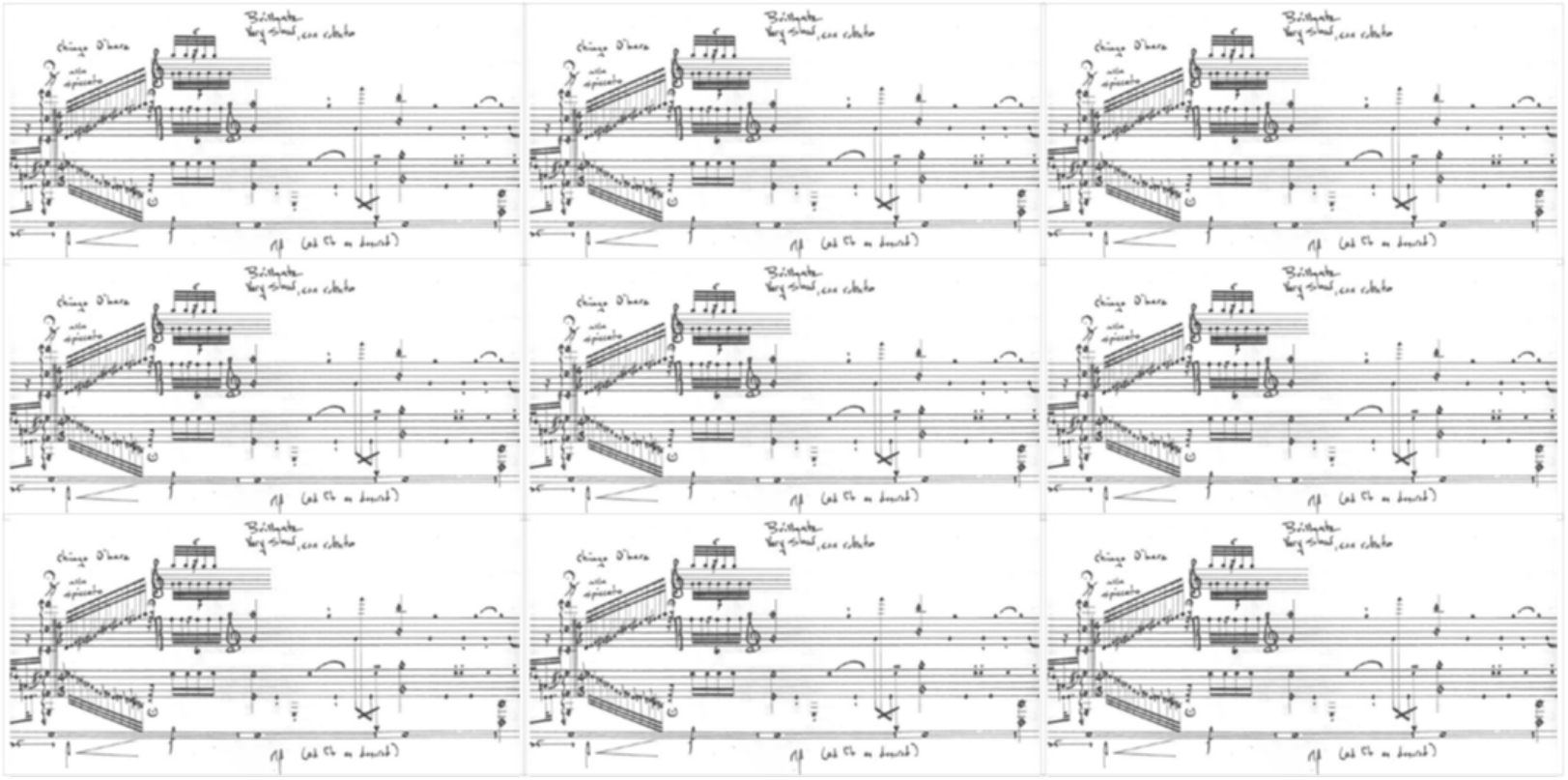
5

ALBERT-LÁSZLÓ BARABÁSI

NETWORK SCIENCE
THE BARABÁSI-ALBERT MODEL



ACKNOWLEDGEMENTS

MÁRTON PÓSFAI
GABRIELE MUSELLA
MAURO MARTINO
ROBERTA SINATRA
SARAH MORRISON
AMAL HUSSEINI
PHILIPP HOEVEL

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?

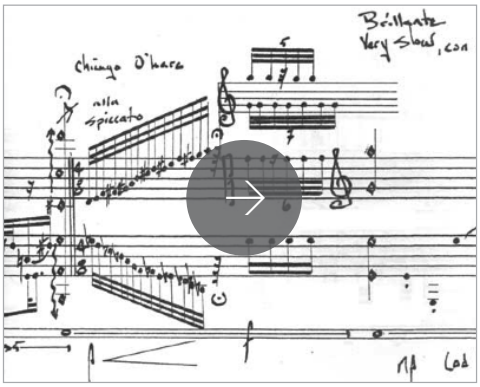
The first question is particularly puzzling given the fundamental differences in the nature, origin, and scope of the systems that display the scale-free property:

- 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.

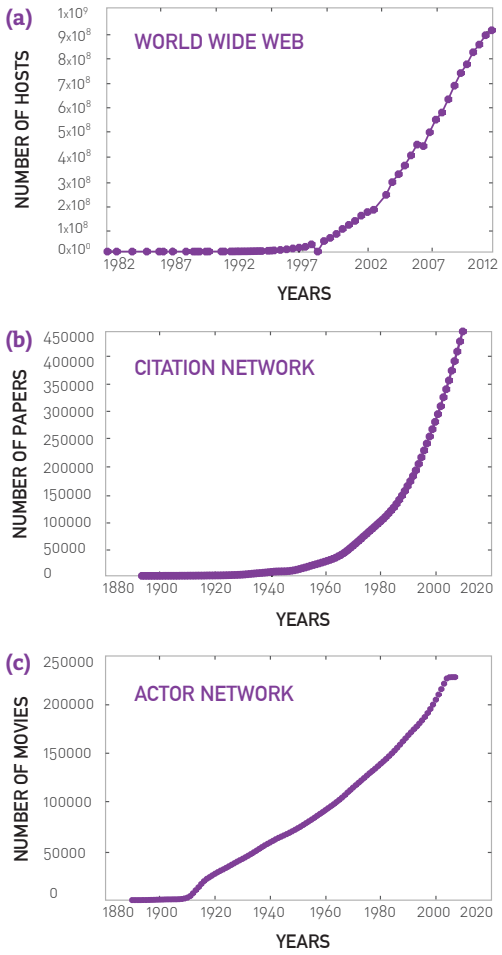


Figure 5.1

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 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.

PREFERENTIAL ATTACHMENT: A BRIEF HISTORY

György Pólya
PÓLYA PROCESS
MATHEMATICIAN

George Kinsley Zipf
WEALTH DISTRIBUTION
ECONOMIST

Herbert Alexander Simon
MASTER EQUATION
POLITICAL SCIENTIST

Robert Merton
MATTHEW EFFECT
SOCIOLOGIST

Albert-László Barabási & Réka Albert
PREFERENTIAL ATTACHMENT
NETWORK SCIENTISTS

George Udmy Yule
YULE PROCESS
STATISTICIAN

Robert Gibrat
PROPORTIONAL GROWTH
ECONOMIST

Derek de Solla Price
CUMULATIVE ADVANTAGE
PHYSICIST

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

Gospel of Matthew

MILESTONES

PUBLICATION DATE

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

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \qquad (5.1)$$

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 γ=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-

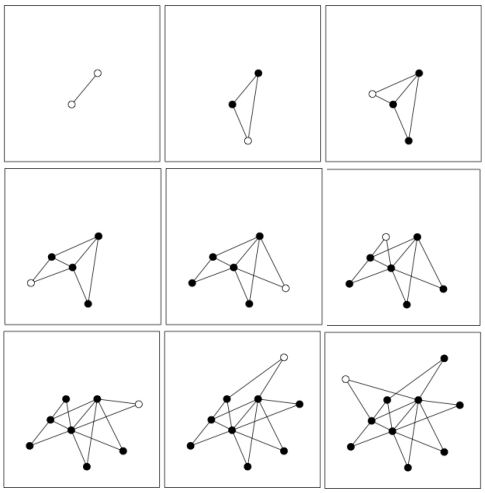
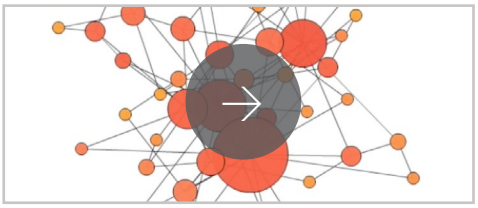


Figure 5.3
Evolution of the Barabási-Albert Model

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.

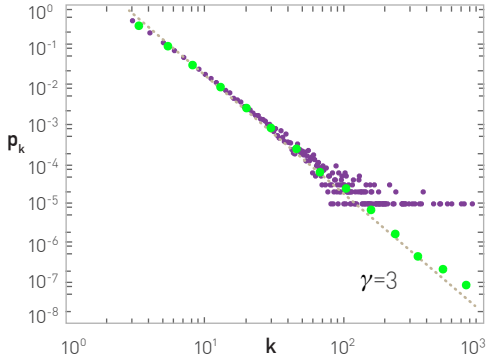


Figure 5.4

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 MATHEMATICAL DEFINITION OF THE BARABÁSI-ALBERT MODEL

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 $k_i/(2t\text{-}1)$, 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.

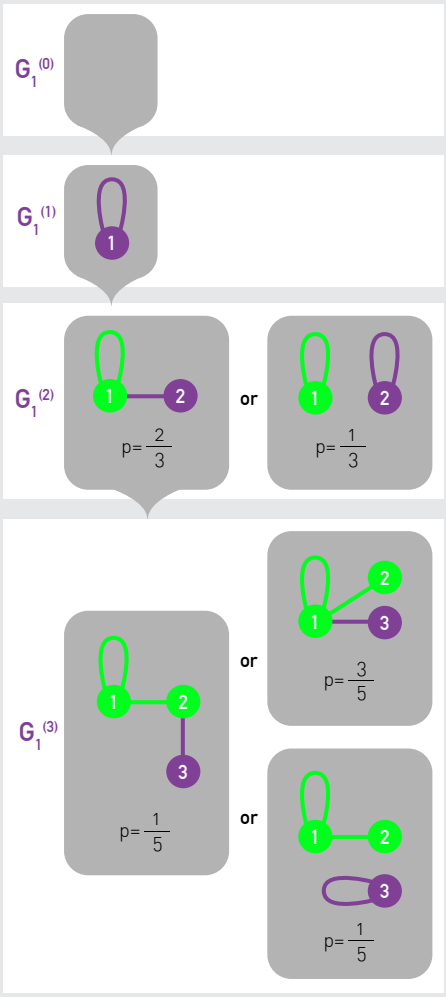


Figure 5.5
The Linearized Chord Diagram (LCD)

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.