THE SEVENTH LINK

Rich Get Richer

ONCE A PROMINENT MERCHANT PORT of the Portuguese empire, Porto to-
day gives the impression of a forgotten city. Built where the slow-moving
Duoro River wends its way to the Atlantic through the steep hills guard-
ing the seashore, it carries the signature of a busy medieval town strategi-
cally located on an easily defensible narrow key. With its magnificent cas-
tles overlooking the river and a rich history of wine making, one might
expect it to be one of the most visited cities in the world. But hidden as it
is in the northwest corner of the Iberian Peninsula, few tourists make the
detour. There are apparently too few fans of the distinctive full-bodied
Porto vintage to awaken this great medieval city from its dreamlike state.

I visited Porto in the summer of 1999, shortly after my students and
I finished our manuscript on the role of power laws on the Web. I was
attending a workshop on nonequilibrium and dynamical systems organ-
ized by two professors of physics at the University of Porto, José Mendes
and Maria Santos. During the summer of 1999 very few people were
thinking about networks, and there were no talks on the subject during
this workshop. But networks were very much on my mind. I could not
help carrying with me on the trip our unresolved questions: Why hubs?
Why power laws?

At that time the Web was the only network mathematically proven
to have hubs. Struggling to understand it, we were searching for its dis-
tinguishing features. At the same time, we wanted to learn more about

the structure of other real networks. Therefore, just before leaving for Porto, I had contacted Duncan Watts, who kindly provided us the data describing the power grid of the western United States and the C. *elegans* topology. Brett Tjaden, the former graduate student behind The Oracle of Bacon Website, now assistant professor of computer science at Ohio University in Athens, Ohio, sent us the Hollywood actor database. Jay Brockman, a computer science professor at Notre Dame, gave us data on a man-made network, the wiring diagram of a computer chip manufactured by IBM. Before I left for Europe, my graduate student Réka Albert and I agreed that she would analyze these networks. On June 14, a week after my departure, I received a long e-mail from her detailing some ongoing activities. At the end of the message there was a sentence added like an afterthought: "I looked at the degree distribution too, and in almost all systems (IBM, actors, power grid), the tail of the distribution follows a power law."

Réka's e-mail suddenly made it clear that the Web was by no means special. I found myself sitting in the conference hall paying no attention to the talks, thinking about the implications of this finding. If two networks as different as the Web and the Hollywood acting community both display power-law degree distribution, then some universal law or mechanism must be responsible. If such a law existed, it could potentially apply to all networks.

During the first break between talks I decided to withdraw to the quiet of the seminary where we were being housed. I did not get far, however. During the fifteen-minute walk back to my room a potential explanation occurred to me, one so simple and straightforward that I doubted it could be right. I immediately returned to the university to fax Réka, asking her to verify the idea using the computer. A few hours later she e-mailed me the answer. To my great astonishment, the idea worked. A simple, rich-get-richer phenomenon, potentially present in most networks, could explain the power laws we spotted on the Web and in Hollywood.

After Porto I returned briefly to Notre Dame before taking off for another month-long trip. It was clear, however, that we could not wait another month to submit our results. We had seven days to write a paper.

The eight-hour flight from Lisbon to New York seemed an ideal opportunity to prepare the first draft. As soon as the plane took off, I pulled out a laptop newly purchased before the Porto trip and frantically started typing. I was just about finished with the introduction when the flight attendant, handing a Coke to the passenger next to me, suddenly poured the entire contents of the glass onto my keyboard. Random letters flickered on the screen of my now useless laptop. But I did finish the paper on the plane, writing it out from beginning to end in longhand. A week later it was submitted to the prestigious journal *Science* only to be rejected after ten days without having undergone the usual peer review process because the editors believed that the paper did not meet the journal's standards of novelty and wide interest. By then I was in Transylvania, visiting my family and friends in the heart of the Carpathian Mountains. Disappointed but convinced that the paper was important, I did something that I had never done before: I called the editor who rejected the paper in a desperate attempt to change his mind. To my great surprise, I succeeded.

1.

The random model of Erdős and Rényi rests on two simple and often disregarded assumptions. First, we start with an inventory of nodes. Having all the nodes available from the beginning, we assume that the number of nodes is fixed and remains unchanged throughout the network's life. Second, all nodes are equivalent. Unable to distinguish between the nodes, we link them randomly to each other. These assumptions were unquestioned in over forty years of network research. But the discovery of hubs—and the power laws that describe them—forced us to abandon both assumptions. The manuscript submitted to *Science* was the first step along this path.

2.

There is one thing about the Web that everybody agrees on: It is growing. Each day new documents are added by individuals detailing their latest hobby or interest; by corporations expanding their online products

and services; by governments increasingly reliant on the Web to disseminate information to citizens; by college professors publishing their lecture notes; by nonprofit organizations trying to reach those who could benefit from their services; and by thousands of dot.com companies designing flashy pages to compete for your wallet. It is estimated that within ten years the Web will host about an exabyte (10^{18}) of information spread across the planet in numerous formats, most of which are presently unknown. While the rate of this explosion will likely taper as the majority of information collected by humanity lands online, so far there are no signs of a slowdown.

With over a billion documents available today, it is hard to believe the Web emerged one node at a time. But it did. Barely a decade ago it had only *one* node, Tim Berners-Lee's famous first Webpage. As physicists and computer scientists started creating pages of their own, the original site gradually gained links pointing to it. This modest Web of a dozen primitive documents was the precursor to the planet-sized self-assemblage the Web is today. Despite its overwhelming dimensions and complexity, it continues to grow incrementally, node by node. This expansion is in stark contrast to the assumption of the network models described so far in this book, which assume the number of nodes in a network is constant over time.

The Hollywood network also started with a tiny core, the actors of the first silent movies back in the 1890s. According to the IMDb.com database, Hollywood had only 53 actors in 1900. With increasing demand for motion pictures, this core slowly expanded, adding a few new faces with each movie. Hollywood experienced its first boom between 1908 and 1914, when the number of actors joining the trade went from under 50 to close to 2,000 a year. A second spectacular boom starting in the 1980s turned moviemaking into the entertainment megaindustry we know today. From a tiny cluster of silent actors grew a gigantic network of over a half-million nodes, and it continues to grow at an incredible rate. In the period of only one year, 1998, as many as 13,209 names of actors appearing for the first time on the wide canvas of the movie screen were added to the IMDb.com database.

Despite their diversity most real networks share an essential feature: growth. Pick any network you can think of and the following will likely be true: Starting with a few nodes, it grew incrementally through the addition of new nodes, gradually reaching its current size. Obviously, growth forces us to rethink our modeling assumptions. Both the Erdős-Rényi and Watts-Strogatz models assumed that we have a fixed number of nodes that are wired together in some clever way. The networks generated by these models are therefore *static*, meaning that the number of nodes remains unchanged during the network's life. In contrast, our examples suggested that for real networks the static hypothesis is not appropriate. Instead, we should incorporate growth into our network models. This was the initial insight we gained while trying to explain the hubs. In so doing, we ended up dethroning the first fundamental assumption of the random universe—its static character.

3.

It is relatively easy to model a growing network. We start from a tiny core and keep adding nodes, one after the other. Let us assume that each new node has two links. Thus, if we start with two nodes, our third node will link to both of them. The fourth node has three nodes from which to choose. How do we pick which two we should link to? For the sake of simplicity, let's follow the lead of Erdős and Rényi and randomly select two of the three nodes and link the new node to them. We can continue this process indefinitely, so that each time we add a new node, we connect it to two randomly selected nodes. The network generated by this simple algorithm, called *Model A*, differs from the random network model of Erdős and Rényi only in its growing nature. This difference, however, is significant. Despite the fact that we choose the links randomly and democratically, the nodes in Model A are not equivalent to each other. We have easily identifiable winners and losers. At each moment all nodes have an equal chance to be linked to, resulting in a clear advantage for the senior nodes. Indeed, apart from some rare statistical fluctuations, the first nodes in Model A will be the richest, since these nodes have had the longest time to collect links.

The poorest node will be the last one to join the system, with two links only, because nobody has had time to link to it yet. Model A was among our first attempts to explain the power laws we observed on the Web and in Hollywood. The computer simulations quickly convinced us that we had not yet found the answer. The degree distribution, the function that distinguishes scale-free networks from random models, decayed too fast, following an exponential. While the early nodes were clear winners, the exponential form predicted that they are too small and there are too few of them. Therefore, Model A failed to account for the hubs and the connectors. It demonstrated, however, that growth *alone* cannot explain the emergence of power laws.

4.

During the 1999 Super Bowl numerous neverheardof.com companies such as OurBeginning.com, WebEx.com, and Epidemic Marketing blew $2 million per advertising spot to bring their name to millions of Americans following the duel between Denver and St. Louis. In one year alone E*Trade spent $300 million promoting itself. AltaVista, one of the most popular search engines, had an advertising budget close to $100 million. America Online, the Goliath of the online world, effectively matched that with $75 million. In 1999 over $3.2 billion was spent on online marketing, about half the amount spent during the same period on cable television advertising, a medium whose history spans over two decades.

What did these companies want to achieve? The answer is simple, if unconventional. Startups and established companies alike had been burning venture capital and hard-earned cash, millions a day, to defeat the random universe of Erdős and Rényi. They knew that we do not link randomly on the Web. They wanted to take advantage of this non-randomness by begging us to link to them.

How do we in fact decide which Websites to link to on the World Wide Web? According to the random network models, we would randomly link to any of the nodes. A bit of reflection as to how we make our choices, however, indicates otherwise. For example, choices of Webpages

with links to news outlets abound. A quick search for "news" on Google returns about 109,000,000 hits. Yahoo's manually ordered directory offers a choice of over 8,000 online newspapers. How do we pick one? The random network models tell us that we select *randomly* from the list. Frankly, I do not think that anybody ever does that. Rather, most of us are familiar with a few major news outlets. Without giving the matter much thought we link to one of them. As a longtime reader of the *New York Times,* it is a no-brainer for me to choose nytimes.com. Others might prefer CNN.com or MSNBC.com. Significantly, however, the Webpages to which we prefer to link are not ordinary nodes. They are hubs. The better known they are, the more links point to them. The more links they attract, the easier it is to find them on the Web and so the more familiar we are with them. In the end we all follow an unconscious bias, linking with a higher probability to the nodes we know, which are inevitably the more connected nodes of the Web. We prefer hubs.

The bottom line is that when deciding where to link on the Web, we follow *preferential attachment:* When choosing between two pages, one with twice as many links as the other, about twice as many people link to the more connected page. While our individual choices are highly unpredictable, as a group we follow strict patterns.

Preferential attachment rules in Hollywood as well. The producer whose job it is to make a movie profitable knows that stars sell movies. Thus casting is determined by two competing factors: the match between the actor and the role, and the actor's popularity. Both introduce the same bias into the selection process. Actors with more links have a higher chance of getting new roles. Indeed, the more movies an actor has made, the more likely it is that he or she will appear again on the casting director's radar screen. This is where aspiring actors have a huge disadvantage, a Catch-22 everybody knows both in and out of Hollywood. You need to be known to get good roles, but you need good roles in order to be known.

The World Wide Web and Hollywood force us to abandon the second important assumption inherent in random networks—their democratic character. In the Erdős-Rényi and Watts-Strogatz models there is no difference between the nodes of a network; thus all nodes are equally likely to get links. The examples just discussed suggest other-

wise. In real networks linking is never random. Instead, popularity is at-tractive. Webpages with more links are more likely to be linked to again, highly connected actors are more often considered for new roles, highly cited papers are more likely to be cited again, connectors make more new friends. Network evolution is governed by the subtle yet unforgiving law of preferential attachment. Guided by it, we unconsciously add links at a higher rate to those nodes that are already heavily linked.

5.

Putting the pieces of the puzzle together, we find that real networks are governed by two laws: *growth* and *preferential attachment*. Each network starts from a small nucleus and expands with the addition of new nodes. Then these new nodes, when deciding where to link, prefer the nodes that have more links. These laws represent a significant departure from earlier models, which assumed a fixed number of nodes that are ran-domly connected to each other. But are they sufficient to explain the hubs and power laws encountered in real networks?

To answer this, in the 1999 *Science* paper we proposed a network model that incorporates both laws. The model is very simple, as growth and preferential attachment naturally lead to an algorithm defined by two straightforward rules (Figure 7.1):

A. *Growth:* For each given period of time we add a new node to the network. This step underscores the fact that networks are assem-bled one node at a time.

B. *Preferential attachment*: We assume that each new node connects to the existing nodes with two links. The probability that it will choose a given node is proportional to the number of links the chosen node has. That is, given the choice between two nodes, one with twice as many links as the other, it is twice as likely that the new node will connect to the more connected node.

Each time we repeat (a) and (b), we add a new node to the net-work. Therefore, node by node we generate a continuously expanding
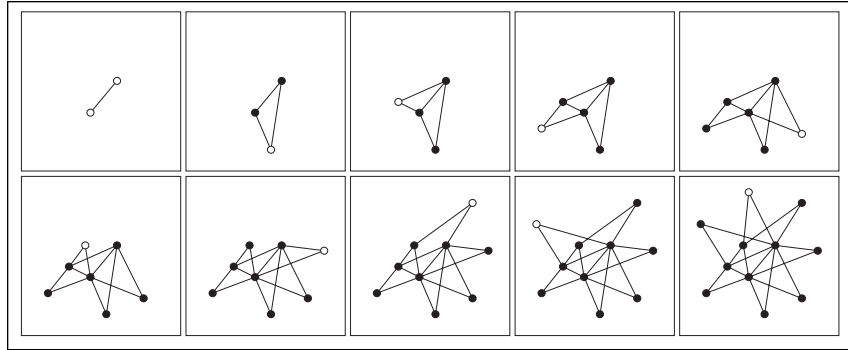
Figure 7.1   The Birth of a Scale-Free Network. *The scale-free topology is a natural consequence of the ever-expanding nature of real networks. Starting from two connected nodes (top left), in each panel a new node (shown as an empty circle) is added to the network. When deciding where to link, new nodes prefer to attach to the more connected nodes. Thanks to growth and preferential attachment, a few highly connected hubs emerge.*

web (Figure 7.1). This model, combining growth and preferential attachment, was our first successful attempt to explain the hubs. Réka's computer simulations soon indicated that it generated the elusive power laws. As the first model to explain the scale-free power laws seen in real networks, it quickly became known as the *scale-free model.*

6.

Why do hubs and power laws emerge in the scale-free model? First, growth plays an important role. The expansion of the network means that the early nodes have more time than the latecomers to acquire links: If a node is the last to arrive, no other node has the opportunity to link to it; if a node is the first in the network, all subsequent nodes have a chance to link to it. Thus growth offers a clear advantage to the senior nodes, making them the richest in links. Seniority, however, is not sufficient to explain the power laws. Hubs require the help of the second law, preferential attachment. Because new nodes prefer to link

to the more connected nodes, early nodes with more links will be selected more often and will grow faster than their younger and less connected peers. As more and more nodes arrive and keep picking the more connected nodes to link to, the first nodes will inevitably break away from the pack, acquiring a very large number of links. They will turn into hubs. Thus preferential attachment induces a *rich-get-richer* phenomenon that helps the more connected nodes grab a disproportionately large number of links at the expense of the latecomers.

This rich-get-richer phenomenon naturally leads to the power laws observed in real networks. Indeed, the computer simulations we performed indicated that the number of nodes with exactly *k* links follows a power law for any value of *k*. The precise value of the degree exponent, the parameter that characterizes the power law distribution, was no longer a mystery either. We were able to calculate it analytically, using a mathematical tool, called a *continuum theory,* that we developed for this purpose. Indeed, thanks to preferential attachment, each node attracts new links at a rate proportional to the number of its current links. Using this simple observation, we were able to propose a simple equation predicting how nodes acquire links as the network expands. The solution allowed us to calculate analytically the degree distribution, confirming that indeed it follows a power law.[1]

Could either growth or preferential attachment alone explain the power laws? Computer simulations and calculations convinced us that both are necessary to generate a scale-free network. A growing network without preferential attachment has an exponential degree distribution, which is similar to a bell curve in that it forbids the hubs. In the absence of growth we are back to the static models, unable to generate the power laws.

7.

Our purpose with the scale-free model was rather modest: to demonstrate that two simple laws of *growth* and *preferential attachment* could solve the puzzle of hubs and power laws. Therefore, the model's great

---

[1.] The degree exponent for the scale free model is $\gamma = 3$, i.e. the degree distribution follows $P(k) \sim k^{-3}$.

influence on subsequent research was a pleasant surprise for us, particularly since it was clear from the beginning that the topology of real networks was shaped by many effects that we had ignored for the purpose of simplicity and transparency. One of the most obvious of these is the fact that, whereas all links present in the scale-free model are added when new nodes join the network, in most networks new links can emerge spontaneously. For example, when I add to my Webpage a link pointing to nytimes.com, I create an *internal link* connecting two old nodes. In Hollywood, 94 percent of links are internal, formed when two established actors work together for the first time. Another feature absent from the scale-free model is that in many networks nodes and links can disappear. Indeed, many Webpages go out of business, taking with them thousands of links. Links can also be rewired, as when we decide to replace our link to CNN.com with a new one pointing to nytimes.com. These and other phenomena frequent in some networks but absent from the scale-free model illustrate that the evolution of real networks is far more complex than the scale-free model predicts. To understand networks in the complex world around us, we would have to incorporate these mechanisms into a consistent network theory and explain their impact on the network structure.

After submitting our paper on the scale-free model, Réka Albert and I started to investigate the effects of processes like internal links and rewiring on the structure of scale-free networks. We were no longer alone, however. A month after our paper's publication in *Science*, I learned of similar work going on in several research laboratories worldwide. Luis Amaral, my longtime collaborator, currently a research professor at Boston University, was in the process of generalizing the scale-free model to include aging, incorporating the possibility that actors stop acquiring links after retirement. Amaral, working together with Gene Stanley and two students, Antonio Scala and Mark Barthélémy, demonstrated that if nodes fail to acquire links after a certain age the size of the hubs will be limited, making large hubs less frequent than predicted by a power law. At the same time, José Mendes and Sergey Dorogovtsev were working independently on a similar problem in Porto; they soon published the first in a string of very influential papers on scale-free networks. Assuming that nodes slowly lose their ability to

attract links as they age, Mendes and Dorogovtsev showed that gradual aging does not destroy the power laws, but merely alters the number of hubs by changing the degree exponent. Paul Krapivsky and Sid Redner, also from Boston University, working with Francois Leyvraz from Mexico, generalized preferential attachment to account for the possibility that linking to a node would not be simply proportional to the number of links the node has but would follow some more complicated function. They found that such effects can destroy the power law characterizing the network.

These were the first of numerous subsequent results obtained by physicists, mathematicians, computer scientists, sociologists, and biologists who scrutinized the scale-free model and its various extensions. Thanks to their efforts, we currently have a rich and consistent theory of network growth and evolution, something that would have been unthinkable just a few years ago. We understand that internal links, rewiring, removal of nodes and links, aging, nonlinear effects, and many other processes affecting network topology can be seamlessly incorporated into an amazing theoretical construct of evolving networks, which contain as a particular case the scale-free model. These processes alter the way networks grow and evolve, inevitably changing the number and the size of the hubs. But in most cases when growth and preferential attachment are simultaneously present, hubs and power laws emerge as well. In complex networks a scale-free structure is not the exception but the norm, which explains its ubiquity in most real systems.

8.

The theory of evolving networks, developed in the past three years, represents a one-way sign in network modeling. By viewing networks as dynamical systems that change continuously over time, the scale-free model embodies a new modeling philosophy. The classic static models starting with Erdős-Rényi sought simply to arrange a fixed number of nodes and links such that the final web conforms to the network being modeled. This process is similar to drawing. Seated in front of a Ferrari, our task is to draw a picture that will allow anyone to recognize the car.

Having a faithful drawing, however, doesn't bring us any closer to understanding the processes that created the car in the first place. For that we need to know how to build one just like the original. This is exactly what the various evolving network models aim to accomplish. They capture how networks are assembled by reproducing the steps followed by nature when it created its various complex systems. If we correctly model the network assembly, our final network should closely match the reality. Thus our goals have shifted from describing the topology to understanding the mechanisms that shape network evolution.

This shift in focus resulted in a dramatic change in the language of networks, as well. The static nature of the classical models had gone unnoticed until we were forced to incorporate growth. Similarly, randomness had not been a problem until the power laws required us to introduce preferential attachment. Understanding that structure and network evolution couldn't be divorced from one another made it difficult to revert to the static models that dominated our thinking for decades. These shifts in thinking created a set of opposites: *static* versus *growing*, *random* versus *scale-free*, *structure* versus *evolution*.

At the end of the previous chapter we came to an important question: Does the presence of power laws imply that real networks are the result of a phase transition from disorder to order? The answer we've arrived at is simple: Networks are not en route from a random to an ordered state. Neither are they at the edge of randomness and chaos. Rather, the scale-free topology is evidence of organizing principles acting at each stage of the network formation process. There is little mystery here, since growth and preferential attachment can explain the basic features of the networks seen in nature. No matter how large and complex a network becomes, as long as preferential attachment and growth are present it will maintain its hub-dominated scale-free topology.

The scale-free model would have remained an interesting academic exercise if there hadn't been several subsequent discoveries. The most important was the realization that most complex networks of scientific and practical importance are scale-free. The Web data was large and detailed enough to convince us that power laws can describe real networks. This realization started an avalanche of discoveries that

continues to this day. As Hollywood, the metabolic network within the cell, citation networks, economic webs, and the network behind language[2] joined the list, suddenly the origins of scale-free topology became important for many scientific fields. The two laws governing network evolution built into the scale-free model offered a good starting point for exploring these diverse systems.

First, power laws gave legitimacy to the hubs. Then the scale-free model elevated the power laws seen in real networks to a mathematically backed conceptual advance. Supported by a sophisticated theory of evolving networks that allows us to precisely predict the scaling exponents and network dynamics, we have reached a new level of comprehension about our complex interconnected world, bringing us closer than ever to understanding the architecture of complexity.

But the scale-free model raised new questions. One in particular kept resurfacing: How do latecomers make it in a world in which only the rich get richer? The quest for the answer took us to a very unlikely place: the birth of quantum mechanics at the beginning of the twentieth century.

[2.] The scale-free nature of language has been shown by various research groups. In this network the nodes are words, and links represent significant cooccurences in texts, or semantic relationships (synonyms, antonyms).