access to two maps of the same network taken at time t and $t+\Delta t$, comparing them allows us to measure the $\Pi(k)$ function. Specifically, we look at nodes that have gained new links thanks to the arrival of the two new green nodes at $t+\Delta t$. The orange lines correspond to links that connect previously disconnected nodes, called *internal links*. Their role is discussed in CHAPTER 6.

(b) In the presence of preferential attachment $\Delta k/\Delta t$ will depend linearly on a node's degree at time t.

(c) The scaling of the cumulative preferential attachment function $\pi(k)$ helps us detect the presence or absence of preferential attachment (**Figure 5.10**).

In practice the obtained $\Delta k_i / \Delta t$ curve can be noisy. To reduce this noise we measure the *cumulative preferential attachment function*

$$\pi(k) = \sum_{k_i=0}^{k} \Pi(k_i). \tag{5.21}$$

In the absence of preferential attachment we have $\Pi(k_i)$=constant, hence, $\pi(k) \sim k$ according to (5.21). If linear preferential attachment is present, i.e. if $\Pi(k_i)=k_i$, we expect $\pi(k) \sim k^2$.

Figure 5.10 shows the measured $\pi(k)$ for four real networks. For each system we observe a faster than linear increase in $\pi(k)$, indicating the presence of preferential attachment. Figure 5.10 also suggests that $\Pi(k)$ can be approximated with

$$\Pi(k) \sim k^\alpha. \tag{5.22}$$

For the Internet and citation networks we have $\alpha \approx 1$, indicating that $\Pi(k)$ depends linearly on k, following (5.1). This is in line with Hypotheses 1 and 2. For the co-authorship and the actor network the best fit provides α=0.9±0.1 indicating the presence of a *sublinear preferential attachment*.

In summary, (5.20) allows us to detect the presence (or absence) of preferential attachment in real networks. The measurements show that the attachment probability depends on the node degree. We also find that while in some systems preferential attachment is linear, in others it can be sublinear. The implications of this non-linearity are discussed in the next section.
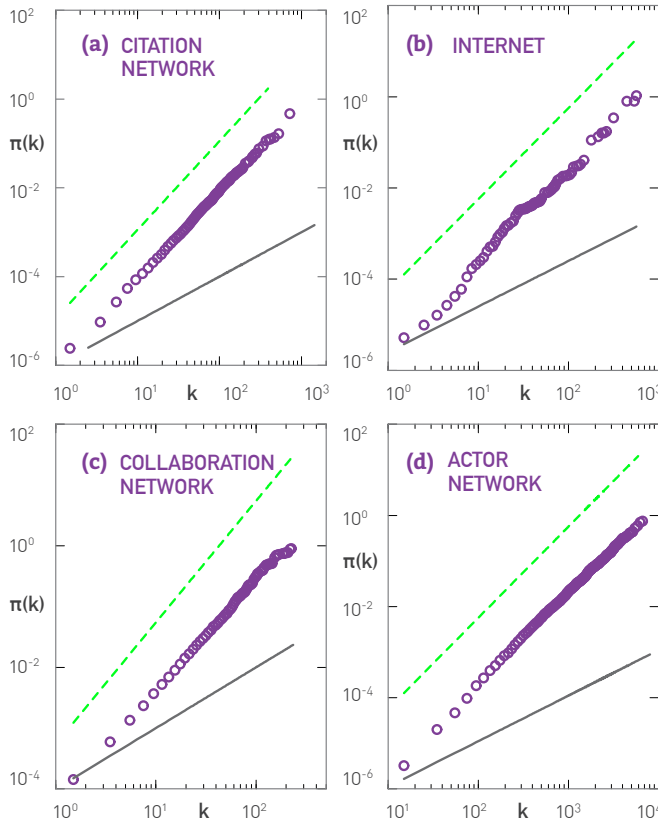
THE ORIGINS OF PREFERENTIAL ATTACHMENT

Given the key role preferential attachment plays in the evolution of real networks, we must ask, where does it come from? The question can be broken to two narrower issues:

Why does $\Pi(k)$ depend on $k?$
Why is the dependence of $\Pi(k)$ linear in $k?$

In the past decade we witnessed the emergence of two philosophically different answers to these questions. The first views preferential attachment as the interplay between random events and some structural property of a network. These mechanisms do not require global knowledge of the network but rely on random events, hence we will call them *local* or *random* mechanisms. The second assumes that each new node or link balances conflicting needs, hence they are preceeded by a cost-benefit analysis. These models assume familiarity with the whole network and rely on optimization principles, prompting us to call them *global* or *optimized* mechanisms. In this section we discuss both approaches.

LOCAL MECHANISMS

The Barabási-Albert model postulates the presence of preferential attachment. Yet, as we show below, we can build models that generate scale-free networks apparently without preferential attachment. They work by *generating* preferential attachment. Next we discuss two such models and derive $\Pi(k)$ for them, allowing us to understand the origins of preferential attachment.

Link Selection Model

The *link selection model* offers perhaps the simplest example of a local mechanism that generates a scale-free network without preferential attachment [16]. It is defined as follows (Figure 5.13):

• *Growth*: At each time step we add a new node to the network.

• *Link Selection*: We select a link at random and connect the new node to
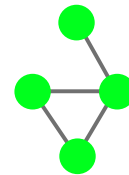
(a) NEW NODE
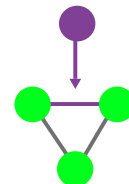
(b)

Figure 5.13
Link Selection Model

(a) The network grows by adding a new node, that selects randomly a link from the network (shown in purple).
(b) The new node connects with equal probability to one of the two nodes at the ends of the selected link. In this case the new node connected to the node at the right end of the selected link.

one of the two nodes at the two ends of the selected link.

The model requires no knowledge about the overall network topology, hence it is inherently local and random. Unlike the Barabási-Albert model, it lacks a built-in $\Pi(k)$ function. Yet, next we show that it generates preferential attachment.

We start by writing the probability q_k that the node at the end of a randomly chosen link has degree k as

$$q_k = Ckp_k .$$ (5.26)

Equation (5.26) captures two effects:

- The higher is the degree of a node, the higher is the chance that it is located at the end of the chosen link.

- The more degree-k nodes are in the network (i.e., the higher is p_k), the more likely that a degree k node is at the end of the link.

In (5.26) C can be calculated using the normalization condition $\Sigma q_k = 1$, obtaining $C=1/\langle k \rangle$. Hence the probability to find a degree-k node at the end of a randomly chosen link is

$$q_k = \frac{kp_k}{\langle k \rangle} .$$ (5.27)

Equation (5.27) is the probability that a new node connects to a node with degree k. The fact that the bias in (5.27) is linear in k indicates that the link selection model builds a scale-free network by generating linear preferential attachment.

Copying Model

While the link selection model offers the simplest mechanism for preferential attachment, it is neither the first nor the most popular in the class of models that rely on local mechanisms. That distinction goes to the *copying model* (Figure 5.14). The model mimics a simple phenomena: The authors of a new webpage tend to borrow links from other webpages on related topics [17, 18]. It is defined as follows:

In each time step a new node is added to the network. To decide where it connects we randomly select a node u, corresponding for example to a web document whose content is related to the content of the new node. Then we follow a two-step procedure (Figure 5.14):

(i) *Random Connection*: With probability p the new node links to u, which means that we link to the randomly selected web document.

(ii) *Copying*: With probability *1-p* we randomly choose an *outgoing link* of node u and link the new node to the link's target. In other words, the new webpage *copies* a link of node u and connects to its target, rather than connecting to node u directly.
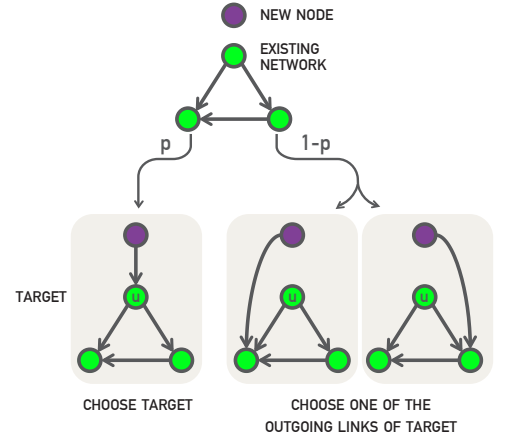

Figure 5.14
Copying Model

The main steps of the copying model. A new node connects with probability p to a randomly chosen target node u, or with probability *1-p* to one of the nodes the target u points to. In other words, with probabilty *1-p* the new node *copies* a link of its target u.

The probability of selecting a particular node in step (i) is *1/N*. Step (ii) is equivalent with selecting a node linked to a randomly selected link. The probability of selecting a degree-*k* node through this copying step (ii) is *k/2L* for undirected networks. Combining (i) and (ii), the likelihood that a new node connects to a degree-*k* node follows

$$\Pi(k) = \frac{p}{N} + \frac{1-p}{2L}k .$$

which, being linear in *k*, predicts a linear preferential attachment.

The popularity of the copying model lies in its relevance to real systems:

- *Social Networks*: The more acquaintances an individual has, the higher is the chance that she will be introduced to new individuals by her existing acquaintances. In other words, we "copy" the friends of our friends. Consequently without friends, it is difficult to make new friends.

- *Citation Networks*: No scientist can be familiar with all papers published on a certain topic. Authors decide what to read and cite by "copying" references from the papers they have read. Consequently papers with more citations are more likely to be studied and cited again.

- *Protein Interactions*: Gene duplication, responsible for the emergence of new genes in a cell, can be mapped into the copying model, explaining the scale-free nature of protein interaction networks [19, 20].

Taken together, we find that both the link selection model and the copying model generate a linear preferential attachment through random linking.

OPTIMIZATION

A longstanding assumption of economics is that humans make rational decisions, balancing cost against benefits. In other words, each individual aims to maximize its personal advantage. This is the starting point of rational choice theory in economics [21] and it is a hypothesis central to modern political science, sociology, and philosophy. As we show below, such rational decisions can lead to preferential attachment [22, 23, 24].

Consider the Internet, whose nodes are routers connected via cables. Establishing a new Internet connection between two routers requires us to lay down a new cable between them. As this is costly, each new link is preceded by a careful cost-benefit analysis. Each new router (node) will choose its link to balance access to good network performance (i.e. proper bandwith) with the cost of laying down a new cable (i.e. physical distance). This can be a conflicting desire, as the closest node may not offer the best network performance.
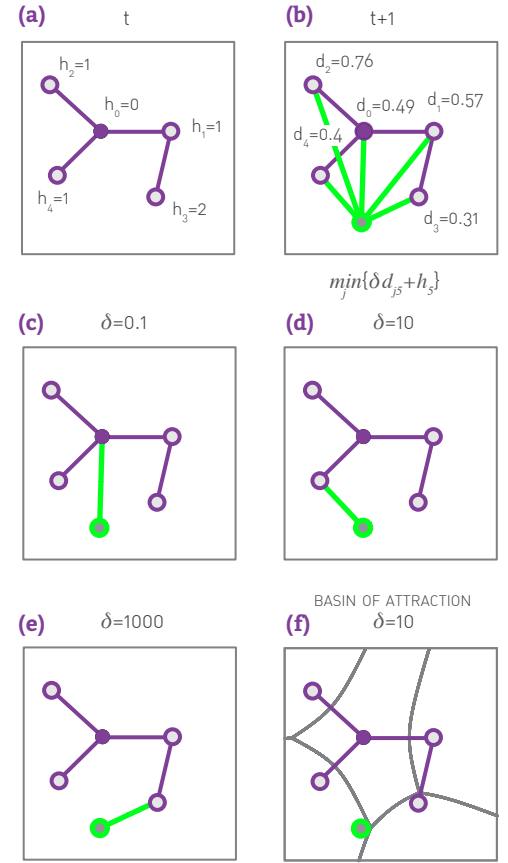

Figure 5.15

Optimization Model

(a) A small network, where the h_j term in the cost function (5.28) is shown for each node. Here h_j represents the network-based distance of node *j* from node *i*=0, designated as the "center" of the network, offering the best network performance. Hence h_0=0 and h_3=2.

(b) A new node (green) will choose the node *j* to which it connects by minimizing C_j of (5.28).

(c)-(e) If δ is small the new node will connect to the central node with h_j =0. As we increase δ, the balance in (5.28) shifts, forcing the new node to connect to more distant nodes. The panels (c)-(e) show the choice of the new green node makes for different values of δ.

(f) The basin of attraction for each node for δ=10. A new node arriving inside a basin will always link to the node at the center of the basin. The size of each basin depends on the degree of the node at its center. Indeed, the smaller is h_j, the larger can be the distance to the new node while still minimizing (5.28). Yet, the higher is the degree of node *j*, the smaller is its expected distance to the central node h_j.

For simplicity let us assume that all nodes are located on a continent with the shape of a unit square. At each time step we add a new node and randomly choose a point within the square as its physical location. When deciding where to connect the new node i, we calculate the cost function [22]

$$C_i = \min_j [\delta d_{ij} + h_j] \qquad (5.28)$$

which compares the cost of connecting to each node j already in the network. Here d_{ij} is the Euclidean distance between the new node i and the potential target j, and h_j is the network-based distance of node j to the *first node* of the network, which we designate as the desireable "center " of the network (Figure 5.15), offering the best network performance. Hence h_j captures the "resources" offered by node j, measured by its distance to the network's center.

The calculations indicate the emergence of three distinct network topologies, depending on the value of the parameter δ in (5.28) and N (Figure 5.15):

Star Network $\delta < (1/2)^{1/2}$
For $\delta = 0$ the Euclidean distances are irrelevant, hence each node links to the central node, turning the network into a star. We have a star configuration each time when the h_j term dominates over δd_{ij} in (5.28).

Random Network $\delta \geq N^{1/2}$
For very large δ the contribution provided by the distance term δd_{ij} overwhelms h_j in (5.28). In this case each new node connects to the node closest to it. The resulting network will have a bounded degree distribution, like a random network (Figure 5.16b).

Scale-free Network $4 \leq \delta \leq N^{1/2}$
Numerical simulations and analytical calculations indicate that for intermediate δ values the network develops a scale-free topology [22]. The origin of the power law distribution in this regime is rooted in two competing mechanisms:

 (i) *Optimization*: Each node has a basin of attraction, so that nodes landing in this basin will always link to it. The size of each basin correlates with h_j of node j at its center, which in turn correlates with the node's degree k_j (Figure 5.15f).

(ii) *Randomness*: We choose randomly the location of the new node, ending in one of the N basins of attraction. The node with the largest degree has largest basin of attraction, hence gains the most new nodes and links. This leads to preferential attachment, as documented in Figure 5.16d.

In summary, we can build models that do not have an explicit $\Pi(k)$ func-

tion built into their definition, yet they generate a scale-free network. As we showed in this section, these work by inducing preferential attachment. The mechanism responsible for preferential attachment can have two fundamentally different origins (Figure 5.17): it can be rooted in random processes, like link selection or copying, or in optimization, when new nodes balance conflicting criteria as they decide where to connect. Note that each of the mechanisms discussed above lead to linear preferential attachment, as assumed in the Barabási-Albert model. We are not aware of mechanisms capable of generating nonlinear preferential attachment, like those discussed in SECTION 5.7.

The diversity of the mechanisms discussed in this section suggest that linear preferential attachment is present in so many and so different systems precisely because it can come from both rational choice and random actions [25]. Most complex systems are driven by processes that have a bit of both. Hence luck or reason, preferential attachment wins either way.
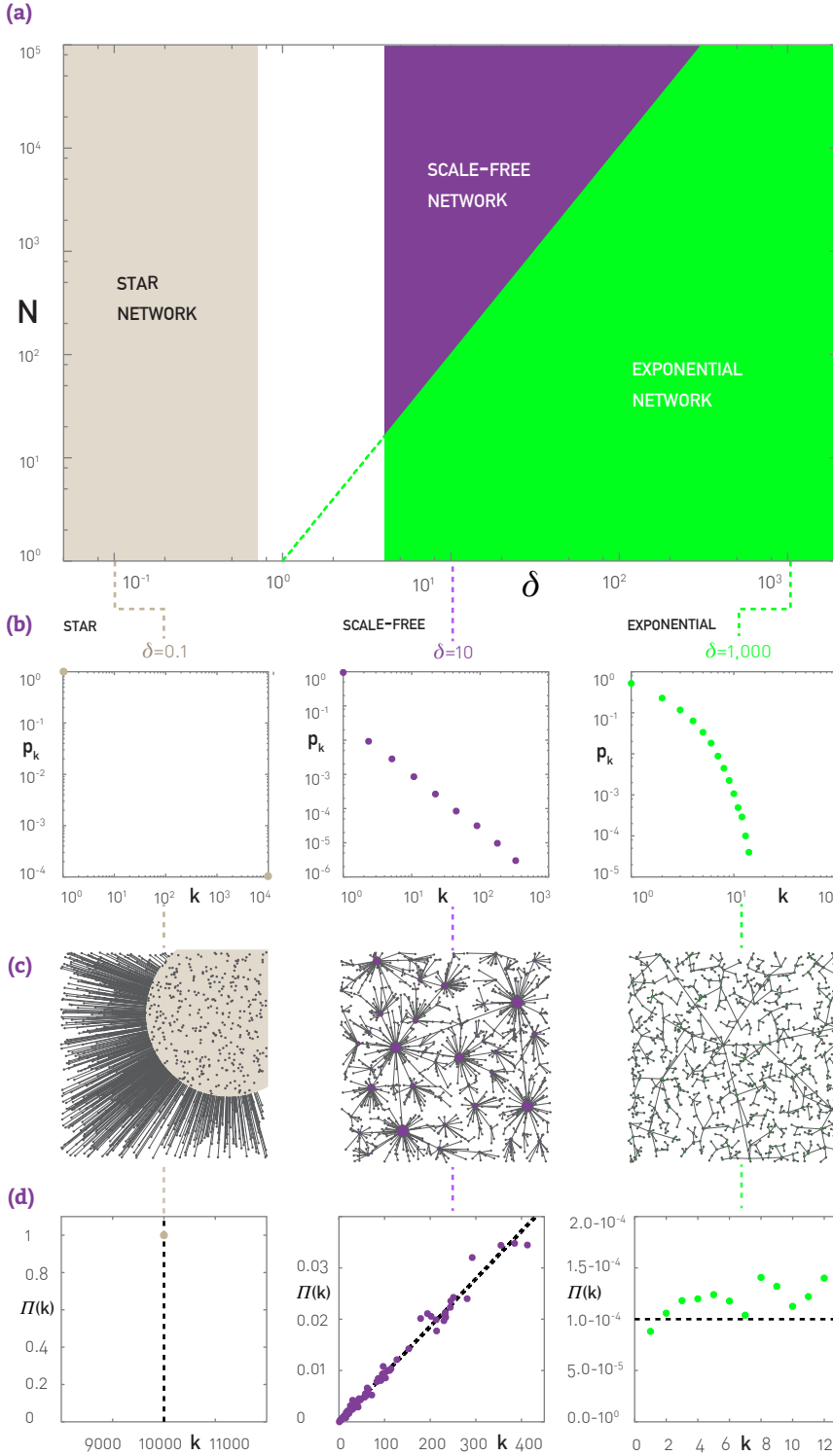
Figure 5.16
Scaling in the Optimization Model

(a) The three network classes generated by the optimization model: star, scale-free, and exponential networks. The topology of the network in the unmarked area is unknown.

The vertical boundary of the star configuration is at $\delta = (1/2)^{1/2}$. This is the inverse of the maximum distance between two nodes on a square lattice with unit length, over which the model is defined. Therefore if $\delta < (1/2)^{1/2}$, for any new node $\delta d_{ij} < 1$ and the cost (5.28) of connecting to the central node is $C_i = \delta d_{ij} + 0$, always lower than connecting to any other node at the cost of $f(i,j) = \delta d_{ij} + 1$. Therefore for $\delta < (1/2)^{1/2}$ all nodes connect to node 0, resulting in a network dominated by a single hub (star-and-spoke network (c)).

The oblique boundary of the scale-free regime is $\delta = N^{1/2}$. Indeed, if nodes are placed randomly on the unit square, then the typical distance between neighbors decreases as $N^{-1/2}$. Hence, if $d_{ij} \sim N^{-1/2}$ then $\delta d_{ij} \geq h_{ij}$ for most node pairs. Typically the path length to the central node h_j grows slower than N (in small-world networks $h_j \sim \log N$, in scale-free networks $h_j \sim \ln\ln N$). Therefore C_i is dominated by the δd_{ij} term and the smallest C_i is achieved by minimizing the distance-dependent term. Note that strictly speaking the transition only occurs in the $N \to \infty$ limit. In the white regime we lack an analytical form for the degree distribution.

(b) Degree distribution of networks generated in the three phases marked in (a) for $N = 10^4$.

(c) Typical topologies generated by the optimization model for selected δ values. Node size is proportional to its degree.

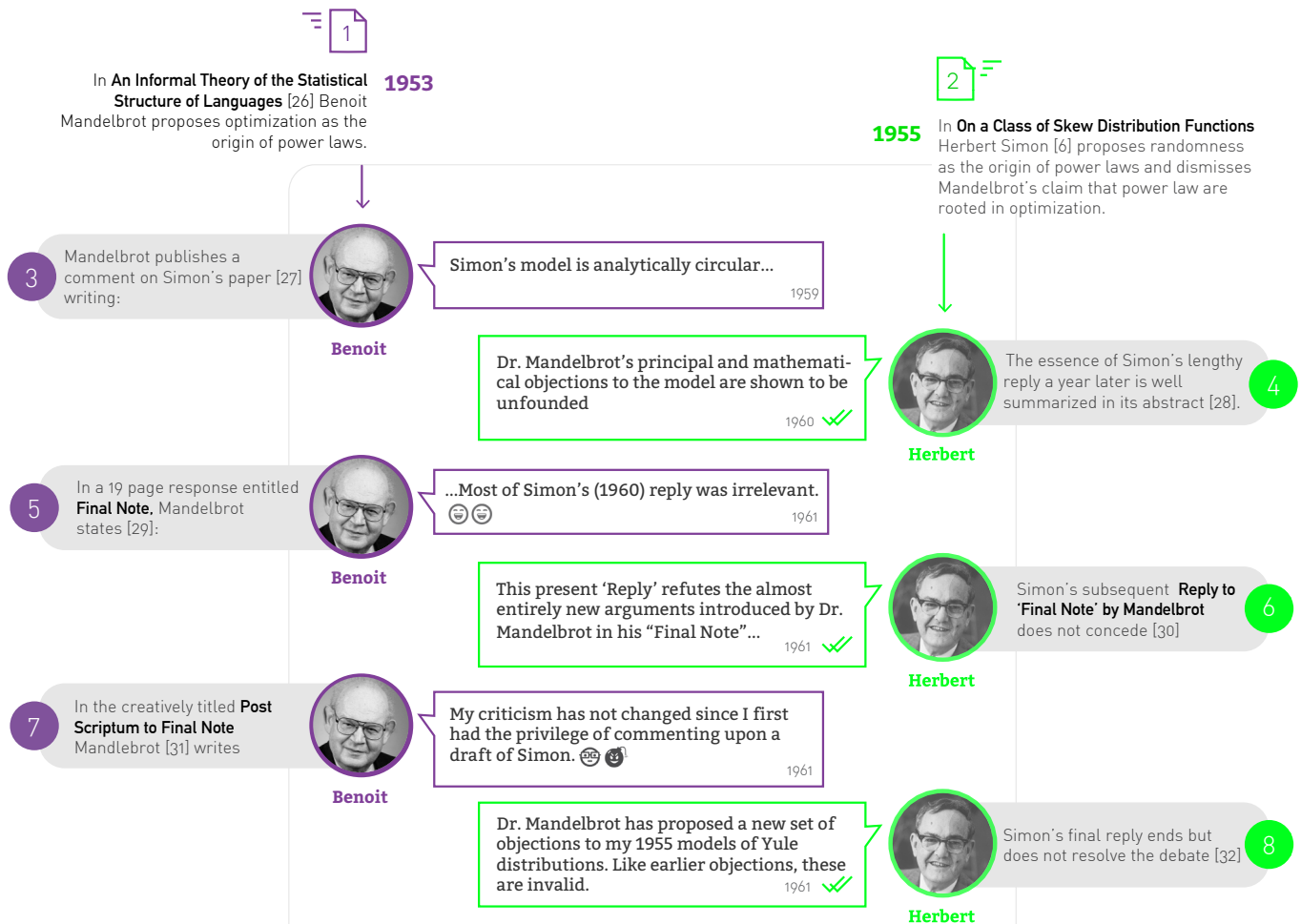**(d)** We used the method described in SECTION 5.6 to measure the preferential attachment function. Starting from a network with $N = 10,000$ nodes we added a new node and measured the degree of the node that it connected to. We repeated this procedure 10,000 times, obtaining $\Pi(k)$. The plots document the presence of linear preferential attachment in the scale-free phase, but its absence in the star and the exponential phases.

FIG. 5.17

LUCK OR REASON: AN ANCIENT FIGHT

The tension between randomness and optimization, two apparently antagonistic explanations for power laws, is by no means new: In the 1960s Herbert Simon and Benoit Mandelbrot have engaged in a fierce public dispute over this very topic. Simon proposed that preferential attachment is responsible for the power-law nature of word frequencies. Mandelbrot fiercely defended an optimization-based framework. The debate spanned seven papers and two years and is one of the most vicious scientific disagreement on record.

In the context of networks today the argument titled in Simon's favor: The power laws observed in complex networks appear to be driven by randomness and preferential attachment. Yet, the optimization-based ideas proposed by Mandelbrot play an important role in explaining the origins of preferential attachment. So at the end they were both right.

1

1953

In **An Informal Theory of the Statistical Structure of Languages** [26] Benoit Mandelbrot proposes optimization as the origin of power laws.

2

1955

In **On a Class of Skew Distribution Functions** Herbert Simon [6] proposes randomness as the origin of power laws and dismisses Mandelbrot's claim that power law are rooted in optimization.

3 Mandelbrot publishes a comment on Simon's paper [27] writing:

Simon's model is analytically circular...

1959

Benoit

Dr. Mandelbrot's principal and mathematical objections to the model are shown to be unfounded

1960

The essence of Simon's lengthy reply a year later is well summarized in its abstract [28]. **4**

Herbert

5 In a 19 page response entitled **Final Note**, Mandelbrot states [29]:

...Most of Simon's (1960) reply was irrelevant. 😑😑

1961

Benoit

This present 'Reply' refutes the almost entirely new arguments introduced by Dr. Mandelbrot in his "Final Note"...

1961

Simon's subsequent **Reply to 'Final Note'** by Mandelbrot does not concede [30] **6**

Herbert

7 In the creatively titled **Post Scriptum to Final Note** Mandlebrot [31] writes

My criticism has not changed since I first had the privilege of commenting upon a draft of Simon. 😎😼

1961

Benoit

Dr. Mandelbrot has proposed a new set of objections to my 1955 models of Yule distributions. Like earlier objections, these are invalid.

1961

Simon's final reply ends but does not resolve the debate [32] **8**

Herbert

DIAMETER AND CLUSTERING COEFFICIENT

To complete the characterization of the Barabási-Albert model we discuss the behavior of the network diameter and the clustering coefficient.

Diameter

The network diameter, representing the maximum distance in the Barabási-Albert network, follows for $m > 1$ and large N [33, 34]

$$\langle d \rangle \sim \frac{\ln N}{\ln \ln N} \, . \tag{5.29}$$

Therefore the diameter grows slower than ln N, making the distances in the Barabási-Albert model smaller than the distances observed in a random graph of similar size. The difference is particularly relevant for large N.

Note that while (5.29) is derived for the diameter, the average distance $\langle d \rangle$ scales in a similar fashion. Indeed, as we show in Figure 5.18, for small N the ln N term captures the scaling of $\langle d \rangle$ with N, but for large $N (\geq 10^4)$ the impact of the logarithmic correction ln ln N becomes noticeable.

Clustering coefficient

The clustering coefficient of the Barabási-Albert model follows (ADVANCED TOPICS 5.C) [35, 36]

$$\langle C \rangle \sim \frac{(\ln N)^2}{N} \, , \tag{5.30}$$

The prediction (5.30) is quite different from the $1/N$ dependence obtained for the random network model (Figure 5.19). The difference comes in the $(\ln N)^2$ term, that increases the clustering coefficient for large N. Consequently the Barabási-Albert network is locally more clustered than a random network.
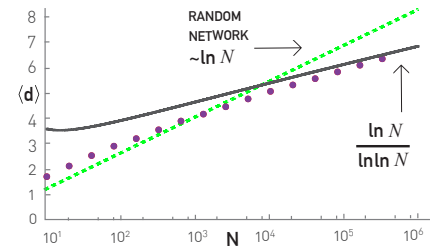

Figure 5.18
Average Distance

The dependence of the average distance on the system size in the Barabási-Albert model. The continuous line corresponds to the exact result (5.29), while the dotted line corresponds to the prediction (3.19) for a random network. The analytical predictions do not provide the exact perfactors, hence the lines are not fits, but indicate only the predicted N-dependent trends. The results were averaged for ten independent runs for $m = 2$.
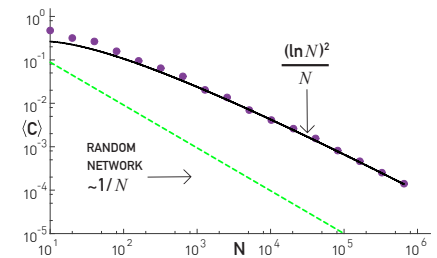

Figure 5.19
Clustering Coefficient

The dependence of the average clustering coefficient on the system size N for the Barabási-Albert model. The continuous line corresponds to the analytical prediction (5.30), while the dotted line corresponds to the prediction for a random network, for which $\langle C \rangle \sim 1/N$. The results are averaged for ten independent runs for $m = 2$. The dashed and continuous curves are not fits, but are drawn to indicate the predicted N dependent trends.

SUMMARY

The most important message of the Barabási-Albert model is that network structure and evolution are inseparable. Indeed, in the Erdős-Rényi, Watts-Strogatz, the configuration and the hidden parameter models the role of the modeler is to cleverly place the links between a *fixed number of nodes*. Returning to our earlier analogy, the networks generated by these models relate to real networks like a photo of a painting relates to the painting itself: It may look like the real one, but the process of generating a photo is drastically different from the process of painting the original painting. The aim of the Barabási-Albert model is to capture the processes that assemble a network in the first place. Hence, it aims to paint the painting again, coming as close as possible to the original brush strokes. Consequently, the modeling philosophy behind the model is simple: *to understand the topology of a complex system, we need to describe how it came into being.*

Random networks, the configuration and the hidden parameter models will continue to play an important role as we explore how certain network characteristics deviate from our expectations. Yet, if we want to explain the origin of a particular network property, we will have to use models that capture the system's genesis.

The Barabási-Albert model raises a fundamental question: Is the combination of growth and preferential attachment the real reason why networks are scale-free? We offered a *necessary* and *sufficient* argument to address this question. First, we showed that growth and preferential attachment are jointly needed to generate scale-free networks, hence if one of them is absent, either the scale-free property or stationarity is lost. Second, we showed that if they are both present, they do lead to scale-free networks. This argument leaves one possibility open, however: Do these two mechanisms explain the scale-free nature of *all* networks? Could there be some real networks that are scale-free thanks to some completely different mechanism? The answer is provided in SECTION 5.9, where we did encountered the link selection, the copying and the optimization models that do not have a preferential attachment function built into them, yet

they do lead to a scale-free network. We showed that they do so by generating a linear $\Pi(k)$. This finding underscores a more general pattern: To date all known models and real systems that are scale-free have been found to have preferential attachment. Hence the basic mechanisms of the Barabási-Albert model appear to capture the origin of their scale-free topology.

The Barabási-Albert model is unable to describe many characteristics of real systems:

- The model predicts $\gamma=3$ while the degree exponent of real networks varies between 2 and 5 (Table 4.2).

- Many networks, like the WWW or citation networks, are directed, while the model generates undirected networks.

- Many processes observed in networks, from linking to already existing nodes to the disappearance of links and nodes, are absent from the model.

- The model does not allow us to distinguish between nodes based on some intrinsic characteristics, like the novelty of a research paper or the utility of a webpage.

- While the Barabási-Albert model is occasionally used as a model of the Internet or the cell, in reality it is not designed to capture the details of any particular real network. It is a minimal, proof of principle model whose main purpose is to capture the basic mechanisms responsible for the emergence of the scale-free property. Therefore, if we want to understand the evolution of systems like the Internet, the cell or the WWW, we need to incorporate the important details that contribute to the time evolution of these systems, like the directed nature of the WWW, the possibility of internal links and node and link removal.

As we show in CHAPTER 6, these limitations can be systematically resolved.