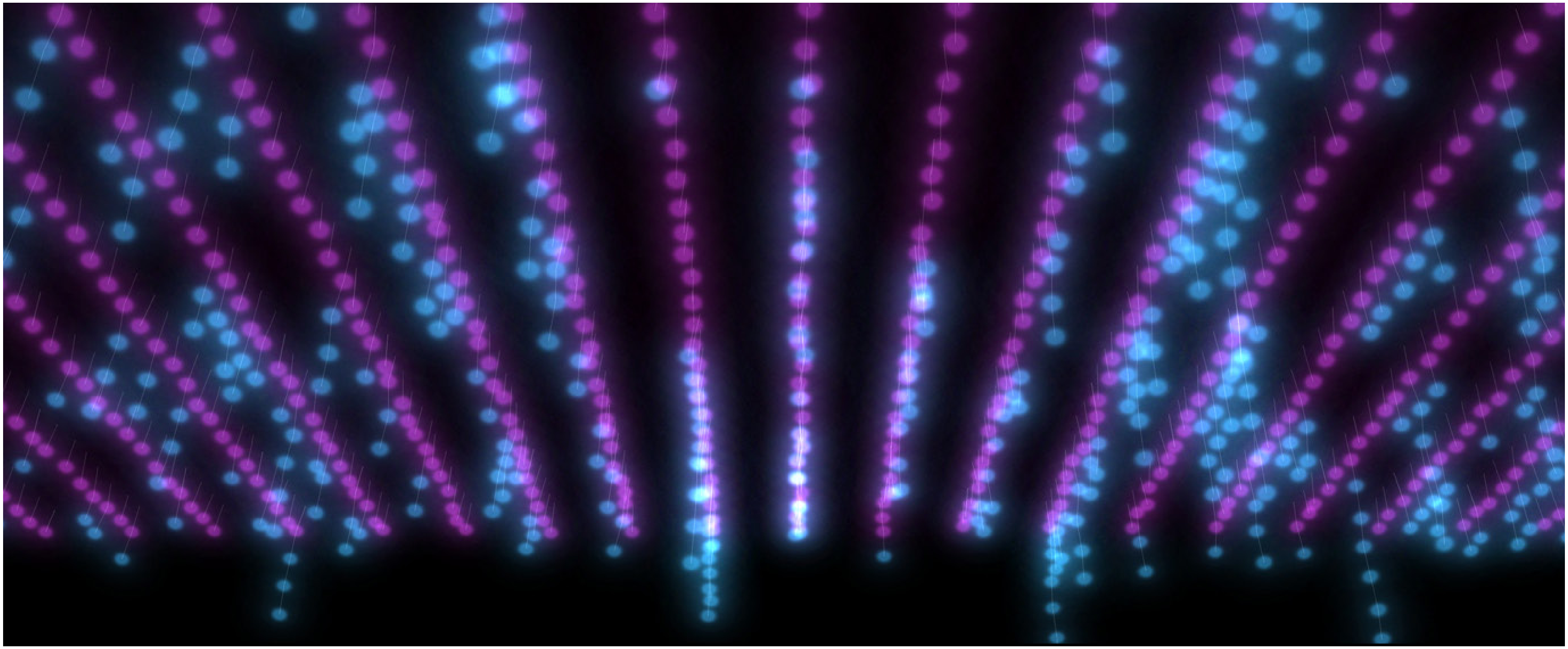
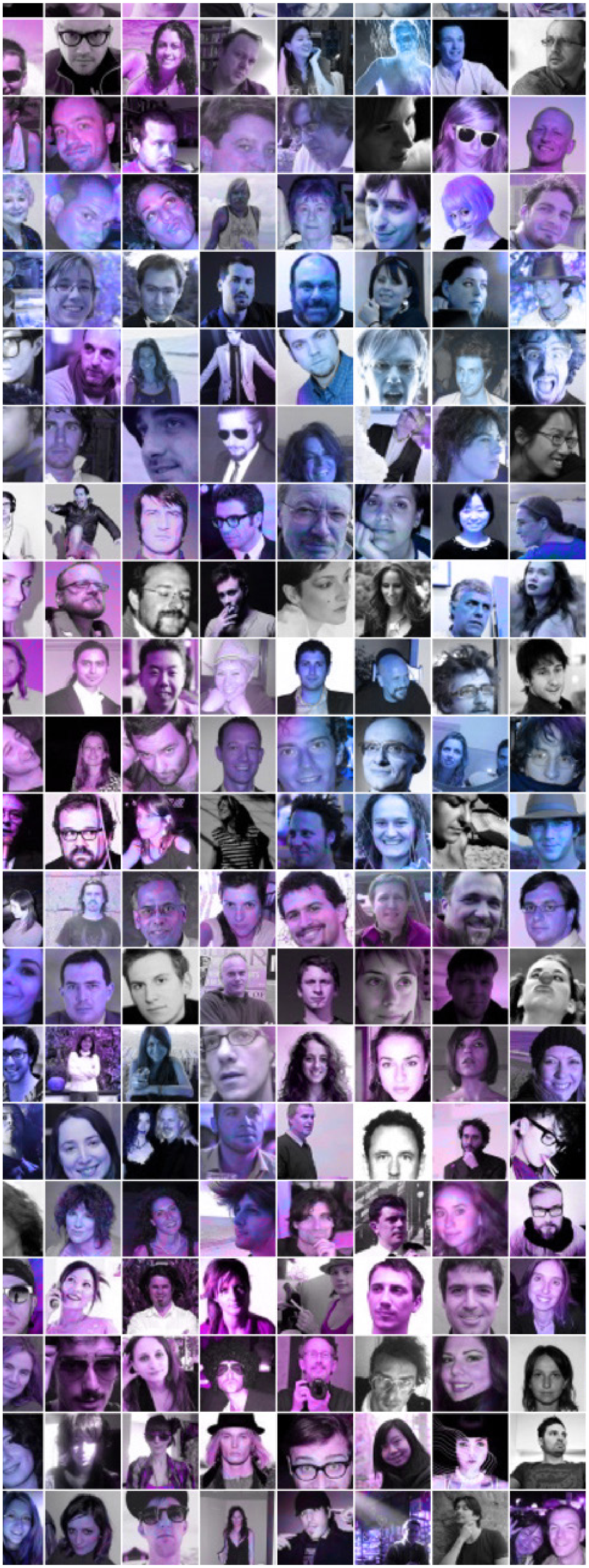


Network Science

Albert-László Barabási

Data Visualization by **Mauro Martino**
Data Analysis by **Márton Pósfai**

PDF Version:
November 2012



CHAPTER 1
INTRODUCTION

Introduction

From Saddam Hussein to network theory

Vulnerability due to interconnectivity

Networks at the heart of complex systems

Two forces helped the emergence of network science

The characteristics of network science

The impact of network science

Scientific impact

Summary

Bibliography

SECTION 1

INTRODUCTION

This book aims to help teach network science to an interdisciplinary audience. Many of the choices I made in presenting the material were guided by the desire to offer an enjoyable, yet systematic introduction to the field. I kept in mind that those entering the field are just as interested in learning about the genesis of the concepts network science introduced, as the tools they can use to study real networks and interpret the obtained results.

Several over-arching themes are present in this book, helping to offer an effective introduction:

(i) Given the empirical roots of network science, there is strong emphasis on empirical data. We have therefore assembled a set of 'canonic' databases, representing networks that are frequently analyzed in network science to test various network characteristics. Whenever possible, we use these datasets to illustrate the tools we introduce.

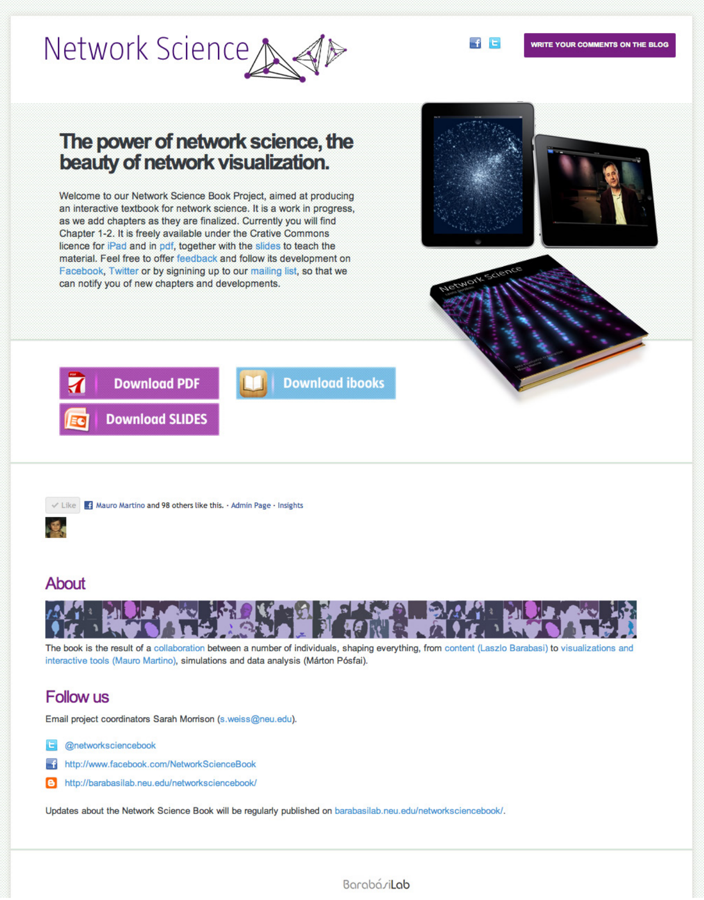
(ii) Given the potential diversity of the students interested in the field that may be familiar with one domain of inquiry but not other, we devote special sections to each dataset. The goal is to offer some degree of familiarity with the range of datasets explored in network science, and through this diversity to learn about the issues pertaining to data collection and curation.

This book is not a finished product but a work in progress. Hence we continue to update it, adding additional chapters as they are finished.

There is a dedicated website to this project (Image 1.1),

http://barabasilab.com/networksciencebook

that contains not only the chapters, but also the slides I used in my classes to teach the material. Those who are interested in teaching any part of the book are welcome to use these slides. The website also offers tools to provide feedback on the material, from comments to suggestions for improvement.



The power of network science, the beauty of network visualization.

Welcome to our Network Science Book Project, aimed at producing an interactive textbook for network science. It is a work in progress, as we add chapters as they are finalized. Currently you will find Chapter 1-2. It is freely available under the Crative Commons licence for iPad and in pdf, together with the slides to teach the material. Feel free to offer feedback and follow its development on Facebook, Twitter or by signining up to our mailing list, so that we can notify you of new chapters and developments.

Mauro Martino and 98 others like this.

About

The book is the result of a collaboration between a number of individuals, shaping everything, from content (Laszlo Barabasi) to visualizations and interactive tools (Mauro Martino), simulations and data analysis (Márton Pósfai).

Follow us

Email project coordinators Sarah Morrison (s.weiss@neu.edu).

@networksciencebook

http://www.facebook.com/NetworkScienceBook

http://barabasilab.neu.edu/networksciencebook/

Updates about the Network Science Book will be regularly published on barabasilab.neu.edu/networksciencebook/.

Image 1.1 http://barabasilab.com/networksciencebook

FROM SADDAM HUSSEIN TO NETWORK THEORY

American forces encountered relatively little military resistance as they took control of Iraq during the invasion that started in March 19, 2003. Yet, many of the regime's high ranking officials, including Saddam Hussein, avoided capture.
Hussein was last spotted kissing a baby in Baghdad some time in April 2003, and then his trace went cold. To aid awareness of the officials they sought, the coalition forces designed a deck of cards, each card engraved with the image of one of the 55 most wanted. It worked. By May 1st 15 men on the cards were captured and by the end of the month another 12 were under custody. Yet, the ace of spades (Image 1.2a), i.e. Hussein himself, remained at large.

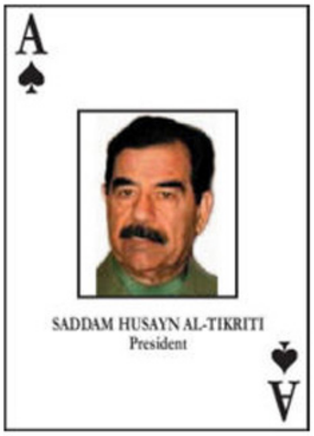


Image 1.2a
The network of Saddam Hussein.

Ace of Spades. One of the 55 cards the US military has handed out to the coalition forces in Iraq, each listing a top official to be captured following the country's 2003 invasion. The card shows the ace of spades, with the image of Saddam Hussein, Iraq's deposed president and dictator, the top prize of the hunt.

Intelligence officials hoped that some of the high ranking officials would surely know Hussein's whereabouts. Yet, it was not to be. This became painfully obvious after the capture of Saddam's trusted personal secretary and the ace of diamonds. Newspapers trumpeted his mid-June capture as the war's biggest feat, as this could lead to Saddam's whereabouts. Yet, the dictator parted ways with his ally soon after the invasion, sending a clear signal to the investigators: relying on the traditional lines of power was of little help in trying to find him. Instead, they decided to turn to a tool that had little presence in military thinking before: network theory [1].

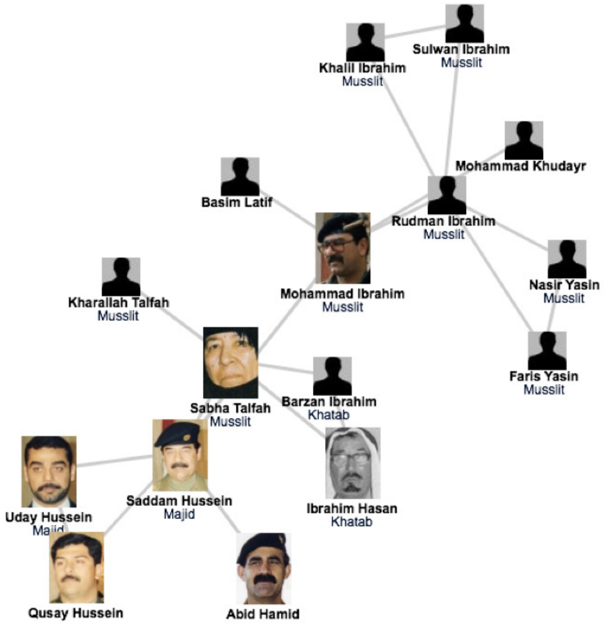


Image 1.2b
The network of Saddam Hussein.

The Social Network. A small region of the social network reconstructed by the US forces in the process of searching for Saddam Hussein. The map represents the relationship between individuals in Saddam's inner circle.

In 2003 network theory was an already burgeoning research field, but the soldiers in the war zone had little access to the exploding advances in this area. Instead, they arrived to it through a healthy dose of common sense and intuition. Col. James Hickey, in charge of a series of raids known as *Operation Desert Scorpion*, wanted to know the relationship between everyone killed or captured. The task fell to Lt. Col. Steve Russell, who was in direct charge of the raids, and Brian Reed, the operations officer under Hickey, who was exposed to social networks during his studies at West Point. Reed started to systematically reconstruct the social network of Saddam's inner circle. He did not rely on government documents and decrees, but rather gossip and family trees. As they meticulously pieced together an extensive diagram of who is related to whom in the Tikrit region, where Saddam was from, they started to use network diagrams to guide the raids. In one of those raids they found over $8 million in US currency, about $1 million in Iraqi currency, jewelry worth over $2 million, rifles, and ammunition. Yet, the biggest prize was Saddam's family photo album, providing the faces of those that the family trusted, filling with intimate details of their growing network diagram.

The maps consistently pointed to two individuals, Rudman Ibrahim and Mohammed Ibrahim (Image 1.2b). Not high in the government hierarchy, they were Saddam's second-level bodyguards, serving as his driver, cook, or mechanic. Yet, Rudman had a heart attack and died within a few hours of his capture, without having a chance to reveal his secrets. Next the investigators turned to their network diagram to identify individuals who could know the whereabouts of Mohammad, dubbed the fat man. He was not a major player in the regime's power structure, hence while Saddam's whereabouts were handled with fear, Mohammed's social ties were not as protected. Sure enough, once they found someone to turn Mohammad Ibrahim in, he revealed the spider hole that hid the dictator at a farm near the Tigris river. The capture of Saddam Hussein illustrates many issues that we will encounter as we delve into network theory:

- It shows the predictive power of networks, allowing even non experts to extract crucial information from them, as the soldiers did using Saddam's social network.

- It underlines the need for accurate maps of the networks we study, and the often heroic difficulties encountered during the mapping process.

- It demonstrates the remarkable stability of these networks: the capture of Hussein was not based on fresh intelligence, but rather on his pre-invasion social links, unearthed from old photos stacked in his family album.

- It shows that the choice of network we focus on makes a huge difference: it took months for the military to realize that the hierarchical network that described the official organization of the Iraqi government was of no use when it came to Saddam Hussein's whereabouts.

In many ways the network building exercise by the US military, deployed to capture Saddam Hussein, was a primitive one driven more by intuition and guesswork than hard science. The purpose of this book is to turn these insights into a robust theory and methodology, so that we can fully and repeatedly unleash their predictive power.

VULNERABILITY DUE TO INTERCONNECTIVITY

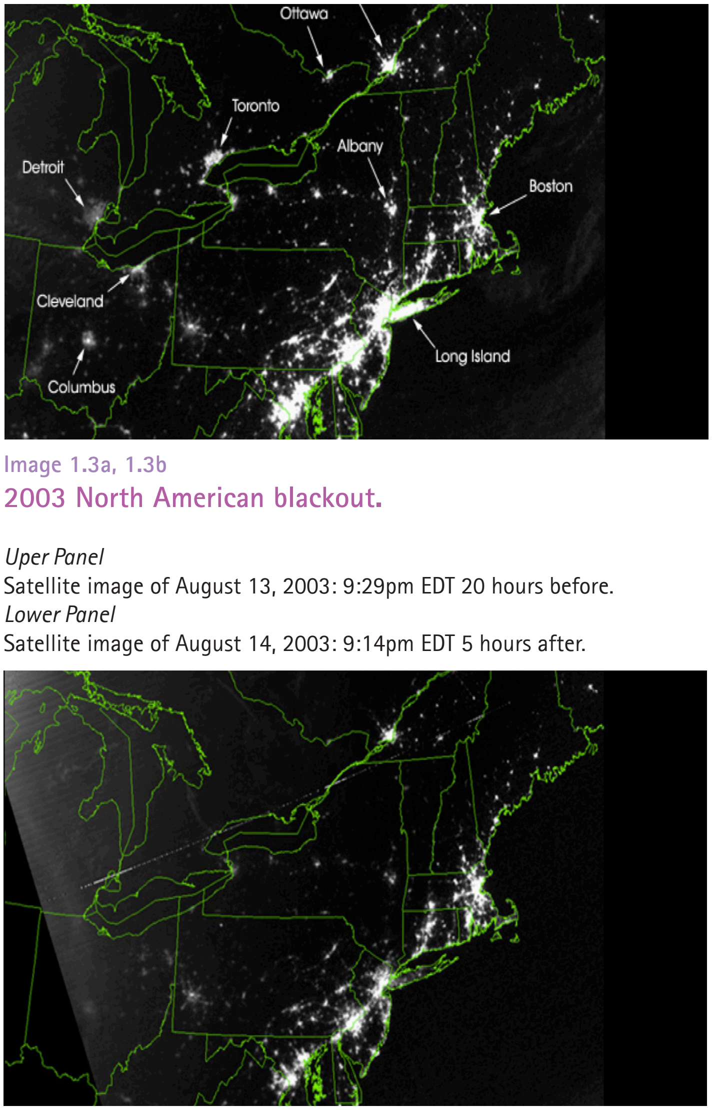


Image 1.3a, 1.3b

2003 North American blackout.

Uper Panel
Satellite image of August 13, 2003: 9:29pm EDT 20 hours before.
Lower Panel
Satellite image of August 14, 2003: 9:14pm EDT 5 hours after.

At a first look the two satellite maps of Image 1.3a/b are indistinguishable: lights shining brightly in highly populated areas, and dark spaces marking vast uninhabited forests and oceans. Yet, upon closer inspection something strange becomes apparent. The light in several regions, Toronto, Detroit, Cleveland, Columbus, Long Island have simply disappeared. This is not a doctored shot from the next Armageddon movie but represents a real image of the US Northeast on August 14, 2003, the night of a blackout that left an estimated 45 million people in eight US states and another 10 million in Ontario without power. It illustrates a much ignored aspect of networks, one that will be an important theme in this book: vulnerability due to interconnectivity.

The 2003 blackout is a typical example of a cascading failure. When a network acts as a transportation system, a local failure shifts loads to other nodes. If the extra load is negligible, the rest of the system can seamlessly absorb it, and the failure remains effectively unnoticed. If the extra load is too much for the neighboring nodes to carry, they will either tip or redistribute the load to their neighbors. Either way, we are faced with a cascading failure, the magnitude of which depends on the network position and capacity of the nodes that have been removed in the first and subsequent rounds. Case in point is electricity: as it cannot be stored, when a line goes down, its power must be shifted to other lines. Most of the time, the neighboring lines have no difficulty carrying the extra load. If they do, they will also tip and redistribute their increased load to their neighbors.

Cascading failures can occur in most complex systems. They take place on the Internet, when traffic is rerouted to bypass malfunctioning routers, occasionally creating denial of service attacks on routers that do not have the capacity to handle extra traffic. We witnessed one in 1997, when the International Monetary Fund pressured the central banks of several Pacific nations to limit their credit. There was a cascading failure behind the 2009-2011 financial meltdown, when the US credit crisis paralyzed the economy of the globe, leaving behind scores of failed banks, corporations, and even bankrupt states. Cascading failures are occasionally our ally, however. The world wide effort to dry up the money supply of terrorist organizations is aimed at crippling terrorist networks, and doctors and researchers hope to induce cascading failures to kill cancer cells.
The Northeast blackout illustrates an important theme of this book: we must understand how the network structure affects the robustness of a complex system. We will therefore develop quantitative tools to assess the interplay between network structure and dynamical processes on networks and their impact on failures. Although such failures may appear chaotic and unpredictable, we will learn that they follow rather reproducible laws that can be quantified and even predicted using the tools of network science.

SECTION 4

NETWORKS AT THE HEART OF COMPLEX SYSTEMS

> *"I think the next century will be the century of complexity."*
>
> Stephen Hawking

com.plex
[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]

1) composed of many interconnected parts; compound; composite: a complex highway system

2) characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery

3) so complicated or intricate as to be hard to understand or deal with: a complex problem

Source: Dictionary.com

Box 1.1

Image 1.4

The subtle networks behind the economy.

A credit card, selected as the 99th object in the popular exhibition by the British Museum, entitled The History of the World in 100 Objects. This card is a vivid demonstration of the interconnected nature of the modern economy, creating subtle linkages that one normally does not even think of. The card was issued in the United Arab Emirates in 2009 by the Hong Kong and Shanghai Banking Corporation, commonly known HSBC, a London based bank. The card functions through protocols provided by VISA, an USA based credit association. Yet, the card adheres to Islamic banking principles, which operates in accordance with Fiqhal-Muamalat (Islamic rules of transactions), most notably eliminating interest or riba. The card is not limited to muslims in the United Arab Emirates, but it is also offered to Muslim minorities in non-Muslim countries, and is used by many non-Muslims who agree with its strict ethical guidelines.

We are surrounded by systems that are hopelessly complicated, from the society, whose seamless functioning requires cooperation between billions of individuals, to communications infrastructures that integrate billions of cell phones with computers and satellites. Our ability to reason and comprehend the world around us is guaranteed by the coherent activity of billions of neurons in our brain. Our very existence is rooted in seamless interactions between thousands of genes and metabolites within our cells. These systems are collectively called complex systems. Given the important role they play in our life, in science and economy, the understanding, mathematical description, prediction, and eventually the control of such complex systems is one of the major intellectual and scientific challenges of the 21*st* century.

The emergence of network theory, at the dawn of the 21st century is a vivid demonstration that science can live up to this challenge. Indeed, *behind each complex system, there is an intricate network that encodes the interactions between the system's components*:

- The network describing the interactions between genes, proteins, and metabolites integrates the processes behind living cells.

- The wiring diagram capturing the connections between neural cells holds the key to our understanding of brain functions.

- The sum of all professional, friendship, and family ties is the fabric of the society.

- The network describing which communication devices interact with each other, capturing internet connections or wireless links, is the heart of the modern communication system.

- The power grid, a network of generators and transmission lines, supplies with energy virtually all modern technology.

- Trade networks maintain our ability to exchange goods and services, being responsible for the material prosperity that an increasing fraction of the world has enjoyed since WWII (Image 1.4). They also play a key role in the spread of financial and economic crises.

Networks are at the heart of some of the most revolutionary technologies of the 21st century, empowering everything from Google to Facebook, CISCO, and Twitter. At the end, networks permeate science, technology, and nature to a much higher degree than may be evident upon a casual inspection. Consequently, it is increasingly clear that *we will never understand complex systems unless we gain a deep understanding of the networks behind them.*

The scientific explosion that network science experienced during the first decade of the 21st century is rooted in the discovery that *despite the apparent differences, the emergence and evolution of different networks is driven by a common set of fundamental laws and reproducible mechanism.* Hence despite the amazing diversity in form, size, nature, age, and scope characterizing real networks, most networks observed in nature, society, and technology are driven by common organizing principles. In other words, once we disregard the nature of the components and their interactions, the obtained networks are more similar than different from each other. In the following sections, we discuss the forces that have led to the emergence of this new research field and its impact on science, technology, and society.

TWO FORCES HELPED THE EMERGENCE OF NETWORK SCIENCE

Why didn't network science emerge two hundred years earlier? The networks it explores are by no means new: metabolic networks date back to the origins of life, with a history of four billion years, and the Internet is over four decades old. Furthermore, many disciplines, from biochemistry to sociology, and brain science, have been dealing with their notion of networks. Graph theory, a prolific subfield of mathematics, has focused on networks since 1735. Why do we dare to call network science the science of the 21st century?

Something special happened at the dawn of the 21st century that transcended individual research fields and catalyzed the emergence of a new discipline (Image 1.5). To understand why this happened only now, and not two hundred years earlier, we need to discuss the forces that have contributed to the emergence of network science.

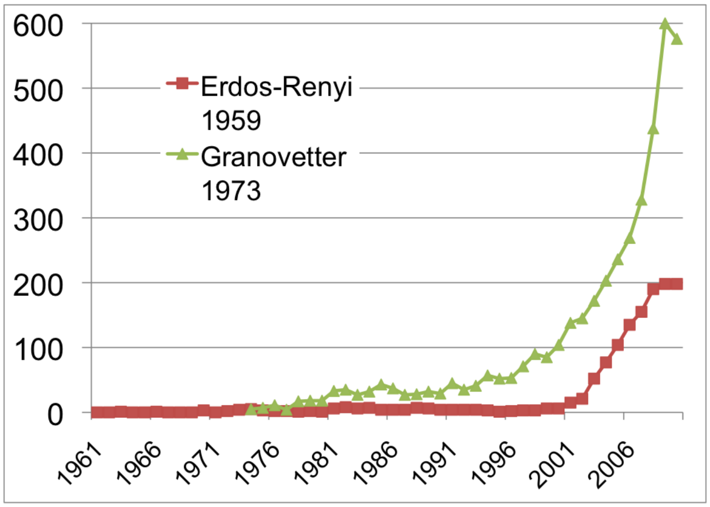


Image 1.5

The emergence of network science.

While the study of networks has a long history from graph theory to sociology, the modern chapter of network science emerged only during the first decade of the 21st century, following the publication of two seminal papers in 1998 [2] and 1999 [3]. The explosive interest in network science is well documented by the citation pattern of two classic network papers, the 1959 paper by Paul Erdős and Alfréd Rényi that marks the beginning of the study of random networks in graph theory [4] and the 1973 paper by Mark Granovetter, the most cited social network paper [5]. Both papers were hardly or only moderately cited before 2000. The explosive growth of citations to these papers in the 21st century documents the emergence of network science, drawing a new, interdisciplinary audience to these classic publications.

The emergence of network maps: To describe the behavior of a system consisting of hundreds to billions of interacting components, we first need a map of the system's wiring diagram. In a social system, this would require knowing the list of your friends, your friends' friends, and so on. In the WWW, this map tells us which webpages link to each other. In the cell, this corresponds to a detailed list of binding interactions and reactions that the genes, proteins, and metabolites participate in. In the past, we either lacked the tools to map these networks out, or it was difficult to keep track of the huge amount of data behind these maps. The emergence of the Internet, offering effective and fast data sharing methods, together with cheap digital storage, fundamentally changed this, allows us to collect, assemble, share, and analyze data pertaining to real networks.

While many of the canonical maps studied today in network science were not collected with the purpose of studying networks (Box 2), we witnessed an explosion of map making at the end of the 1990s. These offered detailed maps of the networks behind numerous complex system, from cell to the economy. Examples include the CAIDA or DIMES project aimed at obtaining an accurate map of the Internet [8]; the hundreds of millions of dollars spent by biologists to systematically map out protein-protein interactions in human cells [6], or the Connectome project of the US National Institute of Health that aims to trace the neural connection in mammalian brains [7].

The universality of network characteristics: It is easy to list the differences between the various networks we encounter in nature or society: the nodes of the metabolic network are tiny molecules and the links are chemical reactions governed by quantum mechanics; the nodes of the WWW are web documents and the links are URLs maintained by computer algorithms; the nodes of the social network are individuals, the links representing family, professionals, friendship, and acquaintance ties. The processes that shape these networks also differ greatly: metabolic networks are shaped by billions of years of evolution; WWW is collectively built by the actions of millions of individuals; social networks are shaped by social norms whose roots go back thousands of years. Given this diversity in size, nature, scope, history, and evolution, one would not be surprised if the networks behind these systems would differ greatly. Yet, a key discovery of network science is that the architecture and the evolution of networks emerging in various domains of science, nature, and technology are rather similar to each other, allowing us to use a common set of mathematical tools to explore these systems. This universality is one of the guiding principle of this book: we will not only seek to uncover specific network properties, but we will aim to understand its origins, encoding the laws that shape network evolution, as well as its consequences in understanding network behavior.

The origins of network maps

Many of the maps studied today by network scientists were not generated with the purpose of studying networks:

- *The list of chemical reactions that take place in a cell were discovered over a 150 year period by biochemists and biologists. In the 1990s they were collected in central databases, offering the first chance to assemble the networks behind a cell.*

- *The list of actors that play in each movie were traditionally scattered in books and encyclopedias. With the advent of the Internet, these disparate data were assembled into a central database by imdb.com, mainly to feed the curiosity of movie aficionados. The database offered the first chance for network scientists to explore the structure of the affiliation network behind Hollywood.*

- *The detailed list of authors of millions of research papers were traditionally scattered in the table of content of thousands of journals, but recently the Web of Science, Google Scholar, and other sites assembled them into comprehensive databases, easing the search for scientific information.*

In the hands of network scientists these databases turned into the first science collaboration maps. Hence, much of the early history of network science relied on the investigators' ingenuity to recognize and extract the networks from existing datasets. Network science changed that: today well-funded research collaborations focus on map making from biology to the Internet.

Box 1.2

SECTION 6

THE CHARACTERISTICS OF NETWORK SCIENCE

Network science is distinguished, not only by its subject matter, but also by its methodology. In the following we briefly discuss the key characteristics of the approach network science adopted to understand complex systems, helping us better understand the domain we are about to embark on.

Interdisciplinary nature: Network science offers a language through which different disciplines can seamlessly interact with each other. Indeed, cell biologists and computer scientists alike are faced with the task of characterizing the wiring diagram behind their system, extracting information from incomplete and noisy datasets, and the need to understand their systems' robustness to failures or deliberate attacks. To be sure, each discipline brings along a different set of technical details and challenges, which are important on their own. Yet, the common character of the many issues various fields struggle with have led to a cross-disciplinary fertilization of tools and ideas. For example, the concept of betweenness centrality that emerged in the social network literature in the 1970s, today plays a key role in identifying high traffic nodes on the Internet; algorithms developed by computer scientists for graph partitioning have found novel applications in cell biology.

Empirical, data driven nature: The tools of network science have their roots in graph theory, a fertile field of mathematics. What distinguishes network science from graph theory is its empirical nature, i.e. its focus on data and utility. As we will see in the coming chapters, we will never be satisfied with developing the abstract mathematical tools to describe a certain network property. Each tool we develop will be tested on real data and its value will be judged by the insights it offers about a system's structure or evolution.

Quantitative and mathematical nature: To contribute to the development of network science, it is essential to master the mathematical tools behind it. The tools of network science borrowed the formalism to deal with graphs from graph theory and the conceptual framework to deal with randomness and seek universal organizing principles from statistical physics. Lately, the field is benefiting from concepts borrowed from engineering, control and information theory, statistics and data mining, helping us extract information from incomplete and noisy datasets.

Computational nature: Finally, given the size of many of the networks we explore, and the exceptional amount of data behind them, network science offers a series of formidable computational challenges. Hence, the field has a strong computational character, actively borrowing from algorithms, database management and data mining. A series of software tools help practitioners with diverse computational skills analyze networks.

THE IMPACT OF NETWORK SCIENCE

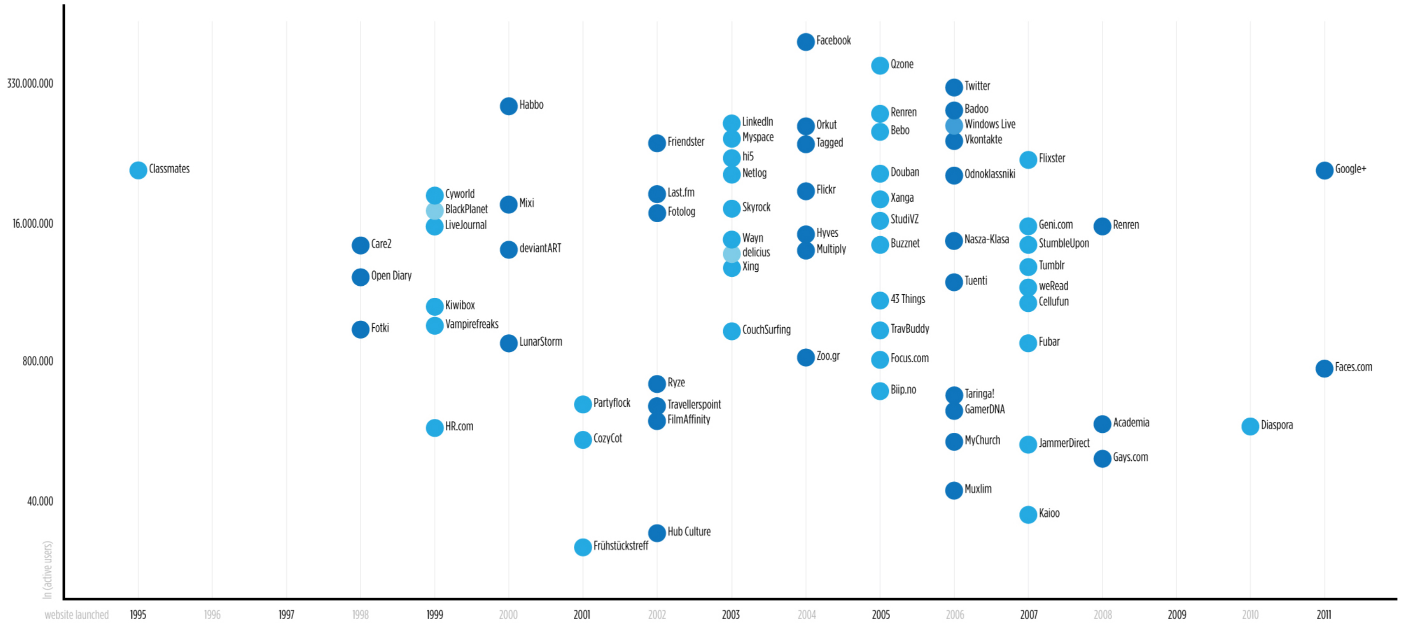


Image 1.6

The rise of social networking.

The popularity of the best known social networks, in terms of the number of users they attracted by the end of 2011 (vertical axis) shown as a function of their founding year (horizontal axis).

The impact of a new research field is measured both by its intellectual achievements as well as by the reach and the potential of its applications. While network science is a young field, its impact is everywhere around us, as we discuss below.

Economic Impact: From web search to social networking.

Some of the most successful companies of the 21st century, from *Google* to *Facebook*, from *Cisco* to *Apple* and *Akamai*, base their technology and business model on networks. Indeed, Google is not only the biggest network mapping operation, building a comprehensive map of the WWW, but its search technology relies on the network characteristics of the Web. Networks have gained particular popularity with the emergence of Facebook, the company with the oft-emphasized ambition to map out the social network of the whole planet. While Facebook was not the first social networking site, it is likely also not the last: an extensive ecosystem of social networking tools, from *Twitter* to *Orkut*, are attracting an impressive number of users (Image 1.6). The tools developed by network science fuel these sites, aiding everything from friend recommendation to advertising.

Health: From drug design to metabolic engineering.

The human genome project, completed in 2001, offered the first comprehensive list of all human genes [9, 10]. Yet, to fully understand how our cells function, and the origin of disease, we need accurate maps that tell us how these genes and other cellular components interact with each other. Most cellular processes, from the processing of food by our cells to sensing changes in the environment, rely on molecular networks. The breakdown of these networks is responsible for most human diseases. This has led to the emergence of network biology, a new subfield of biology that aims to understand the behavior of cellular networks. A parallel movement within medicine, called network medicine, aims to uncover the role of networks in human disease (Image 1.7a/b). Networks are particularly important in drug development. The ultimate goal of network pharmacology is to develop drugs that can cure diseases without significant side effects. This goal is pursued at many levels, from millions of dollars invested to map out cellular networks to the development of tools and databases to store, curate, and analyze patient and genetic data. Several new companies take advantage of these opportunities, from *GeneGo* that aims to collect accurate maps of cellular interactions from scientific literature to *Genomatica* that uses the predictive power behind metabolic networks to identify drug targets in bacteria and humans. Recently most major pharmaceutical companies have made signifi-

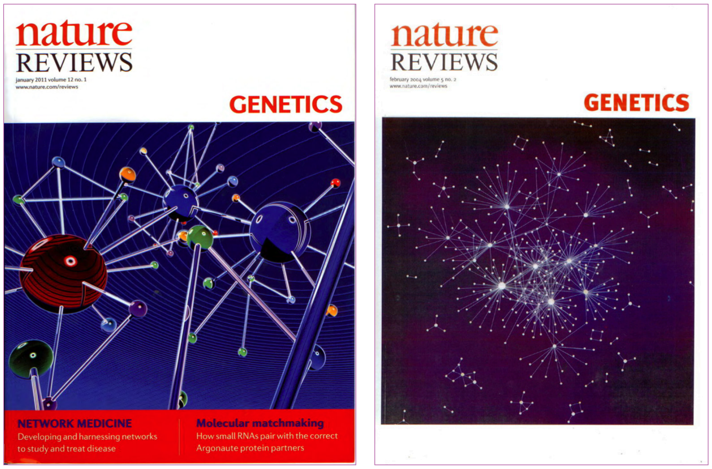

nature
REVIEWS
GENETICS
NETWORK MEDICINE
Developing and harnessing networks to study and treat disease
Molecular matchmaking
How small RNAs pair with the correct Argonaute protein partners
nature
REVIEWS
GENETICS

Image 1.7a, 1.7b

Networks in biology and medicine.

a) The cover of two issues of *Nature Reviews Genetics*, the top review journal in genetics. The cover from 2004, focuses on network biology [11], the cover from 2011 discuses network medicine [12].

b) The prominent role networks play in both cell biology and medical research is illustrated by the fact that the 2004 article on network biology is the second most cited article in the history of Nature Reviews Genetics.

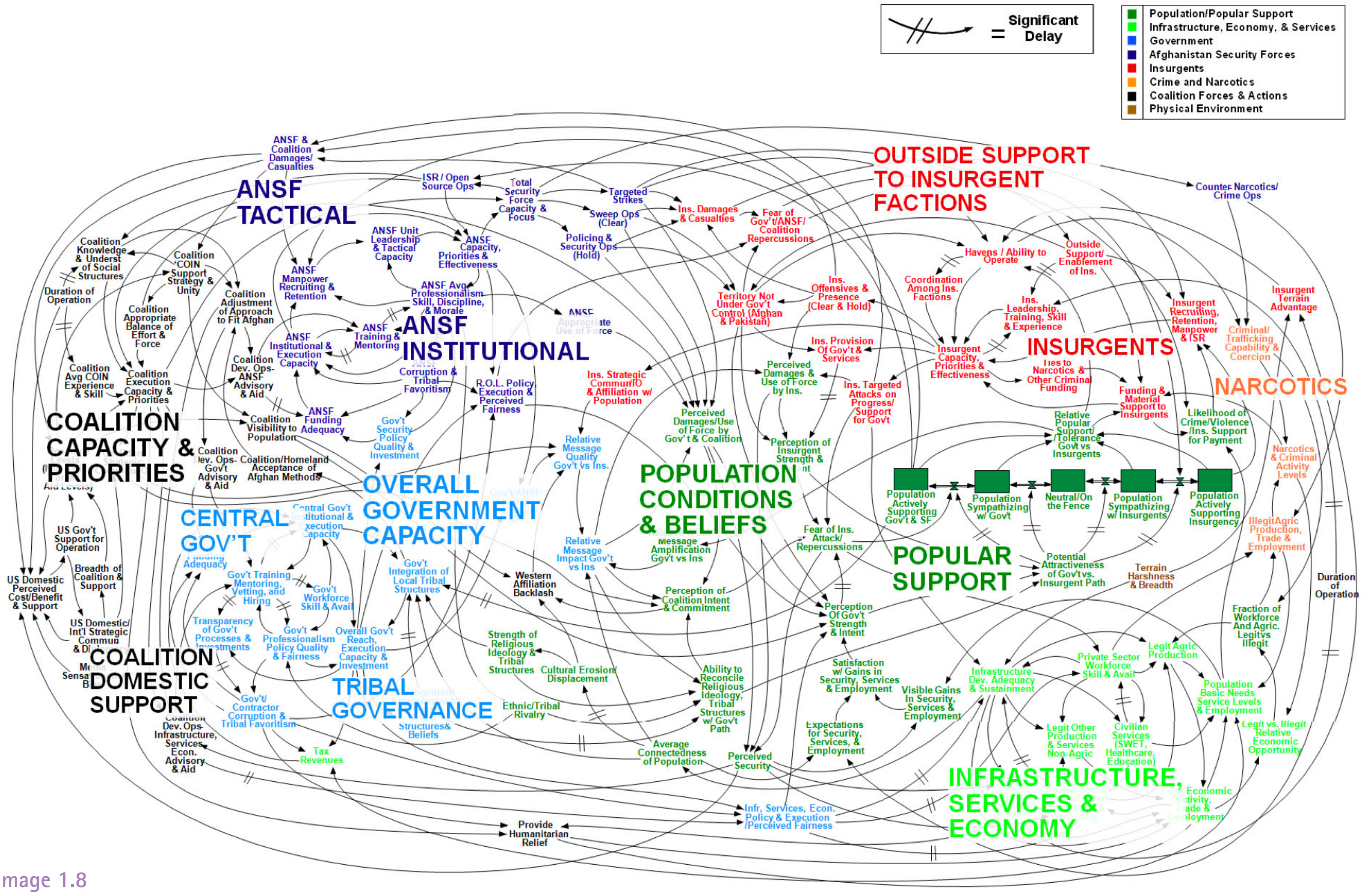

= Significant Delay
🟩 Population/Popular Support
🟩 Infrastructure, Economy, & Services
🟦 Government
🟦 Afghanistan Security Forces
🟥 Insurgents
🟧 Crime and Narcotics
⬛ Coalition Forces & Actions
🟫 Physical Environment
ANSF TACTICAL
ANSF INSTITUTIONAL
COALITION CAPACITY & PRIORITIES
CENTRAL GOV'T
OVERALL GOVERNMENT CAPACITY
TRIBAL GOVERNANCE
COALITION DOMESTIC SUPPORT
OUTSIDE SUPPORT TO INSURGENT FACTIONS
INSURGENTS
NARCOTICS
POPULATION CONDITIONS & BELIEFS
POPULAR SUPPORT
INFRASTRUCTURE, SERVICES & ECONOMY
Population Actively Supporting Gov't & SF
Population Sympathizing w/ Gov't
Neutral/On the Fence
Population Sympathizing w/ Insurgents
Population Actively Supporting Insurgency
Provide Humanitarian Relief
Duration of Operation

Image 1.8

The network behind a military engagement.

This diagram was designed during the Afghan war to portray the American strategy in Afghanistan. While it has been occasionally ridiculed in the press, it portrays well the complexities and the interconnected nature of a military's engagement. *(Image from New York Times)* cant investments in network and systems medicine, seeing it as the path towards future drugs.

Security: Fighting Terrorism.

Terrorism is one of the maladies of the 21st century, absorbing significant resources to combat it worldwide. Network thinking is increasingly present in the arsenal of various law enforcement agencies in charge of limiting terrorist activities. It is used to disrupt the financial network of terrorist organizations, to map terrorist networks, and to uncover the role of their members and their capabilities. While much of the work in this area is classified, several success stories have surfaced. Examples include the use of social networks to capture Saddam Hussein or the capture of the individuals behind the March 11, 2004 Madrid train bombings through the examination of the mobile call network. Network concepts have impacted military doctrine as well, leading to the concept of net-war, aimed at fighting low intensity conflicts and crime waged by terrorist and criminal networks that employ decentralized flexible network structures [13]. One of the first network science programs at the college level was started at West Point, the US Army's military academy. In 2009 the Army Research Lab and the Department of Defense devoted over $300 million to support network science centers across the US.

Epidemics: From forecasting to halting deadly viruses.

While the H1N1 pandemic was not as devastating as it was feared at the beginning of the outbreak in 2009, it gained a special role in the history of epidemics: it was the first pandemic whose course and time evolution was accurately predicted months before the pandemic reached its peak (Image 1.9) [14]. This was possible thanks to fundamental advances in understanding the role of networks in the spread of viruses. Indeed, before 2000 epidemic modeling was dominated by compartment models, assuming that everyone can infect everyone else one word the same socio-physical compartment. The emergence of a network-based framework has fundamentally changed this, offering a new level of predictability in epidemic phenomena.

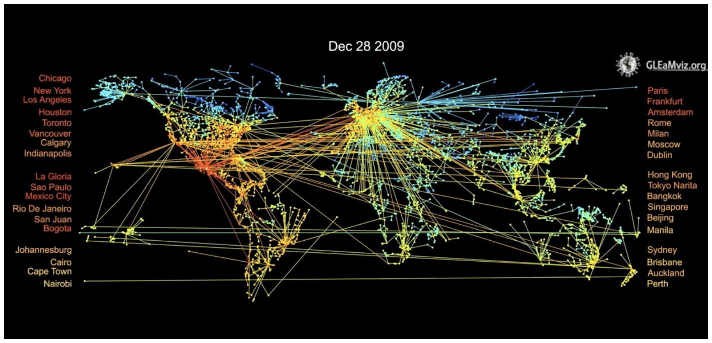


Image 1.9
Predicting the H1N1 epidemic.

The predicted spread of the H1N1 epidemics during 2009, representing the first successful prediction of a pandemic. The project, relying on the details of the worldwide transportation networks, foresee that H1N1 will peak out in October 2009, in contrast with the normal January-February peaks of influenza. This meant that the vaccines planned for November 2009 were too late, which was indeed the case. The success of this project shows the power of network science in facilitating advances in areas affected by networks.

Movie by D.Balcom, B.Gonçalves, H.Hu, and A.Vespignani.

Today epidemic prediction is one of the most active applications of network science [15, 16]. It is the source several fundamental results, covered in this book, that are used to predict the spread of both biological and electronic viruses. The impact of these advances are felt beyond biological viruses. In January 2010 network science tools have predicted the conditions necessary for the emergence of viruses spreading through mobile phones [17]. The first major mobile epidemic outbreak that started in the fall of 2010 in China, infecting over 300,000 phones each day, closely followed the predicted scenario.

Brain Research: Mapping neural network.

The human brain, consisting of hundreds of billions of interlinked neurons, is one of the least understood networks from the perspective of network science. The reason is simple: we lack maps telling us which neurons link to each other. The only fully mapped neural map available for research is that of the *C.Elegans* worm, with only 300 neurons. Should detailed maps of mammalian brains become available, brain research could become the most prolific application area of network science. Driven by the potential impact of such maps, in 2010 the National Institutes of Health has initiated the *Connectome* project, aimed at developing the technologies that could provide an accurate neuron-level map of mammalian brains.

Management: Uncovering the internal structure of an organization.

While traditionally management uses the official chain of command to understand the inner structure of an organization, it is increasingly evident that the informal network, capturing who really communicates with whom, matters even more for the success of a company. Accurate maps of this network can expose lack of communication between key units, can identify individuals who play an outsize role in bringing different departments and products together, and help higher management diagnose diverse organizational issues. Furthermore, there is increasing evidence in the management literature that the position of an employee within this network correlates with his/her productivity [18].

Therefore, several dozen consulting companies have emerged with expertise to map out the true structure of an organization. Established consulting firms, from *IBM* to *SAP*, have added social networking capabilities to their consulting business. These companies offer a host of services, from identifying opinion leaders to preventing employee churn and from identifying optimal groups for a task to modeling product diffusion (Image 1.10a/b/c/d). Hence lately network science tools are increasingly indispensable in management and business, enhancing productivity and boosting innovation within an organization.

Network science can therefore offer a microscope for higher management, helping them improve the company's effectiveness by uncovering the true network behind any organization.

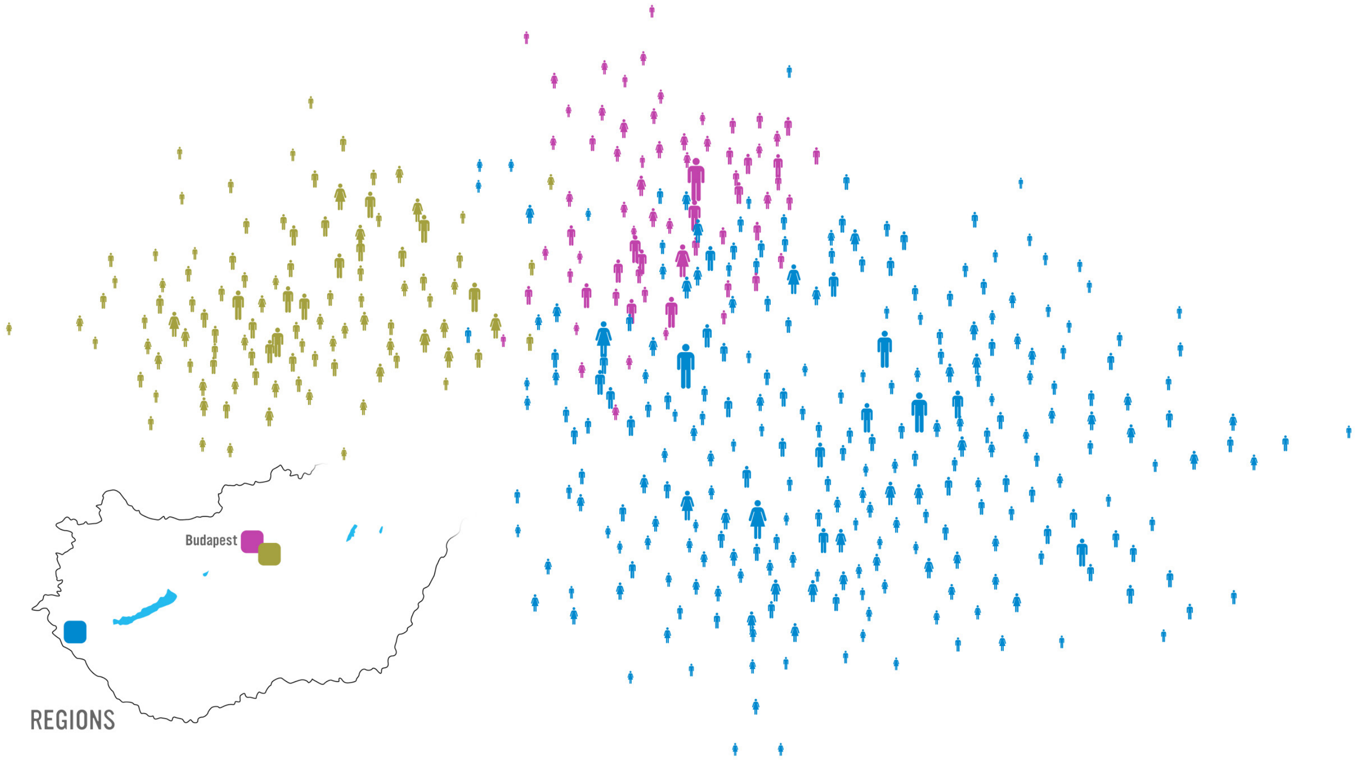


Image 1.10a
Understanding the inner workings of an organization.

The workforce of a Hungarian company with three main locations, one on Budapest, whose employees are shown in purple, and two manufacturing sites outside of the city, shown in yellow and blue. The company had a major internal communication problem: information that reached the workers about the intentions of the higher management often had nothing do to with the management's real plans. Seeking to understand the source of this discrepancy, and looking for ways to embrace information flow within the company, the management turned to Maven 7, a social networking consulting company that applies network science in diverse organizational setting.

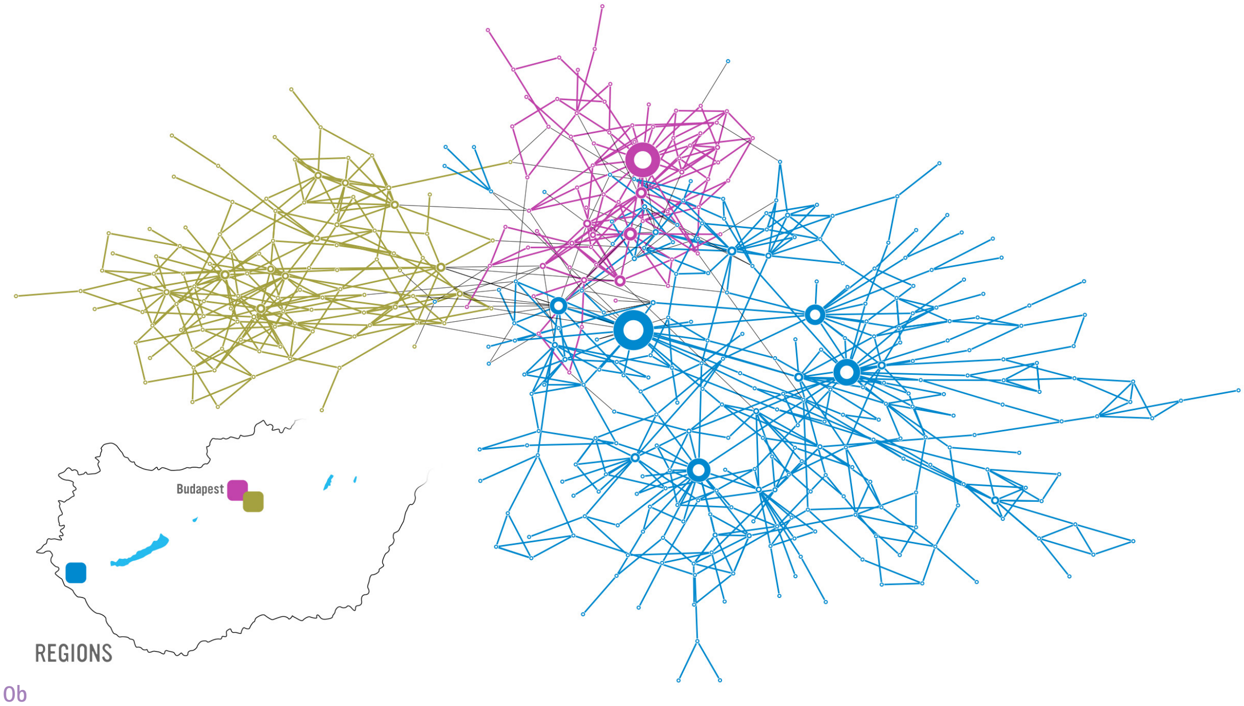


Image 1.10b

Understanding the inner workings of an organization.

Having the list of the workers and their role in the company, together with the official hierarchy is not sufficient to understand how an organization works. For that we need to know who listens to whom, who is asking for advice from whom, eventually uncovering the paths through which knowledge and information travels within the organization. Hence Maven 7 developed an online platform to ask each employee whom do they turn to for advice when it comes to decisions impacting the company, from restructuring to advancement. This allowed them to build the map shown above, where two individuals are connected if one nominated the other as his/her source of information on organizational and professional issues.
The map identifies several highly influential individuals that are the hubs of the organization. The problem was that none of the hubs were part of the leadership.

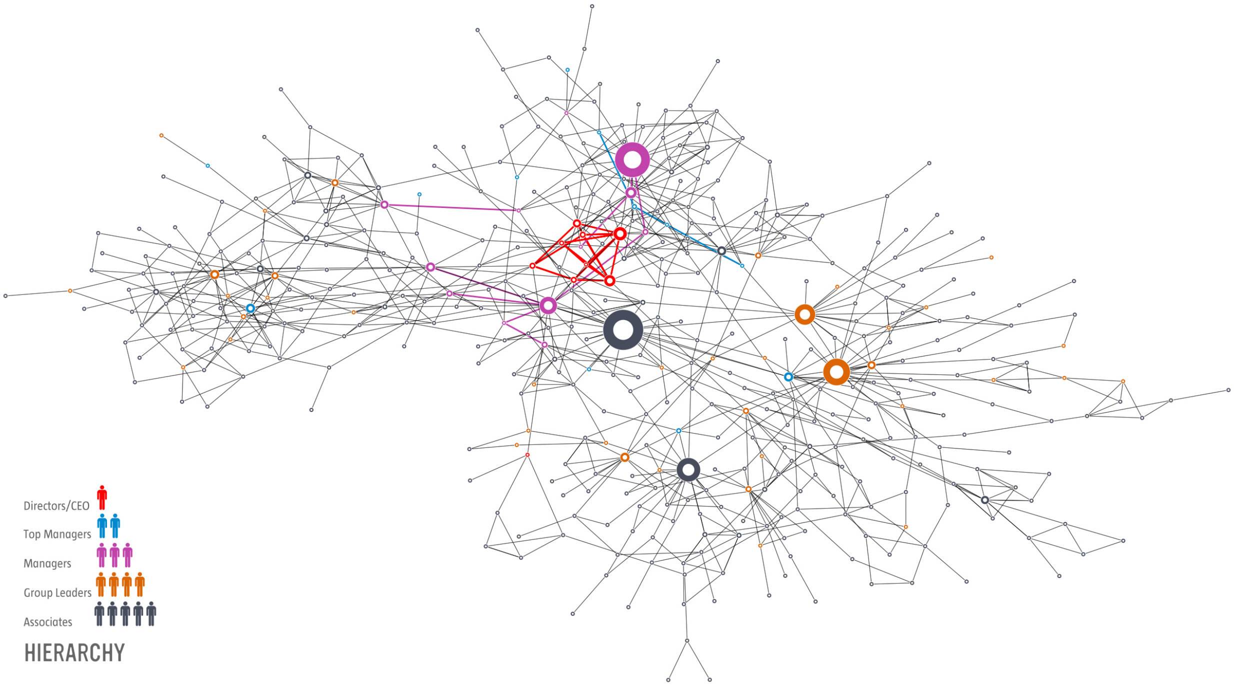


Image 1.10c

Understanding the inner workings of an organization.

The position of the leadership within the company's informal network is illustrated on this map, where we colored the nodes based on their company rank within the company. None of the company directors, including the CEO, shown in red, are hubs. Nor are the top managers, shown in blue. The hubs are managers, group leaders and associates, or workers. The biggest hub, hence the most influential individual, is an associate, shown as a gray node in the center.

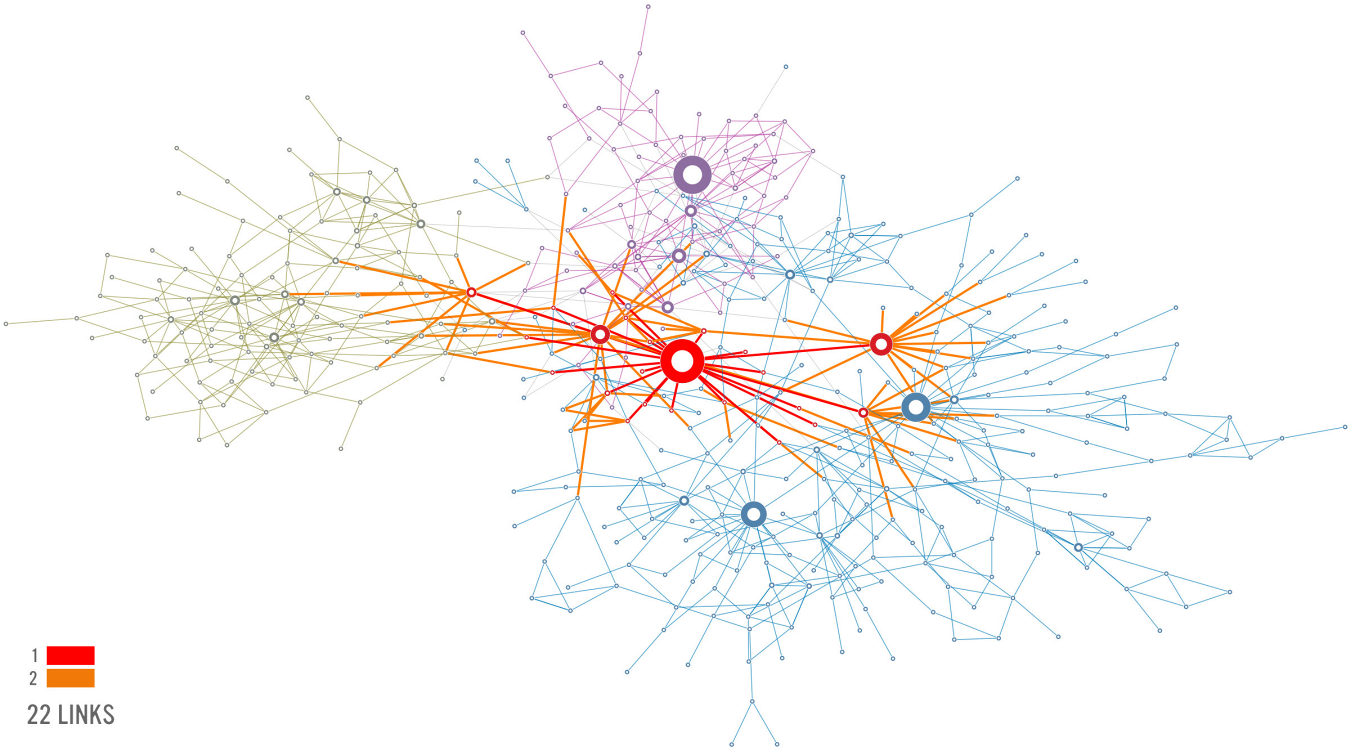


Image 1.10d

Understanding the inner workings of an organization.

The image indicates that a significant fraction of employees are one to two links from the biggest hub. It turns out that he is the safety and environmental expert in the company, whose job is to visit each location and talk with most employees. There is only one part of the company he has no links to: the directors or the top management. With little access to the management and their intentions, he passes on information that he collects along his trail, effectively running a gossip center.

How does one remedy this situation? Fire the biggest hub? He is not the problem and firing him would probably make the problem even more acute. The real issue is that higher management failed to put in place proper channels of communication, leaving behind a structural hole, which was naturally filled by the environmental and safety manager. Offering him and the few other hubs access to the true information can fundamentally change the reliability of information within the company. Network science can therefore offer a potent microscope for higher management, helping them improve the company's effectiveness by uncovering the true network behind an organization.

SCIENTIFIC IMPACT

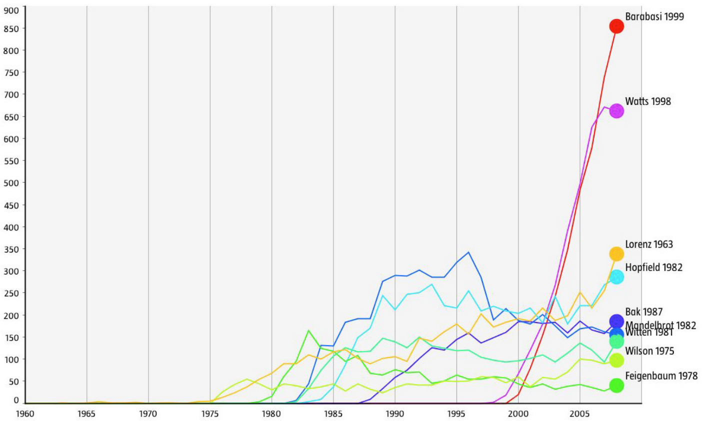


Image 1.11

Complexity and network science.

The impact of network science can be put into perspective by looking at the citation patterns of the most cited papers in complexity. The study of complex systems in the 70s and 80s was dominated by Edward Lorenz's 1963 classic work on chaos [19], Kenneth G. Wilson's renormalization group [20], and Mitchell Feigenbaum's discovery of the bifurcation diagram [21]. In the 1980s the community has shifted its focus on pattern formation, following Benoit Mandelbrot's book on fractals [22] and Thomas Witten and Len Sander's introduction of the diffusion limited aggregation mode [23]. Equally influential was John Hopfield's paper on neural networks [24] and Per Bak, Chao Tang and Kurt Wiesenfeld's paper on self-organized criticality [25]. These papers are continuing to define our understanding of complex systems, each of them writing a separate chapter in modern statistical mechanics. The video compare their citation pattern with the citations of the two most cited papers in this area [2,3].

Nowhere is the impact of network thinking more evident than in the scientific community. The most prominent scientific journals, from *Nature* and *Science* to *Cell* and *PNAS*, have devoted special issues, reviews, or editorials addressing the impact of networks on various topics from biology to social sciences. During the past decade, each year several dozen international conferences, workshops, summer and winter schools have focused exclusively on network science. A successful network science meeting series, called *NetSci*, attracts the field's practitioners since 2005. Several general-interest books, making the bestseller lists in many countries, have brought network science to the public. Most major universities offer network science courses, attracting a diverse student body. Finally, *Science Magazine* has devoted a special issue to networks, marking the ten-year anniversary of the paper that reported the discovery of scale-free networks [3] (Image 1.12).

The relative impact of network science can be put into perspective by looking at the citation patterns of the most cited papers in the area of complex systems (Image 1.11). Each of these papers are citation classics, cumulatively amassing anywhere between 2,000 and 5,000 citations, continuing to gather anywhere between 50 to 300 citations a year. To see how the interest in network science compares to these classic discoveries, in Movie 3 we also show the citation patterns of the two most cited network science papers: the 1998 paper on small-world phenomena by Duncan Watts and Steve Strogatz [2] and the 1999 Science paper reporting the discovery of scale-free networks by Albert-László Barabási and Réka Albert [3]. As one can see, the growth in citations to these papers unparalleled in the area of complex systems.

Image 1.12

Complex systems and networks.

Special issue of Science magazine on Complex Systems and Networks, published on July 24, 2009, marking the 10th anniversary of the 1999 discovery of scale-free networks [3].

Several other metrics indicate that network science is impacting in a defining manner particular disciplines. For example, several research fields witnessed network papers become some of the most cited papers in their leading journals:

- The 1998 paper by Watts and Strogatz in *Nature* on small world phenomena [2] and the 1999 paper by

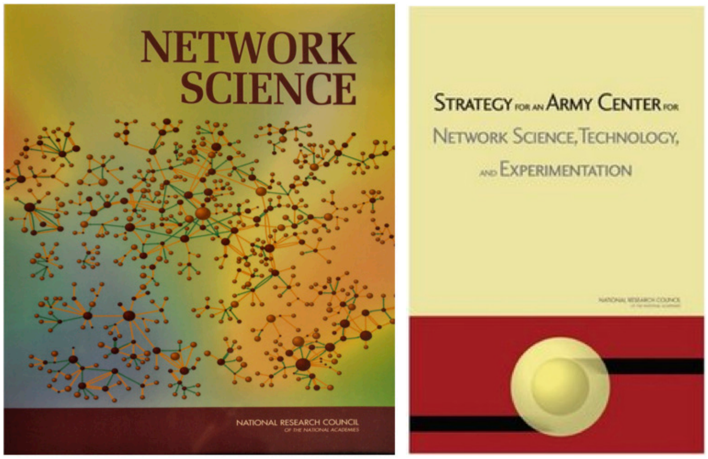

NETWORK SCIENCE
NATIONAL RESEARCH COUNCIL
STRATEGY FOR AN ARMY CENTER FOR NETWORK SCIENCE, TECHNOLOGY, AND EXPERIMENTATION

Image 1.13
National Research Council Reports.

The two National Research Council Reports on network science have not only documented the emergence of a new discipline, but have also explained their long-term impact on a number of research fields, as well as national competitiveness and the military. They have urged dedicated support for the field, leading to the establishment of a series of network science centers in US and the network science program within NSF.

Barabási and Albert in Science on scale-free networks [3] were identified by ISI as the top ten most cited papers in physics during the decade after their publications. Furthermore, currently (2011) the Watts-Strogatz paper is the second most cited of all papers published by *Nature* in 1998, and the Barabási-Albert paper is the most cited paper among all papers published in Science in 1999.

- Four years after its publication, the *SIAM* review of Mark Newman on network science became the most cited paper of any journal published by the *Society of Industrial Mathematics* [26].

- *Reviews of Modern Physics*, published continuously since 1929, is the physics journal with the highest impact factor. Currently the most cited paper of the journal is Chandrasekhar classic 1944 review that summarized the author's work that led to his Nobel in physics, entitled *Stochastic Problems in Physics and Astronomy* [27]. During over 60 years since its publication, the paper gathered over 5,000 citations. Yet, it will soon be taken over by a paper published only in 2001 entitled *Statistical Mechanics of Complex Networks*, the first review of network science [28].

- The paper leading to the discovery that in scale-free networks the epidemic threshold is zero, by Pastor-Satorras and Vespignani [29], is the most cited paper among the papers published in 2001 by *Physical Review Letters*, a position the paper is sharing with a paper on quantum computing.

- The paper by Michelle Girvan and Mark Newman on community discovery in networks [30] is the most cited paper published in 2002 by *Proceedings of the National Academy of Sciences*.

- The 2004 review entitled *Network Biology*, by Barabási and Oltvai [11], is the second most cited paper in the history of *Nature Reviews Genetics*, the top review journal in genetics.

Given this extraordinary response by the scientific community, network science was examined by the National Research Council (NRC), the arm of the US National Academies in charge of offering policy recommendation to the US government. NRC has assembled two panels, resulting in two publications [31], defining the field of network science (Image 1.13). They not only document the emergence of a new research field, but highlight the field's vital importance to national competitiveness and security. Following these reports, the National Science Foundation (NSF) in the US established a network science directorate and a series of network science centers were established by the Army Research Labs.

General Audience

The results of network science have excited the public as well. This was fueled partly by the success of several general audience books, like *Linked: The New Science of Networks* by Albert-László Barabási, *Nexus* by Mark Buchanan, and *Six Degrees* by Duncan Watts, each being translated in many of languages. Newer books, like *Connected* by Nicholas Christakis and James Fowler, were also exceptionally successful (Image 1.15). An award-winning documentary, *Connected*, by Australian filmmaker Annamaria Talas, has brought the field to our TV screen, being broadcasted all over the world and winning several prestigious prizes (Image 1.14). Networks have inspired artists as well, leading to a wide range of network science research inspired art-project, and even an annual symposium series that

Image 1.14
Connected.

The trailer of the award winner document *Connected*, directed by Annamaria Talas, focusing on network science.

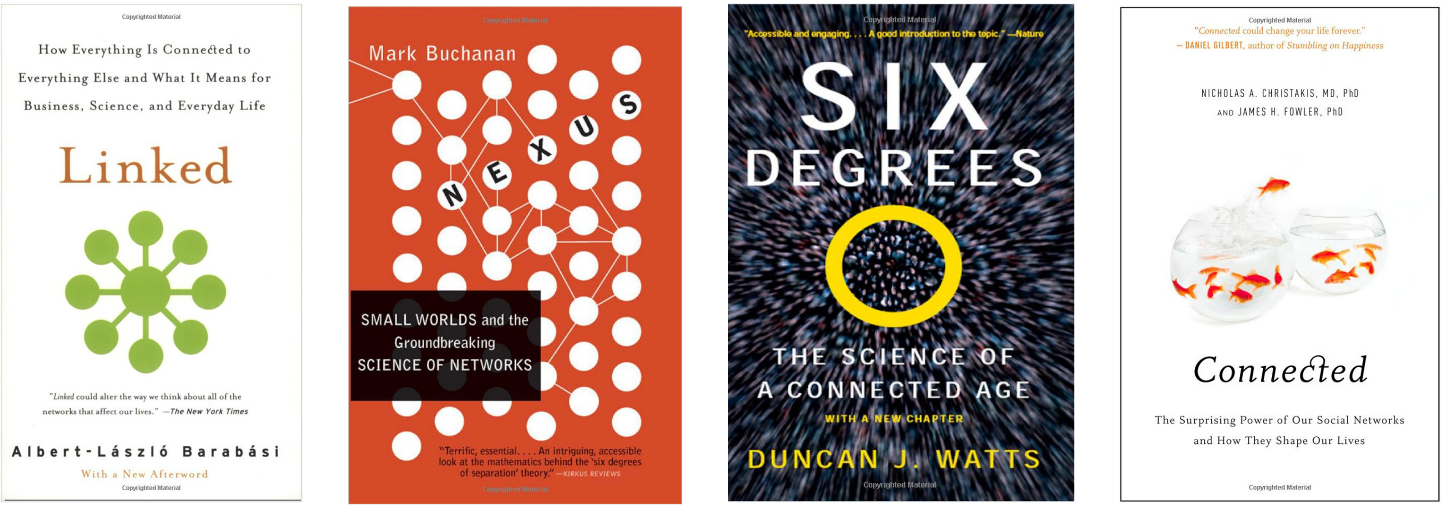

How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life
Linked
"Linked could alter the way we think about all of the networks that affect our lives." —The New York Times
Albert-László Barabási
With a New Afterword
Mark Buchanan
NEXUS
SMALL WORLDS and the Groundbreaking SCIENCE OF NETWORKS
"Terrific, essential. . . . An intriguing, accessible look at the mathematics behind the 'six degrees of separation' theory." —KIRKUS REVIEWS
"Accessible and engaging. . . . A good introduction to the topic." —Nature
SIX DEGREES
THE SCIENCE OF A CONNECTED AGE
WITH A NEW CHAPTER
DUNCAN J. WATTS
"Connected could change your life forever." — DANIEL GILBERT, author of Stumbling on Happiness
NICHOLAS A. CHRISTAKIS, MD, PhD AND JAMES H. FOWLER, PhD
Connected
The Surprising Power of Our Social Networks and How They Shape Our Lives

Image 1.15

Wide impact.

Four widely read books are bringing network science to the public.

brings together, on a yearly basis, artists and scientists [32]. Fueled by successful movies like *The Social Network*, and a series of novels and short stories, from science fiction to novels exploiting the network paradigm, today networks have permeated popular culture.

SUMMARY

While the emergence of the scientific interest in networks was rather sudden, the enthusiasm for the field was responding to the emergence of a wider social awareness of the importance of networks. This is illustrated in Image 1.16, where we show the usage frequency of the words that represent two important scientific revolutions of the past two centuries: evolution, capturing the most common term to refer to Darwin's theory of *evolution*, and *quantum*, the most frequently used term when one refers to *quantum* mechanics. The use of evolution increases only after the 1859 publication of Darwin's *On the Origins of Species*. The word *quantum*, first used in 1902, is virtually absent until the 1920s, when quantum mechanics gains prominence. The use of the word network has increased dramatically following the 1980s. While the word has many uses (as do evolution and *quantum*), its dramatic rise captures the extraordinary awareness of networks in the society at large. Indeed, evolution and quantum mechanics are just as important as core scientific fields, as they are as enabling platforms: the current revolution in genetics is built on evolutionary theory, and quantum mechanics offers a platform for a wide range of advances in contemporary science, from chemistry to wireless communications. In a similar fashion, network science is an enabling science, offering new tools and perspective for a wide range of scientific fields from social networking to drug design. Given the wide importance and impact of networks, we need to develop the tools to study and quantify them. The rest of this book is devoted to this worthy subject.

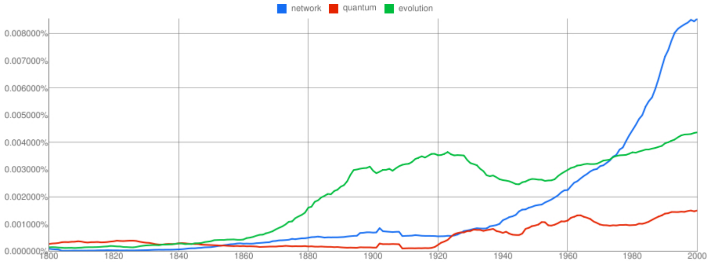


Image 1.16

The rise of networks.

The frequency of the use of the words *evolution* and *quantum* represents the major scientific advances of the 19th and 20th century, namely Darwin's theory of evolution and quantum mechanics. The plot indicates the exploding awareness of networks in the last decades of the 20th century, preparing a fertile ground for the emergence of network science. The plots were generated by using the ngram platform of Google: http://books.google.com/ngrams.

BIBLIOGRAPHY

[1] C. Wilson. *Searching for Saddam: a five-part series on how the US military used social networking to capture the Iraqi dictator.* 2010. www.slate.com/id/2245228/.

[2] D. J. Watts and S .H. Strogatz. *Collective dynamics of 'small-world' networks*. Nature, 393 (440), 1998.

[3] A.-L. Barabási and R. Albert. *Emergence of scaling in random networks*, Science, 286 (509), 1999.

[4] P. Erdős and A. Rényi. *On random graphs*. Publicationes Mathematicae, 6 (290), 1959.

[5] M. S. Granovetter. *The strength of weak ties.* American Journal of Sociology, 78 (1360), 1973.

[6] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, and J. Timm. *An empirical framework for binary interactome mapping*. Nature Methods, 6 (83), 2009.

[7] O. Sporns, G. Tononi, and R. Kötter. *The Human Connectome: A Structural Description of the Human Brain*. PLoS Comput. Biol., 1 (4), 2005.

[8] http://www.caida.org/ http://www.netdimes.org/

[9] International Human Genome Sequencing Consortium. *Initial sequencing and analysis of the human genome*. Nature, 409 (6822), 2001.

[10] J. C. Venter et al.,*The Sequence of the Human Genome*, Science, 291 (1304), 2001.

[11] Z. N. Oltvai and A.-L. Barabási. *Understanding the cell's functional organization*. Nature Reviews Genetics, 5 (101), 2004.

[12] N. Gulbahce, A.-L. Barabási, and J. Loscalzo. *Network medicine: a network-based approach to human disease*. Nature Reviews Genetics, 12 (56), 2011.

[13] J. Arquilla and D. Ronfeldt, *Networks and Netwars: The Future of Terror*, Crime, and Militancy (RAND: Santa Monica, CA), 2001.

[14] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, and A. Vespignani. *Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility*. BMC Medicine, 7 (45), 2009.

[15] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, and J. J. Ramasco, A. Vespignani, *Multiscale mobility networks and the spatial spreading of infectious diseases.* Proc. Natl. Acad. Sci., 106 (21484) 2009.

[16] L. Hufnagel, D. Brockmann, and T. Geisel, *Forecast and control of epidemics in a globalized world.* Proc. Natl. Acad. Sci., 101 (15124), 2004.

[17] P. Wang, M. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. *Understanding the spreading patterns of mobile phone viruses*. Science, 324 (1071), 2009.

[18] L. Wu , B. N. Waber, S. Aral, E. Brynjolfsson, and A. Pentland, *Mining Face-to-Face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1130251

[19] E. N. Lorenz, *Deterministic Non periodic Flow*. J. Atmos. Sci., 20 (130), 1963.

[20] K. G. Wilson, *The renormalization group: Critical phenomena and the Kondo problem*, Rev. Mod. Phys. 47 (773), 1975.

[21] M. J. Feigenbaum, *Quantitative Universality for a Class of Non-Linear Transformations*. J. Stat. Phys. 19 (25), 1978.

[22] B. B. Mandelbrot, *The Fractal Geometry of Nature*. W.H. Freeman and Company. 1982

[23] T. Witten, Jr. and L. M. Sander, *Diffusion-Limited Aggregation*, a Kinetic Critical Phenomenon. Phys. Rev. Lett., 47 (1400), 1981.

[24] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proc. Natl. Acad. Sci., 79 (2554), 1982.

[25] P. Bak, C. Tang, and K. Wiesenfeld. *Self-organized criticality: an explanation of 1/f noise*. Phys. Rev. Lett., 59 (4), 1987.

[26] M. E. J. Newman. *The structure and function of complex networks*, SIAM Review. 45 (167), 2003.

[27] S. Chandrasekhar. *Stochastic Problems in Physics and Astronomy*, Rev. Mod. Phys., 15 (1), 1943.

[28] R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Rev. Mod. Phys., 74 (47), 2002.

[29] R. Pastor-Satorras and A. Vespignani. *Epidemic spreading in scale-free networks*. Phys. Rev. Lett., 86 (3200), 2001.

[30] M. Girvan and M. E. J. Newman. *Community structure in social and biological networks*. Proc. Natl. Acad. Sci., 99 (7821), 2002.

[31] National Research Council, *Network Science*. Washington, DC: The National Academies Press, 2005.

National Research Council. Strategy for an Army Center for Network Science, Technology, and Experimentation . Washington, DC: The National Academies Press, 2007.

[32] M. Schich, R. Malina, and I. Meirelles (Editors), *Arts, Humanities, and Complex Networks* [Kindle Edition], 2012.

CHAPTER 2
GRAPH THEORY

The bridges of Königsberg

Networks and graphs

Degree, average degree and degree distribution

Real networks are sparse

Adjacency Matrix

Weighted and unweighted networks

Bipartite networks

Paths and distances in networks

Connectedness and Components

Clustering coefficient

Case study and summary

Appendix A: Global clustering coefficient

Bibliography

SECTION 1

THE BRIDGES OF KÖNIGSBERG

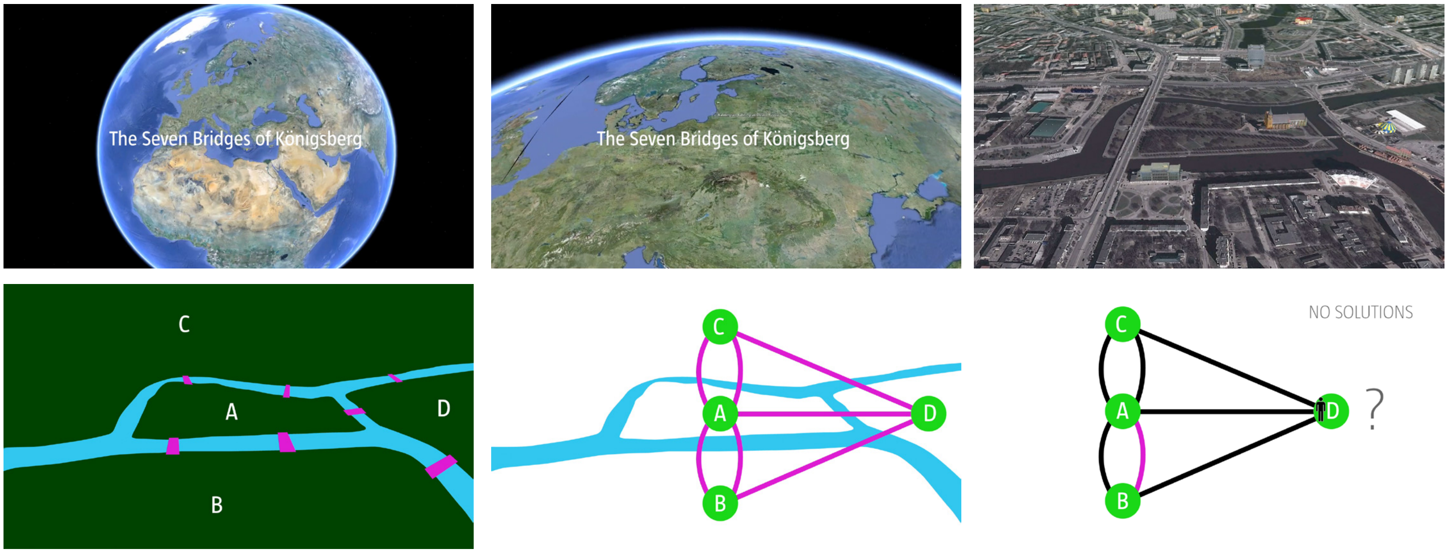

The Seven Bridges of Königsberg
The Seven Bridges of Königsberg
C
A
D
B
C
A
D
B
NO SOLUTIONS
C
A
D
?
B

Image 2.1

The bridges of Königsberg.

From the contemporary map of Königsberg (now Kaliningrad, Russia) to Euler's graph. The graph constructed by Euler consists of four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. Euler showed in 1736 that there is no continuous path that would cross seven the bridges while never crossing the same bridge twice. The people of Königsberg agreed with him, gave up their fruitless search and in 1875 they built a new bridge between B and C, increasing the number of links of these two nodes to four. Now only one node was left with an odd number of links and it became rather straightforward to find the desired path.

Few research fields can trace their birth to a single moment and place in history. Graph theory, the mathematical scaffold behind network science, can. Its roots go back to 1736 to Königsberg, the capital of Eastern Prussia and a thriving merchant city of its time. The trade supported by its busy fleet of ships allowed city officials to build seven bridges across the river Pregel that surrounded the town. Five of these connected the elegant island Kneiphof, caught between the two branches of the Pregel, to the mainland; two crossed the two branches of the river (Image 2.1). This peculiar arrangement gave birth to a contemporary puzzle: Can one walk across all seven bridges and never cross the same one twice? Despite many attempts, no one could find such path. The problem remained unsolved until 1735, when Leonard Euler, a Swiss born mathematician, offered a rigorous mathematical proof that such path does not exist.

Euler represented each of the four land areas separated by the river with letters A, B, C, and D. (Image 2.1). Next he connected with lines each piece of land that had a bridge between them. He thus built a *graph*, whose nodes were pieces of land and *links* were the bridges. Then Euler made a simple observation: if there is a path crossing all bridges, but never the same bridge twice, then nodes with odd number of links must be either the starting or the end point of this path. Indeed, if you arrive to a node with an odd number of links you may eventually have no unused link for you to leave it. A continuous path that goes through all bridges can have only one starting and one end point. Thus such a path cannot exist on a graph that has more than two nodes with an odd number of links. The Königsberg graph had three nodes with an odd number of links, B, C, and D, so no path could satisfy the problem.

Euler's proof was the first time someone solved a mathematical problem by turning it into a graph. For us the proof has two important messages: the first is that some problems become simpler and more treatable if they are represented as a graph. The second is that the existence of the path does not depend on our ingenuity to find it. Rather, it is a property of the graph. Indeed, given the structure of the Königsberg graph, no matter how smart we are, we will never find the desired path. In other words, networks have properties hidden in their structure that limit or enhance their behavior. To fully understand how networks affect the properties of a system, we need to become familiar with graph theory, a branch of mathematics that grew out of Euler's proof, offering a formalism that will be used throughout this book.

SECTION 2

NETWORKS AND GRAPHS

If we want to understand a complex system, we first need a map of its wiring diagram. A network is a catalog of a system's components often called **nodes** or **vertices** and the direct interactions between them, called **links** or **edges** (Box 2.1).

The network representation offers a common language to study systems that may differ greatly in nature, appearance, or scope. Indeed as shown in Image 2.3, three rather different systems have exactly the same network representation.

Networks or graphs?

In the scientific literature the terms network and graph are used interchangeably. Yet, there is a subtle distinction between the two terminologies: the *network*, *node*, and *link* combination often refers to real systems: the WWW is a network of web pages connected by URLs; society is a network of individuals connected by family, friendship or professional ties; the metabolic network is the sum of all chemical reactions that take place in a cell. In contrast, we use the terms graph, vertex, and edge when we talk about the mathematical representation of these networks: we talk about the web graph, the social graph (a term made popular by Facebook), or the metabolic graph. Yet, this distinction is rarely made, so these two terminologies are often used as synonyms of each other.

Network Science	Graph Theory
network	graph
node	vertex
link	edge

Box 2.1

Image 2.3 also introduces two basic network parameters:

Number of nodes, which we denote with N, representing the number of components in the system. We will often call N the *size of the network*.

Number of links, which we denote with L, representing the total number of interactions between the nodes.

The networks shown in Image 2.1 all have $N = 4$ and $L = 4$. To distinguish the nodes, we label them $i = 1, 2, ..., N$. The links are rarely labeled, as they can be identified through the nodes they connect. For example, the (2, 4) link connects nodes 2 and 4.

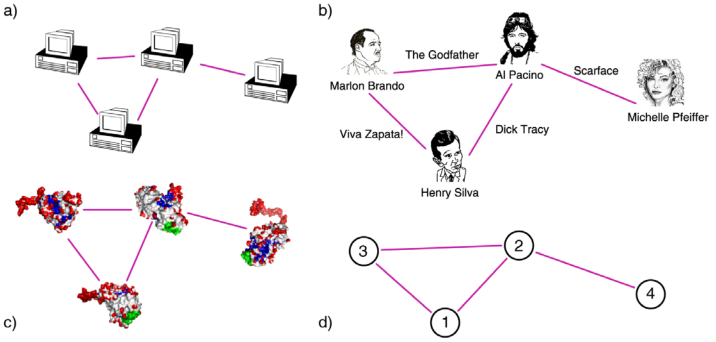


Image 2.3

Real systems of quite different nature can have the same network representation.

In the figure we show a small subset of (a) the *Internet*, where routers (specialized computers) are connected to each other; (b) the *Hollywood actor network*, where two actors are connected if they played in the same movie; (c) a *protein-protein interaction network*, where two proteins are connected if there is experimental evidence that they can bind to each other in the cell. While the nature of the nodes and the links differs widely, each network has the same graph representation, consisting of $N = 4$ nodes and $L = 4$ links, shown in (d).

The links of a network can be *directed* or *undirected*. Some systems have directed links, like the WWW, whose uniform resource locators (URL) point from one web document to the other, or phone calls, where one person calls the other. Other systems display undirected links, like romantic ties: if I date Janet, Janet also dates me, or transmission lines on the power grid, on which the electric current can flow in both directions.

A network is called *directed* (or digraph) if all of its links are directed or *undirected* if all of its links are undirected. Some networks simultaneously have directed and undirected links. For example in the metabolic network some reactions are reversible (i.e. bidirectional or undirected) and others are irreversible, taking place in only one direction (directed).

Throughout this book we will use ten networks to illustrate the tools of network science. These networks, listed in Table 2.1, were selected having diversity in mind, spanning social systems (mobile call graph or email network), collaboration and affiliation networks (science collaboration network, Hollywood actor network), information systems (WWW), technological and infrastructural systems (Internet and power grid), biological systems (protein interaction and metabolic network), and reference networks (citations). They differ widely in their sizes, from as few as N =1,039 nodes and L = 5,802 links in the *E. coli* metabolism, to almost half million nodes and five million links in the citation network. They cover several of the areas where networks are actively applied, representing 'canonical' datasets, often used by researchers in the field of network science to illustrate key network properties. In the coming chapters we will discuss in detail the nature and the characteristics of each of these datasets, turning them into the guinea pigs of our journey to understand complex networks.

NETWORK NAME	NODES	LINKS	DIRECTED/ UNDIRECTED	N	L	‹K›
Internet	routers	Internet Connections	Undirected	192,244	609,066	2.67
WWW	webpages	links	Directed	325,729	1,497,134	4.60
Power Grid	power plants, transformers	cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	subscribers	calls	Directed	36,595	91,826	2.51
Email	email addresses	emails	Directed	57,194	103,731	1.81
Science Collaboration	scientists	co-authorships	Undirected	23,133	186,936	16.16
Actor Network	actors	co-acting	Undirected	212,250	3,054,278	28.78
Citation Network	papers	citations	Directed	449,673	4,707,958	10.47
E. coli Metabolism	metabolites	chemical reactions	Directed	1,039	5,802	5.84
Yeast Protein Interactions	proteins	binding interactions	Undirected	2,018	2,930	2.90

Table 2.1

Network maps and their basic properties.

The basic characteristics of the networks that we use throughout this book to illustrate the use of network science. This table lists the nature of their nodes and links, indicating if links are directed or undirected, the number of nodes *(N)* and links *(L)*, and the network's average degree. For directed networks the average degree equals the average in- and out-degrees as $‹k› = <k_{in}>=<k_{out}>$.

Box 2.2

Choosing the proper network representation.

The choices we make when we represent a complex system as a network will determine our ability to use network science successfully. For example, the way we define the links between two individuals dictates the nature of the questions we can explore:

- By connecting individuals that regularly interact with each other in the context of their work, we obtain the *professional network*, that plays a key role in the success of a company or an institution, and it is of major interest to organizational research.

- By linking friends to each other, we obtain the *friendship network*, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.

- By connecting individuals that have an intimate relationship, we obtain the *sexual network*, of key importance for the spread of sexually transmitted diseases, like AIDS, and of major interest for epidemiology.

- By using phone and email records to connect individuals that call or email each other, we obtain the *acquaintance network*, capturing a mixture of professional, friendship or intimate links, of importance to communications and marketing.

While many links in these four networks overlap (some coworkers may be friends or may have an intimate relationship), these networks are not identical. Other networks may be valid from a graph theoretic perspective, but may have little practical utility. For example, by linking all individuals with the same first name, Johns with Johns and Marys with Marys, we do obtain a well-defined network, yet its utility is questionable. Hence in order to apply network theory to a system, careful considerations must precede our choice of nodes and links, ensuring their significance to the problem we wish to explore.

SECTION 3

DEGREE, AVERAGE DEGREE, AND DEGREE DISTRIBUTION

A key property of each node is its *degree*, representing the number of links it has to other nodes. The degree can represent the number of mobile phone contacts an individual has in the call graph (i.e. the number of different individuals the person has talked to), or the number of citations a research paper gets in the citation network.

We denote with k_i the degree of the i^{th} node in the network. For example, for the undirected networks shown in Image 2.3 we have $k_1=2, k_2=3, k_3=2, k_4=1$.

In an undirected network total number of links, L, can be expressed as the sum of the node degrees:

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i \qquad (1)$$

Here the 1/2 factor corrects for the fact that in the sum (1) each link is counted twice. For example, the link connecting the nodes 2 and 4 in *Image 2.3* will be counted once in the degree of node 1 ($k_2 = 3$) and once in the degree of node 4 ($k_4 = 1$).

Box 2.3

Brief statistics review.

The average, the standard deviation, and the distribution of random variables will play a key role throughout this book.
For a sample of N values $x_1, \ldots , x_N$ we have:

Average (mean value):

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (2)$$

nth moment:

$$\langle x \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n \qquad (3)$$

Standard deviation (fluctuations around the average):

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2} \qquad (4)$$

Distribution of x (probability that a randomly chosen value is a):

$$p = \frac{1}{N}\sum_{i} \delta_{x,x_i} \qquad (5)$$

which yields

$$\sum_{i} p_x = 1\left(\int p_x \, dx = 1\right) \qquad (6)$$

An important property of a network is its *average degree*, which for an undirected network is

$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} \qquad (7)$$

In directed networks we distinguish between *incoming degree*, k_i^{in}, representing the number of links that point node *i*, and *outgoing degree*, k_i^{out} , representing the number of links that point from the node i to other nodes and the *total degree*, k_i , given by

$$k_i = k_i^{in} + k_i^{out} \qquad (8)$$

For example, on the WWW the number of pages a given document points to represents its outgoing degree, k_{out}, and the number of other documents that point to it represents its incoming degree, k_{in}.

The total number of links in a directed network is

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} \qquad (9)$$

The 1/2 factor in Eq. (1), is absent above, as for directed networks the two sums in (9) separately count the outgoing and the incoming degrees.

The average degree of a directed network is

$$(k^{in}) = \frac{1}{N}\sum_{i=1}^{N} k_i^{in} = (k^{out}) = \frac{1}{N}\sum_{i=1}^{N} k_i^{out} = \frac{L}{N} \, . \qquad (10)$$

The *degree distribution*, p_k, provides the probability that a randomly selected node in the network has degree k. Since p_k is a probability, it must be normalized, i.e. $\sum_{k=1}^{\infty} p_k = 1$. For a fixed network of N nodes the degree distribution is the normalized histogram (see Gallery 2.1),

$$p_k = \frac{N_k}{N} \,,$$

where N_k is the number of degree k nodes. Hence the number of degree k nodes can be obtained from the degree distribution as $N_k = N_{pk}$.

The degree distribution has taken a central role in network theory following the discovery of scale-free networks (Barabási & Albert, 1999). Another reason for its importance is that the calculation of most network properties requires us to know p_k. For example, the average degree of a network can be written as

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

We will see in the coming chapters that the precise functional form of p_k determines many network phenomena, from network robustness to the spread of viruses.

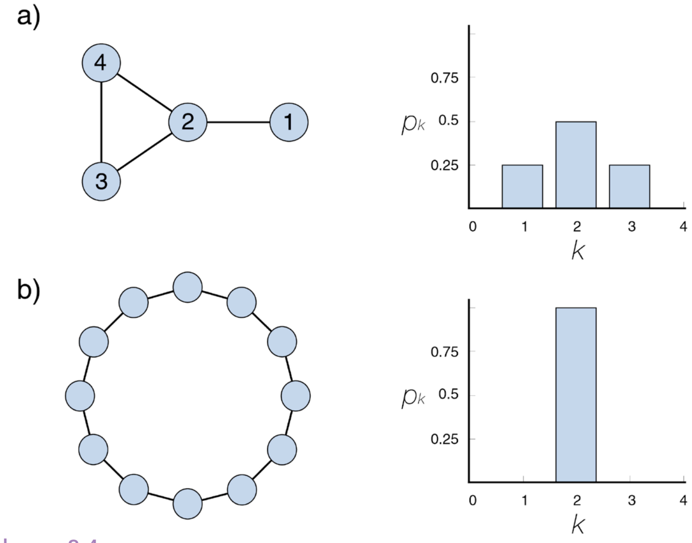


Image 2.4a

Degree distribution.

The degree distribution is defined as the $p_k = N_k/N$ ratio, where N_k denotes the number of k-degree nodes in a network. For the network in (a) we have $N = 4$ and $p_1 = 1/4$ (one of the four nodes has degree $k_1 = 1$), $p_2 = 1/2$ (two nodes have $k_3 = k_4 = 2$), and $p_3 = 1/4$ (as $k_2 = 3$). As we lack nodes with degree k > 3, $p_k = 0$ for any $k > 3$. Panel (b) shows the degree distribution of a one dimensional lattice. As each node has the same degree $k = 2$, the degree distribution is a Kronecker's delta function $p_k = \delta(k - 2)$.

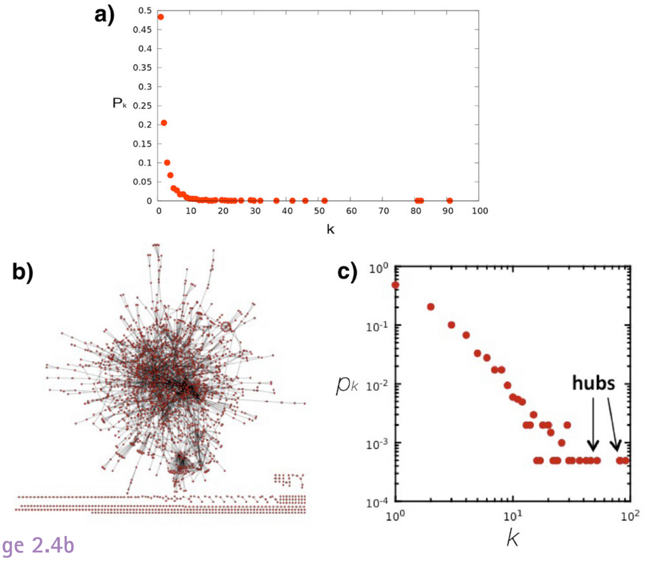


Image 2.4b

In many real networks, the node degree can vary considerably. For example, as the degree distribution (a) indicates, the degrees of the proteins in the protein interaction network shown in (b) vary between $k=0$ (isolated nodes) and $k=92$, which is the degree of the largest node, called a hub. There are also wide differences in the number of nodes with different degrees: as (a) shows, almost half of the nodes have degree one (i.e. p_1=0.48), while there is only one copy of the biggest node, hence $p_{92} = 1/N$=0.0005. (c) The degree distribution is often shown on a so-called log-log plot, in which we either plot log p_k in function of *log k*, or, as we did in (c), we use logarithmic axes.

REAL NETWORKS ARE SPARSE

In real networks the number of nodes (N) and links (L) can vary widely. For example, the neural network of the worm *C. elegans*, the only fully mapped brain of a living organism, has 297 neurons (nodes) and 2,345 synapses (links), while a human brain is estimated to have about a hundred billion (10^{11}) neurons, each with an average of 7,000 synaptic connections. The genetic network of a human cell has about 20,000 genes as nodes; the social network consists of seven billion individuals ($N \simeq 7\times10^9$) and the WWW is estimated to have over a trillion webpages ($N>10^{12}$). These wide differences in size are noticeable in Table 2.1 where we list N and L for several network maps. Some of these maps offer a complete wiring diagram of the system they describe (like the actor network or the *E. Coli* metabolism), others are only samples, representing a subset of a real system's nodes (WWW, mobile call graph).

Table 2.1 indicates that the number of links also varies widely. In a network of N nodes the number of links is between $L = 0$ and L_{max}, where L_{max} is the total number of links present in a complete graph (Image 2.5),

$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2} \qquad (11)$$

a graph in which each node is connected to all other nodes. In real networks L is much smaller than L_{max}, indicating that real networks are sparse. For example, the WWW graph in Table 2.1 has about 1.5 million links. Yet, if the WWW were to be a complete graph, this sample should have $L_{max} \approx 10^{12}$ links according to (11).
Therefore, the web graph has only a 10^{-6} fraction of the links it could have, making it a sparse network. In fact each network in Table 2.1 has only a tiny fraction of the links it could have according to (11). As we will see later sparseness has important consequences on the way we explore and store real networks.

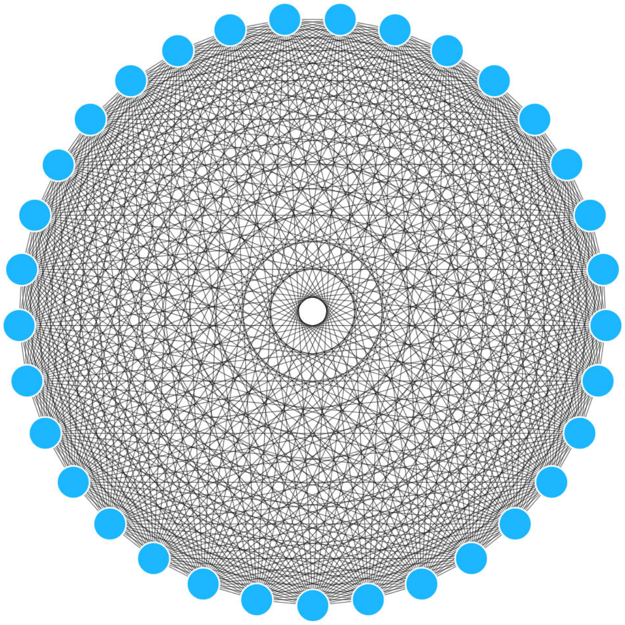


Image 2.5

Complete graph.

The figure shows a complete graph with N = 16 nodes and L_{max} = 120 links, as predicted by Eq. (11). The adjacency matrix of a complete graph is A_{ij} = 1 for all *i, j = 1, ...N* and A_{ii} = 0. The average degree of a complete graph is $\langle k \rangle = N - 1$.

SECTION 5

ADJACENCY MATRIX

A full description of a network requires us to keep track of its links. The simplest way to achieve this is to provide a complete list of the links. For example, the network of Image 2.1 is uniquely described by the list of its four (i, j) links: $\{(1, 2), (1, 3), (2, 3), (2, 4)\}$.

For mathematical purposes we often represent a network through its adjacency matrix. The adjacency matrix of a directed network of N nodes has N rows and N columns, its elements being:

$A_{ij} = 1$ if there is a link pointing from node j to node i

$A_{ij} = 0$ if nodes i and j are not connected to each other.

The adjacency matrix of an undirected network has two entries for each link, e.g. link (1,2) is represented as $A_{12} = 1$ and $A_{21} = 1$. Hence the adjacency matrix of an undirected network is symmetric, i.e. $A_{ij} = A_{ij}$ (Image 2.7).

The degree k_i of node i can be directly obtained from the elements of the adjacency matrix. For undirected networks a node's degree is a sum over either the rows or the columns of the matrix, i.e.

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{i=1}^{N} A_{ij} \quad . \qquad (12)$$

For directed networks the sums over the adjacency matrix' rows and columns provide the incoming and outgoing degrees, respectively

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \qquad k_i^{out} = \sum_{i=1}^{N} A_{ij} \quad . \quad (13)$$

Given that in an undirected network the number of outgoing links equals the number of incoming links, we have

$$2L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_i^{out} = \sum_{ij}^{N} A_{ij} \qquad (14)$$

The number of nonzero elements of the adjacency matrix is $2L$, or twice the number of links. Indeed, an undirected link connecting nodes i and j appears in two entries: $A_{ij} = 1$, a link pointing from node j to node i, and $A_{ji} = 1$, and a link pointing from i to j (Image 2.7).

The sparsity of real networks implies that the adjacency matrices are also sparse. Indeed, a complete network has $A_{ij} = 1$, for all (i, j), i.e. each of its matrix elements are equal to one. In contrast in real networks only a tiny fraction of the matrix elements are nonzero. This is illustrated in Image 2.6, where we show the adjacency matrix of the protein-protein interaction network listed in Table 2.1. One can see that the matrix appears nearly empty. One immediate consequence of the sparseness is that when we store a large network in our computer, it is better to store only the list of links (i.e. elements for which $A_{ij} \neq 0$), rather than full adjacency matrix, as an overwhelming fraction of Aij elements are zero.

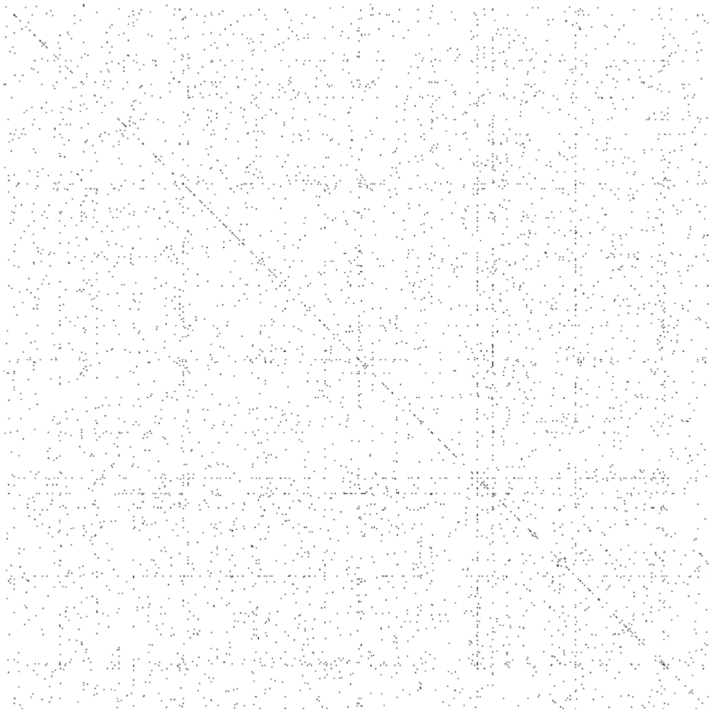


Image 2.6
The adjacency matrix is typically sparse.

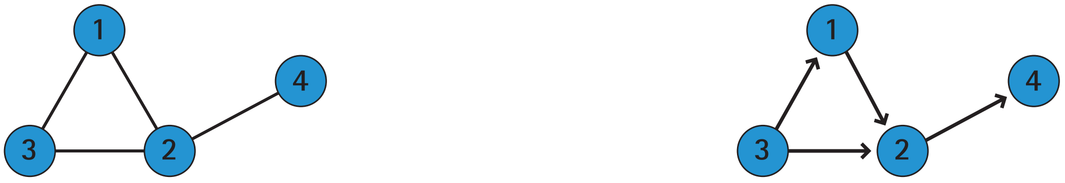
The adjacency matrix of the yeast protein-protein interaction network, consisting of 2,018 nodes, each representing a yeast protein (Table 2.1). A dot is placed on each spot of the adjacent matrix for which $A_{ij} = 1$, indicating the presence of an interaction. There are no dots for $A_{ij} = 0$. The small fraction of dots underlines the sparse nature of the protein-protein interaction network.

Adjacency matrix

$$A_{ij} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$$

Undirected network

Directed network



Undirected network

$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$k_2 = \sum_{j=1}^{4} A_{2j} = \sum_{i=1}^{4} A_{i2} = 3$$

$$A_{ij} = A_{ji} \qquad A_{ii} = 0$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \langle k \rangle = \frac{2L}{N}$$

Directed network

$$A_{ij} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$k_2^{in} = \sum_{j=1}^{4} A_{2j} = 2$$

$$k_2^{out} = \sum_{i=1}^{4} A_{i2} = 1$$

$$A_{ij} \neq A_{ji} \qquad A_{ii} = 0$$

$$L = \sum_{i,j=1}^{N} A_{ij} \qquad \langle k^{in} \rangle = \langle k^{out} \rangle = \frac{L}{N}$$

Image 2.7
The adjacency matrix.

Top: The elements of the adjacency matrix. The adjacency matrix of a directed (left column) and an undirected (right column) network. The figure highlights the fact that the degree of a node (in this case node 2) can be expressed as the sum over the appropriate column or row of the adjacency matrix. It also shows a few basic network characteristics, like the total number of links, (*L*), and average degree, (‹*k*›), expressed in terms of the elements of the adjacency matrix.

SECTION 6

WEIGHTED AND UNWEIGHTED NETWORKS

So far we discussed only networks for which all links have the same weight, i.e. $A_{ij} = 1$. Yet, in many applications we need to study weighted networks, where each link (i, j) has a unique weight w_{ij}. In mobile call networks the weight can represent the total number of minutes two mobile phone users talk with each other on the phone; on the power grid the weight is the amount of current flowing through a transmission line.

For weighted networks the elements of the adjacency matrix carry the weight of the link

$$A_{ij} = w_{ij} \quad . \tag{15}$$

Most networks of scientific interest are weighted, but we can not always measure the appropriate weights, hence we often approximate these networks as unweighted. In this book we predominantly focus on unweighted networks, but we will devote a separate chapter to network characteristics that are unique to weighted networks.

The value of a network: Metcalfe's Law.

Metcalfe's law states that the value of a network is proportional to the square of the number of its nodes, i.e. N^2. Formulated around 1980 in the context of communication devices by Robert M. Metcalfe (Gilder, 1993), the idea behind Metcalfe's law is that the more individuals use a network, the more valuable it becomes. Indeed, the more of your friends use email, the more valuable it is to you as well, as the more individuals you can communicate with.

During the Internet boom of the late 1990s Metcalfe's law was frequently used to offer a quantitative valuation for Internet companies, supporting a "build it and they will come" mentality (Briscoe et al., 2006). It suggested that the value of a service is proportional to the square of the number of its users, in contrast with the cost that grows only linearly. Hence if the service attracts sufficient number of users, it will inevitably become profitable, as N^2 will surpass N at some sufficiently large N. Hence Metcalfe's Law offered credibility to growth over profits, fueling the Internet bubble of 2001.

Metcalfe's law is based on Eq. (11), telling us that if all links of a communication network with N nodes are equally valuable, the total value of the network is proportional to $N(N - 1)/2$, that is, roughly, N^2. If a network has $N = 10$ members, there are $L_{max} = 45$ different possible connections between them. If the network doubles in size to $N = 20$, the number of connections doesn't merely double but roughly quadruples to 190, a phenomenon called network externality in economics.

Two issues limit the validity of Metcalfe's law: (i) most real networks are sparse, which means that only a very small fraction of the links are present. Hence the total value of the network will not grow like N^2, but the growth is often only linear in N. (ii) As the links have weights, not all links are of equal value; some links are used heavily while the vast majority of links are rarely utilized.

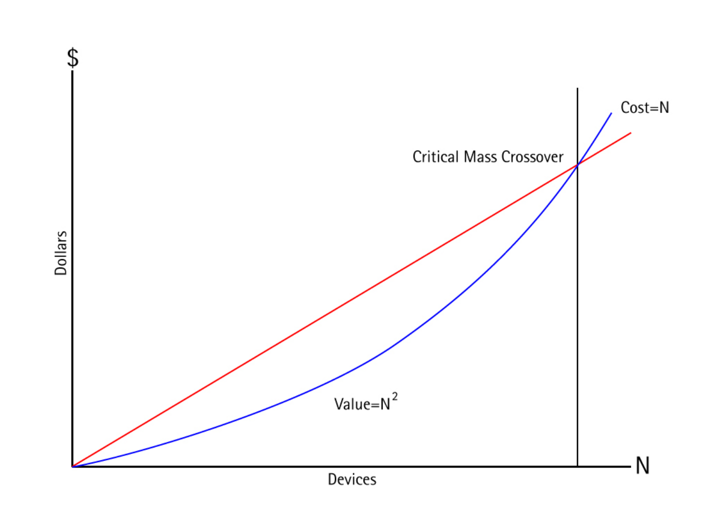


Image 2.8

According to Metcalfe's law the cost of network based services and products increases linearly with the number of nodes (users or devices) while the benefits or income is driven by the number of links L_{max} the technology makes possible, growing like N^2. Hence once the number of devices exceeds some "critical mass crossover", the technology becomes profitable.

Box 2.4

SECTION 7

BIPARTITE NETWORKS

A bipartite graph (or bigraph) is a network whose nodes can be divided into two disjoint sets *U* and *V* such that each link connects a *U*-node to a *V*-node. In other words, if we color the *U*-nodes yellow and the *V*-nodes green, then each link must connect nodes of different colors (Image 2.9a/b).

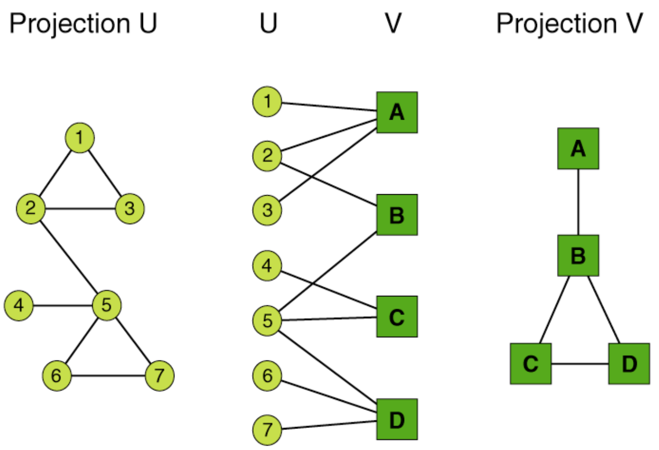


Image 2.9a
Bipartite network.

In a bipartite network we have two sets of nodes, *U* and *V*, so that nodes in the *U*-set connect directly only to nodes in the *V*-set. Hence there are no direct *U-U* or *V-V* links. The figure also shows the two projections we can generate from any bipartite network. Projection *U* is obtained by connecting two *U*-nodes to each other if they link to the same *V*-node in the bipartite representation. Projection *V* is obtained by connecting two *V*-nodes to each other if they link to the same *U*-node in the bipartite network.

We can generate two projections for each bipartite network. The first projection connects two *U*-nodes to each other by a link if they are linked to the same *V*-node in the bipartite representation; the second projection connects the *V*-nodes to each other by a link of they connect to the same *U*-node.

In network theory we encounter numerous bipartite networks. A well-known example is the Hollywood actor network, in which one set of nodes corresponds to movies (*U*), and the other to actors (*V*), a movie being connected to an actor if the actor plays in that movie. In this network one projection corresponds to the actor network, in which two nodes are connected to each other if they played in the same movie; this is the network characterized in Table 2.1. The other projection is the movie network, in which two movies are connected if they share at least one actor in their cast. Another example of bipartite network emerges in medicine, connecting diseases to the genes whose effects can cause or influence the corresponding disease (Image 2.9a/b). Finally, one can also define multipartite networks, like the tripartite recipe-ingredient-compound network described in Image 2.10 a/b.

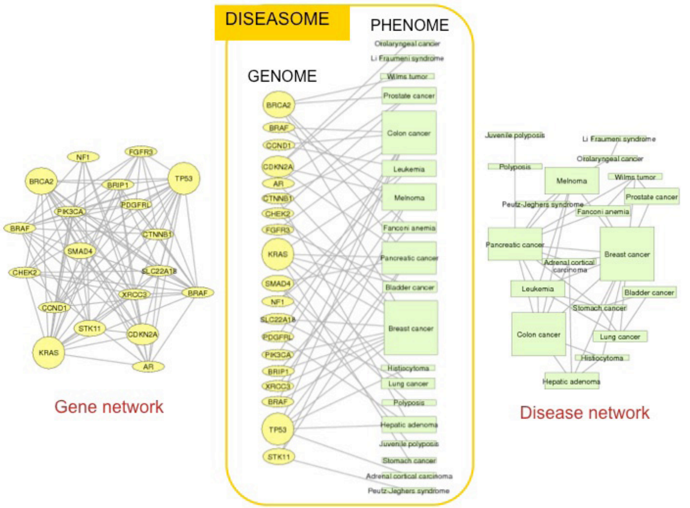


Image 2.9b
Bipartite network.

The *human diseaseome* is a bipartite network, whose nodes are diseases (*U*) and genes (*V*), in which a disease is connected to a gene if mutations in that gene are known to affect the particular disease [4]. One projection of the diseaseome is the *gene network*, whose nodes are genes, two genes being connected if they are associated with the same disease. The second projection is the disease network, whose nodes are diseases, two diseases being connected if the same genes are associated with them, indicating that the two diseases have common genetic origins. The figure shows a subset of the diseaseome, focusing on cancers. The full human diseaseome map, connecting 1,283 disorders via 1,777 shared disease genes. (After [4])

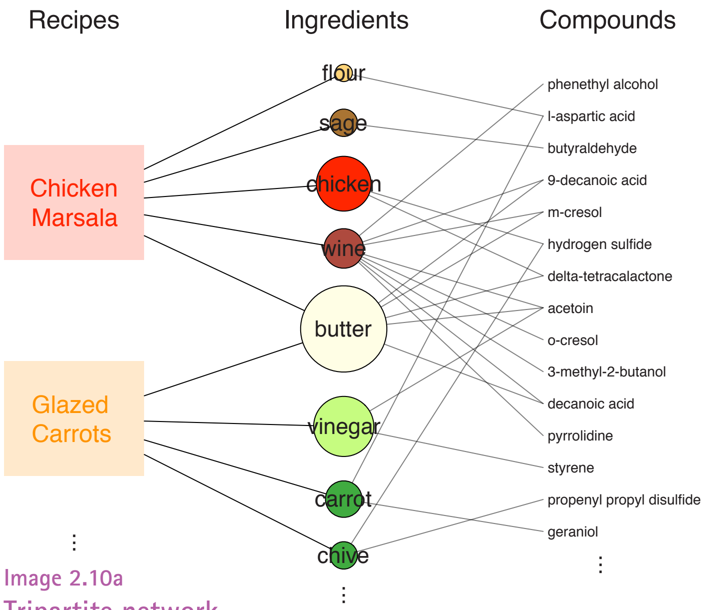


Image 2.10a

Tripartite network.

The tripartite recipe-ingredient-compound network, in which one set of nodes are recipes, like Chicken Marsala, the second set corresponds to the ingredients each recipe has (like flour, sage, chicken, wine, and butter for Chicken Marsala), and the third set captures the flavor compounds, or chemicals that contribute to the taste of a particular ingredient.

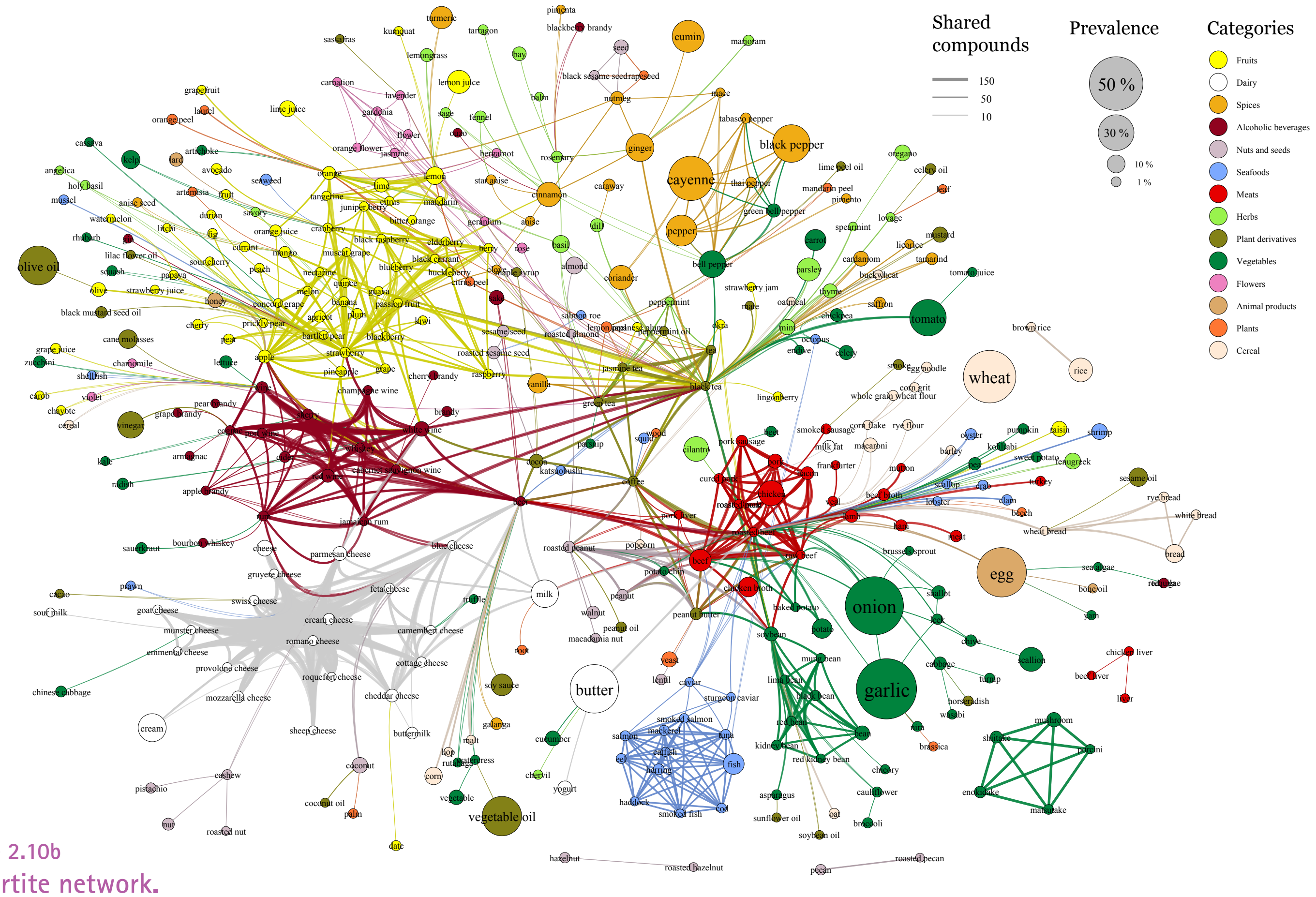


Image 2.10b

Tripartite network.

A projection of the tripartite network, resulting in the ingredient network, often called the flavor network. Each node denotes an ingredient; the node color indicating the food category and node size reflects the ingredient prevalence in recipes. Two ingredients are connected if they share a significant number of flavor compounds, link thickness representing the number of shared compounds between the two ingredients (After [12]).

SECTION 8

PATHS AND DISTANCES IN NETWORKS

In physical systems the components are characterized by obvious distances, like the distance between two atoms in a crystal, or between two galaxies in the universe. In networks distance is a challenging concept. Indeed, what is the distance between two webpages on the WWW, or two individuals who may or may not know each other? The physical distance is not relevant here: two webpages linked to each other could be sitting on computers on the opposite sides of the globe and two individuals, living in the same building, may not know each other. In networks physical distance is replaced by *path length*. A *path* is a route that runs along the links of the network, its length representing the number of links the path contains. A path can intersect itself and pass through the same link repeatedly (Image 2.5). In network science paths play a central role, hence next we discuss some of their most important properties, many more being summarized in Gallery 2.4.

Shortest Path (or geodesic path) between nodes *i* and *j* is the path with fewest number of links (Image 2.5). The shortest path is often called the *distance* between nodes *i* and *j*, and is denoted by d_{ij} , or simply *d*. We can often find multiple shortest paths of the same length d between a pair of nodes (Image 2.5). The shortest path never contains loops or intersects itself.

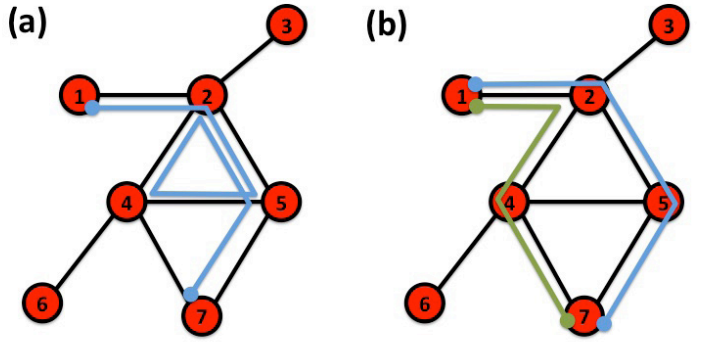


Image 2.11

The adjacency matrix is typically sparse.

(a) A path between nodes i_0 and i_n is an ordered list of *n* links $P_d = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \dots ,(i_n-1, i_n), \}$.The length of this path is d. The path shown in (a) follows the route 1→2→5→4→2→5→7, hence its length is n = 6.
(b) The shortest paths between nodes 1 and 7, representing the distance d_{17}, is the path with the fewest number of links that connect nodes 1 and 7. There can be multiple paths of the same length, as illustrated by the two paths shown in different colors. The network diameter is the largest distance in the network, being $d_{max} = 3$ here.

In an undirected network $d_{ij} = d_{ji}$, i.e. the distance between node *i* and *j* is the same as the distance between node *j* and *i*. In a directed network often dij ≠dji. Furthermore, in a directed network the existence of a path from node *i* to node *j* does not guarantee the existence of a path from *j* to *i*.

In real networks we frequently need to determine the distance between two nodes. For a small network, like the one shown in Image 2.5, this is an easy task. For a network of millions of nodes finding the shortest path between two nodes can be rather time consuming. The length of the shortest path and the number of such paths can be formally obtained from the adjacency matrix (Box 2.5). In prac-

Box 2.5

Number of shortest paths between two nodes.

The number of shortest paths, N_{ij} , between nodes *i* and *j* and the distance d_{ij} between them can be determined directly from the adjacency matrix, A_{ij} .

- $d_{ij} = 1$: If there is a link between *i* and *j*, then $A_{ij} = 1$ ($A_{ij} = 0$ otherwise).

- $d_{ij} = 2$: If there is a path of length two between *i* and *j*, then the product of d elements $A_{ik} A_{kj} = 1$ ($A_{ik} A_{kj} = 0$ otherwise). The number of $d_{ij} = 2$ paths between *i* and *j* is

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik} A_{kj} = \left[A^2\right]_{ij} \qquad (16)$$

where [...]*ij* denotes the *(ij)*th element of a matrix.

- $d_{ij} = d$: If there is a path of length d between *i* and *j*, then $A_{ik} \dots A_{lj} = 1$ ($A_{ik} \dots A_{lj} = 0$ otherwise). The number of paths of length d between i and j is

$$N_{ij}^{(d)} = \left[A^d\right]_{ij} \quad . \qquad (17)$$

Equation (17) holds for both directed and undirected networks and can be generalized to multigraphs as well. The distance between nodes i and j is the path with the smallest d for which $N_{ij}^{(d)} > 0$. Despite the mathematical elegancy of Eq. (17), faced with a large network, it is more efficient to use the breadth-first-search algorithm described in Box 2.6.

tice we most often use the breadth first search (BFS) algorithm discussed in Box 2.6 and Gallery 2.5 to measure the distance between two nodes.

Network diameter: the diameter of a network, denoted by d_{max}, is the maximal shortest path in the network. In other words, it is the largest distance recorded between any pair of nodes. One can verify that the diameter of the network shown in Image 2.5 is $d_{max} = 3$. For larger graphs the diameter can also be determined using the breadth first search algorithm (Box 2.6).

Average path length, denoted by ‹*d*›, is the average distance between all pairs of nodes in the network. For a directed network of *N* nodes, ‹*d*› is given by

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j=1,N} d_{i,j} \qquad (18)$$

For an undirected network we need to multiply the r.h.s. of Eq. (18) by two.

We can use the BFS algorithm to determine the average path length for a large network. For this we first determine the distance between a node and all other nodes in the network using the algorithm described in Box 2.6. We then determine the shortest path between a second node and all other nodes but the first one, a procedure that we repeat for all nodes. The sum of these shortest paths divided by L_{max} provides the average path length.

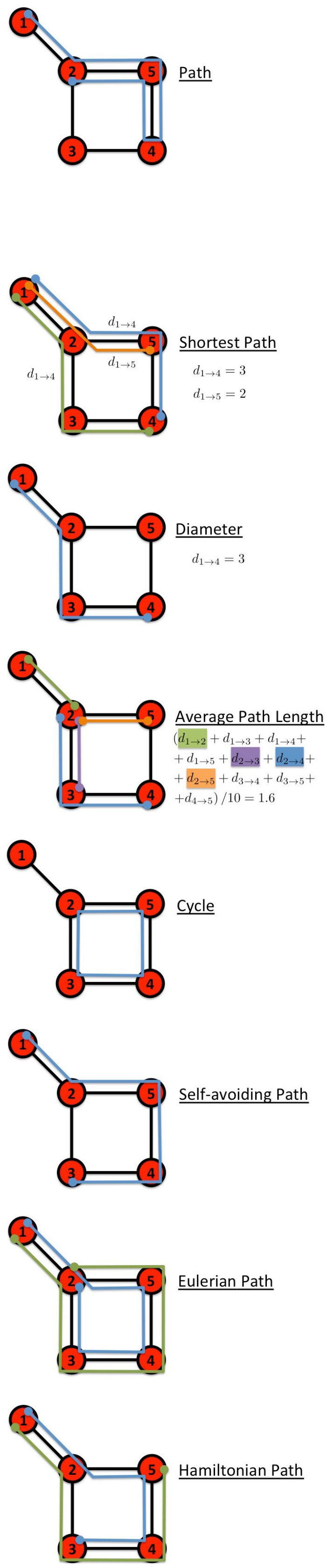


Image 2.12
Pathology.

PATH: A sequence of nodes such that each node is connected to the next node along the path by a link. A path always consists of *n* nodes and *n - 1* links. The length of a path is defined as the number of its links, counting multiple edges multiple times.

SHORTEST PATH (geodesic path, *d*): the path with the shortest distance *d* between two nodes. We will call it the distance between two nodes.

DIAMETER (d_{max}): the longest shortest path in a graph, or the distance between the two furthest away nodes.

AVERAGE PATH LENGTH (‹*d*›): the average of the shortest paths between all pairs of nodes.

CYCLE: a path with the same start and end node.

SELF-AVOIDING PATH: a path that does not intersect itself, i.e. the same node or link does not occur twice along the path.

EULERIAN PATH: a path that traverses each link exactly once.

HAMILTONIAN PATH: a path that visits each node exactly once.

Finding the shortest path: breath first search.

BFS is one of the most frequently used algorithms in network science. Similar to throwing a pebble in a pond and watching the ripples spread from the center, we start from a node and label its neighbors, then the neighbors' neighbors, until we encounter the target node. The number of "*ripples*" needed to reach the target provides the distance. To be specific, the identification of the shortest path between node i and j follows the following steps (Gallery 2.5):

1. Start at node *i*.

2. Find the nodes directly linked to *i*. Label them distance "1" and put them in a queue.

3. Take the first node, labeled n, out of the queue (*n = 1* in the first step). Find the unlabeled nodes adjacent to it in the graph. Label them with *n + 1* and put them in the queue.

4. Repeat step 3 until you find the target node j or there are no more nodes in the queue.

5. The distance between *i* and *j* is the label of *j*. If j does not have a label, then $d_{ij} = \infty$.

The time complexity of the BFS algorithm, representing the approximate number of steps the computer needs to find dij on a network of *N* nodes and *L* links, is O (*N* + *L*). It is linear in *N* and *L* as each node needs to be entered and removed from the queue at most once, and each link has to be tested only once.

Box 2.6

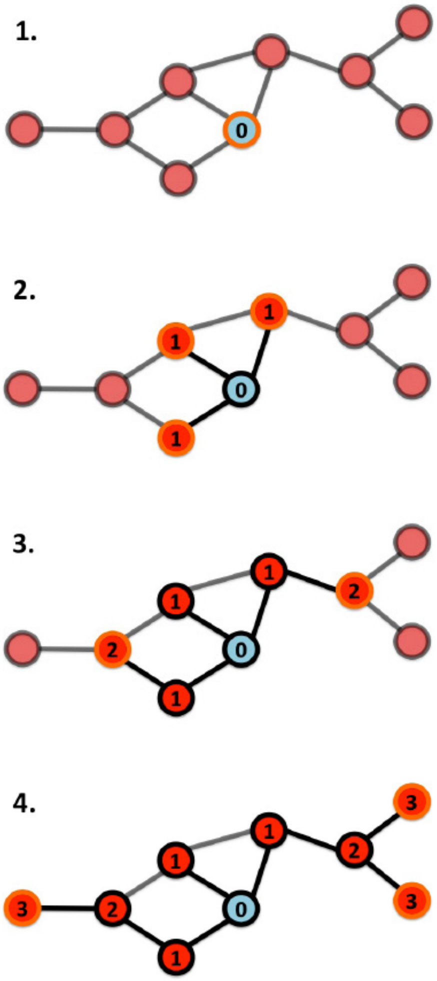


Image 2.13

The BFS algorithm applied to a small network.

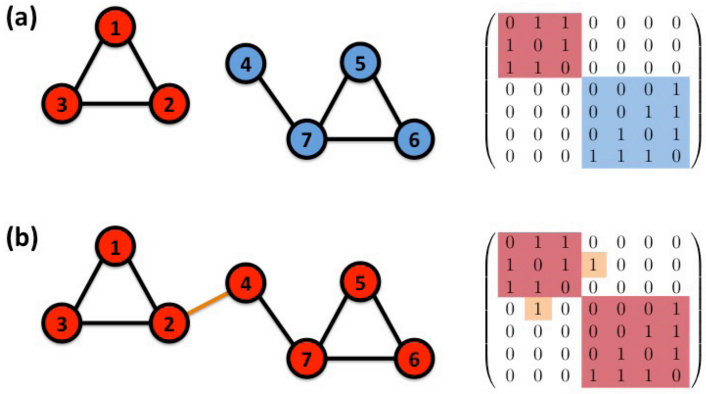
Starting from the orange node, labeled "0", we identify all its neighbors, labeling them "1". Then we label "2" the unlabeled neighbors of all nodes labeled "1", and so on, in each iteration increasing the labels, until no node is left unlabeled. The length of the shortest path or the distance d_{0i} between node 0 and some other node i in the network is given by the label on node i. For example, the distance between node 0 and the leftmost node is $d_{03} = 3$.

SECTION 9

CONNECTEDNESS AND COMPONENTS

The phone would be of limited use as a communication device if we could not call any valid phone number; the email world be rather useless if we could send emails to only certain email addresses, and not to others. From a network perspective this means that the technology behind the phone or the Internet must be capable of establishing a path between any two devices or clients, like your phone and any other phone on the network or between yours and your acquaintance's email address. This is in fact the key utility of most networks: they are built to ensure connectedness. In this section we discuss the graph-theoretic formulation of connectedness.

In an undirected network two nodes *i* and *j* are connected if there is a path between them on the graph. They are disconnected if such a path does not exist, in which case we have $d_{ij} = \infty$. This is illustrated in Image 2.14a, which shows a network consisting of two disconnected clusters. While there are paths between the nodes that belong to the same cluster (for example nodes 4 and 6), there are no paths between nodes that belong to different clusters (for example nodes 1 and 6).



$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Image 2.14

Connected and disconnected networks.

(a) The network consists of two disconnected components, i.e. there is a path between any pair of nodes in the (1,2,3) component, as well in the (4,5,6,7) component. However, there are no paths between nodes that belong to different connected components. The right panel shows the adjacently matrix of the network. If the network consists of disconnected components, the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements of the matrix are contained in square blocks along the diagonal of the matrix and all other elements are zero.

(b) The addition of one link, called a *bridge*, can turn a disconnected network into a single connected component. Now there is a path between every pair of nodes in the network. Consequently the adjacency matrix cannot be written in a block diagonal form.

A network is connected if all pairs of nodes in the network are connected. It is disconnected if there is at least one pair with $d_{ij} = \infty$. Clearly the network shown in Image 2.6a is disconnected, and we call its two subnetwork components (or clusters). A component is a subset of nodes in a network, so that there is a path between any two nodes that belong to the component, but one cannot add any more nodes to it that would have the same property. If a network consists of two components, a properly placed single link can connect them, making the network connected (Image 2.14b). Such a link is called a bridge. In general a bridge is any link that, if cut, disconnects the graph.

While for a small network visual inspection can help us decide if it is connected or disconnected, for a network consisting of millions of nodes connectedness is a challenging question. Several mathematical tools help us identify the connected components of a graph:

- For a disconnected network the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements in the matrix are contained in square blocks along the matrix' diagonal and all other elements are zero (Image 2.14a). Each square block will correspond to a component. We can use the tools of linear algebra to decide if the adjacency matrix is block diagonal, helping us to identify the connected components.

- In practice, for large networks the components are more efficient identified using the breadth first search algorithm (Box 2.7).

Box 2.7

Finding the connected components of a graph.

- 1. Start from a randomly chosen node i and perform a BFS from this node (Box 2.6). Label all nodes reached this way with *n = 1*. By linking friends to each other, we obtain the *friendship network*, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.

- 2. If the total number of labeled nodes equals *N*, then the network is connected. If the number of labeled nodes is smaller than *N*, the network consists of several components. To identify them, proceed to step 3.

- 3. Increase the label $n \longrightarrow n + 1$. Choose an unmarked node *j*, label it with *n*. Use BFS to find all nodes reachable from *j*, label them with *n*. Return to step 2.

SECTION 10

CLUSTERING COEFFICIENT

The local clustering coefficient captures the degree to which the neighbors of a given node link to each other. For a node i with degree k_i the local clustering coefficient is defined as [5].

$$C_i = \frac{2L_i}{k_i(k_i-1)} \qquad (19)$$

where L_i represents the number of links between the k_i neighbors of node *i*. Note that C_i is between 0 and 1:

- $C_i = 0$ if none of the neighbors of node i link to each other;

- $C_i = 1$ if the neighbors of node i form a complete graph, i.e. they all link to each other (Image 2.7).

- In general C_i is the probability that two neighbors of a node link to each other: C = 0.5 implies that there is a 50% chance that two neighbors of a node are linked.

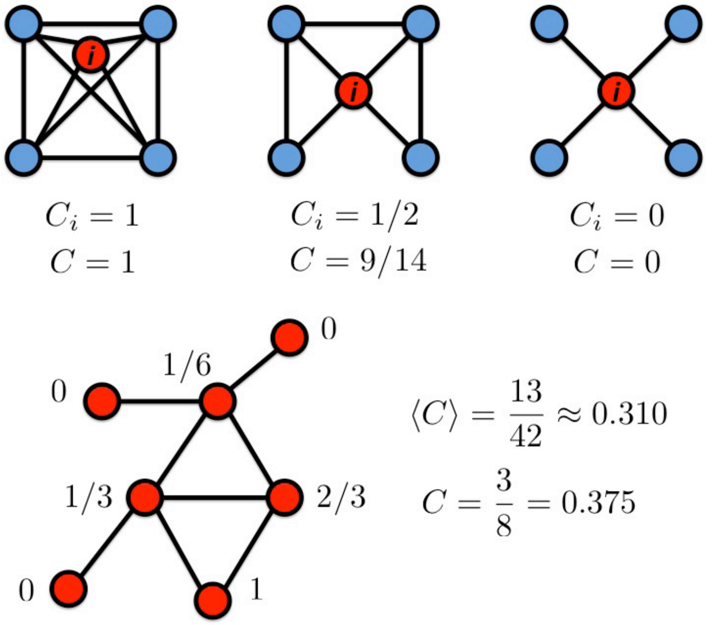
- In summary C_i measures the network's local density: the more densely interconnected the neighborhood of node i, the higher is C_i.

The degree of clustering of a whole network is captured by the *average clustering coefficient*, <C>, representing the average of C_i over all nodes $i = 1, ..., N$ [5],

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i \quad . \qquad (20)$$

In line with the probabilistic interpretation <*C*> is the probability that two neighbors of a randomly selected node link to each other.

While Eq. (19) is defined for undirected networks, the clustering coefficient can be generalized to directed and weighted [6,7,8,9]) networks as well. Note that in the network literature one also often encounters the *global clustering coefficient*, defined in Appendix A.



i
i
i
$C_i = 1$
$C = 1$
$C_i = 1/2$
$C = 9/14$
$C_i = 0$
$C = 0$
0
1/6
0
1/3
2/3
0
1
$\langle C \rangle = \frac{13}{42} \approx 0.310$
$C = \frac{3}{8} = 0.375$

Image 2.15

Clustering Coefficient.

The local clustering coefficient, C_i, of the central node with degree k_i=4 for three different configurations of its neighborhood. The clustering coefficient measures the local density of links in a node's vicinity. The bottom figure shows a small network, with the local clustering coefficient of a node shown next to each node. Next to the figure we also list the network's average clustering coefficient <*C*>, according to Eq. (20), and its global clustering coefficient *C*, declined in Appendix A, Eq. (21). Note that for nodes with degrees k_i=0,1, the clustering coefficient is taken to be zero.

SECTION 11

CASE STUDY AND SUMMARY

The purpose of the crash course in graph theory offered in this chapter was to familiarize us with some of the basic graph theoretical concepts and tools that network science uses. They define a set of elementary network characteristics, summarized in Image 2.16, that will serve as a language through which we can explore real networks. Yet, many of the networks we study in network science consist of hundreds to millions of nodes and links (Table 2.1). To explore them, we need to go beyond the small graphs discussed in Image 2.16 and use the introduced measures to explore large networks. A glimpse of what we are about to encounter is offered in Image 2.17a, where we show the protein-protein interaction network of baker's yeast, whose nodes are proteins, two proteins being connected if there is experimental evidence that they can bind (interact) to each other. The network is obviously too complex to understand its properties through a visual inspection of its wiring diagram. We therefore need to use the tools of network science to characterize its topology.

Let us use the measures we introduced so far to explore some basic characteristics of this network. The undirected network of Image 2.8a has N = 2,018 proteins as nodes and L=2,930 binding interactions as links. Hence the average degree, according to Eq. (7), is $\langle k \rangle$ = 2.90, suggesting that a typical protein interacts with approximately two to three other proteins. Yet, this number is somewhat misleading. Indeed, the degree distribution p_k shown in Image 2.17b indicates that the vast majority of nodes have only a few links. To be precise, in this network 69% of nodes have fewer than three links, i.e. for these $k < \langle k \rangle$. They coexist with a few highly connected nodes, or hubs, the largest having as many as 91 links. Such wide differences in node degrees is a consequence of the network's scale-free property, characterizing many real networks. We will see that the precise shape of the degree distribution determines a wide range of network properties, from the network's robustness to node failures to the spread of viruses.

The breath-first-search algorithm helps us determine the network's diameter, finding d_{max} = 14. We might be tempted to expect wide variations in d, as some nodes are close to each other, others, however, may be quite far. The distance distribution (Image 2.17c), indicates otherwise: pd has a prominent peak around $\langle d \rangle$ =5.61, indicating that most distances are rather short, being in the vicinity of $\langle d \rangle$. Also, p_d decays fast for large $\langle d \rangle$, suggesting that large distances are essentially absent. Instead, the variance of the degrees is σ_d = 1.64, hence we have d= 5.61 ± 1.64, i.e. most path lengths are in the clise vicinity of $\langle d \rangle$. These are manifestations of the small world property, another common feature of real networks, indicating that most nodes are rather close to each other.

The breath first search algorithm will also convince us that the protein interaction network is not connected, but consists of 185 components, shown as isolated clusters in Image 2.17a. The largest, called the giant component, contains 1,647 of the 2,018 nodes; all other components are tiny compared to it. As we will see in the coming chapters, such fragmentation is common in real networks.

The average clustering coefficient of the network is $<C>$ =0.12, which, as we will come to appreciate in the coming chapters, is rather large, indicating a significant degree of local clustering. A further caveat is provided by the dependence of the clustering coefficient on the node's degree, or the $C(k)$ function (Image 2.17d), which indicates that the clustering coefficient of the small nodes is significantly higher than the clustering coefficient of the hubs. This suggests that the small degree nodes are locates in dense local neighborhoods, while the neighborhood of the hubs is much more sparse. This is a consequence of network hierarchy, another widely shared network property.

Finally, a visual inspection reveals an interesting pattern: hubs have a tendency to connect to small nodes, giving the network a hub and spoke character. This is a consequence of degree correlations, which influence a number of network characteristics, from the spread of ideas and viruses in social networks to the number of driver nodes needed to control a network.

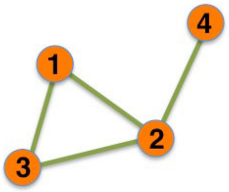
Taken together, Image 2.17 illustrates that the quantities we introduced in this chapter can help us diagnose several key properties of real networks. The purpose of the coming chapters is to study systematically these network characteristics, understanding what they tell us about the behavior of a complex system.

Image 2.16

Graphology.

In network science we encounter many networks distinguished by some elementary property of the underlying graph. Here we summarize the most commonly encountered elementary network types, together with their basic properties, and an illustrative list of real systems that share the particular property. Note that in many real network we need to combine several of these elementary network characteristics. For example the WWW is a directed multi-graph with self-interactions. The mobile call network is directed and weighted, without self-loops.

Undirected



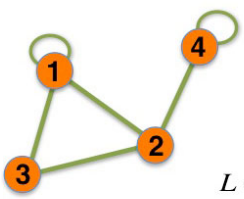
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad <k>= \frac{2L}{N}$$

UNDIRECTED NETWORK: a network whose links do not have a predefined direction. Examples: Internet, power grid, science collaboration networks, protein interactions.

Self-interactions



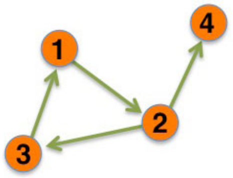
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1,i\neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii} \qquad ?$$

SELF-INTERACTIONS: in many networks nodes do not interact with themselves, so the diagonal elements of adjacency matrix are zero, *$A_{ii} = 0$, $i = 1,...,N$.* In some systems self-interactions are allowed; in such networks, representing the fact that node *i* has a self-interaction. Examples: WWW, protein interactions.

Directed



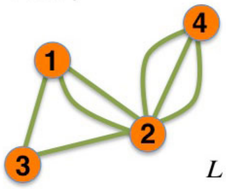
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^{N} A_{ij} \qquad <k>= \frac{L}{N}$$

DIRECTED NETWORK: a network whose links have selected directions. Examples: WWW, mobile phone calls, citation network.

Multigraph
(undirected)



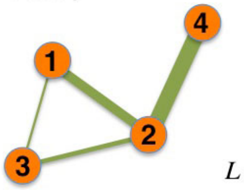
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad <k>= \frac{2L}{N}$$

MULTIGRAPH: in a multigraph nodes are permitted to have multiple links (or parallel links) between them. Hence A_{ii} can have any positive integer.

Weighted
(undirected)



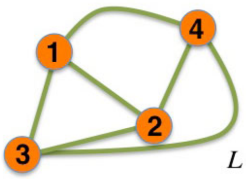
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum^{N} nonzero(A_{ij}) \qquad <k>= \frac{2L}{N}$$

WEIGHTED NETWORK: a network whose links have a predefined weight, strength or fow parameter. The elements of the adjacency matrix are $A_{ij} = 0$ if *i* and *j* are not connected, or $A_{ij} = w_{ij}$ if there is a link with weight wij between them. For unweighted (binary) networks, the adjacency matrix only indicates the presence ($A_{ij} = 1$) or the absence ($A_{ij} = 0$) of a link between two nodes. Examples: Mobile phone calls, email network.

Complete Graph
(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{i\neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \qquad <k>= N-1$$

COMPLETE GRAPH: in a complete graph all nodes are connected to each other; no self-connections.

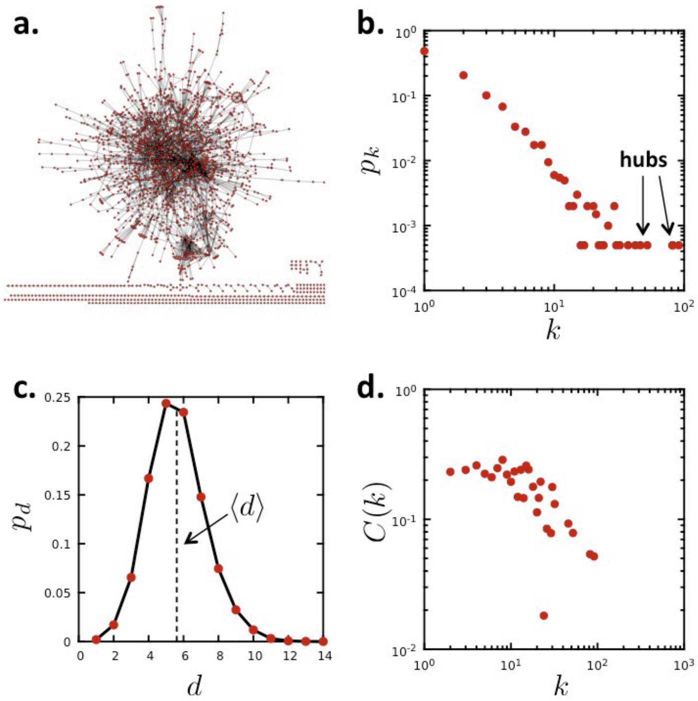


Image 2.16

Characterizing a real network.

(a) The protein-protein interaction (PPI) network of yeast, a network frequently studied not only by biologists, but also by network scientists. The nodes of the network are proteins and links correspond to experimentally documented protein-protein binding interactions. The figure indicates that the network, consisting of *N*=2,018 nodes and *L*=2,930 links, has a giant component that connects 81% of the proteins, several smaller components, and numerous isolated proteins that do not interact with any other node.

(b) The degree distribution, p_k, of the PPI network, providing the probability that a randomly chosen node has degree *k*. As $N_k= Np_k$, the degree distribution provides the number of nodes with degree *k*. The degree distribution indicates that proteins of widely different degrees coexist in the PPI: most nodes have only a few links, a few, however, have dozens of links, representing the hubs of the network.

(c) The distance distribution, pd for the PPI network, providing the probability that two randomly chosen nodes have a distance d between them (shortest path). The dotted line shows the average path length, which is *‹d› =5.61*.

(d) The dependence of the average clustering coefficient on the node's degree, *k*. The *C(k)* function is measured by averaging over the local clustering coefficient of all nodes with the same degree *k*.

SECTION 12

ADVANCED TOPICS: GLOBAL CLUSTERING COEFFICIENT

In the network literature one often encounters the *global clustering coefficient*, which measures the total number of closed triangles in a network. Indeed, L_i in Eq. (19) is the number of triangles that node *i* participates in, as each link between two neighbors of node *i* closes a triangle (Image 2.15). Hence the degree of a network's global clustering is captured by the global *clustering coefficient*, defined as

$$C = \frac{3 \times NumberOfTriangles}{NumberOfConnectedTriples} \qquad (21)$$

where a *connected triplet* consists of three nodes that are connected by two (open triplet) or three (closed triplet) undirected links. For example, an *A, B, C* triangle is made of three triples, *ABC*, *BCA* and *CAB*. In contrast a chain of connected nodes *A, B, C*, in which B connects to *A* and C but *A* does not link to C forms a single open triplet. The factor of three in the denominator of Eq. (21) is due to the fact that each triangle is counted tree times in the triple count. The roots of the global clustering coefficient go back to the social network literature of the 1940s [10,11], hence *C* is often called the *number of transitive triplets*.

Note that the average clustering coefficient <*C*> defined in (20) and the global clustering coefficient defined in (21) are not equivalent.

Indeed, take a network that is a double star consisting of N nodes, where nodes 1 and 2 are joined to each other and to all other vertices, and there are no other links. Then the local clustering coefficient C_i is 1 for i ≥ 3 and $2/(N - 1)$ for *i = 1, 2*. It follows that the average clustering coefficient of the network is $<C> = 1–O(1)$, while the global clustering coefficient gives $C \sim 2/N$. In less extreme networks the definitions will give more comparable values, but they will still differ from each other [13]. For example, in Image 2.15 we have $<C> = 0.31$ while $C = 0.375$.

SECTION 13

BIBLIOGRAPHY

[1] A.-L. Barabási and R. Albert. *Emergence of scaling in random networks*. Science, 286(5439):509–512, 1999.

[2] G. Gilder. *Metcalfe's law and legacy*. Forbes ASAP, 1993.

[3] B. Briscoe, A. Odlyzko, and B. Tilly. *Metcalfe's law is wrong*. IEEE Spectrum, 43(7):34–39, 2006.

[4] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. *The human disease network.* Proceedings of the National Academy of Sciences, 104(21):8685–8690, 2007.

[5] D. J. Watts and S. H. Strogatz. *Collective dynamics of 'small-world' networks*. Nature, 393(6684):440–442, 1998.

[6] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. *The architecture of complex weighted networks*. Proceedings of the National Academy of Sciences of the United States of America, 101(11):3747–3752, 2004.

[7] J. P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. *Intensity and coherence of motifs in weighted complex networks*. Phys. Rev. E, 71:065103, 2005.

[8] B. Zhang and S. Horvath. *A general framework for weighted gene coexpression network analysis*. Statistical Applications in Genetics and Molecular Biology, 4:17, 2005.

[9] P. Holme, S. M. Park, J. B. Kim, and C. R. Edling. *Korean university life in a network perspective: Dynamics of a large affiliation network.* Physica A: Statistical Mechanics and its Applications, 373(0):821–830, 2007.

[10] R. D. Luce and A. D. Perry. *A method of matrix analysis of group structure*. Psychometrika, 14:95–116, 1949.

[11] S. Wasserman and K Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

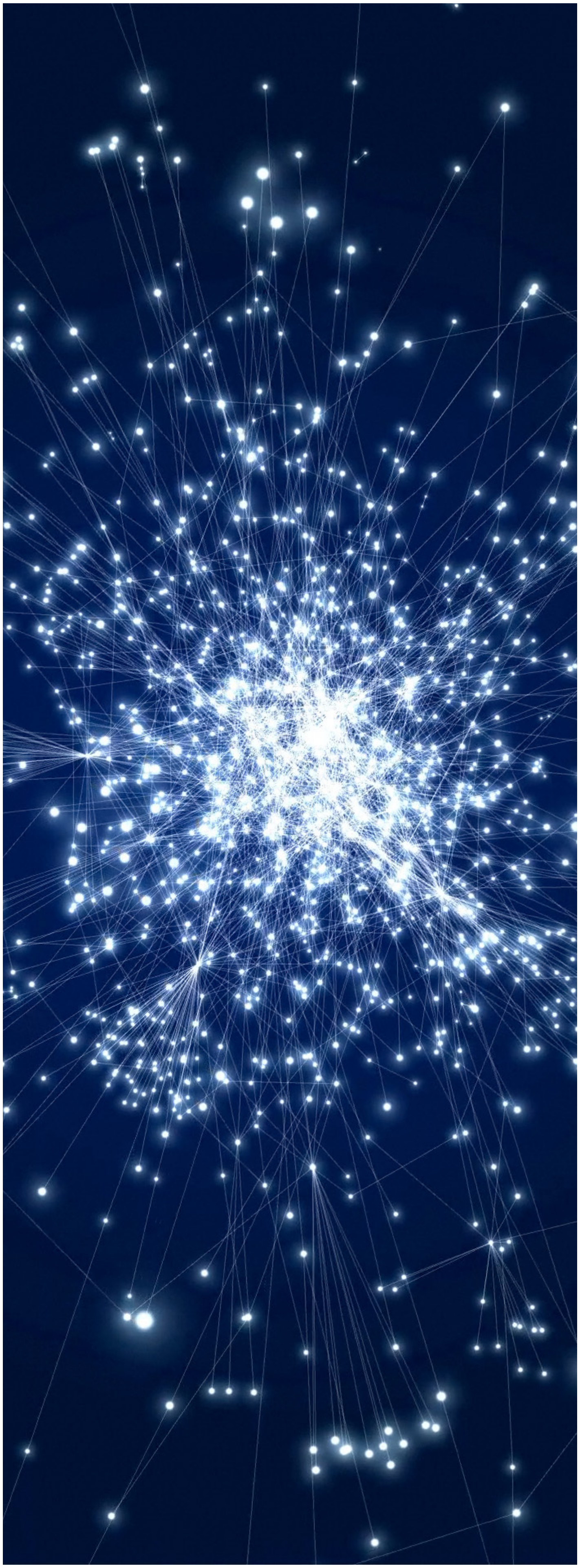
[12] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási, *Flavor network and the principles of food pairing*, Scientific Reports 196, 2011.

[13] B. Bollobás and O. M. Riordan. *Mathematical results on scale-free random graphs*, in Stefan Bornholdt, Hans Georg Schuster, *Handbook of Graphs and Networks: From the Genome to the Internet*, (2003 Wiley-VCH Verlag GmbH & Co. KGaA).

[14] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, *The human disease network*, Proceedings of the National Academy of Sciences 104:21, 8685 (2007).

CHAPTER 3

RANDOM NETWORKS



Introduction

The random network model

The number of links is variable

Degree distribution

Real networks do no not have a Poisson degree distribution

The evolution of a random network

Real networks are supercritical

The small world property

Clustering coefficient

Real networks are not random

Advanced Topics 3.A: Deriving the Poisson degree distribution

Advanced Topics 3.B: The maximum and the minimum degree

Advanced Topics 3.C: Giant component

Advanced Topics 3.D: Component sizes

Advanced Topics 3.E: Supercritical regime

Advanced Topics 3.F: Phase transitions

Advanced Topics 3.G: Correction to small worlds

Bibliography

SECTION 1

INTRODUCTION

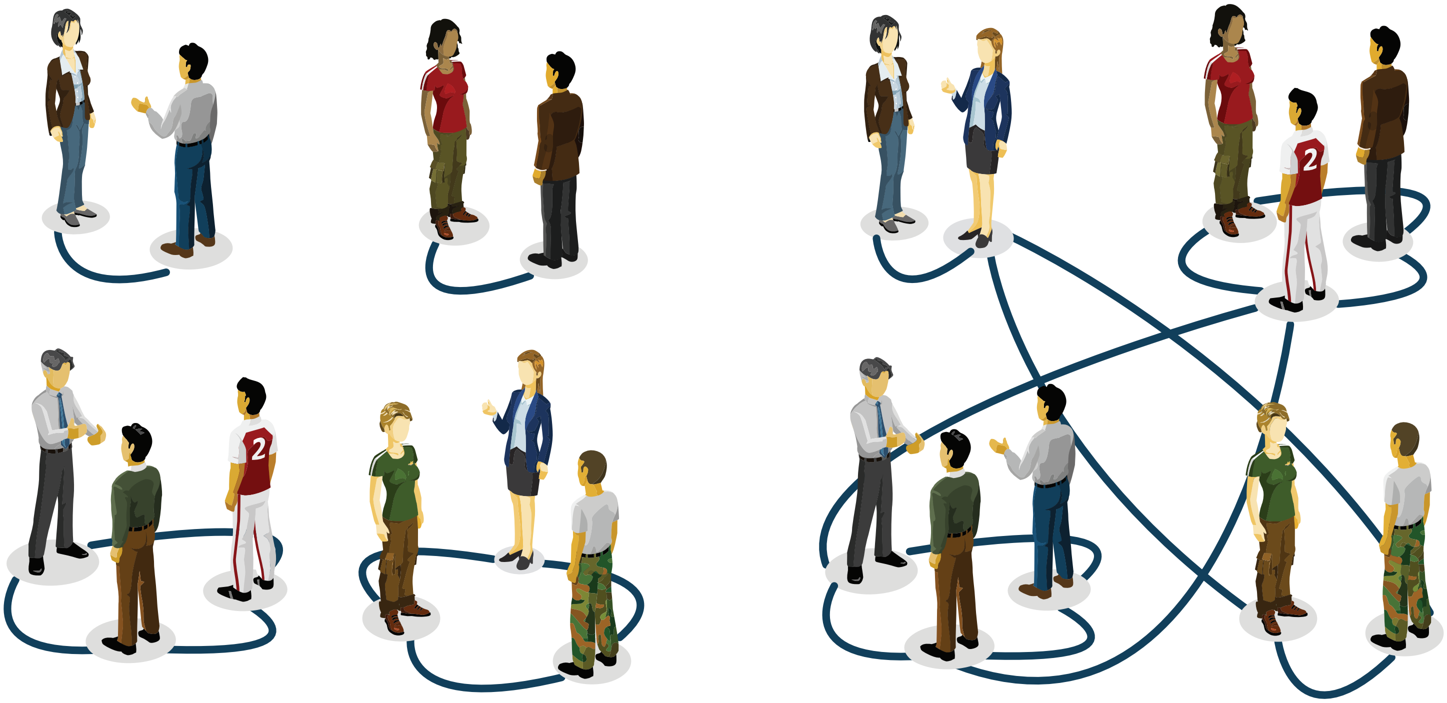


Image 3.1

From a cocktail party to random networks.

The emergence of an acquaintance network through random encounters at a cocktail party.

Imagine organizing a party for a hundred guests who initially do not know each other [1]. Offer them wine and cheese and you will soon have dozens of chatting groups of two to three. Now mention to Mary, one of your guests, that the red wine in the unlabeled dark green bottles is a rare vintage, much better than the one with the fancy red label. If she shares this information only with her acquaintances, you know that your expensive wine is safe, because she only had time to meet a few others in the room. However, the guests will continue to mingle, creating subtle paths between individuals that may still be strangers to each other. For example, while John has not yet met Mary, they have both met Mike, so now there is an invisible path from John to Mary through Mike. As time goes on, the guests will be increasingly interwoven by such intangible links. With that the secret of the unlabeled bottle will be pass from Mary to Mike and from Mike to John, slowly escaping into a rapidly expanding group.

To be sure, when all guests had gotten to know each other, everyone would be pouring the superior wine. But if each encounter took only ten minutes, meeting all ninety-nine others would take about sixteen hours. Thus, you could reasonably hope that a few drops of the better wine would be left for you to enjoy once the party is over.

Yet, you will be wrong. The purpose of this chapter is to show you why. We will see that the party maps into a classic model in network science called the random network model. And random network theory tells us that we do not have to wait until all individuals get to know each other for our expensive wine to be in danger. Rather, soon after each person meets at least one other guest, an invisible network will form that will allow the information to reach most guests. Hence in no time everyone will be drinking the better wine.

SECTION 2

THE RANDOM NETWORK MODEL

An important goal of network science is to build models that accurately reproduce the properties of real networks observed in real systems. Most networks we encounter in nature do not have the comforting regularity of a crystal lattice or the predictable radial architecture of a spider web. Rather, at first inspection most real networks look as if they were spun randomly. Random network theory embraces this apparent randomness by constructing networks that are *truly random*.

From a modeling perspective a network is a relatively simple object, consisting of only nodes and links. The real challenge, however, is to place the links between the nodes in a way to reproduce the complexity and apparent randomness of real systems. In this context the philosophy behind a random network is simple: it assumes that this goal is best achieved by placing the links randomly between the nodes. With that we arrive to the definition of a random network:

A random network consists of N labeled nodes where each node pair is connected with the same probability p.

To construct a random network, denoted with *G(N, p)* (Box 3.1):

1. Start with *N* isolated nodes.

2. Select a node pair, and generate a random number between 0 and 1. If the random number exceeds *p*, connect the selected node pair with a link, otherwise leave them disconnected.

3. Repeat step (2) for each of the *N(N-1)/2* node pairs.

The network obtained through this procedure is called a random graph or a random network. Two mathematicians, Pál Erdős and Alfréd Rényi, have played an important role in understanding the properties of random networks. In their honor a random network is often called the Erdős-Rényi network (Box 3.2).

Box 3.1

Two definitions of random networks.

There are two equivalent ways of defining a random network:

- **G(N,L) model:** *N* labeled nodes are connected with *L* randomly placed links. Erdős and Rényi (Erdős & Rényi, 1959) used this definition in their string of articles on random networks.

- **G(N,p) model:** Each pair of *N* labeled nodes is connected with probability *p*, a model introduced by Gilbert (Gilbert, 1959).

Hence the *G(N,p)* model fixes the probability *p* that two nodes are connected and the *G(N,L)* model fixes the total number of links *L*. While in the *G(N,L)* model the average degree of a node is simply $\langle k\rangle = 2L / N$, other network characteristics are easier to calculate in the *G(N, p)* model. Throughout this book we will explore the *G(N,p)* model, not only for the ease that it allows us to calculate key network characteristics, but also because its construction is closer to the way real systems develop. Indeed, in real networks the number of links is rarely fixed, but we can instead determine the probability that two nodes link to each other.

Box 3.2

A brief history of random networks.

Anatol Rapoport (1911-2007), a Russian immigrant to the United States, was the first to explore the properties of a random network. Trained as a pianist, Rapoport's interests turned to mathematics after realizing that a successful career as a concert pianist would require a wealthy patron. He became interested in mathematical biology at a time when mathematicians and biologists hardly spoke to each other. In a paper written with Ray Solomonoff in 1951 [28], Rapoport demonstrated that if we increase the average degree of a network, we will observe an abrupt transition from a collection of disconnected nodes to a state in which the graph contains a giant component. Despite its pioneering ideas, Rapoport's paper remains relatively unknown.

The study of random networks reached prominence thanks to the fundamental work of Pál Erdős and Alfréd Rényi. In a sequence of eight papers published between 1959 and 1968 [8-15], they merged probability theory and combinatorics with graph theory, establishing random graph theory, a new branch of mathematics [5].

The random network model was independently introduced by Gilbert [18] the same year Erdős and Rényi published their first paper on the subject. Yet, the impact of Erdős and Rényi's work is so overwhelming that they are rightly considered the fathers of random networks.

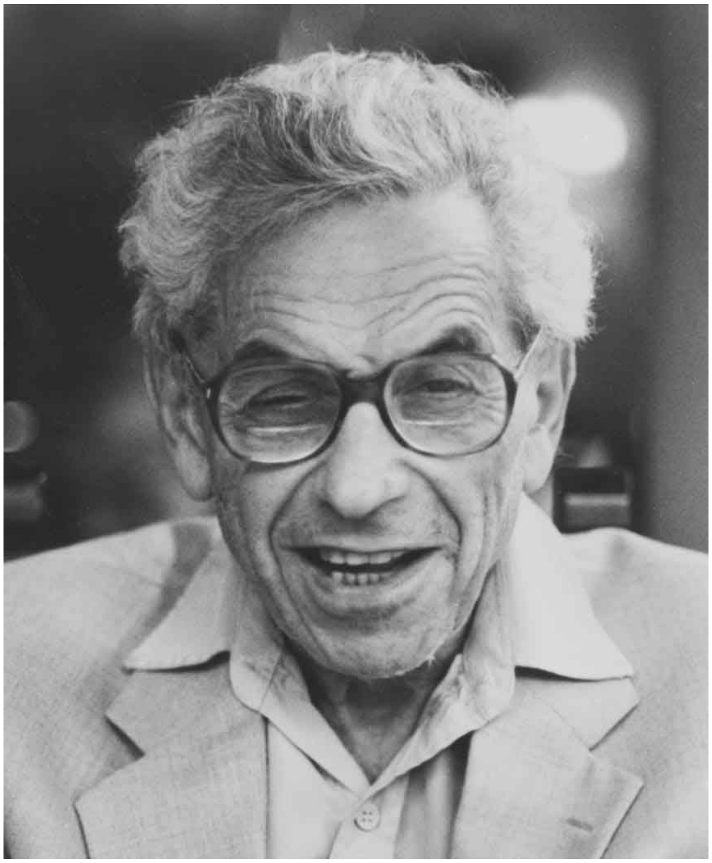


Image 3.2a

Pál Erdős (1913-1996)

Hungarian mathematician known for both his eccentricity and exceptional scientific output, having published more papers than any other mathematician in the history of mathematics. His productivity had its roots in his fondness for collaboration: he co-authored papers with over five hundred mathematicians, inspiring the concept of Erdős number. His legendarily personality and profound professional impact has inspired two biographies [19, 27] and a documentary [7].

Image 3.2b

Alfréd Rényi (1921-1970)

Hungarian mathematician with fundamental contributions to combinatorics, graph theory, and number theory. His impact goes beyond mathematics: the Rényi entropy is widely used in chaos theory and the random network model he co-developed is at the heart of network science. He is remembered through the hotbed of Hungarian mathematics, the Alfréd Rényi Institute of Mathematics in Budapest. He once said, *"A mathematician is a device for turning coffee into theorems"*, a quote often attributed to Erdős.

SECTION 3

THE NUMBER OF LINKS IS VARIABLE

Each random network we generate with the same parameters *N, p* will look slightly different (Image 3.3). Not only the detailed wiring diagram will vary between realizations, but so will the number of links *L*. It is useful, therefore, to determine how many links we expect for a particular realization of a random network with fixed *N* and *p*.

The probability that a random network has exactly *L* links is the product of three terms:

1. The probability that *L* of the attempts to connect the *N(N-1)/2* pairs of nodes have resulted in a link, which is p^L.

2. The probability that the remaining *N(N-1)/2 - L* attempts have not resulted in a link, which is $(1-p)^{N(N-1)/2-L}$

3. A combinational factor, $\left(\begin{pmatrix} N \\ 2 \end{pmatrix} \atop L\right)$ counting the number of different ways we can place *L* links among *N(N-1)/2* node pairs.

Hence the probability that a particular realization of a random graph has exactly *L* links is

$$p_L = \begin{pmatrix} \begin{pmatrix} N \\ 2 \end{pmatrix} \\ L \end{pmatrix} p^L (1-p)^{\frac{N(N-1)}{2}-L}. \qquad (1)$$

As Eq. (1) is a binomial distribution (Box 3.3), the expected number of links in a random graph can be calculated as

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p \frac{N(N-1)}{2}. \qquad (2)$$

Hence ‹*L*› is the product of the probability *p* that two nodes are connected and the number of pairs we attempt to connect, which is $L_{max} = N(N-1)/2$ (Chapter 2).
Using Eq. (2) we obtain the average degree of a random network as

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1). \qquad (3)$$

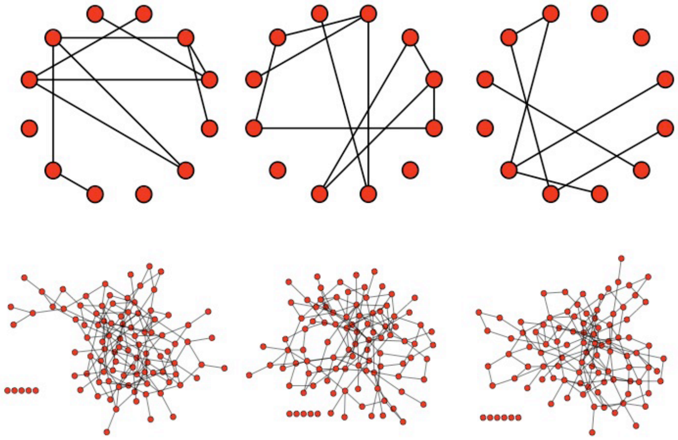


Image 3.3
Random networks are truly random.

Top row: Three realizations of a random network generated with the same parameters *N = 12* and *p =1/6*. Despite the identical parameters, the networks not only look different, but they differ in the number of links they have (*L = 8, 10, 7*) and in the degree of the individual nodes.
Bottom row: Three realizations of a random network with *N = 100* and *p = 1/6*.

Hence ‹*k*› is the product of the probability *p* that two nodes are connected and *(N-1)*, representing the maximum number of links a node can have in a network of size *N*.

In summary the number of links in a random network is not fixed, but varies between realizations. Its expected value is determined by *N* and *p*. If we increase *p* from *p = 0* to *p = 1* the random network becomes denser and the average number of links increase linearly from ‹*L*› = 0 to L_{max} and the average degree of a node increases from ‹*k*› = 0 to ‹*k*› = *N-1*.

Binomial distribution: Mean and variance.

If we toss a fair coin N times, tails and heads should occur with the same probability $p = 1/2$. The binomial distribution provides the probability p_x that we obtain exactly x heads in a sequence of N throws. In general, the binomial distribution describes the number of successes in N independent experiments with two possible outcomes, in which the probability of one outcome is p, and of the other is $1-p$.

The binomial distribution has the form

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}.$$

The mean of the distribution (first moment) is

$$\langle x \rangle = \sum_{x=0}^{N} x p_x = Np. \tag{4}$$

Its second moment is

$$\langle x^2 \rangle = \sum_{x=0}^{N} x^2 p_x = p(1-p)N + p^2N^2, \tag{5}$$

providing its standard deviation as

$$\sigma_x = \left(\langle x^2 \rangle - \langle x \rangle^2 \right)^{\frac{1}{2}} = [p(1-p)N]^{\frac{1}{2}}. \tag{6}$$

Box 3.3

SECTION 4

DEGREE DISTRIBUTION

As Image 3.3 illustrates, in a given realization of a random network some nodes are lucky, gaining numerous links, while others have only a few or no links. These differences are captured by the degree distribution p_k providing the probablity that a randomly chosen node has degree k.

In a random network the probability that node i has exactly k links is the product of three terms [5]:

- The probability that k of its links are present, or p^k.
- The probability that the remaining $(N - 1 - k)$ links are missing, or $(1-p)^{N-1-k}$.
- The number of ways we can select k links from $N - 1$ potential links a node can have, or $\binom{N-1}{k}$.

Hence the degree distribution of a random network follows the binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \qquad (7)$$

The shape of this distribution depends on the system size N and the probability p (Image 3.4). Using the properties of the binomial distribution (Box 3.3), from the degree distribution (7) we can calculate the network's average degree $\langle k \rangle$, recovering Eq. (3). We can also determine the second moment $\langle k^2 \rangle$ and the variance σ_k of the degree distribution (Image 3.4), quantities that will play an important role later.

Most real networks are sparse, hence $\langle k \rangle \ll N$ (Table 3.1, Image 3.4b). In this limit the degree distribution (7) is well approximated by the Poisson distribution (Advanced Topics 3. A)

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}, \qquad (8)$$

which is often called, together with (7), the degree distribution of a random network.

The binomial and the Poisson distribution describe the same quantity, hence they have several common properties (Image 3.4a):

- Both distributions have a peak around $\langle k \rangle$. If we keep N constant and increase p, the network becomes denser, increasing $\langle k \rangle$ and moving the peak to the right.
- The width of the distribution (dispersion) is also controlled by p or $\langle k \rangle$. The denser the network, the wider is the distribution, hence the larger are the differences in the degrees.

As we use the Poisson form in Eq. (8), we need to keep in mind that:

- The exact result for the degree distribution is the binomial form in Eq. (7), thus Eq. (8) represents only an approximation to (7) valid in the $k \ll N$ limit. For most networks of practical importance this condition is easily satisfied.
- The advantage of the Poisson form is that key network characteristics, like $\langle k \rangle$, $\langle k^2 \rangle$ and σ_k , have a much simpler form (Image 3.4a), depending on a single parameter, $\langle k \rangle$.
- The Poisson distribution in Eq. (8) does not explicitly depend on the number of nodes N . Therefore, Eq. (8) predicts that the degree distributions of networks of different sizes but the same average degree $\langle k \rangle$ are indistinguishable from each other (Image 3.4b).

Despite the fact that the Poisson distribution is only an approximation to the degree distribution of a random network, thanks to its analytical simplicity, it is the preferred form for p_k. Hence throughout this book, unless noted otherwise, we will refer to the Poisson form in Eq. (8) as the degree distribution of a random network.

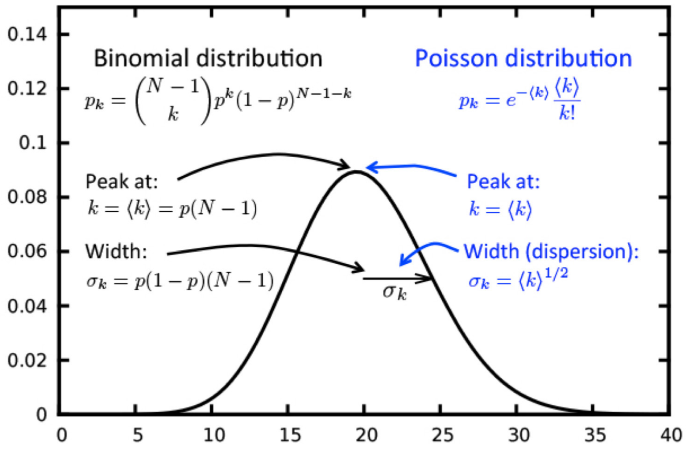


Image 3.4a

Anatomy of a binomial and a Poisson degree distribution.

The exact form of the degree distribution of a random network is the binomial distribution (left). For *N* » ‹*k*›, the binomial can be well approximated by a Poisson distribution (right). As both distributions describe the same quantity, they have the same properties, which are expressed in terms of different parameters: the binomial distribution uses *p* and *N* as its fundamental parameters, while the Poisson distribution has only one parameter, ‹*k*›.

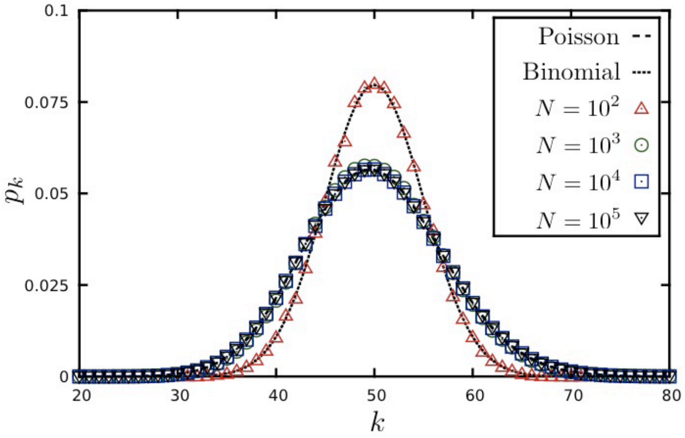


Image 3.4b

Degree distribution is independent of the network size.

The degree distribution of a random network with average degree *‹k› = 50* and sizes *N = 10*^{2} *, 10*^{3} *, 10*^{4}. For *N = 10*^{2} the degree distribution deviates significantly from the Poisson prediction (8), as the condition for the Poisson approximation, *N* » *‹k›*, is not satisfied. Hence for small networks one needs to use the exact binomial form of Eq. (7) (dotted line). For *N = 10*^{3} and larger networks the degree distribution becomes indistinguishable from the Poisson prediction, (8), shown as a continuous line, illustrating that for large *N* the degree distribution is independent of the network size. In the figure we averaged over *1,000* independently generated random networks to decrease the noise in the degree distribution.

REAL NETWORKS DO NO NOT HAVE A POISSON DEGREE DISTRIBUTION

The degree of a node in a random network can vary between *0* and *N-1*, raising an important question: How big are the differences between the node degrees in a particular realization of a random network? That is, can highly connected nodes, or hubs, coexist with small degree nodes? We address answer these questions by estimating the size of the largest and the smallest node in a random network.

Let us assume that the world's social network is described by the random network model. This may not be as far fetched hypothesis as it first sounds: there is significant randomness in whom we meet and whom we choose to become acquainted with. Sociologists estimate that a typical person knows about *1,000* individuals on a first name basis, suggesting that $\langle k \rangle \simeq 1{,}000$. Using the results obtained so far about random networks, we arrive to a number of surprising conclusions about a random society (see Advanced Topics 3.B):

- The most connected individual (the largest degree node) in a random society is expected to have degree $k_{max} = 1{,}185$.

- The least connected individual is expected to have degree $k_{min} = 816$.

- The dispersion of a random network is $\sigma_k = \langle k \rangle^{1/2}$, which for $\langle k \rangle = 1{,}000$ is $\sigma_k = 31.62$. This means that the number of friends of a typical individual should be mainly in the $\langle k \rangle \pm \sigma_k$ range, or between *970* and *1,030*, a rather narrow range.

In other words, in a random society everyone would have a comparable number of friends. We would lack outliers, or highly popular individuals, and no one would be left behind, having only a few friends. This calculation illustrates that in a *large random network the degree of most nodes is in the narrow vicinity of* $\langle k \rangle$ (Box 3.4).

Why hubs are absent in random network.

To understand why hubs are absent in random networks, we turn to the degree distribution (8). We first realize that the *1/k!* term in (8) significantly decreases the chances of observing large degree nodes. Indeed, the Stirling approximation

$$k! \sim \left[\sqrt{2\pi k}\right]\left(\frac{k}{e}\right)$$

allows us rewrite Eq. (8) as

$$p_k = \frac{e^{-\langle k \rangle}}{\sqrt{2\pi k}}\left(\frac{e\langle k \rangle}{k}\right)^k. \qquad (9)$$

For degrees $k > e\langle k \rangle$ the term in the parenthesis is smaller than one, hence for large k both k-dependent terms in (9), i.e. $1/\sqrt{k}$ and $(e\langle k \rangle/k)^k$ decrease rapidly with increasing k. Overall Eq. (9) predicts that in a random network *the chance of observing a hub decreases faster than exponentially.*

Box 3.4

This prediction blatantly conflicts with reality. Indeed, there is extensive evidence of individuals who have considerably more than *1,018* acquaintances. For example, US president Franklin Delano Roosevelt's appointment book had about *22,000* names in it, individuals he met personally [17, 26]. Similarly, a study of the social network behind Facebook has documented numerous individuals with *5,000* Facebook friends, the maximum allowed by the social networking platform [4]. The reason behind these systematic discrepancies can be understood by comparing the degree distribution of real and random networks.

In Image 3.5 we show the degree distribution of three real networks, together with the corresponding Poisson fits. The figure documents considerable differences between the random network predictions and the real data:

- The Poisson form significantly underestimates the number of high degree nodes. For example, according to the random network model the maximum degree for the Internet is expected to be around 20, while the data indicates the existence of nodes with degrees close to 10^3.

- The spread in the degrees of real networks is much wider than expected in a random network. This difference is captured by the dispersion σ_k (Image 3.4a). For example, if the Internet were to be random, we would expect $\sigma_k = 2.52$, while the measurements indicate $\sigma_{internet} = 14.14$, significantly higher than predicted.

These differences are not limited to the networks shown in Image 3.5, but all networks listed in Table 2.1 share this property. Hence the comparison with the real data indicates that the random network model does not capture the degree distribution of real networks. While in a random network most nodes have comparable degrees, forbidding hubs, in real networks we observe a significant number of highly connected nodes and large differences in node degrees. We will resolve these differences in Chapter 4.

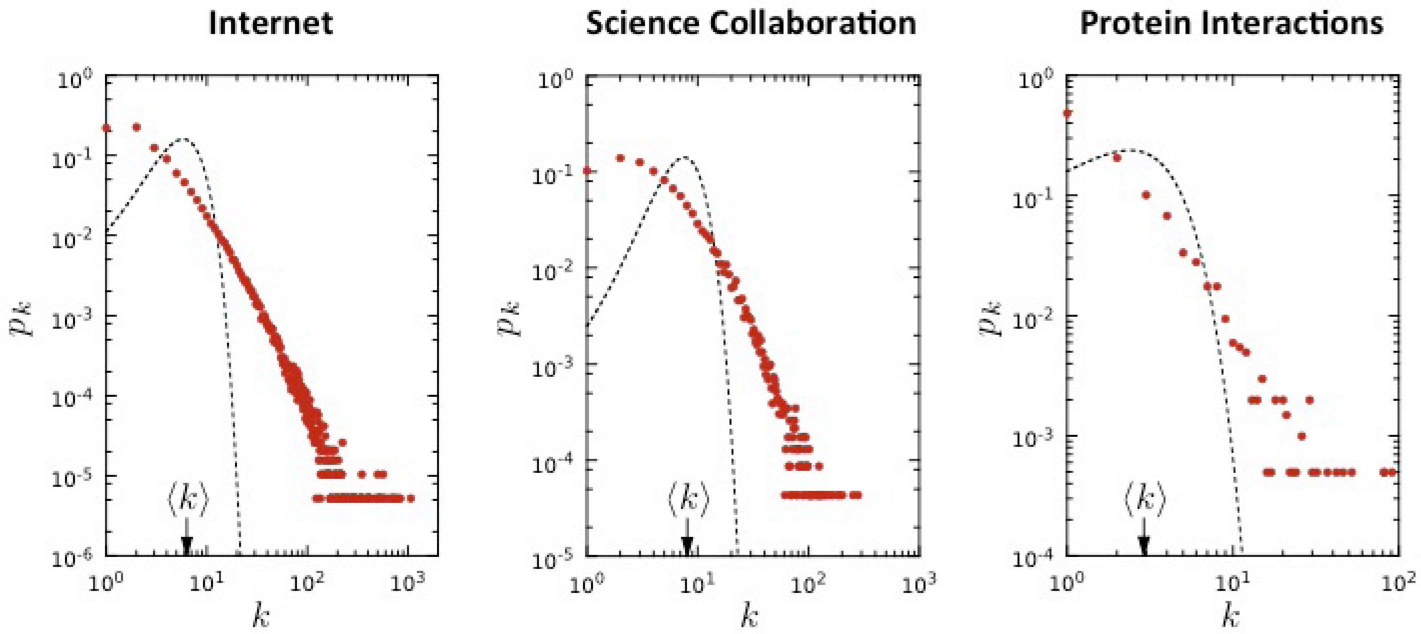


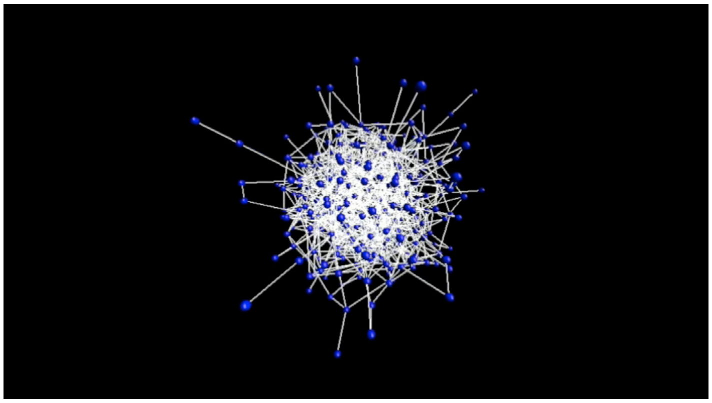
Image 3.5

Degree distribution of real networks.

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

SECTION 6

THE EVOLUTION OF A RANDOM NETWORK



Movie 3.1

Evolution of a random graph.

Changes in the structure of a random graph with increasing *p*, illustrating the absence of a giant component for small *p* and its sudden emergence once *p* exceeds a critical value.

The cocktail party we encountered at the beginning of the chapter captures a dynamical process: starting with *N* isolated nodes, the links are added gradually through random encounters between the guests. Within the random network model this corresponds to a gradual increase of *p*, with striking consequences on the network topology (Movie 3.1). To quantify this process, we first inspect how the size N_G of the *giant component*, which is the largest cluster within the network, varies with ‹*k*›. The two extreme cases are easy to understand:

- For $p = 0$ we have $\langle k\rangle = 0$, hence we observe only isolated nodes. Therefore $N_G = 1$ and $N_G/N \to 0$ for large *N*.

- For $p = 1$ we have $\langle k\rangle = N-1$, hence the network is a complete graph and all nodes belong to a single cluster. Therefore $N_G = N$ and $N_G/N = 1$.

One would expect that the giant component will grow gradually from $N_G = 1$ to $N_G = N$ if we increase ‹*k*› from *0* to *N-1*. Yet, as Image 3.6a indicates, this is not the case: N_G/N remains zero for small ‹*k*›, indicating the lack of a giant component for a range of ‹*k*› values. Once ‹*k*› exceeds a critical value, N_G/N increases rapidly, signaling the emergence of a giant component. Erdős and Rényi in their classical 1959 paper predicted that the *condition for the emergence of the giant component* is

$$\langle k\rangle = 1. \qquad (10)$$

In other words, we have a giant component if and only if when *each node has on average one link* (Advanced Topics 3.C).

The fact that at least one link per node is *necessary* for a giant component is not unexpected. Indeed, for a giant component to exist, each of its nodes must be linked to at least one other node. It is somewhat counterintuitive, however that one link is *sufficient* for its emergence.

If we wish to express Eq. (10) in terms of *p*, using Eq. (3) we obtain

$$p_c = \frac{1}{N-1} \approx \frac{1}{N}, \qquad (11)$$

indicating that the larger a network, the smaller *p* is sufficient for the giant component.

The emergence of the giant component is only one of the important transitions displayed by a random network. Changes in ‹*k*› allow us to distinguish four topologically distinct regimes (Image 3.6), each with its unique characteristics:

(a) Subcritical regime: 0 < ‹*k*› < 1, (p <$\frac{1}{N}$).

For ‹*k*› = *0* the network consists of *N* isolated nodes. Increasing ‹*k*› is equivalent with adding *N‹k› = pN(N-1)/2* links to the network. Yet, given the small number of links in the network in this regime, these links will mainly form clusters of size two (Image 3.6b). Upon increasing ‹*k*› further, some of the new links will join these pairs, forming tiny clusters. While we can designate at any moment the largest such cluster to be the giant component, in this regime the relative size of the largest cluster, N_G/N, remains zero. The reason is that for ‹*k*› < *1* the largest cluster is a tree

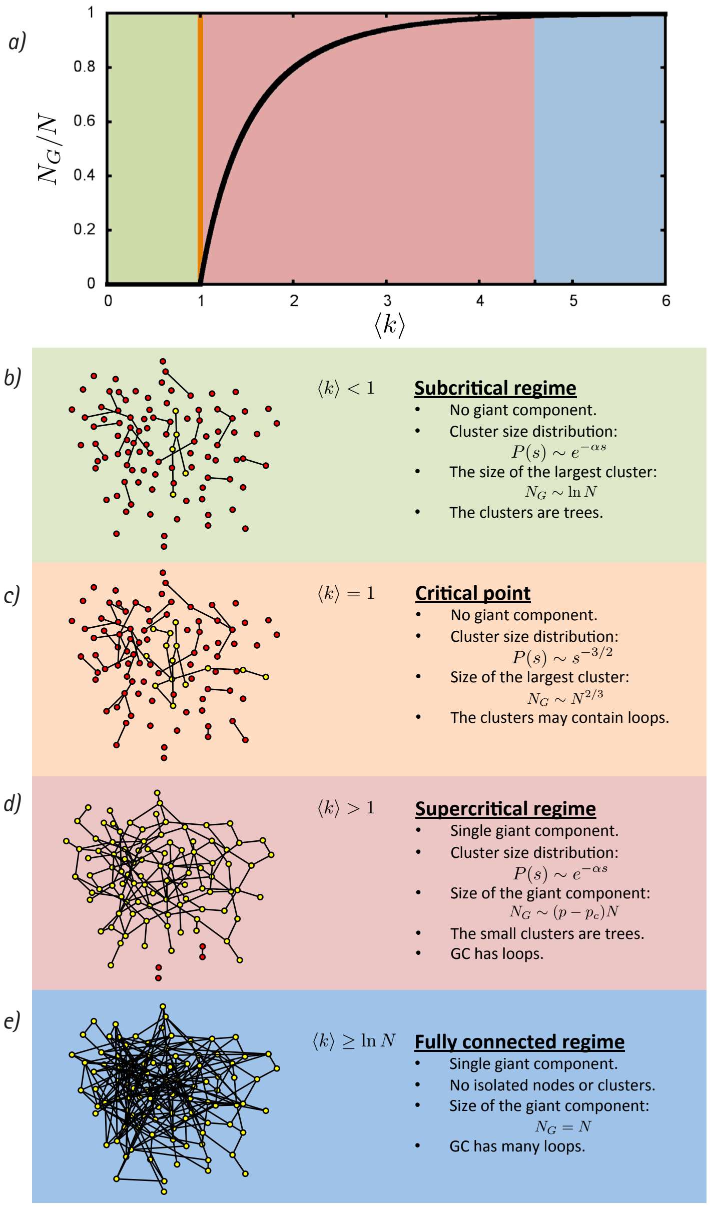


Image 3.6

Evolution of a random network.

(a) The relative size of the giant component in function of the average degree ‹*k*› in the Erdős-Rényi model.
(b)-(e) The main network characteristics in the four regimes that characterize a random network.

with size $N_G \sim \ln N$. Therefore $N_G/N \simeq \ln N/N \to 0$ in the $N \to \infty$ limit, indicating that the largest component is tiny compared to the size of the network.

In summary, in the subcritical regime the network consists of numerous tiny components, whose size follows an exponential distribution. Hence these components have comparable sizes, lacking a clear winner that we could designate as a giant component (Advanced Topics 3.D).

(b) Critical Point: ‹*k*› = 1, (p = $\frac{1}{N}$).

The critical point separates the regime where there is not yet a giant component *(‹k› < 1)* from the regime where there is one *(‹k› > 1)*. While it signals the emergence of the giant component, the relative size of the largest component in this point is still zero (Image 3.6c). Indeed, the calculations indicate that the size of the largest component is $N_G \sim N^{2/3}$, so its relative size decreases as $N_G/N \sim N^{-1/3}$, indicating that N_G is still tiny compared to the network's size.

In absolute terms there is a significant increase in the size of the largest component at ‹*k*› = *1*. For example, for a random network of $N = 7 \times 10^9$ nodes, the size of the globe's social network, for ‹*k*› < *1* the largest cluster is of the order of $N_G \simeq \ln N = \ln(7 \times 10^9) \simeq 22.7$. In contrast at ‹*k*› = *1* we expect $N_G \sim N^{2/3} = (7 \times 10^9)^{2/3} \simeq 3 \times 10^6$, a jump of about five orders of magnitude. Yet, both in the subcritical regime *(‹k› < 1)* and at the critical point *(‹k› = 1) the largest component contains a vanishing fraction of the total number of nodes in the network.*

Therefore most nodes are located in numerous small components, whose size distribution follows Eq. (36), a power law form indicating that components of rather different sizes coexist. These numerous small components are mainly trees, while the giant component may contain loops. Note that many properties of the network at the critical point resemble the properties of a physical system undergoing a phase transition (Advanced Topics 3.F).

(c) Supercritical regime: ‹*k*› > 1, (p > $\frac{1}{N}$).

This regime has the most relevance to real systems, as for the first time we have a giant component that looks like a network. In the vicinity of the critical point the size of the giant component varies as

$$N_G/N \sim \langle k \rangle - 1, \qquad (12)$$

or

$$N_G \sim (p - p_c)N, \qquad (13)$$

where p_c is given by Eq. (11). In other words, the *giant component contains a finite fraction of all nodes in the network*. The further we move from the critical point, a larger fraction of nodes will belong to it. Note that Eq. (12) is valid only in the vicinity of ‹*k*› = *1*, and for large ‹*k*› the dependence between N_G and ‹*k*› is nonlinear (Image 3.6d).
In the supercritical regime there are still numerous isolated components that coexist with the giant component, their size distribution being given by Eq. (35). These small

components are trees, while the giant component contains numerous loops and cycles. The supercritical regime lasts until all nodes are absorbed by the giant component.

(d) Connected regime: $\langle k \rangle \geq \ln N$, $(p \geq \frac{\ln N}{N})$.

For sufficiently large p the giant component will absorb all nodes and components, hence $N_G \simeq N$. In the absence of isolated nodes the network becomes connected. The average degree at which this happens depends on N as (Advanced Topic 3.E)

$$\langle k \rangle \sim \ln N. \qquad (14)$$

Note that when we enter the connected regime the network is still relatively sparse, as $\ln N / N \longrightarrow 0$ for large N. The network turns into a complete graph only at $\langle k \rangle = N - 1$.

In summary, the emergence of a network within the random network model is not a smooth process: the isolated nodes and tiny components observed for small $\langle k \rangle$ organize themselves into a giant component rather suddenly, through a process called phase transition (Advanced Topics 3.F). Along the way we encounter four topologically distinct regimes (Image 3.6). The discussion offered above follows an empirical perspective, fruitful if we wish to compare the observed networks to real systems. A different prospective, leading to it own rich behavior, is discussed in the mathematical literature (Box 3.5).

Network evolution in graph theory.

In the random graph literature it is often assumed that the connection probability $p(N)$ scales as N^z, where z is a tunable parameter between -∞ and 0. The greatest discovery of Erdős and Rényi was that as we vary z, key properties of random graphs appear quite suddenly. To be precise, a graph has a given property Q if the probability of having Q approaches 1 as $N \rightarrow \infty$. That is, for a given probability either almost every graph has the property Q or, almost no graph has it. For example, for z less than $-3/2$ almost all graphs contain only isolated nodes and edges.

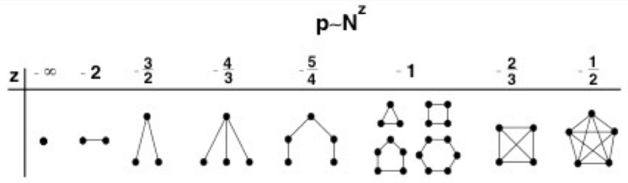


Image 3.7

Evolution of random graph.

The threshold probabilities at which different subgraphs appear in a random graph, as defined by exponent z in the $p(N) \sim N^z$ relationship. For $z < -3/2$ the graph consists of isolated nodes and edges. When z passes -3/2 trees of order 3 appear, while at $z = -4/3$ trees of order 4 appear. At $z = 1$ trees of all orders are present, together with cycles of all orders. Complete subgraphs of order 4 appear at $z =-2/3$, and as z increases further, complete subgraphs of larger and larger order emerge.

Box 3.5

SECTION 7

REAL NETWORKS ARE SUPERCRITICAL

Two predictions of random network theory are of special importance for real networks:

1. Once the average degree exceeds $\langle k \rangle = 1$, a giant component emerges that contains a finite fraction of all nodes. Hence only for $\langle k \rangle > 1$ the nodes organize themselves into a recognizable network.

2. For $\langle k \rangle > lnN$ all components are absorbed by the giant component, resulting in a single connected network.

But, do real networks satisfy the criteria for the existence of a giant component, i.e. $\langle k \rangle > 1$? And will this giant component contain all nodes, i.e. is $\langle k \rangle > lnN$, or do we expect some nodes and components to remain disconnected? These questions can be answered by comparing the measured $\langle k \rangle$ with the theoretical thresholds uncovered above.

Network	N	L	<k>	ln N
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	186,936	8.08	10.04
Actor Network	212,250	3,054,278	28.78	12.27
Yeast Protein Interactions	2,018	2,930	2.90	7.61

Table 3.1

Are real networks connected?

The number of nodes N and links L for several undirected networks, together with $\langle k \rangle$ and lnN. A giant component is expected for $\langle k \rangle > 1$ and all nodes should join the giant component for $\langle k \rangle \geq lnN$. While for all networks $\langle k \rangle > 1$, for most $\langle k \rangle$ is under the lnN threshold.

The measurements indicate that real networks extravagantly exceed the $\langle k \rangle = 1$ threshold. Indeed, sociologists estimate that an average person has around 1,000 acquaintances; a typical neuron is connected to dozens of other neurons, some to thousands; in our cells, each molecule takes part in several chemical reactions, some, like water, in hundreds. This conclusion is supported by Table 3.1, listing the average degree of several undirected networks, in each case finding $\langle k \rangle > 1$. Hence the average degree of real networks is well beyond the $\langle k \rangle = 1$ threshold, implying that they all have a giant component.

Let us now inspect if we have single component (if $\langle k \rangle > lnN$), or we expect the network to be fragmented into multiple components (if $\langle k \rangle < \ln N$). For social networks this would mean that $\langle k \rangle \geq \ln(7 \times 10^9) \simeq 22.7$. That is, if the average individual has more than two dozens acquaintances, then a random society would have a single component, leaving no node disconnected. With $\langle k \rangle \simeq 1{,}000$ this is clearly satisfied. Yet, according to Table 3.1 most real networks do not satisfy this criteria, indicating that they should consist of several disconnected components. This is a disconcerting prediction for the Internet, as it suggests that we should have routers that, being disconnected from the giant component, are unable to communicate with other routers. This prediction is at odd with reality, as these routers would be of little utility.

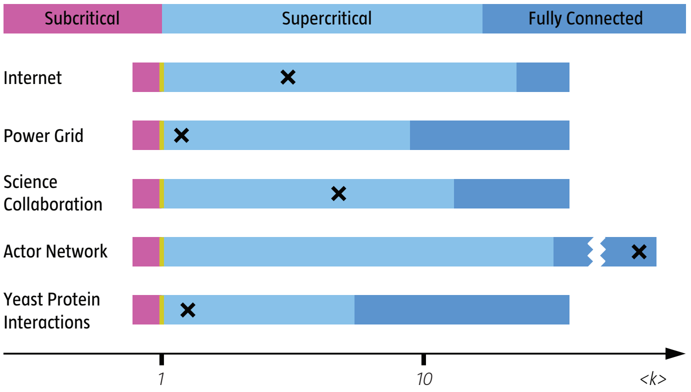


Image 3.8

Most real networks are supercritical.

The four regimes predicted by random network theory, marking with a cross the location of several real networks of Table 3.1. The diagram indicates that most networks are in the supercritical regime, hence they are expected to be broken into numerous isolated components. Only the actor network is in the connected regime, meaning that all nodes are expected to be part of a single giant component. Note that while the boundary between the subcritical and the supercritical regime is always at $\langle k \rangle = 1$, the boundary between the supercritical and the connected regimes is at lnN, hence varies from system to system.

Taken together, we find that most real networks are in the supercritical regime (Image 3.8). This means that these networks have a giant component, but it coexists with many disconnected components and nodes. This is true, however, only if real networks are accurately described by the Erdős-Rényi model, i.e. if real networks are random. In the coming chapters, as we learn more about the structure of real networks, we will understand why real networks can stay connected despite failing the $k > lnN$ criteria.

SECTION 8

SMALL WORLD PROPERTY

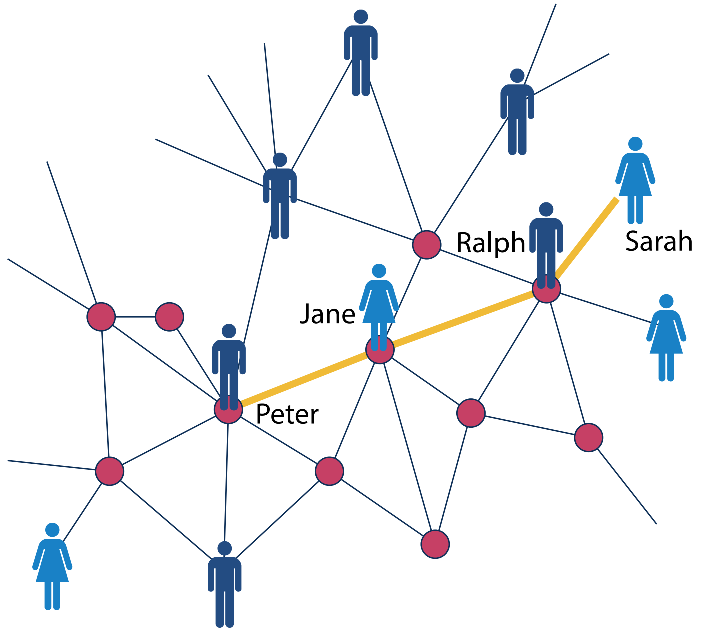


Image 3.9

Six degrees of separation.

According to six degrees of separation any two individuals, anywhere in the world, can be connected through a chain of six or fewer acquaintances. This means that while Sarah does not know Peter, she knows Ralph, who knows Jane and who in turn knows Peter. Hence Sarah is three degrees from Peter. In the language of network science six degrees, also called the small world property, states that the distance between any two nodes in a network is unexpectedly small.

Small world phenomena, also known as *six degrees of separation*, has long fascinated the general public. It states that if you choose any two individuals anywhere on earth, you will find a path of at most six acquaintances between them (Image 3.9). The fact that individuals who live in the same city are only a few handshakes from each other is by no means surprising. The small world concept goes further, however, stating that even individuals who are on the opposite side of the globe are six or fewer hand-shakes from us.

In the language of network science small world phenomena implies that *the distance between two randomly chosen nodes in a network is surprisingly short*. This statement raises two questions:

- What does short (or small) mean, i.e. short compared to what?

- How do we explain the existence of these short distances?

Both of these questions are answered by a simple calculation within the context of random networks. Consider a random network with average degree $\langle k \rangle$. A node in this network has on average:

$\langle k \rangle$ nodes at distance one ($d=1$).
$\langle k \rangle^2$ nodes at distance two ($d=2$).
$\langle k \rangle^3$ nodes at distance three ($d=3$).
...
$\langle k \rangle^d$ nodes at distance d.

For example, if $\langle k \rangle \simeq 1{,}000$, we expect 10^6 individuals at distance two and about a billion individuals, i.e. almost the whole earth's population, at distance three from us.

To be precise, the expected number of nodes up to distance d from our starting node is

$$N(d) \simeq 1 + \langle k \rangle + \langle k \rangle^2 + \ldots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}. \quad (15)$$

Yet, $N(d)$ must not exceed the total number of nodes, N, in the network. Therefore the distances cannot take up arbitrary values. We can identify a maximum distance d_{max} or the network's diameter at which $N(d)$ reaches N. By setting

$$N(d_{max}) \simeq N, \quad (16)$$

and assuming that $\langle k \rangle \gg 1$, we can neglect the (-1) term in both the nominator and denominator of Eq. (15), obtaining

$$\langle k \rangle^{d_{max}} \simeq N. \quad (17)$$

Therefore the diameter of a random network follows

$$d_{max} \propto \frac{\log N}{\log \langle k \rangle}, \quad (18)$$

which represents the *quantitative formulation of the small world phenomena*. The key, however is its interpretation:

- As derived, Eq. (18) predicts the scaling of the network diameter, d_{max}. Yet, for most networks Eq. (18) offers a better approximation to the average distance between two randomly chosen nodes, $\langle d \rangle$, than to d_{max} (Table 3.2). This is because d_{max} is often dominated by a few extreme paths, while $\langle d \rangle$ is averaged over all node pairs, a process that diminishes the fluctuations. Hence typically the small world property is defined by

$$\langle d \rangle \propto \frac{\log N}{\log\langle k \rangle}, \qquad (19)$$

describing the dependence on N and $\langle k \rangle$ of the average distance in a network.

- In general $\log N \ll N$, hence the dependence of $\langle d \rangle$ on $\log N$ implies that the distances in a random network are *orders of magnitude smaller than the size of the network.* Consequently small world phenomena implies that the average path length or the diameter depends logarithmically on the system size. Hence, "small" means that $\langle d \rangle$ is proportional to $\log N$, rather than N or some power of N (Image 3.10).

- The $1/\log \langle k \rangle$ term implies that the denser the network, the smaller is the distance between the nodes.

- In real networks there are systematic corrections to Eq. (18), rooted in the fact that the number of nodes at distance $d > \langle d \rangle$ drops rapidly (Advanced Topics 3.F).

Network Name	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\log N}{\log\langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.59
WWW	325,729	1,497,134	4.60	11.27	93	8.32
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	186,936	8.08	5.35	15	4.81
Actor Network	212,250	3,054,278	28.78	-	-	-
Citation Network	449,673	4,707,958	10.47	11.21	42	5.55
E Coli Metabolism	1,039	5,802	5.84	2.98	8	4.04
Yeast Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Table 3.2

Six degrees of separation.

The average distance $\langle d \rangle$ and the maximum distance d_{max} of the ten networks explored in this book. The last column provides $\langle d \rangle$ predicted by Eq. (19), indicating that it offers a reasonable approximation to $\langle d \rangle$. Yet, the agreement is not perfect - we will see in the next chapter that for many real networks Eq. (19) needs to be adjusted. For directed networks we list the average out-degree $\langle k_{out} \rangle$ and the path lengths are measured only along the direction of the links.

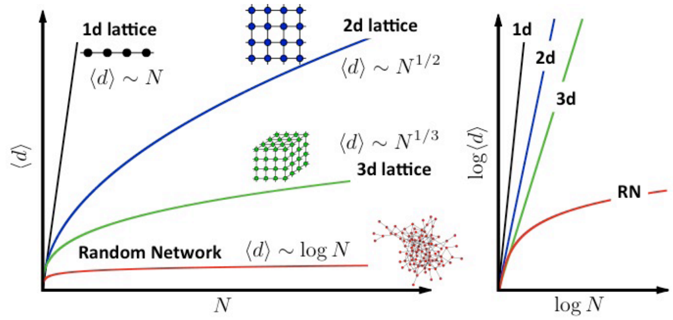


Image 3.10

Why are small worlds surprising?

Much of our intuition about distance is based on our experience with regular lattices, which do not display the small world phenomenon. Indeed,

- For a one-dimensional lattice (a line of length *N*) the diameter and the average path length scale linearly with *N*: d_{max}~‹*d*› ~*N*.
- For a square lattice d_{max}~‹*d*› ~ $N^{1/2}$.
- For a cubic lattice d_{max}~‹*d*› ~ $N^{1/3}$.
- In general, for a *d*-dimensional lattice we have d_{max} ~ ‹*d*› ~ $N^{1/d}$.

Such polynomial dependence predicts a much faster increase with *N* than Eq. (19), indicating that in regular lattices the path lengths are significantly longer than in a random network. The figure shows the predicted *N*-dependence of ‹*d*› for regular and random networks on a linear (left) and on a log-log (right) scale. If the social network would form a regular 2d lattice, where each individual knows only its nearest neighbors, the average distance between two individuals would be roughly *$(7\times10^9)^{1/2}$ = 83,666*. Even if we correct for the fact that a person has about 1,000 acquaintances, not four, the average separation will be orders of magnitude larger than predicted by Eq. (19).

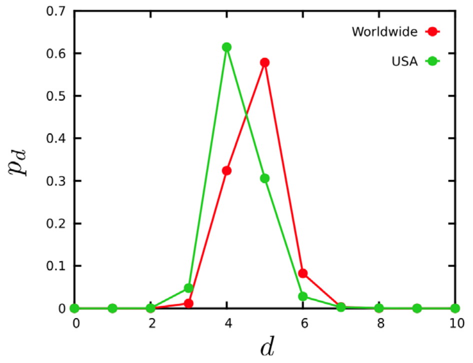


Image 3.11

Six degrees? Facebook finds only four.

Milgram's experiment could not detect the true distance between his study's participants, as he lacked an accurate map of the full social network. Today Facebook has the most extensive social network map ever assembled. Using Facebook's social graph of May 2011, consisting of 721 million active users and 68 billion symmetric friendship links, the average distance between the users was 4.74. The figure shows the distance distribution, p_d, for all pairs of Facebook users worldwide (full dataset) and within the US only. Therefore, instead of 'six degrees' researchers detected only 'four degrees of separation' [4], closer to the prediction of Eq. (20) than to Milgram's six degrees [23]. Using Facebook's *N* and *L* Eq. (19) predicts the average degree to be approximately 3.90, not far from the reported four degrees.

Let us illustrate the implications of Eq. (19) for social networks. Using *N*≃ *7 ×10⁹ and ‹k›≃10³*, we obtain

$$\langle d\rangle = \frac{\ln 7\times10^9}{\ln(10^3)} = 3.28. \qquad (20)$$

Therefore, all individuals on Earth should be within three to four handshakes of each other, about a half of "six degrees". The estimate (20) is probably closer to the real value given by Eq. (7) than the frequently quoted six degrees (Image 3.11).

While discovered in the context of social systems, the small world property applies beyond social networks. In Table 3.2 we compare the prediction of Eq. (19) with the average path length ‹*d*› for several real networks, finding that despite the diversity of these systems and the significant differences between them in terms of *N* and ‹*k*›, Eq. (19) offers a reasonable approximation to the empirically observed ‹*d*›.

The small world property has not only ignited the public's imagination, but plays an important role in network science as well. It affects most network characteristics, from the spread of ideas in social networks to search on networks. The small world phenomena can be reasonably well understood in the context of the random network model: it is rooted in the fact that the number of nodes at distance *d* from a node increases exponentially with *d*. While in the coming chapters we will see that in real networks we encounter systematic deviations from Eq. (19), forcing us to replace it with more accurate predictions, the intuition offered by the random network model on the origin of the phenomenon remains valid.

A BRIEF HISTORY OF SIX DEGREES

Image 3.12

Frigyes Karinthy (1887-1938)

Hungarian writer, journalist and playwright, the first to describe the small world property. He remains one of the most popular writers in Hungary. English translation of *Chains*, the 1929 short story describing the small world phenomena, is available in [25].

Image 3.13

Stanley Milgram (1933-1984)

American social psychologist known for his experiments on obedience and authority. He designed and carried out the small world experiment in 1967 as part of his Harvard dissertation.

Box 3.6a

The first description of small world phenomena goes back to a 1929 story collection entitled *Minden másképpen van* (Everything is Different) by the Hungarian writer **Frigyes Karinthy** [21]. In *Láncszemek* (Chains), a short story in the volume, Karinthy suggests that one could name any person among earth's one and a half billion inhabitants (estimated population in 1929) and through at most five acquaintances, one of which he knew personally, he could link to him. To demonstrate his thesis Karinthy links a Nobel Prize winner to himself, noting that the Nobelist must know King Gustav, the Swedish monarch who hands out the Nobel Prize, who in turn is a consummate tennis player and occasionally plays with a tennis champion who is one of Karinthy's good friends. Remarking that finding a chain of acquaintances to celebrities, like a Nobelist, is easy, he next links a worker in Ford's factory to himself: *"The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Árpád Pásztor, someone I not only know, but to the best of my knowledge a good friend of mine."*

Box 3.6b

The first experimental study of small world phenomena took place four decades after Karinthy, in 1967, when Stanley Milgram turned the idea into an experiment probing the structure of social networks [23]. Milgram chose a stock broker in Boston and a divinity student in Sharon, Massachusetts as "targets". Randomly selected residents of Wichita, Kansas and Omaha, Nebraska received a letter containing a short summary of the study's purpose, a photograph, the name, address and information about the target person. They were asked to forward the letter to a friend, relative or acquaintance, who is more likely to know the target person. Milgram wrote in 1969: *"I asked a person of intelligence how many steps he thought it would take, and he said that it would require 100 intermediate persons, or more, to move from Nebraska to Sharon."* Yet, within a few days the first letter arrived, passing through only two links. Eventually 42 of the 160 letters made it back, some requiring close to a dozen intermediates. These completed chains allowed Milgram to determine the number of individuals required to get the letter to the target. He found that the median number of intermediates was 5.5, a relatively small number and remarkably close to Karinthy's 1929 insight.

A BRIEF HISTORY OF SIX DEGREES

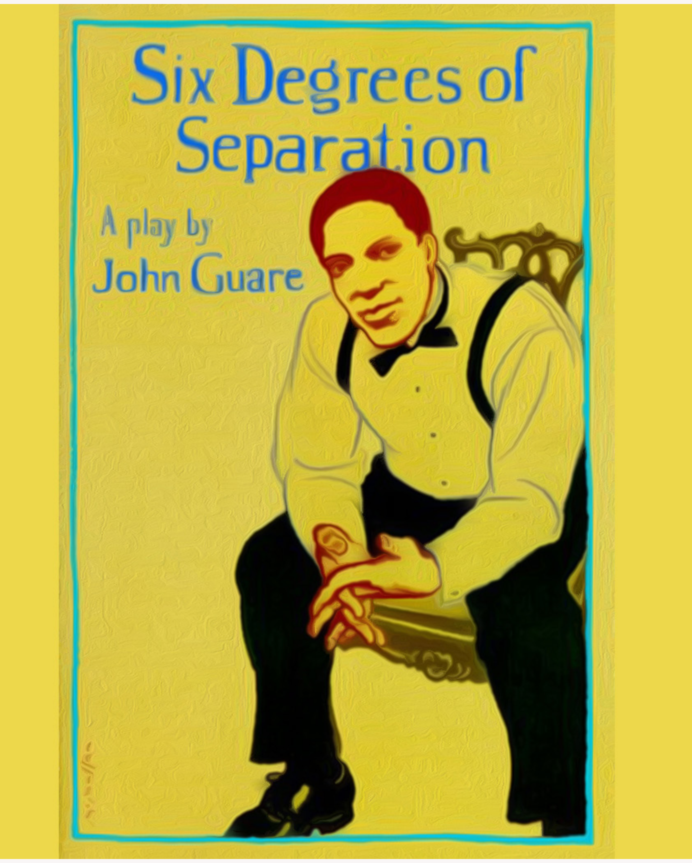


Image 3.14

Six Degrees of Separation.

Cover of John Guareís Six Degrees of Separation play, that helped turn six degrees into a catch phase of popular culture.

The phrase “six degrees of separation” was introduced in 1991 by the playwright **John Guare**, who used it as the title of his Broadway play, later turned into a movie. The play’s lead character, Ousa, musing about the world’s interconnectedness, tells her daughter:

“Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. It’s not just the big names. It’s anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am bound to everyone on this planet by a trail of six people. It’s a profound thought. How every person is a new door, opening up into other worlds.”

Milgram’s study was confined to the United States, linking individuals in Wichita and Omaha to Boston. Guare, however, with the sweep of a writer’s imagination, generalized six degrees to the whole planet, bringing it closer in spirit to Karinthy’s 1929 description. As more people watch movies than read sociology papers, Guare’s version prevailed in popular thought.

Box 3.6c

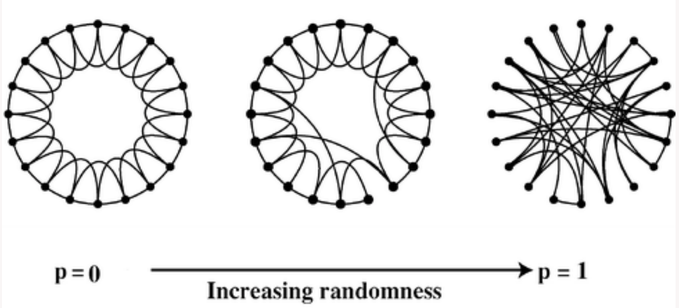


Image 3.15

Watts-Strogatz model.

The model starts from a ring of nodes, each node connected to their immediate and next neighbors, a configuration in which each node has clustering coefficient $C = 3/4$ (left, $p = 0$). With probability p each link is rewired to a randomly chosen node. For small p the network maintains a high average clustering coefficient but the random long-range links drastically decrease the distances between the nodes, inducing the small world effect (middle). For large p (right, $p = 1$) the network turns into a random network. (After [30]).

A new wave of interest in small worlds emerged following the 1998 study of Duncan Watts and Steven Strogatz, applied mathematicians working at Cornell [30]. They analyzed three real systems, the actor network of Hollywood, the neural network of the worm *C. elegans*, and the North American power grid, in each case finding that the average distance between the nodes is comparable to the random network prediction Eq. (19). Hence they found that the small world property applies to networks appearing in natural and technological systems as well. Watts and Strogatz also noted that these networks have a much higher clustering coefficient than expected for a random network, prompting them to propose a model to account for the coexistence of small path lengths and large clustering (Image 3.15). The model’s properties are discussed in detail in the chapter devoted to social networks.

Box 3.6d

SECTION 9

CLUSTERING COEFFICIENT

The local clustering coefficient C_i captures the density of links in node *i*'s immediate neighborhood: $C = 0$ means that there are no links between *i*'s neighbors; $C = 1$ implies that each of the i's neighbors link to each other (Sect. 2.10). To calculate C_i for a node in a random network we need to estimate the expected number of links L_i between the node's k_i neighbors. In a random network the probability that two of *i*'s neighbors link to each other is p. As there are $k_i(k_i - 1)/2$ possible links between the k_i neighbors of node *i*, the expected value of L_i is

$$\langle L_i \rangle = p\frac{k_i(k_i-1)}{2}.$$

Thus the local clustering coefficient of a random graph is

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i-1)} = p = \frac{\langle k \rangle}{N}. \qquad (21)$$

Equation (21) makes two predictions:

(a) For fixed ‹*k*›, the larger the network, the smaller is a node's clustering coefficient. Consequently the network's average clustering coefficient <*C*> is expected to decrease as $1/N$.

(b) The local clustering coefficient of a node is independent of the node's degree.

To test the validity of Eq. (21) we plot <*C*>/‹*k*› in function of *N* for several undirected networks (Image 3.16a). We find that <*C*>/‹*k*› does not decrease as N^{-1}, but it is largely independent of *N*, in violation of Eq. (21) . In Image 3.16b–d we also show the dependency of *C* on the node's degree k_i for three real networks, finding that *C(k)* systematically decreases with the degree, again in violation of Eq. (21) .

Taken together, we find that the random network model does not capture the local clustering of real networks. Instead real networks have a much higher clustering coefficient than expected for a random network of similar *N* and *L*, and high-degree nodes tend to have a smaller clustering coefficient than low-degree nodes.

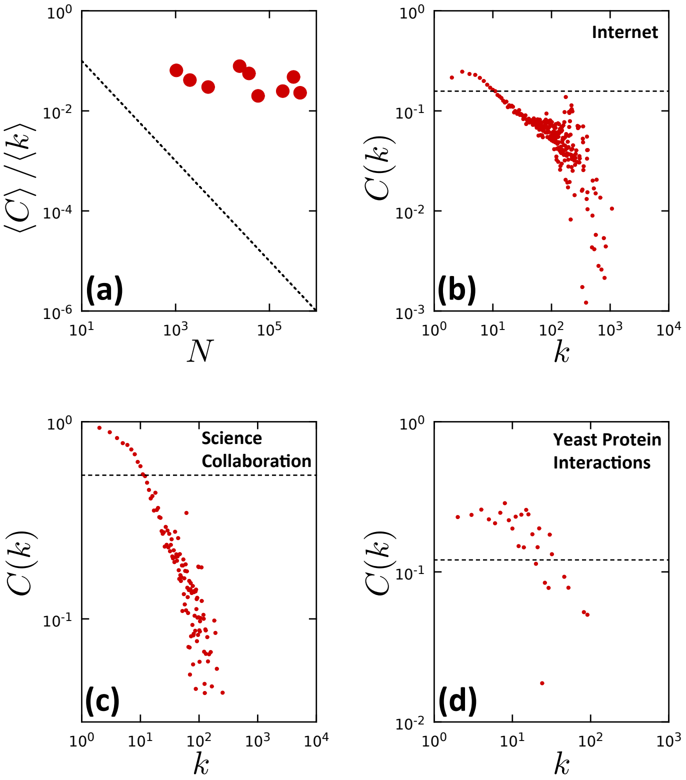


Image 3.16

Clustering in real networks.

(a) Comparison between the average clustering coefficient of real networks and the prediction Eq. (21) for random networks. Each circle corresponds to a network from Table 3.2. Directed network were made undirected to calculate *C*. The dashed line corresponds to Eq. (21), predicting that for random networks the average clustering coefficient should decrease as N^{-1}. In contrast, for real networks ‹*C*› has only a weak dependence on *N*.

(b)-(d) The dependence of the local clustering coefficient, *C(k)*, on the node's degree for *(b)* the Internet, *(c)* science collaboration network and *(d)* protein interaction network. *C(k)* is measured by averaging the local clustering coefficient of all nodes with the same degree *k*. The dashed line corresponds to the prediction of Eq. (21) of the random network model, for which *C(k)* is independent of *k*. In many real networks, the clustering coefficient decreases with *k*.

SECTION 10

REAL NETWORKS ARE NOT RANDOM

For about four decades following its introduction in 1959 the random network model has dominated mathematical approaches to complex networks. The model suggests that if a network is not as regular as a square lattice, we should describe it as random. With that it equated complexity with randomness. We must therefore ask:

Do we really believe that real networks are random?

The answer is clearly no. The interactions between our proteins are governed by the strict laws of biochemistry so for the cell to function its chemical architecture can not be random. Similarly, in a random society an American student would be more likely to have among his friends Chinese factory workers than one of her classmates. In reality we suspect the existence of a deep order behind most complex systems. That order must be reflected in the structure of the network that describes their architecture, resulting in systematic deviations from a pure random configuration.

The degree to which random networks describe (or fail to describe) real systems must not be decided by epistemological arguments, but by a systematic quantitative comparison. This is possible because random network theory makes a number of quantitative predictions that can be tested on real networks:

Degree distribution: The degrees of a random network follow a binomial distribution, well approximated by a Poisson distribution in the $k \ll N$ limit. Yet, as shown in Image 3.5, the Poisson distribution fails to capture the degree distribution of real networks. Instead in real systems we have more highly connected nodes than the random network model could account for.

Connectedness: Random network theory predicts that for $\langle k \rangle > 1$ we should observe a giant component, a condition satisfied by all networks we examined. Most networks do not satisfy the $\langle k \rangle > \ln N$ condition, which implies that these networks should be broken into isolated clusters (Table 3.1). Some networks are indeed fragmented, most are not.

Average path length: Random network theory predicts that the average path length scales as $\langle d \rangle \sim \log N / \log \langle k \rangle$, a prediction that captures the order of magnitude of the path lengths. Hence the random network model can account for the fundamental features of small world phenomena.

Clustering coefficient: In a random network the local clustering coefficient is independent of the node's degree and $\langle C \rangle$ depends on the system size as $1/N$. In contrast, measurements indicate that for real networks C decreases with the node degrees and is largely independent of the system size (Image 3.16).

Taken together, it appears that the small world phenomena is the only property reasonably explained by the random network model. All other network characteristics, from the degree distribution to the clustering coefficient, are significantly different in real and random networks. In fact, the more we learn about real networks, the more we will arrive at the startling conclusion that *we do not know of any real network that is accurately described by the random network model.*

This conclusion begs a legitimate question: If real networks are not random, why did we devote a full chapter to the random network model? The answer is simple: the model serves as a fundamental reference as we try to understand the properties of real networks. Each time we observe some network property we will have to ask if it could have emerged by chance. For this we turn to the random network model as a guide: if the property is present in the model, it means that randomness can account for it. If the property is absent in random networks, it may represents some signature of order, requiring a deeper explanation. So, the random network model may be the wrong model for most real systems, yet, *it remains quite relevant for network science* (Box 3.8).

Box 3.7

Random networks and network science.

The lack of agreement between random and real networks raises an important question: how could a theory survive so long given its poor agreement with reality? The answer is simple: random network theory was never meant to serve as a model of real systems. True Erdős and Rényi did write in their first paper [9] that "This may be interesting not only from a purely mathematical point of view. In fact, the evolution of graphs may be considered as a rather simplified model of the evolution of certain communication nets (railways, road or electric network systems, etc.) of a country or some unit." Yet, this is the only mention of the potential practical value of their approach. The subsequent development of random graphs was driven by inherent mathematical challenges.

It is tempting to follow Thomas Kuhn and view network science as a paradigm change from random graphs to a theory of real networks [22]. In reality, there was no network paradigm before the end of 1990s. This period is characterized by a lack of interest in the problem, without systematic attempts to compare the properties of real networks with graph theoretical models. The work of Erdős and Rényi has gained prominence outside mathematics only after the emergence of network science (see Image 3.17).

Network theory does not lessen the contributions of Erdős and Rényi, but demonstrates the unintended importance of their work. When we point out the disrepacies between the predictions of the random network model and real networks, we do so only to offer a proper ground on which we can understand the properties of real systems.

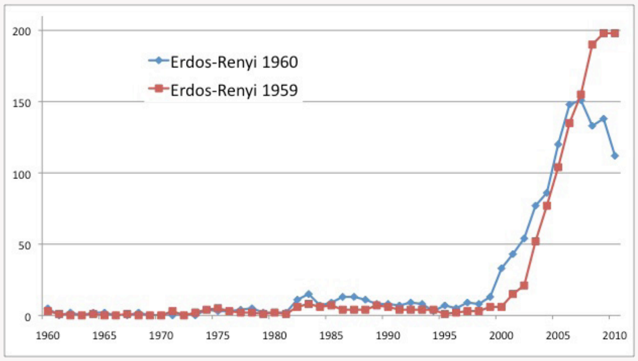


Image 3.17

Network science and random networks.

While today we perceive the Erdős-Rényi model as the cornerstone of network theory, the model was hardly known outside a small segment of mathematics. This is illustrated by the yearly citations of the first two papers by Erdős and Rényi, published in 1959 and 1960. For four decades after their publication the papers gathered less than 10 citations per year. The number of citations exploded after the first papers on scale-free networks [2, 3, 20] have turned Erdős and Rényi's work into the reference model of network theory.

SECTION 11

SUMMARY: THE FIRST LAW OF NETWORKS

Network science has distilled a small number of fundamental organizing principles that govern the structure and evolution of real networks. We call these *network laws* as just like the laws of physics, they encode generic principles obeyed by many real networks. A network property quantifies as a law if

(A) it has a unique quantitative, testable and falsifiable formulation;

(B) it is obeyed by a large number of real networks;

(C) it does not emergence by chance, hence it cannot be explained within the context of the random network model.

The results of this chapter allow us to formulate the fist of these laws:

> THE FIRST LAW: SMALL WORLD PROPERTY
> **In complex networks there are short distances between any pair of nodes.**

Evidence for the first law is provided in Sect. 3.8. To recap in the context of the criteria A-C:

A. Equation (19) offers the quantitative formulation of the First Law, predicting that the average distance between two randomly chosen nodes scales as a logarithm of the system size. Hence node-to-node distances are small compared to the network size.

B. Table 3.2 offers evidence that most real networks obey the first law.

C. As the small world property is present in random networks, the First Law apparently fails criterion C. Yet, we will see in the next chapter that in real networks distances are different from those expected in random networks, forcing us to modify Eq. (19).

At a glance: Random networks

- *Definition*: *N* nodes, where each node pair is connected with probability *p*.

- *Average degree:* $\langle k \rangle = p(N-1)$

- *Average number of links:* $\langle L \rangle = \frac{p(N-1)}{2}$

- *Degree distribution:* $p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$.

For sparse networks *(k « N), P_k* has the Poisson form

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}.$$

- *Giant component (N_G)*:

‹k› < 1: no giant component *($N_G \sim \ln N$)*

1 < ‹k› < ln*N*: one giant component and disconnected clusters

$$\left(N_G \sim N^{\frac{2}{3}} \right)$$

‹k› > ln*N*: all nodes join the giant component $N_G \sim (p - p_i)N$

- *Average distance:* $\langle d \rangle \propto \frac{\log N}{\log \langle k \rangle}$,

- *Clustering coefficient:* $C = \frac{\langle k \rangle}{N}$.

Box 3.8

ADVANCED TOPICS 3.A: DERIVING THE POISSON DEGREE DISTRIBUTION

We start from the exact binomial distribution (7)

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \qquad (22)$$

that characterizes a random graph, and we rewrite the first term on the r.h.s. as

$$\binom{N-1}{k} = \frac{(N-1)(N-1-1)(N-1-2)...(N-1-k+1)(N-1-k)!}{k!(N-1-k)!} = \frac{(N-1)^k}{k!} \qquad (23)$$

The last term of Eq. (22) can be simplified as

$$\ln[(1-p)^{(N-1)-k}] = (N-1-k)\ln(1-\frac{\langle k\rangle}{N-1})$$

and using the series expansion

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - ..., \forall \, |x| \leq 1$$

we obtain

$$\ln[(1-p)^{N-1-k}] \cong (N-1-k)\frac{\langle k\rangle}{N-1} = -\langle k\rangle(1-\frac{k}{N-1}) \cong -\langle k\rangle,$$

which is valid if N » k, representing the small degree approximation at the heart of this derivation. Therefore the last term of Eq. (22) becomes

$$(1-p)^{(N-1)-k} = e^{-\langle k\rangle}. \qquad (24)$$

Combining Eqs. (22), (23), and (24) we obtain the Poisson form of the degree distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} = \frac{(N-1)^k}{k!} p^k e^{-\langle k\rangle}$$

$$= \frac{(N-1)^k}{k!} \left(\frac{\langle k\rangle}{N-1}\right)^k e^{-\langle k\rangle},$$

or

$$p_x = e^{-\langle k\rangle} \frac{\langle k\rangle^k}{k!}. \qquad (25)$$

ADVANCED TOPICS 3.B: MAXIMUM AND MINIMUM DEGREES

To determine the expected degree of the *largest node* in a random network, called the *network's upper cutoff*, we define the degree k_{max} such that in a network of N nodes we have at most one node with degree higher than k_{max}. Mathematically this means that the area behind the Poisson distribution p_k for $k \geq k_{max}$ should be approximately one (Image 3.18). Since the area is given by $1 - P(k_{max})$, where $P(k)$ is the cumulative degree distribution of p_k, the network's largest node satisfies:

$$N\left[1-P(k_{max})\right] \approx 1. \qquad (26)$$

We write ≃ instead of =, because k_{max} is an integer, so in general the exact equation does not have a solution. For a Poisson distribution

$$1-P(k_{max}) = 1 - e^{-\langle k\rangle}\sum_{k=0}^{k_{max}}\frac{\langle k\rangle^k}{k!} = e^{-\langle k\rangle}\sum_{k=k_{max}+1}^{\infty}\frac{\langle k\rangle^k}{k!} \approx e^{-\langle k\rangle}\frac{\langle k\rangle^{k_{max}+1}}{(k_{max}+1)!}, \qquad (27)$$

where in the last term we approximate the sum with its largest (leading) term.

For $N = 10^9$, and $\langle k\rangle = 1{,}000$ corresponding to roughly the size and average degree of the globe's social network, we obtain $k_{max} = 1{,}185$, indicating that a random network lacks extremely popular individuals, or hubs.

We can use a similar argument to calculate the degree of the smallest node k_{min}, or the *natural smallest cutoff*. Indeed, by requiring that there should be at most one node with degree smaller than k_{min} we can write

$$NP(k_{min}) \approx 1. \qquad (28)$$

If $P(0) > 1$ the equation has no solution and $k_{min} = 0$. For the *ER* network we have

$$P(k_{min}) = e^{-\langle k\rangle}\sum_{k=0}^{k_{min}}\frac{\langle k\rangle^k}{k!} \qquad (29)$$

Solving Eq. (28) with $N = 10^9$ and $\langle k\rangle = 1{,}000$ we obtain $k_{min} = 816$.

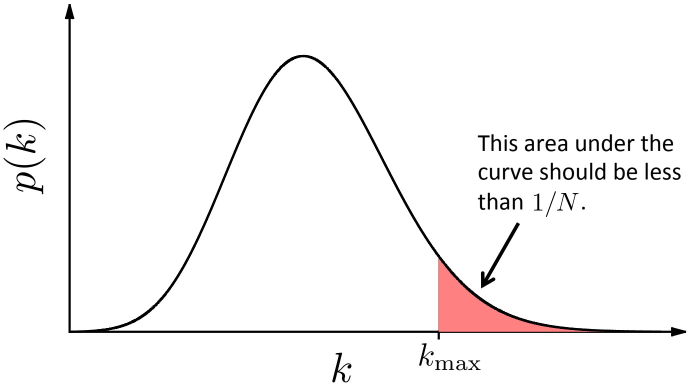


Image 3.18

Approximating the minimum and the maximum degree.

The maximum degree k_{max} is chosen so that there is at most one node whose degree is higher than k_{max}. This is often called the *natural upper cutoff* of a degree distribution. To calculate it, we need to set k_{max} such that the area under the degree distribution p_k for $k \geq k_{max}$ is exactly equal to $1/N$, hence this area multiplied by N, capturing the total number of nodes expected in the regime, is exactly one. We follow a similar argument to determine k_{min}, or the expected smallest degree.

ADVANCED TOPICS 3.C: GIANT COMPONENT

Our aim here is to reproduce the argument, put forward independently by Solomonoff and Rapoport [28], and by Erdős and Rényi [8], on the emergence of giant component at *‹k›= 1* (see also [24]).

Let us denote with $u = 1 - N_G / N$ the fraction of nodes that are not in the giant component *(GC)*, whose size we take to be N_G . If node *i* is part of the *GC*, it must link to another node *j*, which is also part of the *GC*. Hence if *i* is not part of the *GC*, that could happen for two reasons:

- There is no link between *i* and *j* (probability for this is $1- p$).
- There is a link between *i* and *j*, but *j* is not part of the *GC* (probability for this is pu).

Therefore the total probability that *i* is not part of the *GC* via node *j* is $1 - p + pu$. The probability that *i* is not linked to the *GC* via *any other node* is therefore $(1 - p + pu)^{N - 1}$, as there are $N - 1$ nodes that could serve as a potential links to the *GC* for node *i*. As *u* is the fraction of nodes that do not belong to the *GC*, for any *p* and *N* the solution of the equation

$$u = (1 - p + pu)^{N-1} \qquad (30)$$

provides the size of the giant component via $N_G = N(1 - u)$. Using $p = \langle k \rangle / (N - 1)$ and taking the log of both sides, for $\langle k \rangle \ll N$ we obtain

$$\ln u \simeq (N-1)\ln\left[1 - \frac{\langle k \rangle}{N-1}(1-u)\right]. \qquad (31)$$

Taking an exponential of both sides leads to $u = exp[- \langle k \rangle(1 - u)]$. If we denote with *S* the fraction of nodes in the giant component, $S = N_G / N$, then $S = 1 - u$ and Eq. (31) provides

$$S = 1 - e^{-\langle k \rangle S}. \qquad (32)$$

This simple looking equation provides the size of the giant component *S* in function of *‹k›* (Image 3.19). Yet, Eq. (32) does not have a closed solution. We can solve it graphically by plotting the right hand side of Eq. (32) as a function of *S* for various values of *‹k›*. To have a nonzero solution, the obtained curve must intersect with the dotted diagonal, representing the left hand side of Eq. (32). For small *‹k›* the two curves intersect each other only for $S = 0$, indicating that for small *‹k›*, the size of the giant component is zero. Only when *‹k›* exceeds a threshold value, does a non-zero solution emerge.

To determine the value of *‹k›* at which we start having a nonzero solution we take a derivative of Eq. (32), as the phase transition point is when the r.h.s. of Eq. (32) has the same derivative as the l.h.s. of Eq. (32), i.e.

$$\frac{d}{dS}\left(1 - e^{-\langle k \rangle S}\right) = 1,$$

$$\langle k \rangle e^{-\langle k \rangle S} = 1 \, . \qquad (33)$$

Setting $S = 0$, we obtain that the phase transition point is at $\langle k \rangle = 1$.

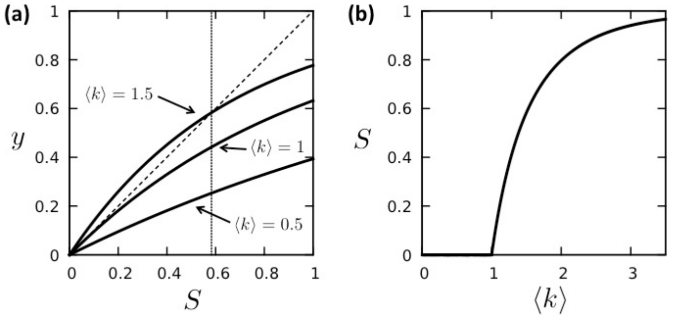


Image 3.19

Graphical solution for the size of the giant component.

(a) The three curves in the left panel show *y = 1-exp[-‹k› S]* for various *‹k›*. The diagonal dashed line corresponds *y = S*, and the intersection of the dotted and continuous lines provides the solution to Eq. (32), *S = 1-exp[-‹k›S]*. For the bottom curve there is only one intersection, at *S = 0*, indicating the absence of a giant component. The top curve a solution at *S = 0.583...* (vertical dashed line). The middle curve is precisely at the threshold between the regime where a non-zero solution for S exists and the regime where there is only the solution *S = 0*.
(b) The size of the giant component in function of *‹k›* as predicted by Eq. (32) [24].

ADVANCED TOPICS 3.D: COMPONENT SIZES

In Image 3.5 we focused only on the size of the giant component, leaving an important question open: how many smaller components do we expect for a given $\langle k\rangle$, and what is their expected sizes? The aim of this section is to discuss these topics.

Component size distribution: For a random network the probability that a randomly chosen node belongs to a component of size s (different from the giant component G) is [24]

$$p_s \sim \frac{(s\langle k\rangle)^{s-1}}{s!}e^{-\langle k\rangle s}. \qquad (34)$$

Replacing $\langle k\rangle^{s-1}$ with $exp[(s-1)\ ln\langle k\rangle]$ and using the Stirling-formula

$$s! \approx \sqrt{2\pi s}\left(\frac{s}{e}\right)^s$$

for large s we obtain

$$p_s \sim s^{-3/2}e^{-(\langle k\rangle-1)s+(s-1)\ln\langle k\rangle}. \qquad (35)$$

Therefore the component size distribution has two contributions: a slowly decreasing power law term $s^{-3/2}$ and a rapidly decreasing exponential term $e^{-(\langle k\rangle-1)s+(s-1)ln\langle k\rangle}$. Given that an exponential dominates for large s, Eq. (35) predicts that large components are prohibited. The only exception is at the critical point, $\langle k\rangle=1$, where all terms in the exponential cancel, hence p_s follows the power law

$$p_s \sim s^{-3/2}. \qquad (36)$$

As a power law decreases relatively slowly, at the critical point we expect to observe clusters of widely different sizes, a property consistent with the behavior of a system during a phase transition (Advanced Topics 3.E). These predictions are supported by numerical simulations in Image 3.20, that shows p_s for three $\langle k\rangle$ values.

Average component size: The calculations also indicate that the average component size (once again, excluding the giant component) follows [24]

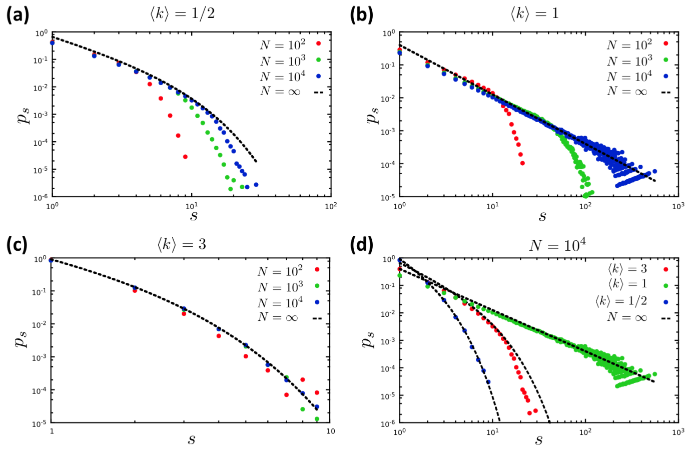


Image 3.20

Component size distribution.

Component size distribution in a random network, p_s, excluding the giant component. *(a)-(c)* shows p_s for different $\langle k\rangle$ values and *N*, indicating that p_s converges for large *N* to the prediction (34). In *(d)* we show the results for $N = 10^4$, plotting together p_s for different $\langle k\rangle$. The plot clearly shows that while for $\langle k\rangle < 1$ and $\langle k\rangle > 1$ the p_s has a exponential form, right at the critical point $\langle k\rangle = 1$ the distribution follows the power law (36). The dotted line in each image correspond to the theoretical prediction (35). The first numerical study of the component size distribution in random networks was carried out in 1998, preceeding the exploding interest in complex networks.

$$\langle s\rangle = \frac{1}{1-\langle k\rangle+\langle k\rangle N_G/N}. \qquad (37)$$

For $\langle k\rangle < 1$ we lack a giant component ($N_G = 0$), hence Eq. (37) becomes

$$\langle s\rangle = \frac{1}{1-\langle k\rangle}, \qquad (38)$$

which diverges when the average degree approaches the critical point $\langle k\rangle = 1$. Therefore as we approach the critical point, the clusters are becoming bigger, signaling the emergence of the giant component at $\langle k \rangle = 1$. Once again, numerical simulation support these predictions for large N (Image 3.21).

To determine the average component size for $\langle k \rangle > 1$ using Eq. (37), we need to first determine the size of the giant component. This can be done in a self-consistent manner, obtaining that the average cluster size decreases for $\langle k \rangle > 1$, as most of the clusters are gradually absorbed by the giant component.

Note that Eq. (37) predicts the size of the component to which a randomly chosen node belongs to. This is a biased measure, as the chance of belonging to a larger cluster is higher than the chance of belonging to a smaller one. The bias is linear in the cluster size, s. If we correct for this bias, we obtain the average size of the small components that we would get if we were to inspect each cluster one by one and measuring their average size [24]

$$\langle s' \rangle = \frac{2}{2 - \langle k \rangle + \langle k \rangle N_G / N}. \qquad (39)$$

Image 3.21 again offers numerical support for Eq. (39).

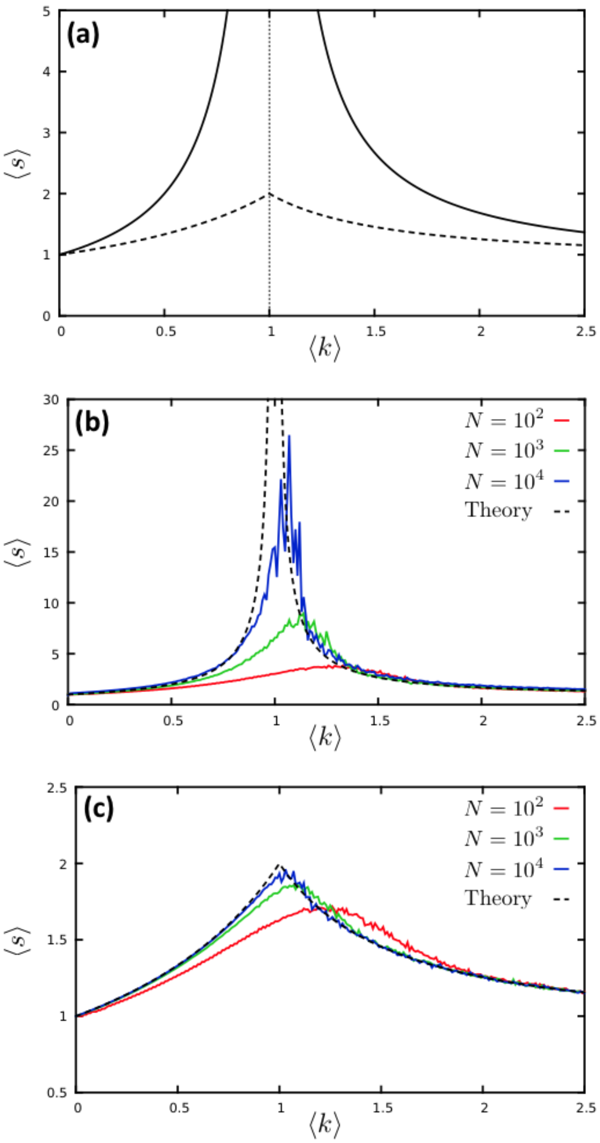


Image 3.21

Average component size.

a. Upper curve: the average size <*s*> of a component to which a randomly chosen node belongs to as predicted by Eq. (39). Lower curve: the overall average size <*s'*> of a component as predicted by Eq. (37). The dotted vertical line marks $\langle k \rangle = 1$ (Redrawn after Newman, 2010).

b. The average cluster size in a network measured in by numerical simulations, where we picked a node in the network and determined the size of the cluster it belongs to. This measure is biased, as each component of size *s'* will be counted *s'* times. The larger *N* becomes, the more closely the numerical data follows the prediction of Eq. (37). As predicted, <*s*> diverges at the $\langle k \rangle = 1$, critical point, supporting the existence of a phase transition in the system (Advanced Topics 3.F).

c. The average cluster size in a network, where we corrected for the bias in *(b)* by selecting each component only once.The larger *N* becomes, the more closely the numerical data follows the prediction of Eq. (39).

SECTION 16

ADVANCED TOPICS 3.E: SUPERCRITICAL REGIME.

To determine the value of ‹*k*› at which most nodes became part of the giant component, we calculate the probability that a randomly selected node does not have a link to the giant component, which is $(1-p)^{N_G} \approx (1-p)^N$, as in this regime $N_G \simeq N$. The expected number of such isolated nodes is

$$I_N = N(1-p)^N = N\left(1-\frac{N \cdot p}{N}\right)^N \approx Ne^{-Np}, \qquad (40)$$

where we used $(1-\frac{x}{n})^n \approx e^{-x}$, an approximation valid for large *n*. If we make *p* sufficiently large, we arrive to the point where only one node remains disconnected from the giant component. At this point $I_N = 1$, hence according to Eq. (40) *p* needs to satisfy $Ne^{-Np} = 1$. Consequently, the value of *p* at which we are about to enter the fully connected regime is

$$p \sim \frac{\ln N}{N}, \qquad (41)$$

which leads to Eq. (14) in terms of ‹*k*›.

SECTION 17

ADVANCED TOPICS 3.F: PHASE TRANSITIONS.

The emergence of the giant component at $\langle k \rangle = 1$ in the random network model is reminiscent of a *phase transition*, a much studied phenomenon in physics and chemistry [29]. Consider two examples:

i. *Water-Ice Transition* (Image 3.22a): At high temperatures the H_2O molecules engage in a diffusive dance, forming small groups and then breaking apart to group up with other molecules. If cooled, at $0^\circ C$ the molecules suddenly form a perfectly ordered ice crystal.

ii. *Magnetism* (Image 3.22b): In ferromagnetic metals like iron at high temperatures the spins point in randomly chosen directions. Under some critical temperature T_c, however, all atoms orient their spins in the same direction and the metal becomes a magnet.

The freezing of a liquid and the emergence of magnetization are examples of phase transitions, representing transitions from disorder to order. Indeed, relative to the perfect order of the crystalline ice, liquid water is rather disordered. Similarly, the randomly oriented spins in a ferromagnetic take up the highly ordered common orientation under T_c.

Many properties of a system undergoing a phase transition are universal, that is, they are the same in a wide range of systems, from magma freezing into rock to a ceramic material turning into a superconductor. Furthermore, near the phase transition point, called the *critical point*, many quantities of interest follow power-laws. The phenomena observed near the critical point $\langle k \rangle = 1$ in a random network in many ways is similar to such a phase transition:

- The similarity between Image 3.6a and the magnetization diagram of Image 3.22b is not accidental: they both show transition from disorder to order, manifested as the emergence of a giant component as $\langle k \rangle$ exceeds $\langle k \rangle = 1$ in a random network.

- As we approach the freezing point, ice crystals of widely different sizes are observed, and so are domains of atoms with spins pointing in the same direction. The size distribution of the ice crystals or magnetic domains follows a power law. Similarly, while for $\langle k \rangle < 1$ and $\langle k \rangle > 1$ the cluster sizes follow an exponential distribution, in a random network right at the phase transition point, p_s follows a power law given by Eq.(36), implying the coexistence of components of widely different sizes.

- At the critical point the average size of the ice crystals or of the magnetic domains diverges, assuring that the whole system turns into a single frozen ice crystal or that all spins point in the same direction. Similarly in a random network the average cluster size $<s>$ diverges as we approach $\langle k \rangle = 1$ (Advanced Topics 3.D).

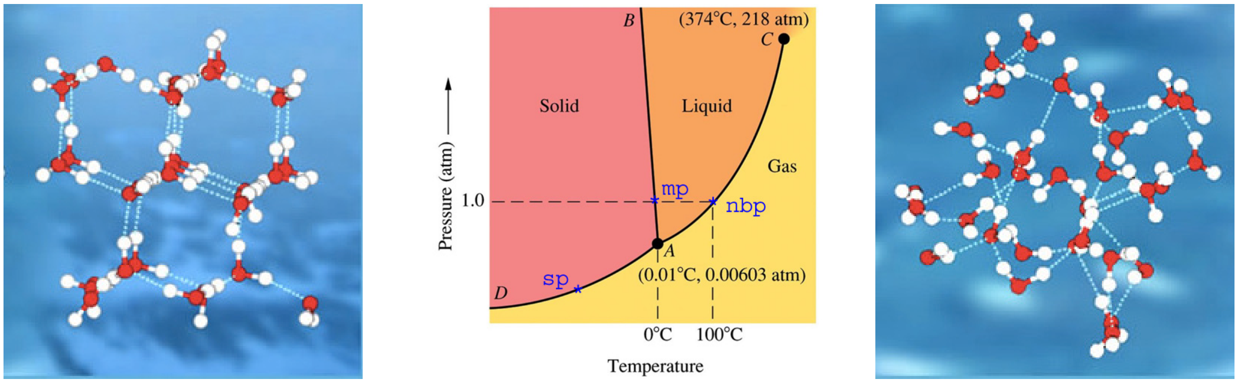


Image 3.22a

Water-Ice phase transition.

The hydrogen bonds that hold the water molecules together (dotted lines) are weak, constantly breaking up and re-forming, maintaining partially ordered local structures (left panel). The temperature-pressure phase diagram indicates (center panel) that by lowering the temperature, the water undergoes a phase transition, moving from a liquid (orange) to a frozen solid (red). In the solid phase each water molecule binds rigidly to four other molecules, forming an ice lattice (right panel). After http://www.lbl.gov/Science-Articles/Archive/sabl/2005/February/ water-solid.html; phase diagram after http://stevengoddard.wordpress.com/2010/09/02/the-ideal-world-phase-diagrams-part-deux/

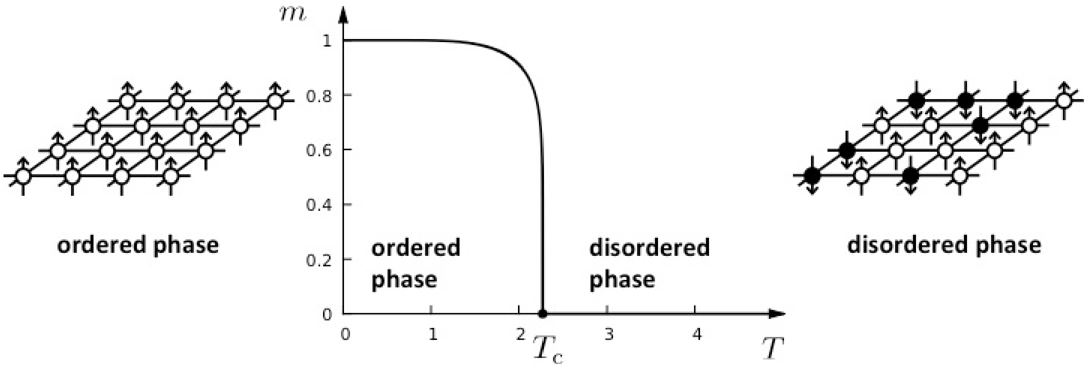


Image 3.22b

Magnetic phase phase transition.

In magnetic materials the magnetic moments of the individual atoms (spins) can point in two different directions. At high temperatures they choose randomly their direction (right panel), hence the system's total magnetization, m = $\Delta M / N$, where ΔM is the number of up spins minus the number of down spins, is zero. The phase diagram (middle panel) indicates that by lowering the temperature X, the system undergoes a phase transition at $T = Tc$ when a nonzero magnetization emerges, hence $m = M / N$ converges to one. In this ordered phase all spins point in the same direction (left panel).

ADVANCED TOPICS 3.G: CORRECTION TO SMALL WORLDS

Equation (18) offers only a rough approximation to the network diameter, valid for very large N and small d. Indeed, as soon as $\langle k\rangle^d$ approaches the system size N the $\langle k\rangle^d$ scaling must break down, as we are hitting the boundary of the network and there are not enough nodes to continue the $\langle k\rangle^d$ expansion. Such finite size effects result in corrections to Eq. (18).

For a random network with average degree $\langle k\rangle$, the network diameter is better approximated by (Fernholz & Ramachandran, 2007)

$$d_{\max} = \frac{\ln N}{\ln\langle k\rangle} + \frac{2\ln N}{\ln[-W(\langle k\rangle \exp - \langle k\rangle)]}, \quad (42)$$

where the Lambert W-function W(z) is the principal inverse of $f(z) = z\ exp(z)$. The first term on the r.h.s is Eq. (18), while the second is the correction that depends on the average degree. The correction increases the diameter, accounting for the fact that when we approach the network's diameter the number of modes must grow slower than $\langle k\rangle$. The magnitude of the correction becomes more obvious if we consider the various limits of Eq. (42).

In the $\langle k\rangle \rightarrow 0$ limit, i.e. when the network approaches the phase transition point, we can determine the Lambert W-function and the diameter becomes

$$d_{\max} = 3\frac{\ln N}{\ln\langle k\rangle}. \quad (43)$$

Hence in the moment when the giant component emerges the network diameter is three times our prediction (18). This is due to the fact that at the critical point $\langle k\rangle = 1$ the network has a tree-like structure, consisting of long chains with hardly any loops, a configuration that significantly increases d_{max} .

In the $\langle k\rangle \rightarrow \infty$ limit, corresponding to a very dense network, Eq. (42) becomes

$$d_{\max} = \frac{\ln N}{\ln\langle k\rangle} + \frac{2\ln N}{\langle k\rangle} + \ln N\left(\frac{\ln\langle k\rangle}{\langle k\rangle^2}\right). \quad (44)$$

Hence if $\langle k\rangle$ increases, the second and the third terms vanish and the solution (42) converges to the result (18).

BIBLIOGRAPHY

[1] Barabási, A.-L. (2003). *Linked: The new science of networks.* New York: Plume Books, 1 edition.

[2] Barabási, A.-L. & Albert R. (1999). *Emergence of scaling in random networks.* Science, 286:509-512.

[3] Barabási, A.-L., Albert, R., and Jeong, H. (1999). *Meanfield theory for scale-free random networks.* Physica A: Statistical Mechanics and its Applications, 272:173-187.

[4] Backstrom, L., Boldi, P., Rosa, M., Ugander, J. & Vigna, S. (2011). *Four degrees of separation*. CoRR, abs/1111.4570.

[5] Bollobás, B. (2001). *Random Graphs.* Cambridge University Press.

[6] Christensen, K., Donangelo, R., Koiller, B., and Sneppen, K. (1998). *Evolution of Random Networks*. Physical Review Letters, 81:2380-2383.

[7] Csicsery, G. P. (1993). *N is a Number: A Portait of Paul Erdős*.

[8] Erdős, P. & Rényi, A. (1959). *On random graphs, I*. Publicationes Mathematicae (Debrecen), 6:290-297.

[9] Erdős, P. & Rényi, A. (1960). *On the evolution of random graphs.* Publ. Math. Inst. Hung. Acad. Sci., 5:17-61.

[10] Erdős, P. & Rényi, A. (1961a). *On the evolution of random graphs.* Bull. Inst. Internat. Statist., 38:343-347.

[11] Erdős, P. & Rényi A. (1961b), *On the Strength of Connectedness of a Random Graph, Acta Math.* Acad. Sci. Hungary 12: 261–267.

[12] Erdős, P. & Rényi, A. (1963). *Asymmetric graphs.* Acta Mathematica Acod. Sci. Hungarica, 14(3-4):295-315.

[13] Erdős, P. & Rényi, A. (1966). *On random matrices.* Publ. Math. Inst. Hung. Acad. Sci., 8:455-461.

[14] Erdős, P. & Rényi, A. (1966). *On the existence of a factor of degree one of a connected random graph.* Acta Math. Acad. Sci. Hungar., 17:359-368.

[15] Erdős, P. & Rényi, A. (1968). *On random matrices II.* Studia Sci. Math. Hung., 13:459-464.

[16] Fernholz, D. & Ramachandran, V. (2007). *The diameter of sparse random graphs.* Random Structures and Algorithms, 31(4):482-516.

[17] Freeman, L. C. & Thompson, C. R. (1989). *Estimating Acquaintanceship.* Volume, pg. 147-158, in The Small World, Edited by Manfred Kochen (Ablex, Norwood, NJ)

[18] Gilbert, E. N. (1959). *Random graphs.* The Annals of Mathematical Statistics, 30:1141-1144.

[19] Hoffman, P. (1998). *The Man Who Loved Only Numbers: The Story of Paul Erdős and the Search for Mathematical Truth.* Hyperion Books.

[20] Jeong, H., Albert, R. & Barabási, A. L. (1999). *Internet: Diameter of the world-wide web*. Nature, 401:130-131.

[21] Frigyes K. , *"Láncszemek," in Minden másképpen van* (Budapest: Atheneum Irodai es Nyomdai R.-T. Kiadása, 1929), 85–90. English translation is available in (Newman, Barabási, and Watts, 2006).

[22] Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.

[23] Milgram, S. (1967). *The Small World Problem*. Psychology Today, 2: 60-67.

[24] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, 1 edition.

[25] Newman, M., Barabási, A. L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

[26] Rosenthal, H. (1960). *Acquaintances and contacts of Franklin Roosevelt.* Unpublished thesis. Massachusetts Institute of Technology.

[27] Schechter, B. (1998). *My Brain is Open: The Mathematical Journeys of Paul Erdős*. Simon & Schuster.

[28] Solomonoff, R. & Rapoport, A. (1951). *Connectivity of random nets*. Bulletin of Mathematical Biology, 13:107–117.

[29] Stanley, H. E. (1987). *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press.

[30] Watts, D. J. & Strogatz, S. H. (1998). *Collective dynamics of 'small-world' networks*. Nature 393: 409–10.

CHAPTER 4

THE SCALE-FREE PROPERTY

Introduction

Power laws and scale-free networks 1

Hubs 2

The meaning of scale-free 3

Universality 4

Ultra-small property 5

The role of the degree exponent 6

Generating networks with a pre-defined degree distribution 7

Summary

ADVANCED TOPICS 4.A
Power laws

ADVANCED TOPICS 4.B
Plotting a power-law degree distribution

ADVANCED TOPICS 4.C
Estimating the degree exponent

Bibliography

Figure 4.0 (next page)
Art and Networks: Tomás Saraceno

Tomás Saraceno creates art work inspired by spider webs and neural networks. Trained as an architect, Saraceno deploys theoretical frameworks and insights from engineering, physics, chemistry, aeronautics, and materials science usingnetworks as a source of inspiration and metaphor. The image shows his work displayed in the Miami Art Museum, an example the artist's take on a complex network.

INTRODUCTION

As difficult it is to overstate the importance of the World Wide Web in our daily life, it is equally hard to exaggerate the role the Web played in the development of network theory. It aided the discovery of a number of fundamental network properties and became a standard testbed for many network measures. As its name states, the WWW is a “web” whose nodes are documents and the links are the uniform resource locators (URLs) that allow us to move with a click from one web document to the other. With an estimated size of over one trillion documents ($N \simeq 10^{12}$), the Web is the largest network humanity has ever built. It exceeds in size even the human brain ($N \simeq 10^{11}$ neurons).

We can use a software called a crawler to map out the Web’s wiring diagram. A crawler can start from any web document, identifying the links (URLs) on it. Next it downloads the documents these links point to and identifies the links on these documents, and so on. This process iteratively returns a local map of the Web. Search engines like Google or Bing operate such crawlers that constantly index new documents, along the way providing a detailed map of the WWW.

The first map of the WWW obtained with the explicit goal of understanding the structure of the network behind it was generated by Hawoong Jeong at University of Notre Dame. He mapped out the nd.edu domain [1], consisting of about 300,000 documents and 1.5 million links. The purpose of the map in Fig. 4.1 was to compare the properties of the Web graph to the random network model. Indeed, in 1998 there were reasons to believe that the WWW could be well approximated by a random network. The content of each document reflects the personal and professional interests of its creator, from individuals to organizations. Given the diversity of these interests, the links on these documents might appear to point to randomly chosen documents. A quick look at the map in Fig. 4.1 supports this view: there is a high degree of randomness behind the Web’s wiring diagram. Yet, a closer inspection reveals some puzzling differences between this map and a random network. In a random network highly connected nodes, or hubs, are effectively forbidden.

In contrast in Fig. 4.1 numerous small-degree nodes coexist with a few hubs, nodes with an exceptionally large number of links. The purpose of this chapter is to show that these hubs are not unique to the Web, but we encounter them in many real networks. They represent a signature of a deeper organizing principle that we call the scale-free property.

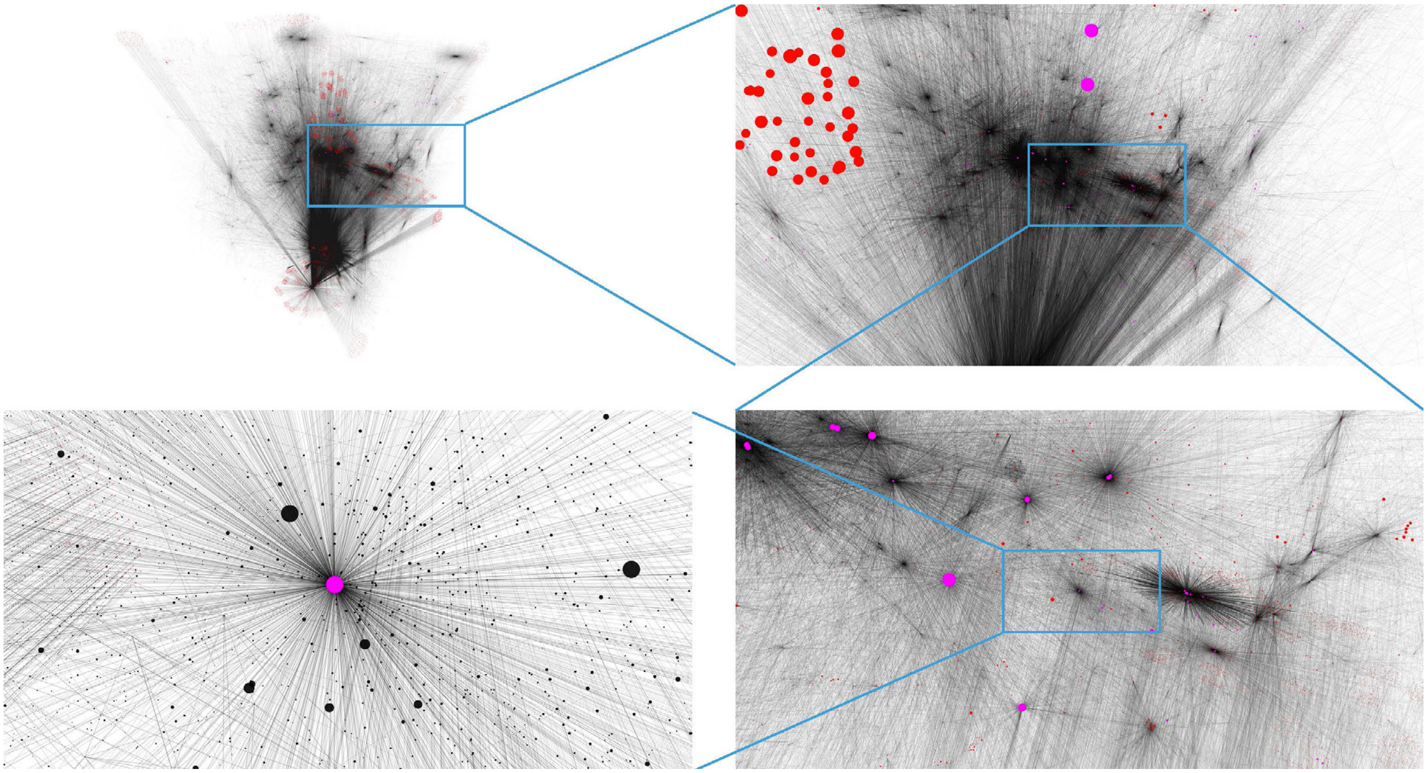


Figure 4.1
The topology of the WWW

A visualization of the web sample that led to the discovery of the scale-free property. The sequence of images shows an increasingly magnified local region of the network. The first panel displays all 325,725 nodes, offering a global view of the full dataset. Nodes with more than 50 links are shown in red and nodes with more than 500 links in purple. The increasingly magnified closeups reveal the presence of a few highly connected nodes, called hubs, that accompany scale-free networks (Image by M. Martino).

POWER LAWS AND SCALE-FREE NETWORKS

If the WWW were to be a random network, its degrees should follow a Poisson distribution. Yet, as Fig. 4.1 indicates, the Poisson form offers a poor fit for the WWW's degree distribution. Instead we find that on a log-log scale the data points form an approximate straight line, suggesting that the degree distribution of the WWW is best approximated with

$$p_k \sim k^{-\gamma}. \quad (4.1)$$

Eq. 4.1 is called a *power law* distribution and the exponent γ is its *degree exponent*. If we take a logarithm of Eq. 4.1, we obtain

$$\log p_k \sim -\gamma \log k. \quad (4.2)$$

Therefore, if Eq. 4.1 holds, logp_k is expected to depend linearly on log$k$, the slope of this line being the degree exponent γ, as observed in Fig. 4.2.

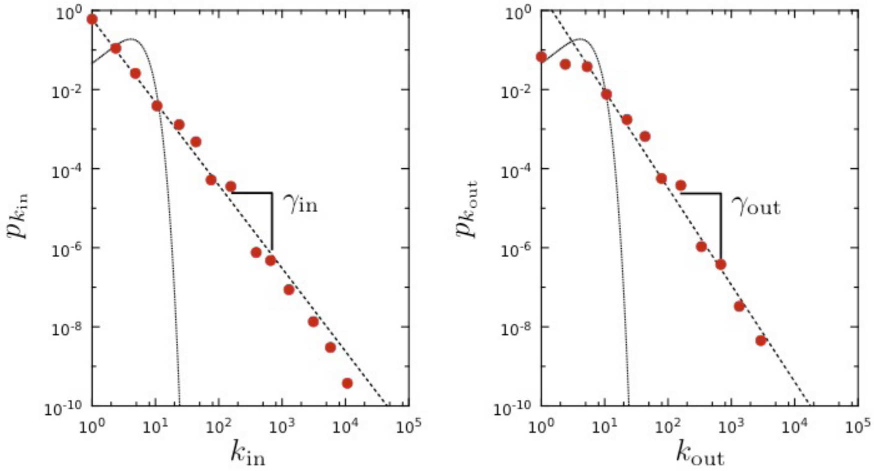


Figure 4.2
The degree distribution of the WWW

The incoming (left panel) and outgoing (right panel) degree distribution of the WWW sample mapped in the 1999 study of Albert *et al.* [1]. The degree distribution is shown on double logarithmic axis (log-log plot), in which a power law is expected to follow a straight line. The symbols correspond to the empirical data and the dotted line corresponds to the power-law fit, with degree exponents γ_{in}= 2.1 and γ_{out} = 2.45. The degree distribution predicted by a Poisson function with average degree $\langle k_{in}\rangle$ = $\langle k_{out}\rangle$ = 4.60, representing the observed values for the WWW sample, is shown as a dotted line.

As the WWW is a directed network, each document is characterized by an out-degree k_{out}, representing the number of links that point from a document to other documents, and an in-degree k_{in}, representing the number of other documents that point to a given document. We must therfore distinguish two different degree distributions: the probability that a randomly chosen document points to k_{out} other web documents, or $p_{k_{out}}$, and the probability that a randomly chosen node has k_{in} other web documents pointing to it, or $p_{k_{in}}$. In the case of the WWW both $p_{k_{in}}$ and $p_{k_{out}}$ can be approximated by a power law

$$p_{k_{in}} \square k^{-\gamma_{in}} \tag{4.3}$$

$$p_{k_{out}} \square k^{-\gamma_{out}} \tag{4.4}$$

where γ_{in} and γ_{out} are the degree exponents for the in- and out-degrees, respectively Fig. 4.2. In general γ_{in} can differ from γ_{out}. For example, for the WWW sample of Fig. 4.1 we have $\gamma_{in} \simeq 2.1$ and $\gamma_{out} \simeq 2.45$. The empirical evidence discussed above leads to the concept of a scale-free network [2]: *Networks whose degree distribution follows a power law are called scale-free networks.* As Fig. 4.2 indicates, for the WWW the power law persists for almost four orders of magnitude, prompting us to call the network behind the Web scale-free. In this case the scale-free property applies to both in and out-degrees. To explore the consequences of the scale-free property, we have to define the power-law distribution in more precise terms. For this we introduce the discrete and the continuum formalisms used throughout this book.

DISCRETE FORMALISM

As node degrees are always positive integers, $k = 0, 1, 2, 3, \ldots, N$, the discrete formalism captures the probability p_k that a node has exactly k links

$$p_k = Ck^{-\gamma}. \tag{4.5}$$

The constant C is determined by the normalization condition

$$\sum_{k=1}^{\infty} p_k = 1. \tag{4.6}$$

Using Eq. 4.4 we obtain, $C\sum_{k=1}^{\infty} k^{-\gamma} = 1$, hence

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\zeta(\gamma)}, \tag{4.7}$$

where $\zeta(\gamma)$ is the Riemann-zeta function. Thus for $k > 0$ the discrete power-law distribution has the form

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}. \tag{4.8}$$

Note that Eq. 4.8 diverges at k=0. We therefore need to separately specify $p_{0,}$ representing the fraction of nodes that have no links to other nodes (isolated nodes).

CONTINUUM FORMALISM

In analytical calculations it is often convenient to assume that the degrees can take up any positive real value. In this case the power-law degree distribution is written as:

$$p(k) = Ck^{-\gamma}. \tag{4.9}$$

Using the normalization condition:

$$\int_{k_{min}}^{\infty} p(k)dk = 1 \tag{4.10}$$

we obtain the constant:

$$C = \frac{1}{\int_{K_{min}}^{\infty} k^{-\gamma} dk} = (\gamma - 1)K_{min}^{\gamma -1}. \tag{4.11}$$

Therefore in the continuum formalism the degree distribution has the form:

$$p(k) = (\gamma - 1)k_{min}^{\gamma -1}\, k^{-\gamma}. \tag{4.12}$$

Here k_{min} is the smallest degree for which the power law Eq. 4.8 holds. Note that p_k encountered in the discrete formalism has a precise meaning: it provides the probability that a randomly selected node has degree k. In contrast, only the integral of $p(k)$ encountered in the continuum formalism has a physical interpretation:

$$\int_{k_1}^{k_2} p(k)dk \tag{4.13}$$

provides the probability that a randomly chosen node has degree between k_1 and k_2. In summary, networks whose degree distribution follows a power law are called scale-free networks. If a network is directed, the scale-free property can apply separately to the in- and the out-degrees.

To mathematically study the properties of scale-free networks, we can use the discrete or the continuum formalism. Note, however, that the scale-free property is independent of the formalism we use to describe the degree distribution.

BOX 4.1

The 80/20 rule and the top one percent

Vilfredo Pareto, a 19th century economist, noticed that in Italy a few wealthy individuals earned most of the money, while the majority of the population earned rather small amounts. He connected this disparity to the observation that incomes follow a power law, representing the first known report of a power-law distribution [3]. His finding entered the popular literature as the 80/20 rule: roughly 80 percent of money is earned by only 20 percent of the population.

The 80/20 emerges in many areas, like management, stating that 80 percent of profits are produced by only 20 percent of the employees or that 80 percent of decisions are made during 20 percent of meeting time.

They are present in networks as well: 80 percent of links on the Web point to only 15 percent of webpages; 80 percent of citations go to only 38 percent of scientists; 80 percent of links in Hollywood are connected to 30 percent of actors [4]. Typically all quantities obeying the 80/20 rule follow a power law distribution.

During the 2009 economic crisis power laws have gained a new meaning: the Occupy Wall Street Movement highlighted the fact that in the US 1% of the population earns a disproportionate 15% of the total US income. This 1% effect, a signature of a profound income disparity, is again a natural consequence of the power-law nature of the income distribution.

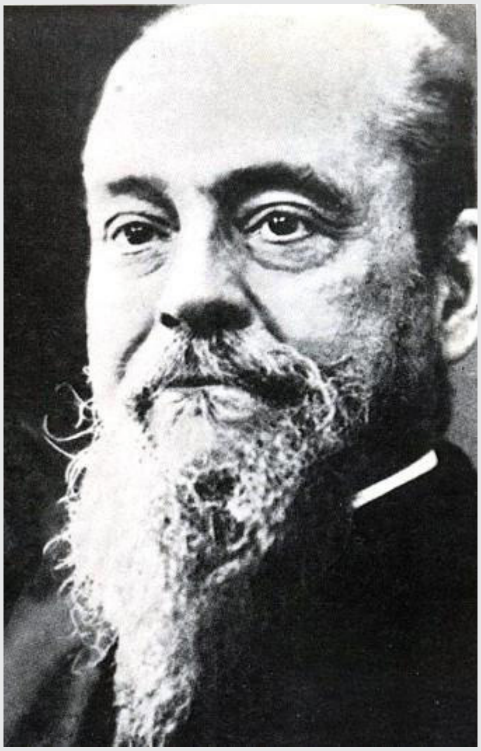


Figure 4.3
Vilfredo Federico Damaso Pareto (1848 – 1923)

Italian economist, political scientist, and philosopher, who had important contributions to our understanding of income distribution and to the analysis of individuals choices. A number of fundamental principles are named after him, like Pareto efficiency, Pareto distribution (another name for a power-law distribution), the *Pareto principle* (or 80/20 law).

HUBS

The main difference between a random and a scale-free network comes in the tail of the degree distribution, representing the high-k region of p_k. Fig. 4.4 compares a power law with a Poisson function, indicating that:

- For small k the power law is above the Poisson function, hence a scale-free network has a large number of small degree nodes that are virtually absent in a random network.

- For k the vicinity of $\langle k \rangle$ the Poisson distribution is above the power law, indicating that in a random network most nodes have degree $k \simeq \langle k \rangle$.

- For large k the power law is again above the Poisson curve. The difference is particularly visible if we show p_k on a log-log plot Fig. 4.4b, indicating that the probability of observing a high-degree node, or hub, is several orders of magnitudes higher in a scale-free than in a random network.

Let us use the WWW to illustrate the properties of the high-k regime. The probability to have a node with $k \simeq 100$ is about $p_{100} \simeq 10^{-30}$ in a Poisson distribution while it is about $p_{100} \simeq 10^{-4}$ if p_k follows a power law. Consequently, if the WWW were to be a random network with

$$N_{k>100} = \sum_{k=101}^{\infty} \frac{(4.6)^k}{k!} e^{-4.6} \simeq 10^{18}, \langle k \rangle \simeq 4.6 \qquad (4.14)$$

and $N \simeq 10^{12}$ Table 4.1, we would expect nodes with more than 100 links, or effectively none. In contrast, given the WWW's power law degree distribution, with $\gamma_{in} = 2.1$, we have $N_{k>100} = 10^9$ nodes with degree k >100.

HUBS

All real networks are finite. The size of the WWW is estimated to be $N \simeq 10^{12}$ nodes; the size of the social network is the Earth's population, about $N \simeq 7 \times 10^9$. These numbers are huge, but finite. Other networks pale in comparison: the genetic network in a human cell has approximately 20,000 genes while the metabolic network of the *E. Coli* bacteria has only about a

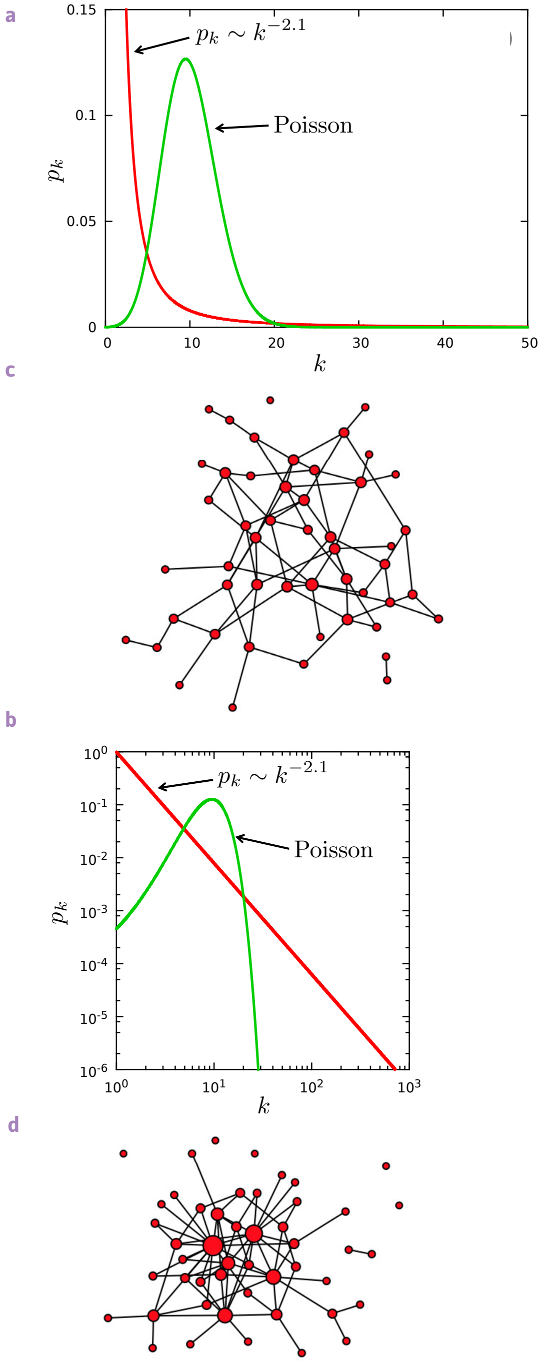


Figure 4.4
Poisson vs. power-law distributions

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

thousand metabolites. This prompts us to ask: how does the network size affect the size of its hubs?

For an arbitrary degree distribution p_k we can calculate the expected maximum degree, k_{max}, often called natural cutoff. It represents the expected size of the largest hub.

It is instructive to perform the calculation first for the exponential distribution $p_k = Ce^{-\lambda k}$. Assuming that the network's minimum degree is k , the normalization condition

$$\int_{k_{min}}^{\infty} p(k)dk = 1 \qquad (4.15)$$

provides $C = \lambda e^{\lambda k_{min}}$. To calculate k_{max} we assume that in a network of N nodes we expect at most one node in the (k_{max}, ∞) regime. In other words the probability to observe a node whose degree exceeds k_{max} is 1 / N:

$$\int_{k_{max}}^{\infty} p(k)dk = \frac{1}{N}. \qquad (4.16)$$

Equation Eq. 4.14 yields

$$k_{max} = k_{min} + \frac{\ln N}{\lambda}. \qquad (4.17)$$

As ln N is a slow function of the system size, Eq. 4.17 tells us that the maximum degree will not be very different from k_{min}. For a Poisson degree distribution the calculation is a bit more involved, but the obtained dependence of k_{max} on N is even slower than the logarithmic dependence predicted by Eq. 4.17.

For a scale-free network, according to Eq. 4.16 and Eq. 4.17 the natural cutoff follows

$$k_{max} \sim k_{min}\, N^{\frac{1}{\gamma - 1}}. \qquad (4.18)$$

Hence the larger a network, the larger is the degree of its biggest hub. The polynomial dependence of k_{max} on N implies that in a large scale-free network there can be orders of magnitude differences in size between the smallest node, $k_{min,}$ and the biggest hub, k_{max} Fig. 4.5 .

To illustrate the difference in the maximum degree of an exponential and a scale-free network let us return to the WWW sample of Fig. 4.1 consisting of $N \simeq 3 \times 10^5$ nodes. As $k_{min} = 1$, if the degree distribution were to follow an exponential, Eq. 4.17 predicts that the maximum degree should be k_{max} $\simeq$ 13. In a scale-free network of similar size and γ=2.1, Eq. 4.18 predicts k_{max} $\simeq$ 85,000, a remarkable difference. Note that the largest in-degree of this WWW map of Fig. 4.1 is 10,721, which is comparable to the predicted k_{max}.

This reinforces our conclusion that in a random network hubs are forbidden, while in scale-free networks they occur naturally.

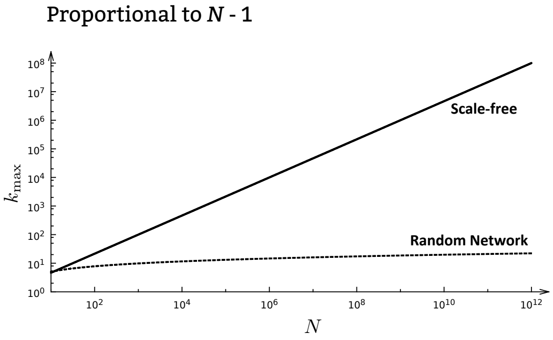


Figure 4.5
Hubs are large in scale-free networks

The expected degree of the largest node (natural cutoff) in scale-free and random networks with the same average degree $\langle k \rangle$= 3. For the scale-free network we chose γ = 2.5. For comparison, we also show the linear behavior, $k_{max} \sim N - 1$, expected for a complete network. Overall, hubs in a scale-free network are several orders of magnitude larger than the biggest node in a random network with the same N and $\langle k \rangle$.

In summary the key difference between a random and a scale-free network comes in the different shape of the Poisson and of the power-law function: in a random network most nodes have comparable degrees and hence hubs are forbidden. Hubs are not only tolerated, but are expected in scale-free networks Fig. 4.5.

The more nodes a scale-free network has, the larger are its hubs. The hubs grow polynomially with the network size, hence their size can be considerable in large networks. In contrast in a random network the size of the largest node grows logarithmically or slower with N, implying that hubs will be tiny even in a very large network.

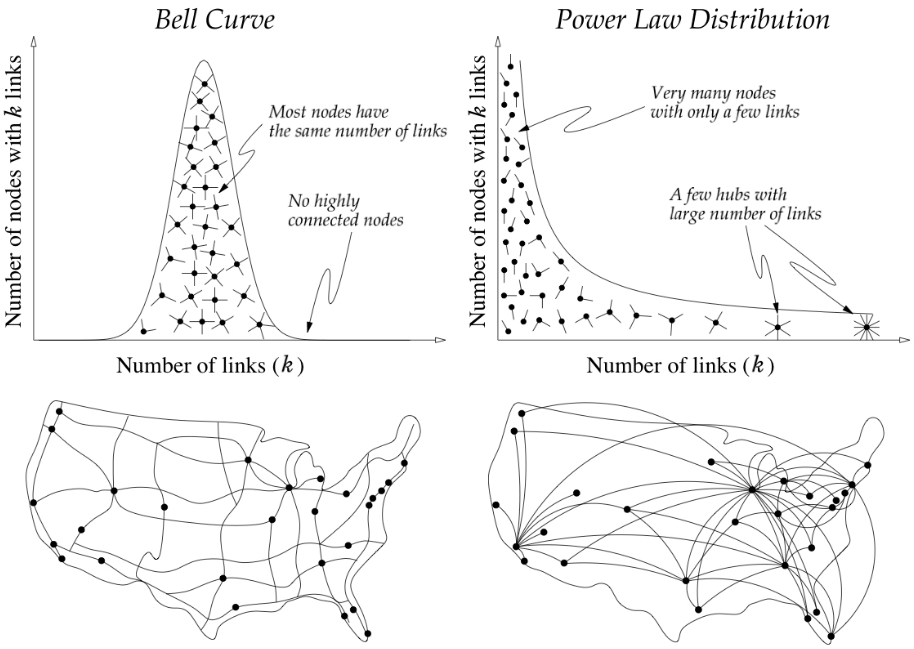

Bell Curve
Power Law Distribution
Number of nodes with k links
Most nodes have the same number of links
No highly connected nodes
Number of links (k)
Number of nodes with k links
Very many nodes with only a few links
A few hubs with large number of links
Number of links (k)

Figure 4.6

Random versus scale-free networks

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

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

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

After [4].

SECTION 4.4

THE MEANING OF SCALE-FREE

What is behind the “scale-free” name? The term is rooted in a branch of statistical physics called the theory of phase transitions SECTION 3.F, that extensively explored power laws in the 1960s and 1970s. To best understand the meaning of the scale-free term, we need to familiarize ourselves with the moments of the degree distribution. The n^{th} moment of the degree distribution is defined as:

$$k^n = \sum_{k_{min}}^{\infty} k^n p_k = \int_{k_{min}}^{\infty} k^n p(k)dk. \tag{4.19}$$

The lower moments have important interpretation:

- n=1: the first moment is the average degree, $\langle k \rangle$.
- n=2: the second moment, $\langle k^2 \rangle$, provides the variance $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$, measuring the spread in the degrees. Its square root, σ, is the standard deviation.
- n=3: the third moment, $\langle k^3 \rangle$, determines the skewness of a distribution, telling us how symmetric is p_k around the average $\langle k \rangle$. Symmetric distributions have zero skewness. For a scale-free network the n^{th} moment of the degree distribution is

$$k^n = \int_{k_{min}}^{k_{max}} k^n p(k)dk = C\frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n-\gamma+1}. \tag{4.20}$$

While typically k_{max} is fixed, the degree of the largest hub, k_{max}, increases with the system size, following Eq. 4.18.

Hence to understand the behavior of $\langle k^n \rangle$ we need to take the asymptotic limit $k_{max} \to \infty$ in Eq. 4.20, probing the properties of very large networks. In this limit Eq. 4.20 predicts that the value of $\langle k^n \rangle$ depends on the interplay between n and γ:

- If $n - \gamma + 1 \leq 0$ then the first term on the r.h.s. of Eq. 4.20, $k_{max}^{n-\gamma+1}$, goes to zero as k_{max} increases. Therefore all moments that satisfy $n \leq \gamma - 1$ will be finite.

- If $n-\gamma+1 \geq 0$ then $\langle k_n \rangle$ goes to infinity as $k_{max} \to \infty$. Therefore all moments satisfying $n \geq \gamma - 1$ diverge.

For most real scale-free networks the degree exponent γ is between 2 and 3 Table 4.1. Hence for these in the $N \to \infty$ limit the first moment $\langle k \rangle$ is finite, but the second and higher moments, $\langle k^2 \rangle$, $\langle k^3 \rangle$, go to infinity. This divergence helps us understand the origin of the "scale-free" term:

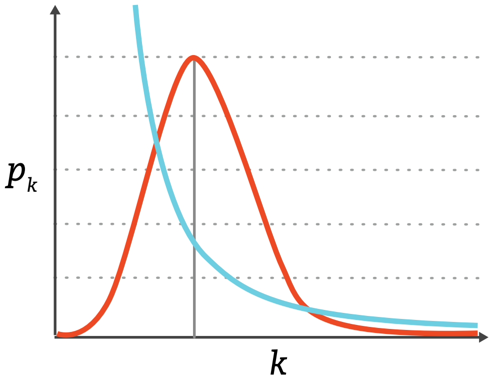
- If the degrees follow a normal distribution, then the degree of a randomly chosen node is

$$k = \langle k \rangle \pm \sigma_k \tag{4.21}$$

For a random network with a Poisson degree distribution $\sigma_k = \sqrt{\langle k \rangle}$, which is always smaller than $\langle k \rangle$. Hence the degrees are in the range $k = \langle k \rangle \pm \langle k \rangle^{1/2}$, indicating that nodes in a random network have comparable degrees. Therefore the average degree $\langle k \rangle$ serves as the "scale" of a random network.

- For a network with a power-law degree distribution and $\gamma < 3$ the first moment is finite but the second moment is infinite. The divergence of $\langle k^2 \rangle$, and hence of σ_k for large N indicates that the fluctuations around the average could be arbitrary large. That is, when we randomly choose a node, we do not know what to expect, as the chosen node's degree could be tiny or arbitrarily large. Hence networks with $\gamma < 3$ do not have a meaningful internal scale. They are "scale-free" Fig. 4.7. For example the average degree of the WWW sample is $\langle k \rangle = 4.60$ Table 4.1. Given that $\gamma \simeq 2.1$, the second moment diverges, which means that our expectation for the in-degree of a randomly chosen WWW document is $\langle k \rangle = 4.60 \pm \infty$ in the $N \to \infty$ limit. That is, a randomly chosen webpage could easily yield a document of degree one or two, as 74.02% of nodes have in-degree less than $\langle k \rangle$. Yet, it could also yield a node with hundreds of millions of links, like google.com or facebook.com.

Strictly speaking $\langle k^2 \rangle$ diverges only in the $N \to \infty$ limit. Yet, the divergence is relevant for finite networks as well. To illustrate this, Table 4.1 and Figure 4.8 show the standard deviation $\sigma = \sqrt{\langle k^2 \rangle - \langle k \rangle^2}$ for ten real networks. For most of these networks σ is significantly larger than $\langle k \rangle$, documenting large variations in node degrees. For example, the degree of a randomly chosen node in the studied WWW sample is $k_{in} = 4.60 \pm 39.05$, indicating once again that the average is not informative in this case. In summary, the scale-free name captures the lack of an internal scale, a consequence of the fact that nodes with widely different degrees coexist. This feature distinguishes scale-free networks from lattices, in which all nodes have exactly the same degree ($\sigma = 0$), or from random networks, whose degrees vary in a narrow range ($\sigma_k = \langle k \rangle^{1/2}$). As we will see in the coming chapters, this divergence is the origin of some of the most interesting properties of scale-free networks, from their robustness to random failures to the anomalous spread of viruses.



Random network
Randomly chosen node: $k = \langle k \rangle \pm \langle k \rangle^{1/2}$
Scale: $\langle k \rangle$

Scale-free network
Randomly chosen node: $k = \langle k \rangle \pm \infty$
$\langle k \rangle$ is meaningless as 'scale'

Figure 4.7
Scale-free networks lack an internal scale

For any bounded distribution (e.g. a Poisson or a Gaussian distribution) the degree of a randomly chosen node will be in the vicinity of $\langle k \rangle$. Hence $\langle k \rangle$ serves as the network's scale. In a scale-free network the second moment diverges, hence the degree of a randomly chosen node can be arbitrarily different from $\langle k \rangle$. As a scale-free network lacks an intrinsic scale, is it scale-free.

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

Table 4.1
The characteristics of several real network

The table shows the standard deviation of the degree distribution $\sigma = \sqrt{\langle k^2 \rangle - \langle k \rangle^2}$ (σ_{in} and σ_{out} for directed networks) for our ten reference networks. It indicates that for most networks σ is much larger than $\langle k \rangle$, consequence of their scale-free nature. It also lists the estimated degree exponent, γ, for each network, determined using the procedure discussed in **ADVANCED TOPICS 4.A**. The stars next to the reported values indicate the statistical confidence for a particular fit to the degree distribution. That is, * means that the fit shows statistical confidence for a power-law $k^{-\gamma}$ fit; while ** marks datasets that display statistical confidence for a $\sigma_k = \sqrt{\langle k^2 \rangle - \langle k \rangle^2}$ fit. Those with no stars do not show statistical confidence for any of the two forms; the reasons for this are discussed later in the next chapter and in **ADVANCED TOPICS 4.C**. Note that the power grid is not considered scale-free. For this network a degree distribution of the form $e^{-\lambda k}$ offers a statically significant fit.

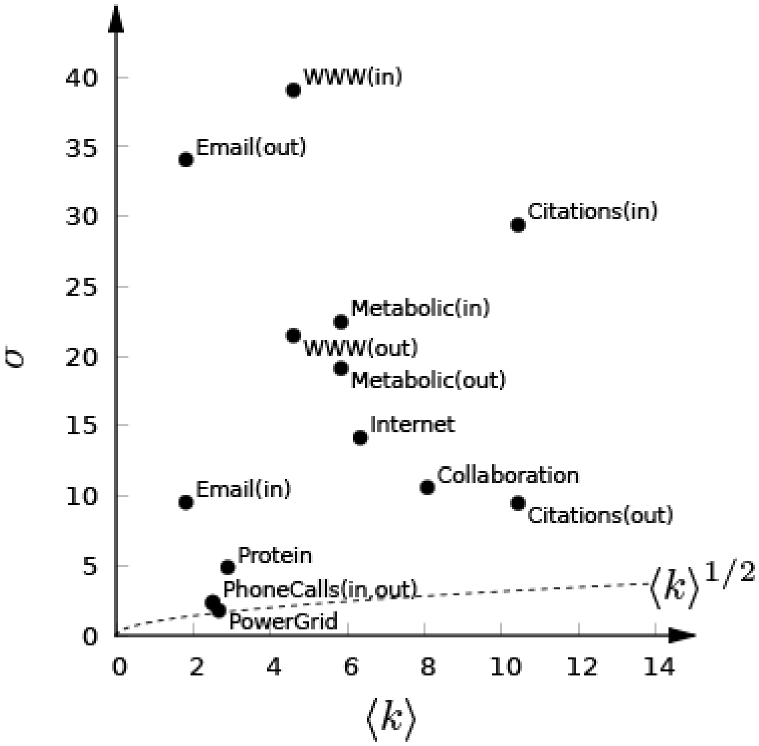


Figure 4.8
Standard deviation is large in real networks

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

SECTION 4.5

UNIVERSALITY

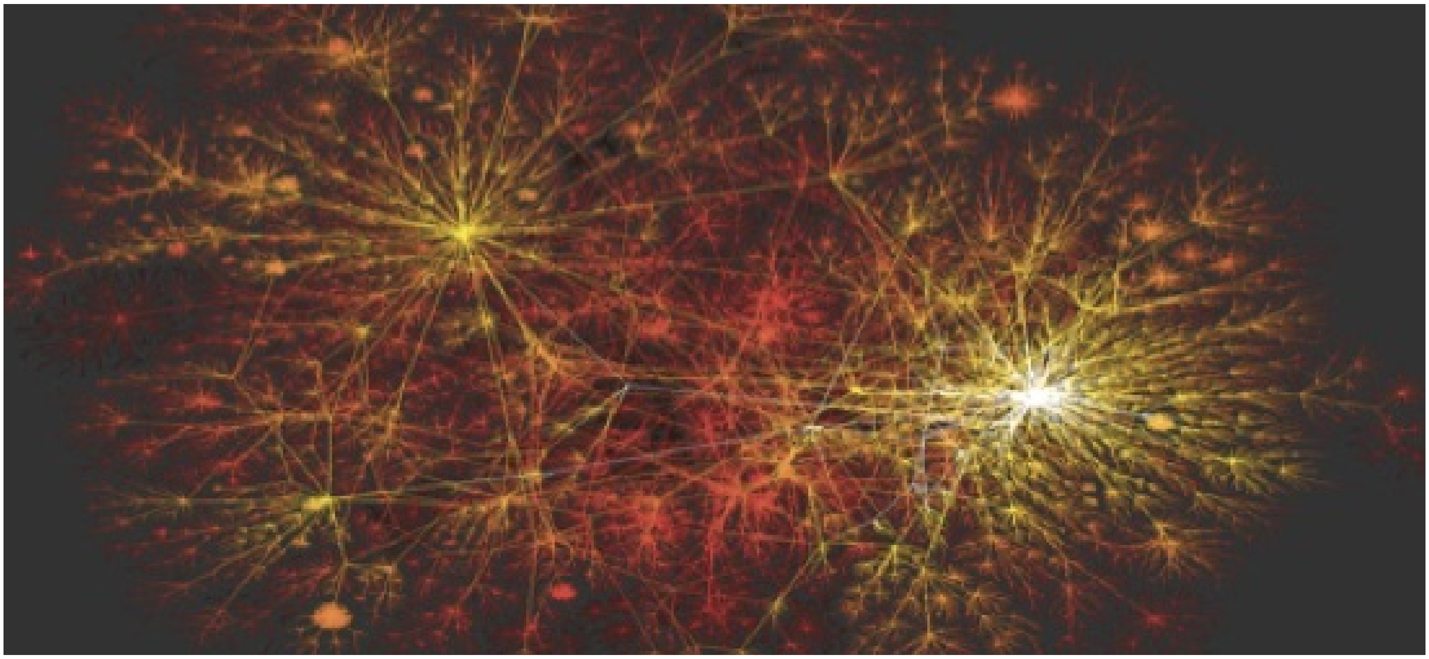


Figure 4.9
The topology of the Internet

An iconic representation of the Internet topology at the beginning of the 21st century. The image was produced by CAIDA, an organization based at University of California in San Diego, devoted to collect, analyze, and visualize Internet data. The map offers a visual demonstration of the Internet's scale-free nature: a few highly connected hubs hold together numerous small nodes.

While the terms 'WWW' and 'Internet' are often used interchangeably in the popular press, they refer to rather different systems. The WWW is an information network, with Web documents as nodes and URLs as links. In contrast the Internet is an infrastructural network, whose nodes are routers and links correspond to physical connections, like copper or optical cables.

This difference has important consequences: while the cost of linking to a web document residing on the same computer or on a different continent is the same, establishing a direct Internet link between routers in Boston and Budapest would require us to lay a new cable between the two continents, which would be prohibitively expensive. Despite these differences, the degree distribution of both networks is well approximated by a power law [1, 5, 6]. We have discussed the scale-free property of the WWW in the previous sections. The signatures of the Internet's scale-free nature are visible in Fig. 4.9, showing that a few high-degree routers hold together a large number of routers with only a few links.

In the past decade many real networks of major scientific, technological and societal importance were found to display the scale-free property. This is illustrated in Fig. 4.10, where we show the degree distribution of an infrastructural network (Internet), a biological network (protein-protein interactions) and a professional affiliation network (Hollywood actors). For each network the degree distribution significantly deviates from a Poisson distribution, being better approximated with a power law.

The diversity of the systems that share the scale-free property is remarkable. Indeed, the WWW is a man-made network with a history of little more than two decades, while the protein interaction network is the product of four billion years of evolution. In some of these networks the nodes are molecules, in others they are computers. It is this diversity that prompts us to call the scale-free property a universal network characteristics.

From the perspective of a researcher, a crucial question is the following: how do we establish the scale-free nature of a network? One one end, a quick look at the degree distribution will immediately reveal whether the network could be scale-free: in scale-free networks we observe orders of magnitude differences between the degrees of the smallest and the largest nodes. In contrast most nodes have comparable degrees in a random network. Yet, as the value of the degree exponent plays an important role in predicting various network properties, we need tools to fit the p_k distribution and to estimate γ. This prompts us to address several issues:

PLOTTING THE DEGREE DISTRIBUTION

The degree distributions shown in this chapter are all plotted on a double logarithmic scale, often called a log-log plot. The main reason is that when nodes with widely different degrees coexist, a linear plot is unable to display them all. We also use logarithmic binning to obtain the clean-looking degree distributions shown throughout this book, ensuring that each datapoint has proper statistical significance. The practical tips for plotting a network's degree distribution are discussed in ADVANCED TOPICS 4.B.

MEASURING THE DEGREE EXPONENT

A quick estimate of the degree exponent is often obtained by fitting a straight line to p_k on a log-log plot.Yet, this approach can be affected by systematic biases, resulting in an incorrect γ. The statistical tools available to estimate γ are discussed in ADVANCED TOPICS 4.C. We used these tools to determine the degree exponents listed in Table 4.1.

THE SHAPE OF p_k FOR REAL NETWORKS

Most degree distributions observed in real networks display clear deviations from a pure power law. These can be attributed to data incompleteness or data collection biases, but the deviations also carry important information about processes that contribute to the emergence of a particular network. In ADVANCED TOPICS 4.B we discuss some of these deviations, and in CHAPTER 6 we explore their origins.

Since the discovery of the scale-free nature of the WWW, an amazing number of real networks of major scientific and technological interest have been found to be scale-free Fig. 4.10 from biological to social and even linguistic networks. This does not mean that all networks are scale-free. Indeed, many important networks, from the power grid to networks observed in materials science BOX 4.2 do not display the scale-free property.

Yet, the prevalence of the scale-free property have prompted the research community to devote special attention to this class of networks. Uncovering the reasons why some networks are scale-free while others are not, and understanding the consequences of the scale-free property, help us better understand real networks.

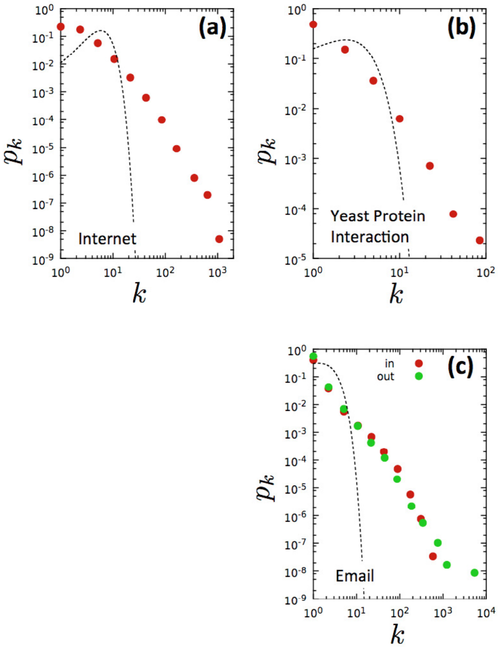


Figure 4.9b

Many real networks are scale-free

The degree distribution of three of the networks listed in **Table 4.1**.

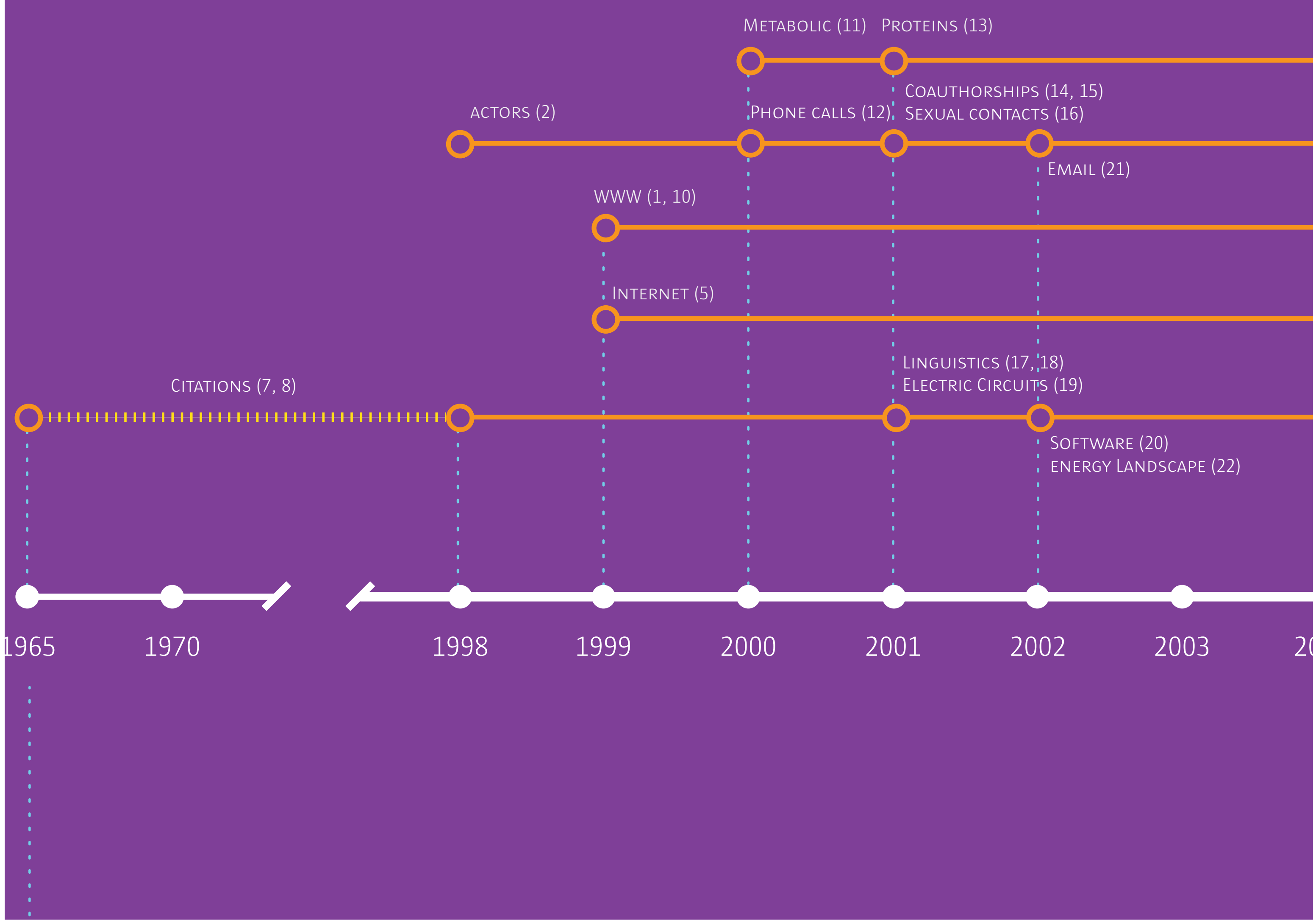
(a) The degree distribution of the Internet at the router level.
(b) The degree distribution of the protein-protein interaction network of yeast.
(c) The degree distribution of the email network of a European university.

In each panel, the dotted line shows the Poisson distribution with the same ⟨k⟩ as the real network, indicating that the random network model cannot account for the observed p_k.

SCALE-FREE HISTORY

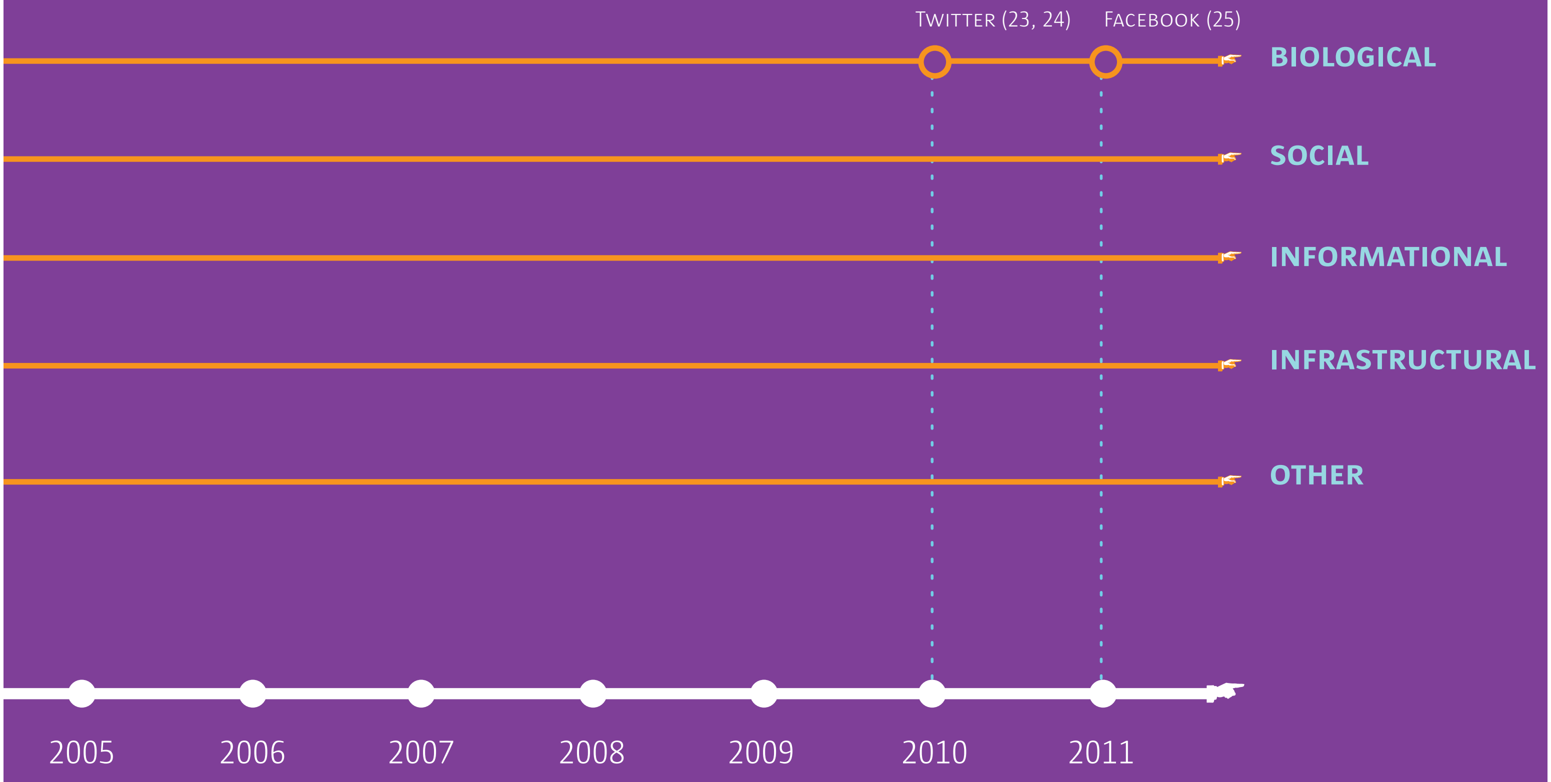
THE TIMELINE OF THE DISCOVERIES REPORTING THE SCALE-FREE NATURE OF VARIOUS REAL NETWORKS

FIG. 4.10



Many biological, social, and technological networks display the scale-free property. The figure shows the timeline of the discoveries reporting the scale-free nature of various real networks. While there is a clear burst of reports following the 1999 discovery of scale-free networks, in hindsight it is clear that several early papers have reported characteristics that are consistent with what we call today a scale-free topology. For example, Etel de Solla Price reported in 1965 that citations to scientific papers follow a power-law distribution [7], a property independently discovered by Redner in 1998 [8]. This is a consequence of the scale-free nature of citation networks.

A common feature of these early works is that they viewed the observed quantities as scalar events, not as a manifestation of some network phenomena. It wasn't until the 1999 that it was understood that power laws are also a fundamental network property. Indeed, Barabási and Albert, in their 1999 Science paper argued that "we expect that the scale-invariant state observed in all systems for which detailed data has been available to us is a generic property of many complex networks, with applicability reaching far beyond the quoted examples." The 'scale-free network' term was also first used in 1999 [2, 9].



BOX 4.2

Not all network are scale-free

The ubiquity of the scale-free property does not mean that all real networks are scale-free. Indeed, several important networks do not share this property:

- Networks appearing in material science, like the network describing the bonds between the atoms in crystalline or amorphous materials, where each node has exactly the same degree.

- The neural network of the *C. elegans* worm.

- The power grid, consisting of generators and switches connected by transmission lines.

For the scale-free property to emerge the nodes need to have the capacity to link to an arbitrary number of other nodes. These links do not need to be simultaneous: we do not constantly chat with each of our acquaintances and a protein in the cell does not simultaneously bind to each of its potential interaction partners. In general the scale-free property is absent in systems that have a limitation in the number of links a node can have, as such limitations limit the size of the hubs. As illustrated in the image, such limitations are common in materials, explaining why they cannot develop a scale-free topology.

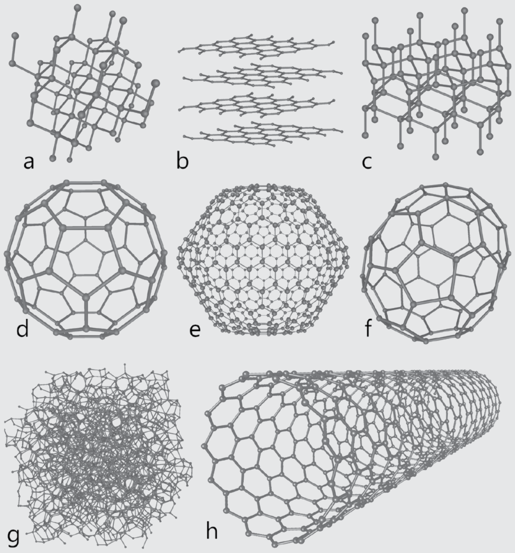


Figure 4.11
The material network

A carbon atom can share only four electrons with other atoms, hence no matter how we arrange these atoms relative to each other, in the resulting network a node can never have more than four links. Hence, hubs are forbidden and the scale-free property cannot emerge. The figure shows several carbon allotropes, each characterized by a different "network", resulting in materials with different physical characteristics, like (a) diamond; (b) graphite; (c) lonsdaleite; (d) C60 (buckminsterfullerene); (e) C540 (a fullerene) (f) C70 (another fullerene); (g) amorphous carbon; (h) single-walled carbon nanotube.

Source: http://www.thenanoage.com/Figures/Eight_Allotropes_of_Carbon.png

SECTION 4.6

ULTRA-SMALL PROPERTY

The presence of hubs in scale-free networks raises an interesting question: how do hubs affect the small world property?

Figure 4.4 suggests that they do: airlines build hubs precisely to decrease the number of hops between two airports. The calculations support this expectation, finding that *distances in a scale-free network are either smaller or equal to the distances observed in an equivalent random network.* The precise dependence of the average distance $\langle d \rangle$ on the system size N and the degree exponent γ are captured by the expression [26, 27].

$$d \sim \begin{array}{ll} \text{const.} & \text{if } \gamma = 2, \\ \dfrac{\ln\ln N}{\ln(\gamma - 1)} & \text{if } 2 < \gamma < 3, \\ \dfrac{\ln N}{\ln\ln N} & \text{if } \gamma = 3, \\ \ln N & \text{if } \gamma > 3. \end{array} \qquad (4.22)$$

In the following we discuss the behavior of $\langle d \rangle$ in the four regimes predicted by Eq. 4.22, Fig. 4.12:

ANOMALOUS REGIME $\gamma = 2$

According to Eq. 4.19 for $\gamma = 2$ the degree of the biggest hub grows linearly with the system size, i.e. $k_{max} \sim N$. This forces the network into a hub and spoke configuration in which all nodes are at a short distance from each other. In this regime the average path length does not depend on N.

ULTRA-SMALL WORLD $2 < \gamma < 3$

As several real networks have degree exponent between two and three Table 4.1, this regime is of particular practical interest. Eq. 4.22 predicts that the average distance increases as $\ln\ln N$, a significantly slower dependence than the $\ln N$ we derived earlier for random networks. We call networks in this regime ultra-small, as the hubs radically reduce the path length [27]. They do so by linking to a large number of small-degree nodes, creating short distances between them.

To see the implication of the ultra-small property let us consider again the social network with $N \simeq 7\text{x}10^9$. If the society were to be random, the N-dependent term is $\ln N = 22.66$. In contrast for a scale-free network the N-dependent term is $\ln\ln N = 3.12$ according to Eq. 4.22, supporting our conclusion that hubs radically shrink the distance between the nodes.

CRITICAL POINT $\gamma = 3$

This value is of particular theoretical interest, as the second moment of the degree distribution does not diverge any longer, prompting us to call $\gamma = 3$ the "critical point." At this critical point the $\ln N$ dependence encountered for random networks returns. Yet the calculations indicate the presence of a double logarithmic correction $\ln\ln N$ [27, 28], which shrink slightly the distances compared to a random network of similar size.

SMALL WORLD $\gamma > 3$

In this regime $\langle k^2 \rangle$ is finite and the average distance follows the small world result derived for random networks. While hubs continue to be present, for $\gamma > 3$ they are not sufficiently large and numerous to have a significant impact on the distance between the nodes.

Taken together, Eq. 4.22 indicates that the more pronounced the hubs are, the more effectively they shrink the distances between the nodes. This conclusion is supported by Fig. 4.11a, which shows the scaling of the average path length for scale-free networks with different γ.

The figure indicates that while for small N the distances in the four regimes are comparable, for large N the differences are remarkable. Further support for this conclusion is provided by the path length distribution for scale-free networks with different γ and N Fig. 4.11b-d. For $N = 10^2$ the path length distributions largely overlap, indicating that at this size differences in γ result in insignificant differences in the path length. For $N = 10^6$, however, p_d observed for different γ are well separated. Fig. 4.11d also shows that the larger the degree exponent, the larger are the distances between the nodes. In summary the scale-free property has two effects on network distances:

- Shrinks the average path lengths.

- Changes the dependence of $\langle d \rangle$ on the system size, as predicted by Eq. 4.21. The smaller γ, the shorter are the distances between the nodes.

Therefore, most scale-free networks of practical interest are not only "small", but are "ultra-small". This is a consequence of the hubs, that act as bridges between the many small nodes. Only for $\gamma > 3$ we recover the small-world property encountered in random networks Fig. 4.12.

BOX 4.3

WE ARE ALWAYS CLOSE TO THE HUBS

Frigyes Karinthy in his 1929 short story [30] that introduced the small world concept writes that "it's always easier to find someone who knows a famous or popular figure than some run-the-mill, insignificant person".

In other words, we are typically closer to hubs than to less connected nodes. This effect is particularly pronounced in scale-free networks as shown in the figure below. The implications are obvious: there are always short paths linking us to famous individuals like well known scientists or to the president of the United States, as they are hubs with an exceptional numbers of acquaintances. It also means that many of the shortest paths go through these hubs.

In contrast with this expectation, recent measurements designed to replicate the six degrees concept in the online world find that the paths that individuals used to reach their target node involve rather few hubs [31]. That is, individuals involved in successful chains (those that reached their target) were less likely to send a message to a hub than individuals involved in incomplete chains. The reason may be self-imposed, we perceive hubs as busy, hence we contact them only in real need. We therefore avoid them in online experiments of no perceived value to us.

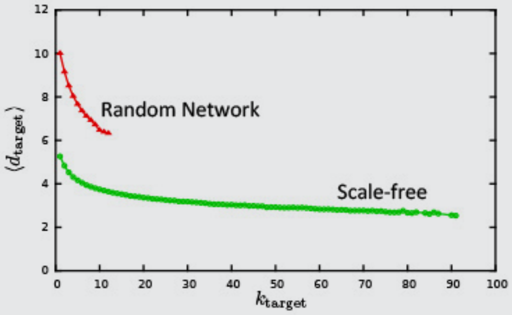


Figure 4.11b
Closing on the hubs

The distance $\langle d_{target} \rangle$ of a node with degree $k \simeq \langle k \rangle$, to a target node with degree k_{target} in a random and a scale-free network. In scale-free networks our distance to the hubs is shorter than in random networks. The figure also documents that in a random network the largest-degree nodes are considerably smaller and hence the path lengths are visibly longer than in a scale-free network. Both networks have $\langle k \rangle$ = 2 and N = 1,000 and for the scale-free network γ = 2.5.

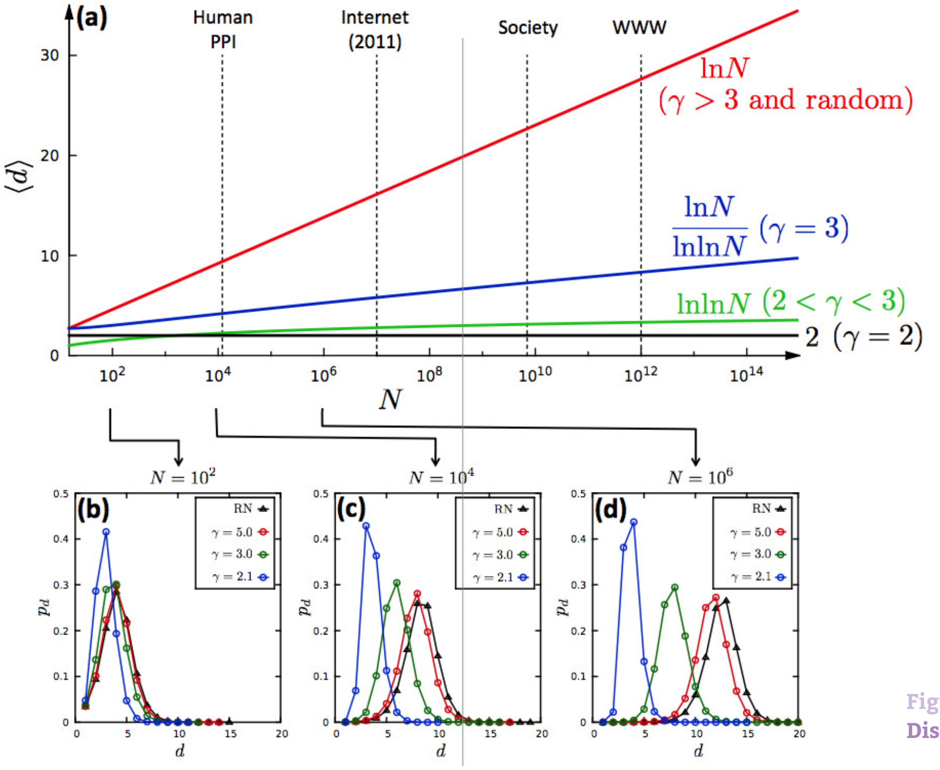


Figure 4.12
Distances in scale-free networks

(a) The scaling of the average path length in the four scaling regimes characterizing a scale-free network: ln$N$ (scale-free networks with γ > 3 and random networks), ln$N$/lnln$N$ (γ = 3) and lnln$N$ (2 < γ< 3). The dotted lines mark the approximate size of several real networks of practical interest. For example, given their modest size, in biological networks the differences in the node to node distances are relatively small in the four regimes. The differences become quite relevant for networks of the size of the social network or the WWW. For these the small-world formula considerably underestimates the real value of $\langle d \rangle$.

(b)(c)(d) Distance distribution for networks of size $N = 10^2, 10^4, 10^6$, illustrating that while for small N (= 10^2) the distance distributions is not too sensitive to γ, for large N (= 10^6) p_d and $\langle d \rangle$ changes visibly with γ. As (d) shows, the smaller γ, the shorter are the distances between the nodes. The networks were generated using the static model [29] with $\langle k \rangle = 3$.

SECTION 4.7

THE ROLE OF THE DEGREE EXPONENT

Many properties of a scale-free network depend on the value of the degree exponent γ. A close inspection of Table 4.1 indicates that:

- γ varies from system to system, prompting us to explore how the properties of a network change with γ

- For many real systems the degree exponent is between 2 and 3, prompting us to ask: why don't we see systems with $\gamma < 2$ and why are so few systems with $\gamma > 3$? To address these questions next we discuss how the properties of a scale-free network change with γ Fig. 4.13

ANOMALOUS REGIME ($\gamma \leq 2$)

According to Eq. 4.18, for $\gamma < 2$ the exponent $1/(\gamma - 1)$ is larger than one, hence the fraction of links connected to the largest hub grows faster than the size of the network. This means that for sufficiently large N the degree of the largest hub must exceed the total number of nodes in the network, running out of nodes to connect to. Similarly, for $\gamma < 2$ the average degree $\langle k \rangle$ diverges in the $N \rightarrow \infty$ limit. These odd predictions are only two of the many anomalous features of scale-free networks in this regime. They represent signatures of a deeper problem: large scale-free network with $\gamma < 2$, that lack self-loops or multi-links, cannot exist BOX 4.4. Hence one needs to inspect with caution any research reporting networks with $\gamma < 2$. Such networks can only exist if the hubs have many self-loops or if multiple links can connect the same pair of nodes.

SCALE-FREE REGIME ($2 < \gamma < 3$)

In this regime the first moment $\langle k \rangle$ of the degree distribution is finite but the second and higher moments diverge as $N \rightarrow \infty$. Consequently scale-free networks in this regime are ultra-small (see SECTION 4.6). Eq. 4.18 predicts that k_{max} grows with the size of the network with exponent $1/(\gamma - 1)$, which is smaller than one. Hence the market share of the largest hub, k_{max}/N, representing the fraction of nodes that connect to it, decreases as $k_{max}/N \sim N^{(2-\gamma)/(\gamma-1)}$.

As we will see in the coming chapters, many interesting features of scale-free networks, from their robustness to failures to anomalous spreading phenomena, are linked to this regime.

RANDOM NETWORK REGIME (γ > 3)

According to Eq. 4.20 for $\gamma > 3$ both the first and the second moments are finite. For all practical purposes the properties of a scale-free network in this regime are difficult to distinguish from the properties a random network of similar size. For example Eq. 4.21 indicates that the average distance between the nodes converges to the small-world formula derived for random networks.The reason is that for large γ the degree distribution p_k decays sufficiently fast to make the hubs smaller and less numerous. The larger γ, the smaller are the hubs (see Eq. 4.18), hence the more indistinguishable is the structure and the behavior of a scale-free network from that of a random network.

Table 4.1 and Fig. 4.12 also indicate that there are fewer networks with $\gamma > 3$, prompting us to ask: does this imply that networks with $\gamma > 3$ cannot exist? A quick calculation indicates that they may exist, but it is hard to distinguish them from a random network. To document the presence of a power-law degree distribution we ideally need 2-3 orders of magnitude of scaling, which means that k_{max} should be at least 10^2 - 10^3 times larger than k_{min}. By inverting Eq. 4.18 we can calculate the network size necessary to observe the desired scaling regime between k_{min} and k_{max}, obtaining

$$N \gg \frac{k_{\max}}{k_{min}}^{\gamma - l} . \tag{4.23}$$

For example, in order to document the scale-free nature of a network with $\gamma = 5$ with $k_{min} \sim 1$ and $k_{max} \simeq 10^2$, according to Eq. 4.23 the size of the network must exceed $N \gg 10^8$. There are very few network maps of this size available for research. Therefore, there may be many real networks with exponent larger than 3, but given their limited size, it is difficult to obtain convincing evidence of their scale-free nature. Hence they are mistakenly classified as networks with an exponential degree distribution.

In summary, we find that the behavior of scale-free networks depends on the value of the degree exponent γ. Theoretically the most interesting regime is $2 < \gamma < 3$, where scale-free networks are ultra-small and $\langle k^2 \rangle$ diverges. Interestingly, many networks of practical interest, from the WWW to protein interaction networks, are in this regime.

BOX 4.4

SCALE-FREE NETWORK WITH□$\gamma < 2$ DO NOT EXIST

To see why networks with $\gamma < 2$ are problematic, we need to attempt to build one. A degree sequence that can be turned into simple graph (i.e. a graph lacking multilinks or self-loops) is called graphical [32]. Yet, not all degree sequences are graphical: if for example the number of stubs is odd, then we will always have an unmatched stub, as shown in Fig. 4.13b.

The graphicality of a degree sequence can be tested with an algorithm proposed by Erdős and Gallai [32, 33, 34, 35]. If we apply the algorithm to scale-free networks we find that the number of graphical degree sequences drops to zero for $\gamma < 2$. Hence degree distributions with $\gamma < 2$ cannot be turned into a network. Indeed, for networks in this regime the largest hub grows faster than N. If we do not allow self-loops and multi-links, then the degree of the largest hub cannot exceed $N - 1$.

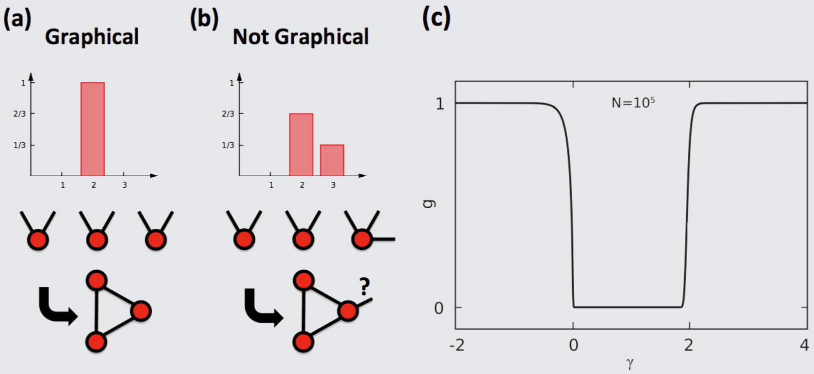


Figure 4.13
Networks with $\gamma < 2$ are not graphical

(a-b) Two degree distributions and the corresponding degree sequences. The difference is limited to the degree of a single node. While we can build a network consistent with the degree distribution (a), it is impossible to build one from (b), as one stub always remains unmatched. Hence (a) is graphical, while (b) is not.

(c) Fraction of networks with a given γ that are graphical. A large number of degree sequences with degree exponent γ and $N = 10^5$ were generated, testing the graphicality of each network.

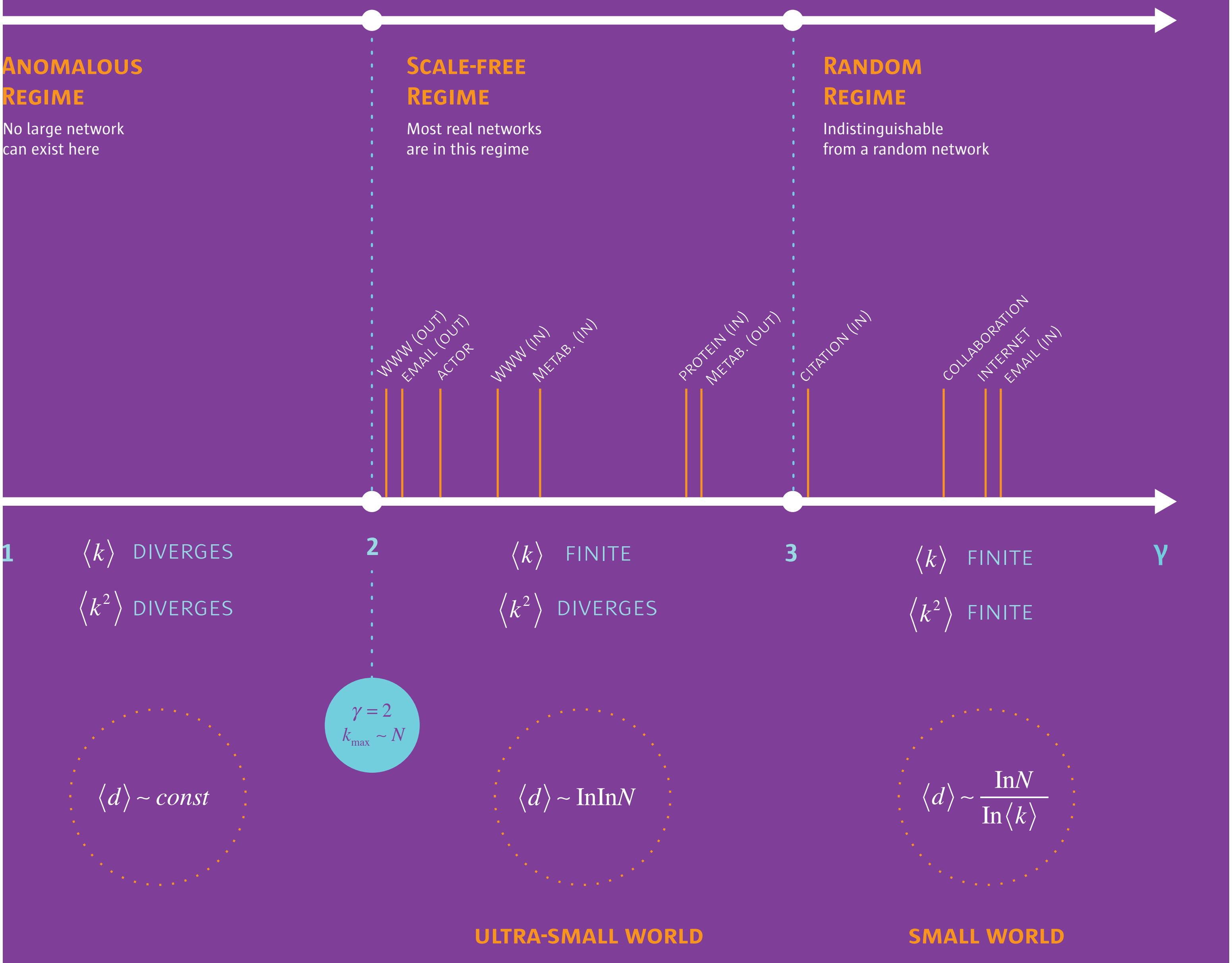
While virtually all networks with $\gamma > 2$ are graphical, it is impossible to find graphical networks with $0 < \gamma < 2$.

DEPENDENT PROPERTIES

A SUMMARY OF THE γ DEPENDENT PROPERTIES OF SCALE-FREE NETWORKS

FIG. 4.14

The degree exponents shown in the figure were taken from Table 4.1. Note that not all listed γ values show statistical significance, as we lack the proper fitting function. Case in point are the Internet and the email datasets, for which earlier studies reported γ < 3. To determine the precise value of γ, we need proper models, a topic discussed in Chapter 6.


ANOMALOUS REGIME
No large network can exist here
SCALE-FREE REGIME
Most real networks are in this regime
RANDOM REGIME
Indistinguishable from a random network
WWW (OUT)
EMAIL (OUT)
ACTOR
WWW (IN)
METAB. (IN)
PROTEIN (IN)
METAB. (OUT)
CITATION (IN)
COLLABORATION
INTERNET
EMAIL (IN)
1
2
3
γ
$\langle k \rangle$ DIVERGES
$\langle k^2 \rangle$ DIVERGES
$\langle k \rangle$ FINITE
$\langle k^2 \rangle$ DIVERGES
$\langle k \rangle$ FINITE
$\langle k^2 \rangle$ FINITE
$\gamma = 2$
$k_{\max} \sim N$
$\langle d \rangle \sim const$
$\langle d \rangle \sim \text{InIn}N$
$\langle d \rangle \sim \frac{\text{In}N}{\text{In}\langle k \rangle}$
ULTRA-SMALL WORLD
SMALL WORLD

SECTION 4.8

GENERATING NETWORKS WITH A PRE-DEFINED DEGREE DISTRIBUTION

The Erdős-Rényi model generates networks with a Poisson degree distribution. The empirical results discussed in this chapter indicate, however, that the degree distribution of most real networks significantly deviates from a Poisson form. This raises an important question: how do we generate networks with an arbitrary p_k? In the following we discuss the three most frequently used algorithms for this purpose.

CONFIGURATION MODEL

The configuration model helps us build a network with a pre-defined degree sequence Fig. 4.15a. In the obtained network each node has a pre-defined degree k_i, but otherwise the network is wired randomly. Consequently the obtained network is often called a random network with a pre-defined degree sequence. By repeatedly applying this procedure to the same degree sequence we can generate different networks with the same p_k Fig. 4.14, panels (2a)-(2c). A couple of a caveats to consider:

- The probability to have a link between nodes of degree k_i and k_j is

$$p_j = \frac{k_i k_j}{2L-1} \tag{4.24}$$

Indeed, a stub starting from node i can connect to $2L$ - 1 other stubs. Of these, k_j are attached to node j. So the probability that a particular stub is connected to a stub of node j is k_j /(2L - 1). As node i has k_i stubs, it will have k_j attempts to link to j, resulting in Eq. 4.24.

- The obtained network will contain self-edges and multi-edges. We can choose to reject stub pairs that lead to these, but if we do so, we may not be able to complete the network. Rejecting self- or multi-edges means that not all possible matchings appear with equal probability. Hence Eq. 4.24 will not be valid any longer, making analytical calculations difficult. The number of self- and multi-edges goes to zero for large networks, so in most cases we do not need to exclude them [39]. The configuration model is frequently used in analytical calculations, as Eq. 4.24 and its inherently random character helps us calculate numerous network measures.

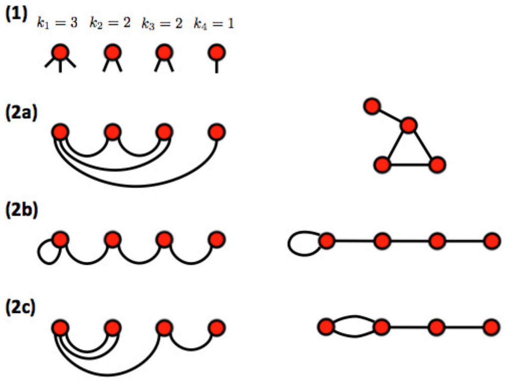


Figure 4.15a
The configuration model

The configuration model allows us to build a network where each node has some pre-defined degree [37, 38]. It consists of the following steps:

(1) Degree sequence: Assign a degree to each node, represented as stubs or half-links. The degree sequence is either generated analytically from a preselected pk distribution BOX 4.5, or it is extracted from the adjacency matrix of a real network. We must start from an even number of stubs, otherwise we will be left with unpaired stubs.

(2) Network assembly: Randomly select a stub pair and connect them. Then randomly choose another pair from the remaining $2L$ - 2 stubs and connect them. This procedure is repeated until all stubs are paired up. Depending on the order in which the stubs were chosen, we obtain different networks. Some networks include cycles (2a), others self-edges (2b) or multi-edges (2c). Yet, the expected number of self- and multi-edges goes to zero in the $N \to \infty$ limit.

DEGREE PRESERVING RANDOMIZATION

As we explore the properties of a real network, we often need to ask if a certain network property is predicted by its degree distribution alone, or if it represents some additional property not contained in p_k. To answer this question we need to generate networks that are wired randomly, but whose p_k is identical to the original network.

This can be achieved through the degree-preserving randomization [40] described in Fig. 4.14. The idea behind the algorithm is simple: we randomly choose two links in the network and swap them, so that the degree of each of the four involved nodes in the swap remains unchanged. Hence, hubs will stay hubs and small-degree nodes will retain their small degree, but the wiring diagram of the generated network will be randomized. Note that degree-preserving randomization is different from full randomization, where we swap links without preserving the node degrees Fig. 4.14. Complete randomization turns any network into an Erdős-Rényi network, hence independent of the original p_k, the randomized version will have a Poisson degree distribution.

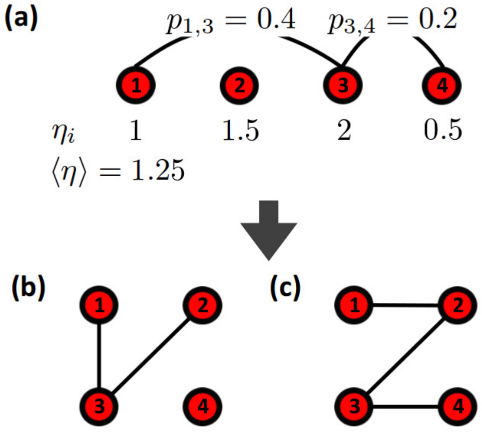


Figure 4.15b
Hidden parameter model

We start with N isolated nodes and assign to each node a "hidden parameter" η_i, which can be randomly selected from a $\rho(\eta)$ distribution or it is provided by a deterministic sequence $\{\eta_i\}$. We next connect each node pair with probability

$$p(\eta_i, \eta j) = \frac{\eta_i \eta j}{(\eta)N}.$$

For example, the figure shows the probability to connect nodes (1,3) and (3,4). After connecting the nodes, we end up with the networks shown in (b) or (c), representing two independent realizations generated by the same hidden parameter sequence (a). The expected number of links in the obtained network is

$$L = \frac{1}{2}\sum_{N}^{i,j}\frac{\eta_i\eta_j}{\langle\eta\rangle N} = \frac{1}{2}\langle\eta\rangle N.$$

Just like in the random network model, L will differ from network to network, following a bounded distribution. If we wish to control precisely the average degree $\langle k\rangle$ we can add the L links to the network one by one. The end points i and j of each link are then chosen randomly with a probability proportional to η_i and η_j, following. In this case we connect i and j only if they were not connected previously.

BOX 4.5

GENERATING A DEGREE SEQUENCE WITH POWER-LAW DISTRIBUTION

The degree sequence of an undirected network is a non-increasing sequence of the node degrees. For example, the degree sequence of each of the networks shown in Fig. 4.15a is {3, 2, 2, 1}. As Fig. 4.15a illustrates, the degree sequence in general does not uniquely identify a graph. There can be multiple graphs with the same degree sequence. We often need to generate a degree sequence from a pre-defined degree distribution. Our purpose here is to provide the tools to achieve this. We start from an analytically pre-defined degree distribution, like $p_k \sim k^{-\gamma}$, shown in panel (a). Our goal is to generate a degree sequence $\{k_1, k_2, ..., k_N\}$ of N degrees that follow the distribution p_k. We start by calculating the complementary cumulative distribution function

$$D(k) = \sum_{k' \geq k} p_{k'}, \tag{4.25}$$

shown in (b). $D(k)$ is between 0 and 1, and the step size at any k equals p_k. Therefore, to generate a sequence of N random numbers following a pre-defined p_k distribution, we generate N random numbers r_i, i = 1, ... , N, chosen from the (0, 1) interval. For each r_i we use the plot in (b) to assign a degree k_i. The obtained $k_i = D^{-1}(r_i)$ set will follow the desired p_k distribution. Note that the degree sequence assigned to a p_k is not unique - we can generate multiple sets of $\{k_1, ..., k_N\}$ sequences compatible with the same p_k.

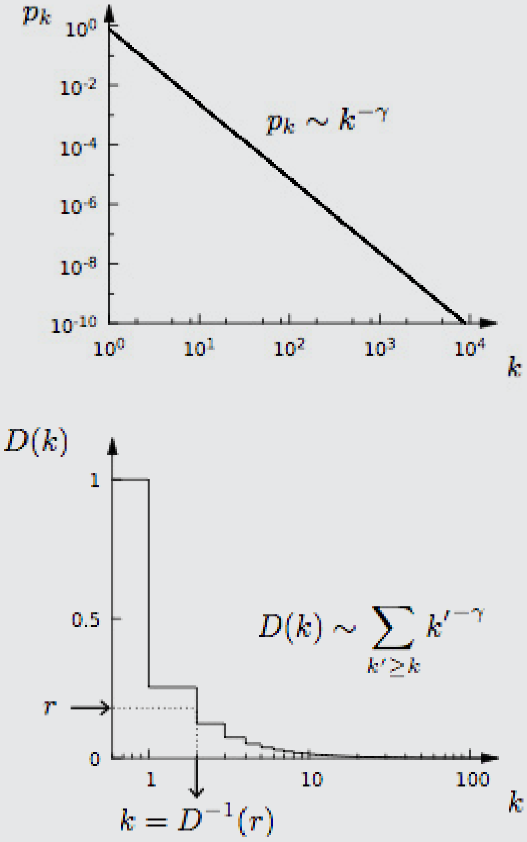


Figure 4.16

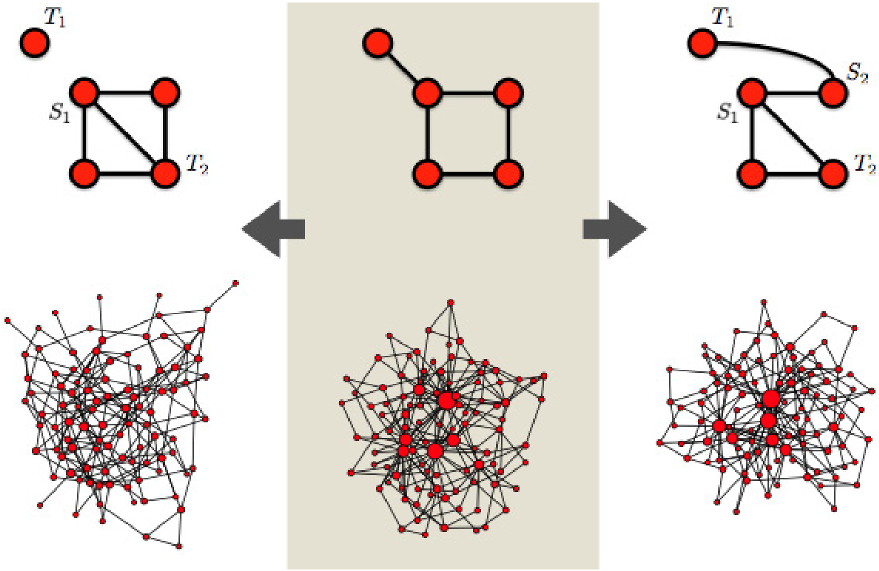


Figure 4.17

Degree preserving randomization

Two randomization methods are used to generate random references to a given network [40]. Full randomization generates a random (Erdős–Rényi) network with the same *N* and *L* as the original network. For this we select randomly a source node (S_1) and two target nodes, where the first target is linked directly to the source node (T_1) and the second target is unconnected to it (T_2). We then rewire the S_1-T_1 link, turning it into an S_1-T_2 link. As a result the degree of the target nodes T_1 and T_2 changes. We perform this procedure once for each link in the network.

Degree-preserving randomization generates a network in which each node has exactly the same degree as in the original network, but the network's wiring diagram has been randomized. We select two source (S_1, S_2) and two target nodes (T_1, T_2), such that initially there is a link between S_1 and T_1, and a link between S_2 and T_2. We then swap the two links, creating an S_1-T_2 and an S_2-T_1 link. This swap leaves the degree of each node unchanged.We repeat this process until we rewire at least once each link.

Bottom panels: Starting from a scale-free network (middle panel), full randomization eliminates the hubs and turns the network into a random network (left panel). In contrast, degree-preserving randomization leaves the hubs in place and hence the network remains scale-free (right panel).

HIDDEN PARAMETER MODEL

In their most general (and most useful) form the configuration and the rewiring model generate loops and multi-links. Loops and multi-links are absent, however, from many real networks. We can use the hidden parameter model, described in Fig. 4.15b, to generate networks with a pre-defined p_k but without multi-links and self-loops [41, 42, 43]. In the model we start from N isolated nodes and assign each node i a hidden parameter η_i, chosen from a distribution $\rho(\eta)$. The nature of the network generated by the hidden parameter model depends on the selection of a $\{\eta_i\}$ hidden parameter sequence.

There are two ways to generate the appropriate hidden parameters:

(i) η_i can be a sequence of N random number chosen from a pre-defined $\rho(n)$ distribution. In this case the degree distribution of the obtained network is

$$p_k = \int \frac{e^{-\eta}\eta^k}{k!} p(\eta) d\eta. \qquad (4.26)$$

(ii) η can come from a deterministic sequence $\{\eta_1, \eta_2, \ldots, \eta_3\}$. In this case the degree distribution of the obtained network is

$$p_k = \frac{1}{N} \sum_j \frac{e^{e-\eta_j} {\eta_j}^k}{k!}. \qquad (4.27)$$

The hidden parameter model offers a particularly simple method to generate a scale-free network. Indeed, using

$$\eta_i = c / i^a, i = 1, \ldots, N. \qquad (4.28)$$

as the sequence of hidden parameters, according to Eq. 4.27 the obtained network will have the degree distribution

$$p_k \sim k^{-(1+1/a)} \qquad (4.29)$$

for large k. We can use $\langle \eta \rangle$ to tune $\langle k \rangle$ as Eq. 4.26 and Eq. 4.27 imply $\langle k \rangle = \langle \eta \rangle$. The three methods discussed above for creating networks with a pre-defined p_k raise the following question: how do we decide which one to use? Our choice depends on whether we start from a degree sequence $\{k_i\}$ or a degree distribution p_k and whether we can tolerate self-loops and multiple links between two nodes. The decision tree involved in this choice is provided in Fig. 4.18.

In summary, the configuration model, degree-preserving randomization and the hidden parameter model are attractive because they generate networks with a pre-defined degree distribution and allow us to analytically calculate several network properties.

We will turn to these each time we explore if a certain network property is a consequence of the network's degree distribution, or it represents some emerging property of the modeling network BOX 4.6. Yet, these models also have a number of limitations:

- These algorithms do not tell us why a network has a certain degree distribution

- Several important network characteristics, present in real networks, from clustering to degree correlations, are lost during randomization

Hence the networks generated by these algorithms are a bit like a photograph of a painting: at first look they appear to be the same as the original. But upon closer inspection we realize that many details, from the texture of the canvas to the brush strokes, are lost.

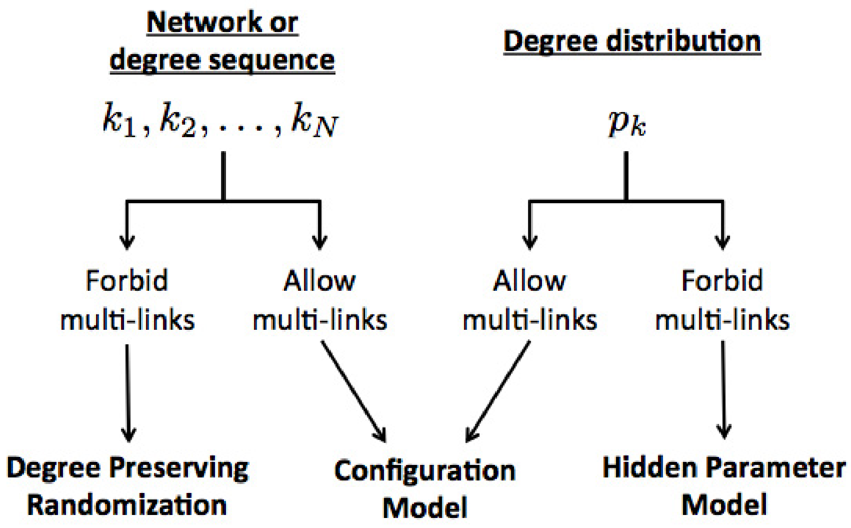


Figure 4.18
Choosing the proper generative model

The choice of the appropriate generative model depends on our starting point as well as our tolerate towards self-loops and multi-links. If we start from the analytical form of the degree distribution, p_k, then the goal is to generate networks whose degree distribution is consistent with p_k. In this case if we allow self-loops and multi-links, we should use the configuration model; if we forbid them, then the hidden parameter model is a better choice.

If we start from a real network or known degree sequence our goal is often to generate networks with the degree sequence identical to the original network. Again if we allow self-loops and multi-links, the configuration model is an appropriate choice; if we wish to forbid them, we can use degree-preserving randomization.

BOX 4.6

TESTING THE SMALL-WORD PROPERTY

A common practice in the network literature is to compare the distances observed in a real network to the small-world formula Eq. 4.19 from CHAPTER 3. Yet, Eq. 4.19 was derived for random networks, while most real networks do not have a Poisson degree distribution. If the network is scale-free, then Eq. 4.22 offers the appropriate formula. That, however, provides only the scaling of the distance with *N*, and not its absolute value. Hence instead of trying to fit the average distance, we often ask the following question: are the distances observed in the real network comparable with the distances observed in a randomized network with the same degree distribution? We can use degree preserving randomization to answer this. We illustrate the procedure on the protein interaction network (PIN) of yeast.

(i) Original p_k: we start by measuring the distance distribution p_d of the original network, obtaining $\langle d \rangle$ = 5.61 (red curve).

(ii) Full randomization: next we generate a random network with the same *N* and *L* as the original network. The obtained p_d (blue curve) is visibly shifted to the right, providing $\langle d \rangle$ = 7.13, much larger than the original $\langle d \rangle$ = 5.61. It is tempting to conclude that the protein interaction network is affected by some unknown organizing principle that keeps the distances shorter than expected in a random configuration. The result (iii) shows that this would be a flawed conclusion, as the difference is explained by the degree distribution.

(iii) Degree preserving randomization: as the original network is scale-free, the proper random reference is a network with the same degree distribution as the original. Hence we determine pd after degree-preserving randomization, finding that it is comparable to the original p_d (green curve).

This indicates that a random network overestimates the distances between the nodes, as it is missing the hubs presented in the original network. The network obtained by degree preserving randomization preserves these hubs, and its distances are comparable to the original network. This example illustrates the importance of choosing the proper random reference frame.

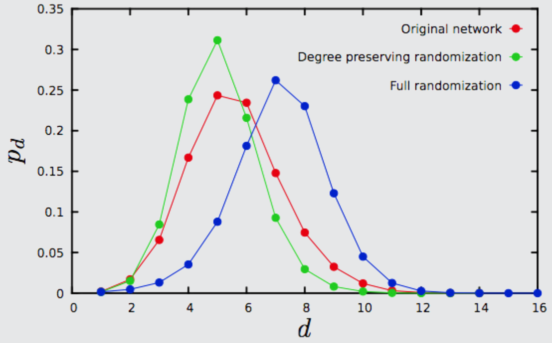


Figure 4.19
Randomizing real networks

The distance distribution p_d (red symbols) between each node pair in *S. Cerevisae* protein-protein interaction network Table 4.1. The purple symbols provide the path-length distribution obtained under full randomization, which turns the original network into an Erdős-Rényi network with the same *N* and *L* as the original network Fig. 4.17.

The green symbols correspond to p_d of the network obtained after degree-preserving randomization, which keeps the degree of each node unchanged.

We have: $\langle d \rangle$=5.61±1.64 (original), $\langle d \rangle$=7.13 ± 1.62 (full randomization), $\langle d \rangle$=5.08 ± 1.34 degree-preserving randomization.

SECTION 4.9

SUMMARY

There are two main reasons why the scale-free property played a key role in the emergence of network science.

First, many networks of scientific and practical interest, from the WWW to the cell, are scale-free.

Second, once the hubs, that accompany the scale-free property, are present, they have an enormous impact on the system's behavior. The ultra-small property offers a first hint of the hubs's impact on a network's properties; we will encounter many more in the coming chapters.

As we continue exploring the consequences of the scale-free property, we must keep in mind that the power-law form Eq. 4.1 is rarely seen in this pure form in real systems. The reason is simple: a host of processes affect the topology of real networks, which also influence the shape of the degree distribution. We will discuss these processes in the coming chapters. The diversity of these processes and the complexity of the resulting p_k confuses those who approach these networks through the narrow perspective of the quality of fit to a pure power law. Instead the scale-free property tells us that we must distinguish between two rather different classes of networks:

- Bounded networks are networks whose degree distribution decrease exponentially or faster for high k. Examples of p_k in this class include the Poisson, Gaussian, or the simple exponential distribution. The Erdős-Rényi network is the best known example of the networks belonging to this class. Bounded networks lack outliers, consequently most nodes have comparable degrees. Real networks in this class include highway networks, the power grid or the atomic networks observed in crystalline or amorphous materials.

- Unbounded networks are networks whose degree distribution has a fat tail in the high-k region. Networks with a power-law degree distribution Eq. 4.1 offer a representative example of this class. A common property of these networks is that the node degrees span several orders of magnitude, differences that are difficult to explain using a bounded distribution. Outliers, or exceptionally high-degree nodes, are not only allowed but expected in these networks. Networks in this class include the WWW, the Internet, the protein interaction networks, and many social and online networks. While it would be desirable to fit and statistically validate the precise form of the degree distribution, often it is sufficient to decide the class to which a given network belongs: bounded or unbounded (see ADVANCED TOPICS 4.A). If the degree distribution is bounded, the random network model offers a reasonable starting point to understand its topology. If the degree distribution is unbounded, a scale-free network offers a better approximation.

In summary, to understand the properties of real networks, it is often sufficient to remember that in scale-free networks a few highly connected hubs coexist with a large number of small nodes. In contrast in random networks most nodes have comparable degrees and hubs are absent. The presence or absence of the hubs plays an important role in the system's behavior. The purpose of this chapter was to explore the basic characteristics of scale-free networks. We are left, therefore, with an important question: why are networks scale-free? The next chapter will provide the answer. Keeping up with the framework established in the previous chapter, the results discussed in this chapter allow us to formulate our next network law:

The Second Law: scale-free property

Many real networks are characterized by a fat-tailed degree distribution. This means that many small-degree nodes are held together by a few hubs.

Let us recap the validity of this law in the context of the three criteria established in CHAPTER 3:

A. **Quantitative formulation**: Eq. 4.1 offers the quantitative formulation of the Second Law, indicating that the degree distribution of such networks can be approximated by a power law.

B. **Universality**: as discussed in SECTION 4.5, the scale-free property is a common feature of many real networks, from the WWW to the protein interaction network in the cell.

C. **Non-random origins**: the scale-free property represents a dramatic deviation from the Poisson degree distribution characterizing random networks, hence it can not be explained in the context of the random network model.

BOX 4.7

At a glance
Scale-free networks

DEGREE DISTRIBUTION

Discrete form:

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

Continuous form:

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

SIZE OF THE LARGEST HUB

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

MOMENTS OF p_k

$2 < \gamma < 3$: $\langle k \rangle$ finite, $\langle k^2 \rangle$ diver□es□when□$N \rightarrow \infty$.

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

DISTANCES IN A SCALE-FREE NETWORK

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

SECTION 4.10

ADVANCED TOPICS 4.A
POWER LAWS

Power laws have a convoluted history in natural and social sciences, being interchangeably called fat-tailed, heavy-tailed, long-tailed, Pareto, or Bradford distributions. They also have a series of close relatives, like log-normal, Weibull, or Lévy distributions. The purpose of this section is to discuss the properties of some of the most frequently encountered distributions in network science and their relationship to the power law function discussed in this chapter.

Many quantities in nature, from the height of individuals to the probability of being in a car accident, follow bounded distributions. A common property of these is that pk decays either exponentially (e^{-x}), or faster than exponentially (e^{-x^2/σ^2}) for high x. Consequently events with high x are extremely rare, the largest expected x being unable to exceed some upper value x_{max} that is not too different from $\langle x \rangle$ (it is "bounded"). The high-x regime is often called the tail of the distribution, and given the absence of numerous events in the tail, these distributions are also called thin tailed. Well known examples of such bounded distributions are the Poisson, Gaussian (normal), or the exponential distribution Table 4.2.

In contrast the terms "fat tailed","heavy tailed","long tailed", or "unbounded" refer to p_k whose decay at large x is slower that exponential. In these distributions one often encounters events characterized by very large x values, unusually called outliers or rare events. The power-law distribution of Eq. 4.1 represents the best known example of such unbounded distributions. In the following we will discuss the basic properties of the most commonly encountered bounded and unbounded distributions in network science Table 4.2.

BOUNDED DISTRIBUTIONS (EXPONENTIALS)

Analytically the simplest bounded distribution is the exponential distribution $e^{-\lambda x}$. Within network science the most prominent bounded distribution is the Poisson distribution, capturing the degree distribution of a random network. Outside of network science the most frequently encountered member of this class is the normal (Gaussian) distribution.

A common property of bounded distributions comes in their tail: for high *x* they decay exponentially or faster. Consequently, the expected largest *x* obtained after we draw *N* numbers from a bounded p_x grows as $x_{max} \sim \log(N)$ or slower. This means that outliers, representing unusually high *x*-values, are rare. They are so rare that they are effectively forbidden, meaning that they do not occur with any meaningful probability. Instead, most events drawn from a bounded distribution are not too far from $\langle x \rangle$.

UNBOUNDED DISTRIBUTIONS (POWER LAWS)

An instantly recognizable feature of an unbounded distribution is that the magnitude of the events *x* drawn from it vary widely, spanning several orders of magnitude. The most prominent member of this class is the power-law distribution discussed in **SECTION 4.2**. Its relevance to networks is provided by several factors:

- Many quantities occurring in networks science, like degrees, link weights and betweenness centrality, follow a power-law distribution in many real and model networks.

- The power-law form is analytically predicted by some of the most fundamental network models **CHAPTER 5**.

In contrast with bounded distributions, in unbounded distributions the size of the largest event after *N* trials scales as $x_{max} \sim N^{\zeta}$ where ζ is some integer related to the exponent γ characterizing the p_x distribution. As N^{ζ} grows fast, rare events or outliers occur with a noticeable frequency, often dominating the properties of the system.

CROSSOVER DISTRIBUTION (LOG-NORMAL, STRETCHED EXPONENTIAL)

Several functions interpolate between bounded and unbounded distributions. This means that depending on their parameters, they can be used to fit unbounded distributions, but technically speaking they are bounded, as their tail for large *x* decays exponentially or faster. In the following we discuss the properties of the most frequently encountered crossover distributions.

A power law with exponential cut-off is often used in network theory to fit the degree distribution of real networks. Its density function has the form:

$$p_k = Cx^{-\gamma} e^{-\lambda x} \tag{4.30}$$

$$C = \frac{\lambda^{1-\gamma}}{\Gamma(1-\gamma, \lambda x_{\min})}, \tag{4.31}$$

where $x > 0$ and $\gamma > 0$. The analytical form of **Eq. 4.30** directly captures its crossover nature: it combines a power-law term, a key component of unbounded distributions, with an exponential term, responsible for its bounded tail. We can explore its crossover characteristics by taking the logarithm of **Eq. 4.30**,

$$\ln p_x = \ln C - \gamma \ln x - \lambda x. \tag{4.32}$$

For $x \ll 1/\lambda$ the second term on the r.h.s dominates, suggesting that the distribution follows a power law with exponent γ. Once $x \geq 1/\lambda$,the λx term overcomes the ln$x$ term, resulting in an exponential cutoff for high x.

Stretched exponential (Weibull distribution) is similar to Eq. 4.30 except that we have a fractional power law in the exponential. Its density function has the form

$$p_x = Cx^{\beta-1}e^{-\lambda x^{\beta}} \tag{4.32}$$

$$C = \beta x^{-\beta} \exp \left(x_{\min} / \lambda\right)^{\beta} \quad . \tag{4.33}$$

In most applications x varies between 0 and $+\infty$. In Eq. 4.32 β is the *stretching exponent*, determining the properties of p_x:

- For $\beta = 1$ we recover a simple exponential function

- If β is between 0 and 1, the graph of log p_x versus x is "stretched", meaning that it spans several orders of magnitude in x. This is the regime where a stretched exponential is difficult to distinguish from a pure power law. The closer β is to 0, the more similar is p_x to the power law x^{-1}

- By taking a logarithm of Eq. 4.32,

$$\ln p_x \sim (\beta - 1)\ln x - \lambda x^{\beta}, \tag{4.34}$$

  we can see why the stretched exponential is often used to approximate a power law distribution. Indeed, for small β and not too large x the function will be indistinguishable from a power law withslope (β-1). For large x the term λx^{β} becomes dominant, generating an exponential cutoff in p_x.

- If $\beta > 1$ we observe a "compressed" exponential function, meaning that x varies in a very narrow range.

- For $\beta = 2$ Eq. 4.32 reduces to the normal distribution.

As we will see in CHAPTERS 5 and 6, several important network models predict a streched exponential degree distribution.

A *log-normal distribution* (*Galton or Gibrat distribution*) emerges if ln$x$ follows a normal distribution. Typically a variable follows a log-normal distribution if it is the product of many independent positive random numbers.We encounter log-normal distributions in finance, representing the compound return from a sequence of trades, where the compound return is the product of the individual trades. The probability density function of a log-normal distribution is

$$p_x = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \tag{4.35}$$

Hence a log-normal is like a normal distribution except that its variable in the exponential term is not x, but $\ln x$. To understand why a log-normal is occasionally used to fit a power law distribution, let us take the logarithm of Eq. 4.35,

$$\ln p_x = \ln \frac{1}{\sqrt{2\pi}\sigma} - \ln x - \frac{(\ln x - \mu)^2}{2\sigma^2} \tag{4.36}$$

If $\ln x \ll \mu$ then the last term is negligible and the distribution follows a power law with slope -1 due to the second term $\ln x^{-1}$.

Therefore, for distributions that appear to follow a power law with slope -1, a log-normal function will likely offer a reasonable fit. For large σ the log-normal distribution may resemble power laws with other exponents too (see dashed line in Fig. 4.20 with slope 2.5). Note that for reasons that are discussed in BOX 4.4, a degree distribution with γ=1 is forbidden in most real networks, hence log-normal distributions are rarely used to approximate a network's degree distribution. In summary, in most areas where we encounter fat-tailed distributions, there is an ongoing debate about the form of the distribution that offers the best fit to the data. Common candidates include a simple power law, a stretched exponential, or a log-normal function. In many systems it is impossible to distinguish these distribution based on empirical data only. Hence as long as there is empirical data to be fitted, the debate surrounding the best fit will never die out.

The debate can be best resolved by developing accurate mechanistic models, which analytically predict the expected degree distribution.We will see in the coming chapters that the distributions that are analytically predicted by network theory are the Poisson, simple exponential, stretched exponential, and power law. The remaining distributions in Table 4.2 are occasionally used to fit the degrees of some networks, despite the fact that we lack theoretical backing to support their relevance for network science.

NAME	p_x	C_i	$\langle x\rangle$	$\langle x^2\rangle$
Exponential (continuous)	$e^{-\lambda x}$	$\lambda e^{\lambda x_{\min}}$	$\lambda^{-1} + x_{\min}$	$\frac{(\lambda x_{\min}+1)^2+1}{\lambda^2}$
Exponential (discrete)	$e^{-\lambda x}$	$(1-e^{-\lambda})e^{-\lambda x_{\min}}$	$(e^{\lambda}-1)+x_{\min}$	$\frac{e^{\lambda}+1}{(e^{\lambda}-1)^2}+\frac{2x_{\min}}{e^{\lambda}-1}+x_{\min}^2$
Poisson	$\mu^x / x!$	$\left[e^{\mu}-\sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$	$\mu-e^{-\mu}\sum_{x=0}^{x_{\min}-1}\frac{\mu^x}{x!}x$	$\mu^2+\mu-e^{-\mu}\sum_{x=0}^{x_{\min}-1}\frac{\mu^x}{x!}x^2$
Power law (continuous)	$x^{-\alpha}$	$(\alpha-1)x_{\min}^{\alpha-1}$	$\begin{cases} x_{\min}\frac{\alpha-1}{\alpha-2} & \text{if } \alpha>2 \\ \infty & \text{if } \alpha\le 2\end{cases}$	$\begin{cases} x_{\min}^2\frac{\alpha-1}{\alpha-3} & \text{if } \alpha>3 \\ \infty & \text{if } \alpha\le 3\end{cases}$
Power law (discrete)	$x^{-\alpha}$	$1/\zeta(\alpha,x_{\min})$	$\begin{cases} \frac{\zeta(\alpha-1,x_{\min})}{\zeta(\alpha,x_{\min})} & \text{if } \alpha>2 \\ \infty & \text{if } \alpha\le 2\end{cases}$	$\begin{cases} \frac{\zeta(\alpha-2,x_{\min})}{\zeta(\alpha,x_{\min})} & \text{if } \alpha>3 \\ \infty & \text{if } \alpha\le 3\end{cases}$
Power law with cutoff (exponential)	$x^{-\alpha}e^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$	$\lambda^{-1}\frac{\Gamma(2-\alpha,\lambda x_{\min})}{\Gamma(1-\alpha,\lambda x_{\min})}$	$\lambda^{-2}\frac{\Gamma(3-\alpha,\lambda x_{\min})}{\Gamma(1-\alpha,\lambda x_{\min})}$
Stretched exponential	$x_X^{\beta-1}e^{-\lambda x^{\beta}}$	$\beta\lambda e^{\lambda x_{\min}^{\beta}}$	$\lambda^{-1/\beta}e^{\lambda x_{\min}^{\beta}}\,\Gamma(1/\beta+1,\lambda x_{\min}^{\beta})$	$\lambda^{-2/\beta}e^{\lambda x_{\min}^{\beta}}\,\Gamma(2/\beta+1,\lambda x_{\min}^{\beta})$
Log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[erfc\left(\frac{\ln x-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$	$e^{\mu+\sigma^2/2}\frac{1+\text{erf}\left[\frac{\mu+\sigma^2-\ln x_{\min}}{\sqrt{2}\sigma}\right]}{1-\text{erf}\left[\frac{-\mu+\ln x_{\min}}{\sqrt{2}\sigma}\right]}$	$e^{2(\mu+\sigma^2)}\frac{1+\text{erf}\left[\frac{\mu+2\sigma^2-\ln x_{\min}}{\sqrt{2}\sigma}\right]}{1-\text{erf}\left[\frac{-\mu+\ln x_{\min}}{\sqrt{2}\sigma}\right]}$
Gaussian	$\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{1}{2\pi\sigma^2}}$	μ	$\sigma^2+\mu^2$

Table 4.2
Distributions in network science

The table lists several frequently encountered distributions in network science. For each distribution we show the density function p_x, the appropriate normalization constant C such that

$$\int_{x=x_{\min}}^{\infty} Cf(x)\,dx = 1$$

for the continuous case or

$$\sum_{x=x_{\min}}^{\infty} Cf(x) = 1$$

for the discrete case. Given that $\langle x\rangle$ and $\langle x^2\rangle$ play an important role in network theory, we list the analytical form of these two quantities for each distribution. As many of these distributions diverge at $x = 0$, $\langle x\rangle$ and $\langle x^2\rangle$ are calculated assuming that there is a small cutoff xmin in the system. In networks x_{min} often corresponds to the smallest positive degree, k_{min-1}, or could reflect the smallest degree k_{min} for which the appropriate distribution offers a good fit.

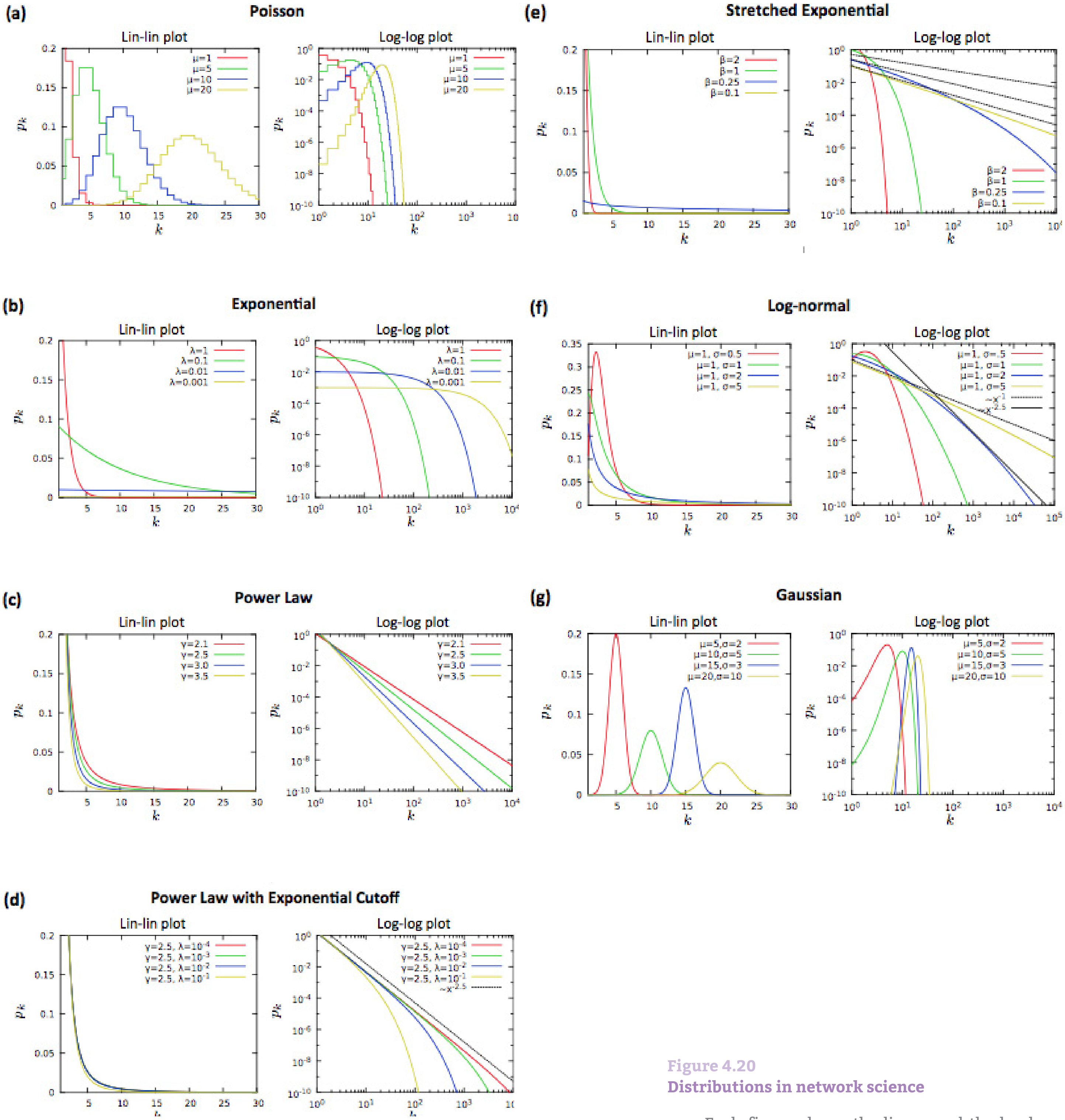


Figure 4.20
Distributions in network science

Each figure shows the linear and the log-log plot for the most frequently encountered distributions in network science. For definitions see **Table 4.2**.

ADVANCED TOPICS 4.B
PLOTTING A POWER-LAW DEGREE DISTRIBUTION

Plotting the degree distribution is an integral part of analyzing the properties of a network. The process starts with obtaining N_k, the number of nodes with degree k. This can be provided by direct measurement or by a model. From N_k we determine $p_k = N_k /N$. The question is, how to plot p_k to best extract its properties.

USE A LOG-LOG PLOT

In a scale-free network numerous nodes with one or two links coexist with a few hubs, representing nodes with thousands or even millions of links. Using a linear k-axis will compress the numerous small degree nodes in the small-k region, rendering them invisible. Similarly, as there are orders of magnitude differences in p_k for $k = 1$ and for some large k, if we plot p_k on a linear vertical axis, its value for large k will appear to be zero (see Fig. 4.21). The use of a log-log plot avoids these problems. We can either use logarithmic axes, with powers of 10 (used throughout this book) or we can plot $\log p_k$ in function of $\log k$ (equally correct, but slightly harder to read). Note that points with N_k =0 or (p_k =0) are not shown on a log-log plot as $\log 0=-\infty$.

AVOID LINEAR BINNING

The most flawed method (yet frequently seen in the literature) is to simply plot $p_k = N_k/N$ on a log-log plot Fig. 4.21b.This is called linear binning, as each bin has the same size $\Delta k = 1$. For a scale-free network linear binning results in an instantly recognizable plateau at large k, consisting of numerous data points that form a horizontal line Fig. 4.21b. This plateau has a simple explanation: as typically we have only one copy of each high degree node, for high k we either have N_k=0 (no node with degree k) or N_k=1 (a single node with degree k). Consequently linear binning will either give p_k=0, not visible on a log-log plot, or $p_k = 1/N$, which effectively applies to all hubs, generating a plateau at $p_k = 1/N$. This plateau affects our ability to estimate the degree exponent γ. For example, if we attempt to fit a power law to the data shown in Fig. 4.21b using linear binning, the fit provides γ that is quite different from real value γ=2.5. The reason is that under linear binning we have a large number of nodes in small k bins, hence in this regime we can confidently fit p_k. We have too few nodes in the large k bins for

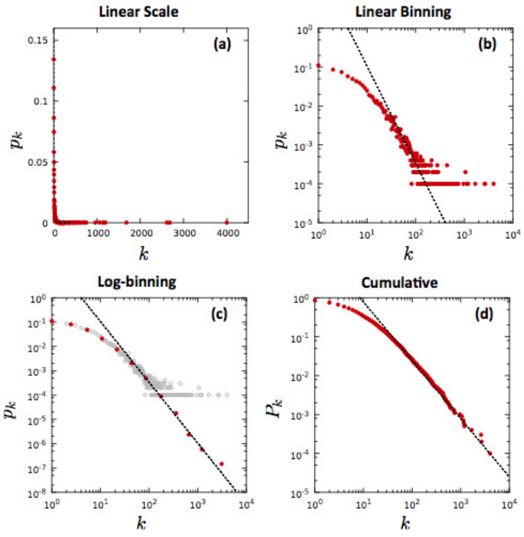


Figure 4.21
Plotting degree distributions

(a-d) The degree distribution of the form $p_k \sim (k + k_0)^{-\gamma}$, with k_0=10 and γ=2.5, plotted using the three procedures described in the text:

(a) linear binning. It is impossible to see the distribution on a lin-lin scale. This is the reason why we always use log-log plots for scale-free networks.

(b-d): The degree distribution shown on a log-log plots using (b) linear binning, (c) logarithmic binning, and (d) plotting the cumulative distribution.

a proper statistical estimate of p_k, hence the plateau biases our fit.Yet, it is precisely this high-*k* regime that plays a key role in determining γ. Increasing the bin size will not solve this problem. It is therefore recommended to avoid linear binning for fat tailed distributions.

USE LOGARITHMIC BINNING

Logarithmic binning aims to correct for the non-uniform sampling observed for linear binning. For log-binning we let the bin sizes increase with the degree, making sure that each bin has a comparable number of nodes. For example, we can choose the bin sizes to be multiples of 2, so that the first bin has size b_0=1 , containing all nodes with *k*=1; the second has size b_1=2, containing nodes with degrees *k*=2, 3; the third bin has size b_2=4 containing nodes with degrees *k*=4, 5, 6, 7. In general, the *n*th bin has size 2^{n-1} and contains all nodes with degrees k=2^{n-1}, 2^{n-1}+1, ..., 2^{n-1}-1. Note thatthe bin size can increase with arbitrary increments, $b_n = c_n$, where $c > 1$. The degree distribution is given by $p_{\langle k_n \rangle}=N_n/b_n$, where N_n is the number of nodes found in the bin n of size b_n, and $\langle k_n \rangle$ is the average degree of the nodes in bin b_n. The logarithmically binned p_k is shown in Fig. 4.21c. Note that now the scaling extends into the high-*k* plateau, previously invisible under linear binning. This indicates that logarithmic binning extracts useful information from the high degree nodes as well BOX 4.8.

USE CUMULATIVE DISTRIBUTION

Another way to extract information from the tail of p_k is to plot the cumulative distribution

$$p_x = \sum_{q=k}^{\infty} P_q, \qquad (4.37)$$

which again enhances the statistical significance the high-degree region. If p_k follows the power law, then the cumulative distribution will scale as

$$p_x \sim k^{-\gamma+1}. \qquad (4.38)$$

The cumulative distribution will again eliminate the plateau observed for linear binning and leads to an extended scaling region Figure 4.21d, allowing for a more accurate estimate of the degree exponent.

In summary, plotting the degree distribution to fully extract its features requires special attention. Mastering the tools of the process can help us better explore the properties of real networks BOX 4.9.

BOX 4.8

The impact of log-binning

To illustrate the rationale for log-binning. we compare three binning strategies: linear binning, log-binning, and variable bins, when the bin lengths were chosen such that each bin contains exactly the same number of events. As the figure shows, for logarithmic binning the bin sizes decrease exponentially with the bin number.

Indeed, choosing the bin sizes to vary between 2^{n-1} and 2^{n} , we obtain that the number of events in each bin decreases as $2^{-(\gamma-1)n}$. Yet, the bin size in case of linear binning decreases even faster, effectively running out of events.

The impact of log-binning is most visible in (b) where we show the obtained degree distribution. As one can see, both the variable binning and the linear binning considerably limits the scaling regime compared to the log-binning strategy.

Note that to compare the three methods we set the total number of bins to 10 in all cases.

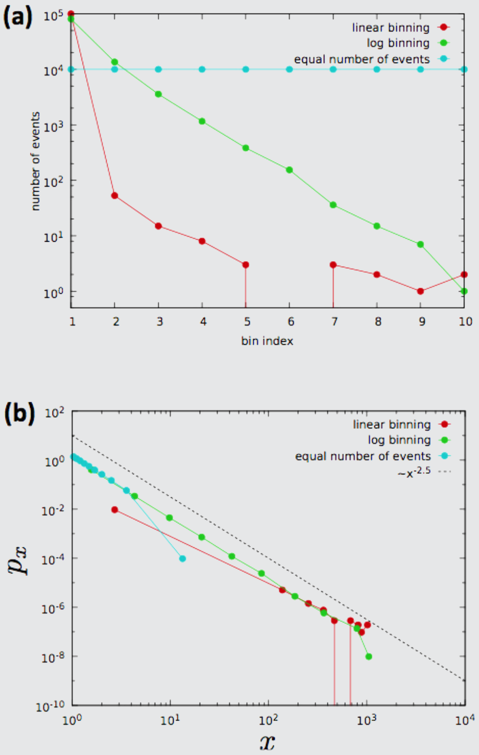


Figure 4.22

BOX 4.9

The degree distribution of real networks

In real systems we rarely observe a degree distribution that follows a pure power law. Instead, for most real systems p_k has the shape shown schematically in (a), with some recognizable features:

- Low-degree saturation is a common deviation from the power law behavior. Its signature is a flattened p_k for $k < k_{sat}$ .This indicates that we have fewer small degree nodes than expected for a pure power law. The origin of the saturation will be explained in CHAPTER 6.

- High-degree cutoff appears as a rapid drop in p_k for $k > k_{cut}$, indicating that we have fewer high-degree nodes than expected in a pure power law. This also limits the size of the largest hub, making it smaller than predicted by Eq. 4.23. High-degree cutoffs emerge if there are inherent limitations in the number of links a node can have. For example, in social networks individuals have difficulty maintaining a meaningful relationship with an exceptionally large number of acquaintances.

Given the widespread presence of such cutoffs we often fit the degree distribution to

$$p_x = a(k + k_{sat})^{-\gamma} \exp\left(-\frac{k}{k_{cut}}\right) . \qquad (4.39)$$

where k_{sat} accounts for the degree saturation, and the exponential term accounts for the high-k cutoff. To extract the full extent of the scaling we plot

$$p_x = p_x \exp\left(\frac{k}{k_{cut}}\right) \qquad (4.40)$$

in function of $\tilde{k} = k + k_{\text{min}}$.According to Eq. 4.40 $\tilde{p} \sim \tilde{k}^{-\gamma}$, correcting for the two cutoffs, as shown in (b). One occasionally encounters the claim that the presence of low-degree or high-degree cutoffs implies that the network is not scale-free. This is a misunderstanding of the scale-free property: most properties of scale-free networks are insensitive to the low-degree saturation. Only the high- degree cutoff affects the system's properties by limiting the divergence of the second moment $\langle k^2 \rangle$. The presence of such cutoffs means that additional phenomena take place in the system, that need to be understood.

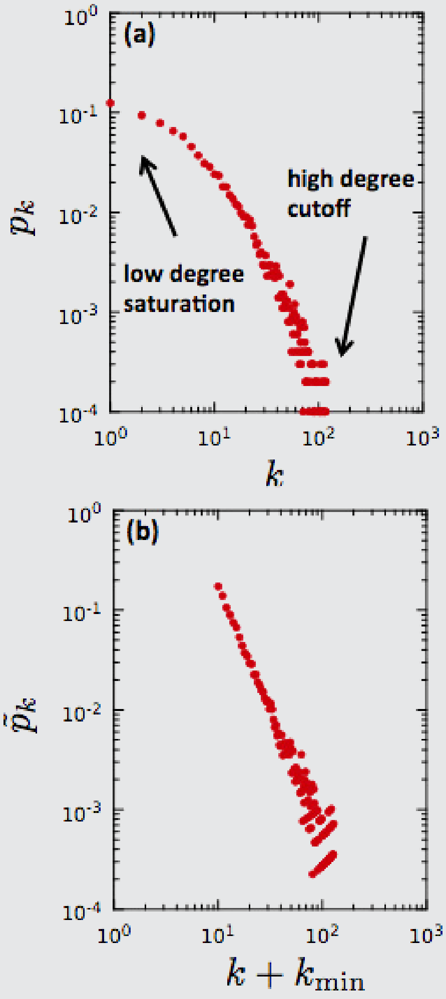


Figure 4.23
Rescaling the degree distribution

(a) The frequently observed form of a degree distribution in real data, characterized by low and high degree cutoffs.

(b) By plotting the rescaled $\tilde{p}_k$ in function of $(k + k_{min})$, as suggested by Eq. 4.39, the degree distribution follows a power law for all degrees.

SECTION 4.12

ADVANCED TOPICS 4.C ESTIMATING THE DEGREE EXPONENT

As discussed in SECTION 4.7, the properties of scale-free networks depend on γ, raising the need to accurately estimate the degree exponent γ. We face several difficulties, however, when we try to fit a power law to real data. The most important one is the fact that the scaling is rarely valid for the full range of the degree distribution.

Rather we observe so called small- and high- degree cutoffs BOX 4.9, denoted by *k* and *k* , within which we can min max observe a clear scaling region. Note that k_{min} and k_{max} are different from K_{min} and K_{max}, which correspond to the smallest and largest degrees in a network. Here we focus on estimating the small degree cutoff K_{min}, as the high degree cutoff can be approximated in a similar fashion. Before implementing this procedure, the reader is advised to consult the discussion on systematic problems provided at the end of this section.

FITTING PROCEDURE

As the degree distribution typically comes as a list of positive integers k=0, 1, 2 , ..., k_{max}, we aim to estimate γ from a discrete set of data points. We follow [44] and the algorithmic tools to perform the fits are available at *http://tuvalu.santafe.edu/~aaronc/powerlaws/.* We use the degree distribution of citation networks to illustrate the procedure. The network consists of N=384,362 nodes, each representing a research paper published between 1890 and 2009 in the family of journal published by the American Physical Society. The network has L=2,353,984 links, each representing a citation from a published research paper to some other publication in the dataset (outside citations are ignored). See [45] for an overall characterization of the full dataset Figure 4.24a. The steps of the fitting process are:

1. Pick a value of k_{min} between k_{min} and k_{max}. Estimate the value of the degree exponent corresponding to this k_{min} using

$$\gamma = 1 + N \sum_{i=1}^{N} \ln \frac{k_i}{K_{min} - \frac{1}{2}}^{-1} . \tag{4.41}$$

2. With the obtained (γ, k_{min}) parameter pair assume that the degree distribution has the form

$$p(k) = \frac{1}{\zeta(\gamma, K_{min})} k^{-\gamma}, \qquad (4.42)$$

hence the associated cumulative distribution function (CDF) is

$$P(k) = 1 - \frac{\zeta(\gamma, k)}{\zeta(\gamma, K_{min})}. \qquad (4.43)$$

3. Use the Kormogorov-Smirnov test to determine the maximum distance D between the CDF of the data $S(k)$ and the fitted model provided by Eq. 4.43 with the selected (γ, k_{min}) parameter pair,

$$D = max_{k \geq K_{min}} |S(k) - P(k)|. \qquad (4.44)$$

Eq. 4.44 identifies the degree for which the difference D between the empirical distribution $S(k)$ and the fitted distribution Eq. 4.43 is the largest.

4. Repeat steps (1-3) by scanning the whole k_{min} range from k_{min} to k_{max}. We aim to identify the k_{min} value for which D provided by Eq. 4.44 is minimal. To illustrate the procedure, we plot D in function of k_{max} for the citation network Fig. 4.24b. The plot indicates that D is minimal for k_{min}= 49, and the corresponding γ estimated by Eq. 4.41, representing the optimal fit, is γ=2.79. The standard error for the obtained degree exponent is

$$\sigma_{\gamma} = \frac{1}{\sqrt{N \frac{\zeta \ (\gamma, K_{\mathrm{min}})}{\zeta(\gamma, K_{\mathrm{min}})} - \frac{\zeta \ (\gamma, K_{\mathrm{min}})}{\zeta(\gamma, K_{\mathrm{min}})}^{2}}} \qquad (4.45)$$

which implies that the best fit is for exponent γ ± σ_{γ}. For the citation network we obtain σ_{γ}=0.003, hence γ=2.79 (3).

Note that Eq. 4.45 represents an approximation, but typically the results provided by it is within 1% of the real value as long as k_{min} >6. Furthermore, in order to obtain a reasonable estimate for γ, we need N >50. Smaller datasets should be treated with caution.

GOODNESS-OF-FIT

Just because we obtained a (γ, k_{min}) pair that represents an optimal fit to our dataset, does not mean that the power law itself is a good model for the studied distribution.We therefore need to use a goodness-of-fit test, which generates a ρ-value that quantifies the plausibility of the power law hypothesis. The most often used procedure [12] consists of the following steps:

(i) Use the cumulative distribution Eq. 4.43 to estimate the KS distance between the real data and the best fit, that we denote by D^{real}. This is step 3 above, taking the value of D for k_{min} that offered the best fit

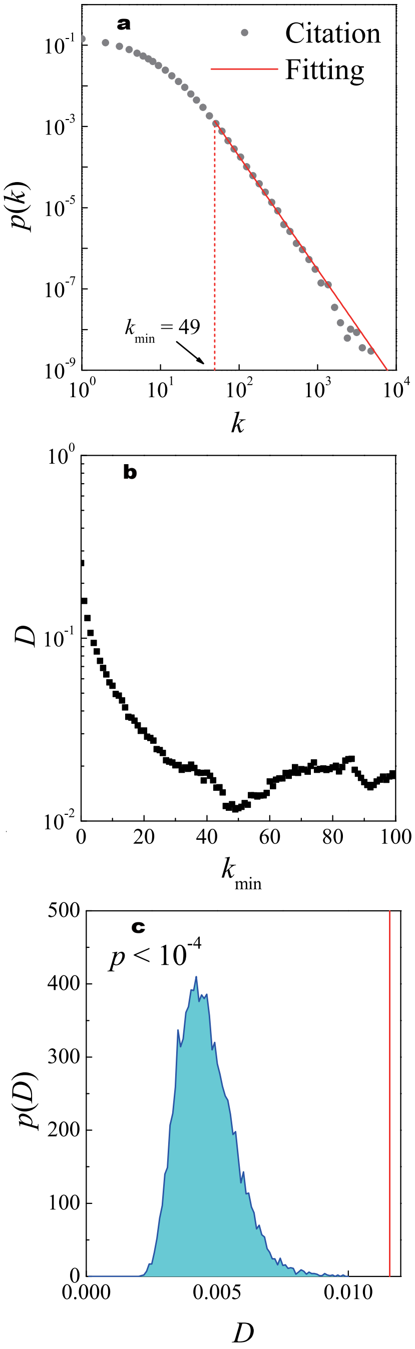


Figure 4.24
Maximum likelihood estimation

(a) The degree distribution p_k of the citation network, where the straight line represents the best based on model Eq. 4.39.

(b) The values of Kormogorov-Smirnov test vs. kmin, where the red lines indicate the minimum value of D and the corresponding k_{min}.

(c) ρ ($D^{\mathrm{synthetic}}$) for M=10,000 synthetic data, where the red line corresponds to the D value from the citation network (a-b).

to the data. For the citation data we obtain D^{real} = 0.01158 for k_{min}= 49 Fig. 4.24.

(ii) Use Eq. 4.42 to generate a degree sequence of N degrees (i.e. the same number of random numbers as the number of nodes in the original dataset) and substitute the obtained degree sequence for the empirical data, determining $D^{\text{synthetic}}$ for this hypothetical degree sequence. Hence $D^{\text{synthetic}}$ represents the distance between a synthetically generated degree sequence, consistent with our degree distribution, and the real data.

iii. The goal is to see if the obtained $D^{\text{synthetic}}$ is comparable to D^{real}. For this we repeat step (ii) M times ($M \gg 1$), and each time we generate a new degree sequence and determine the corresponding $D^{\text{synthetic}}$, eventually obtaining the $p_{D^{synthetic}}$ distribution. Plot $p_{D^{synthetic}}$ and show as a vertical bar D^{real} Fig. 4.24c. If Dreal is within the $p_{D^{synthetic}}$ distribution, it means that the distance between the model providing the best fit and the empirical data is comparable with the distance expected from random degree samples chosen from the best fit distribution. Hence the power law is a reasonable model to the data. If, however, D^{real} falls outside the ($p_{D^{synthetic}}$) distribution, then the power law is not a good model - some other function is expected to describe the original p_k.

While the distribution shown in Figure 4.20 may be in some cases useful to offer a visual illustration, in general is better to assign a p-number to the fit, given by

$$p = \int_{D}^{\infty} P_{D^{synthetic}} dD^{synthetic} \ . \qquad (4.46)$$

The closer p is to 1, the more likely that the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if ρ is small, the model is not a plausible fit to the data.

Typically, the model is accepted if $p > 1\%$. For the citation network we obtai $p < 10^{-4}$, indicating that a pure power law is not a suitable model for the original degree distribution.This outcome is somewhat surprising, as the power-law nature of citation data has been documented repeatedly since 1960s [7, 8]. This failure offers a lesson on the limitation of the blind application of the fitting procedures.

FITTING REAL DISTRIBUTIONS

To correct the problem, we note that the fitting model Eq. 4.44 eliminates all the data points with $k < k_{min}$. As the citation network is fat tailed, choosing k_{min} = 49 forces us to discard over 96% data points. Yet, there is statistically useful information in the $k < k_{min}$ regime, that is ignored by the previous fit.We therefore introduce an alternate model that resolves this problem.

As we discussed in BOX 4.9, the degree distribution of many real networks, like the citation network, can not be described by a pure power law, but has the form

$$p_k = \frac{1}{\sum\limits_{k=1} (k + k_{sat})^{-\gamma} e^{-k/k_{cut}}} (k + k_{sat})^{-\gamma} e^{-k/k_{cut}} \qquad (4.47)$$

and the associated CDF is

$$p_k = \frac{1}{\sum\limits_{k=1} (k + k_{sat})^{-\gamma} e^{-k/k_{cut}}} \sum_{k=1}^{k} (k + k_{sat})^{-\gamma} e^{-k/k_{cut}}, \qquad (4.48)$$

where k_{sat} and k_{cut} correspond to low-k saturation and the large-k cutoff, respectively. The difference between our earlier procedure and Eq. 4.47 is that we now do not discard the points that deviate from a pure power law, but we use a function that may offer a better fit to the whole degree distribution, from k_{min} to k_{max}.

Our goal is to find the fitting parameters k_{sat}, k_{cut}, and γ of the model Eq. 4.47, which we achieve through the following steps:

A. Pick a value for k_{sat} and k_{cut} between k_{min} and k_{max}. Estimate the value of the degree exponent γ using the steepest descend method that maximizes the log-likelihood function

$$\log \mathcal{L}(\gamma \mid k_{min}, k_{cut}) = \sum_{i=1}^{N} \log p(k_i \mid \gamma, k_{min}, k_{cut}). \qquad (4.49)$$

That is, for fixed ($k_{sat,}$ k_{cut}) we vary γ until we find the maximum of L. The steepest descent method provides γ (k_{min}, k_{cut}) for which Eq. 4.48 is maximal.

B. With the obtained γ(k_{sat}, k_{cut}) assume that the degree distribution has the form. Calculate the Kormogorov Smirnov parameter D Eq. 4.47 between the cumulative degree distribution (CDF) of the original data and the fitted model provided by Eq. 4.47.

C. Change ksat and kcut, and repeat steps (1-3), scanning with k_0 from k_{min}= 0 to k_{max} and with k_{cut} from k_{min}= k_0 to k_{max}. The goal is to identify k and k values for which D is minimal. We illustrate this by plotting D in function of k_{sat} for serval k_{cut} values in Fig. 4.25a for our citation sample. The ($k_{sat,}$, k_{cut}) for which D is minimal, and the corresponding γ is provided by Eq. 4.41, will represent the optimal parameters of the fit. For our dataset the optimal fit is obtained for k_{sat}= 12 and k_{cut}= 5691, providing the degree exponent γ = 3.028. We find that now D for the real data is within the generated $p(D)$ distribution Fig.4.25c, and the associated p-value is 69%.

SYSTEMATIC FITTING ISSUES

The procedure described above may offer the impression that determining the degree exponent is a cumbersome but straight forward process. In reality the existing fitting methods have some well known limitations:

1. A pure power law is really an idealized distribution that emerges in its form (1) only in simple models CHAPTER 5. In reality, a whole range of processes contribute to the topology of real networks, affecting the precise shape of the degree distribution. These processes will be described in CHAPTER 6. If p_k does not follow a pure power law, the methods described above, designed to fit a power law to the data, will inevitably fail to detect statistical significance. That does not necessarily mean that the network is not scale-free (but it could also mean that). Most often it means that we have not yet gained a proper understanding of the precise form of the degree distribution, hence we are fitting the wrong functional form of p_k to the dataset.

2. The statistical tools used above to test the goodness of the fit rely on the Kolmogorov-Smirnov criteria, which measures the maximum distance between the fitted model and the dataset. If all data points follow a perfect power law, but a single point for some reason deviates from the curve, we will loose the fit's statistical significance. In real systems there are numerous reasons for such local deviations, that have little impact on the system's overall behavior. Yet, removing these "outliers" could be seen as data manipulation; if kept, however, one cannot detect the statistical significance of the power law fit. A good example is provided by the actor network, whose degree distribution follows a power law for most degrees. There is a single outlier, at k = 1,287, thanks to the 1956 movie, *Around the World in Eighty Days*.

   This is the only movie, where IMDB lists all the uncredited extras in the cast. Hence the movie appears to have 1,288 actors. The second largest movie in the dataset has only 340 actors. Since the extras are only listed for this movie, each of them have links only to the 1,287 extras that played in the same movie, leading to a local peak in p_k at k=1,287. Thanks to this peak, the degree distribution, fitted to a power law fails to pass the Kolmogorov-Smirnov criteria. Indeed, as indicated in Table 4.3, neither the pure power law fit, nor a power law with high-degree cutoff offers a statistically significant fit.

3. Thanks to the issues discussed above, the methodology described above often predicts a small scaling regime, forcing us to remove a huge fraction of the nodes (often as many as 99%, see Table 4.4), to obtain a statistically significant fit. Once plotted next to the original dataset, the obtained fit can be at times ridiculous, even if the method indicates statistical significance. The bottom line, estimating the degree exponents is still not an exact science. We continue to lack methods that would estimate the statistical significance of a proper fit in a manner that would be acceptable to a practitioner. The blind application of the tools describe above often leads to either fits that obviously do not capture the trends in the data, or to a false rejection of the power-law hypothesis. An important improvement will be provided by our ability to derive the expected form of the degree distribution, discussed in CHAPTER 6.

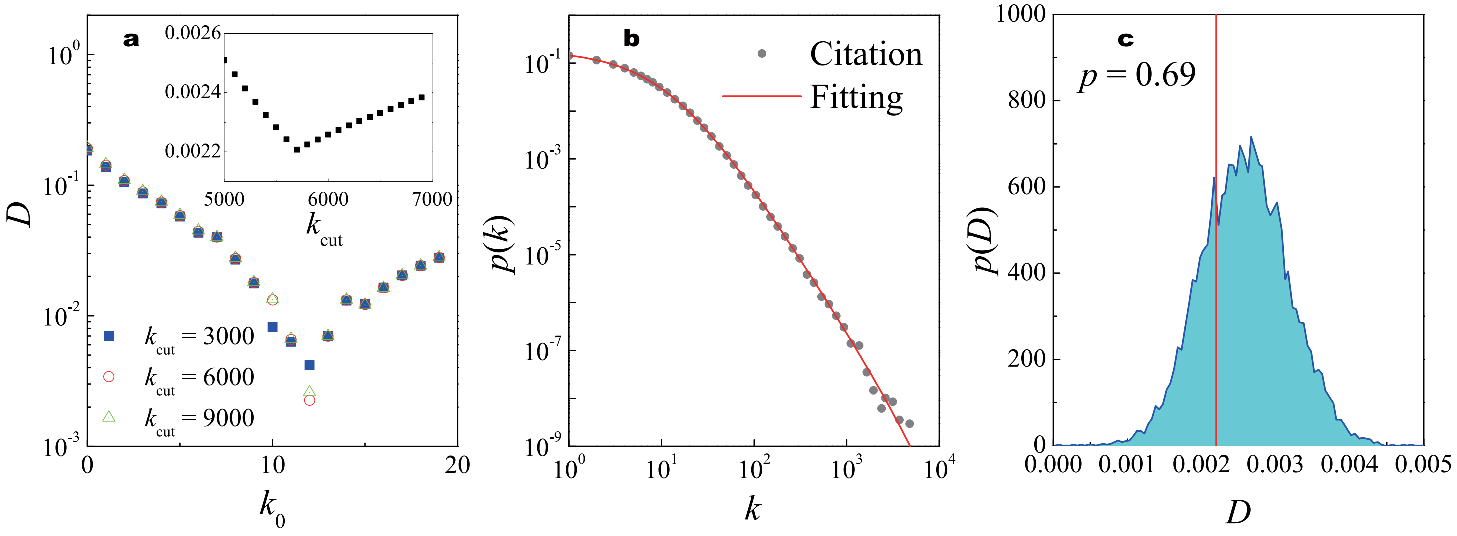


Figure 4.25

Estimating the scaling parameters for citation networks

(a) The Kormogorov-Smirnov parameter D vs. k_0 for k_{cut} = 3.000, 6.000, 9.000, respectively, showing that k_{sat}= 12 corresponds to the minimal D. Inset: D vs. k_{cut} for k_{sat}= 12, indicating that k_{cut} =5.691 minimizes D.

(b) Degree distribution p_k where the straight line represents the best fitting estimated from (a).

(c) $pD^{\text{synthetic}}$ for M = 10.000 synthetic data, where the red line corresponds to the D value from the citation network (a-b).

NETWORK NAME	λ	k_{min}	P-VALUE	PERCENTAGE
Power Grid	0.5174		0.91	12%

Table 4.3

Exponential Fitting

For the power grid a power law does not offer a statistically significant fit as the underlying network is not scale-free. We used the fitting procedure described in this section to now fit the exponential function $e^{-\lambda k}$ to the degree distribution of the power grid, obtaining a statistically significant fit in this case. The table shows the obtained λ parameters, the kmin over which the fit is valid, the obtained p-value, and the percentage of data points included in the fit.

NETWORK NAME	$k^{-\gamma};[k_{\min},\infty]$				$(k+k_{sat})^{-\gamma}\,e^{-k/k_{cut}}$			
	γ	$k_{\min}$	P-VALUE	PERCENT	γ	k_{sat}	k_{cut}	P-VALUE
Internet	3.42	72	0.13	0.6%	3.55	88	500	0.00
WWW-ND (in)2	.001		0.00	100%	1.970		660	0.00
WWW-ND (out)2	.317		0.00	15%2	.828		8500	0.00
Power Grid	4.00	50	.001	2%	8.561	91	40	.00
Mobile Phone Calls (in)	4.69	90	.342	.6%6	.951	51	00	.00
Mobile Phone Calls (out)	5.01	11	0.77	1.7%	7.23	15	10	0.00
Email-PRE (in)3	.438	80	.11	0.2%	2.27	08	500	0.00
Email-PRE (out)2	.033		0.00	1.2%2	.550		8500	0.00
Science Collaboration3	.352	50	.0001	5.4%	1.501	71	20	.00
Actor Network2	.12	54	0.00	33%	--		-0	.00
Citation Network (in)2	.79	51	0.00	3.0%	3.03	12	5691	0.69
Citation Network (out)	4.00	19	0.00	14%-	0.16	51	00	.00
E.coli Metabolism (in)	2.43	30	.00	57%	3.851	91	20	.00
E.coli Metabolism (out)	2.90	50	.00	34%	2.56	15	10	0.00
Yeast Protein Interactions	2.897		0.67	8.3%2	.952		90	0.52
WWW-stanford (in)2	.15	30	.00	44.9%	2.86492	4	222	0.00
WWW-stanford (out)3	.976	20	.000	.6%3	.96102	17	128	0.00
Email-PNAS (in)2	.811	90	.0005	22.2%	0.54	02	50	.00
Email-PNAS (out)2	.272	60	.929	.3%0	.920		36	0.00

Table 4.4

Fitting parameters for real networks

The estimated degree exponents and the appropriate fit parameters for several networks studied in this book. We implemented two fitting strategies, the first aiming to fit a pure power law in the region ($k_{min,}$ ∞) and the second fits a power law with saturation and exponential cutoff to the whole dataset. In the table we show the obtained γ exponent and k_{min} for the fit with the best statistical significance, the *p*-value for the best fit and the percentage of the data included in the fit. In the second case we again show the exponent γ, the two fit parameters, k_{sat} and k_{cut}, and the *p*-value of the obtained fit. Note that $p > 0.01$ is considered to be statistically significant.

BIBLIOGRAPHY

[1] H. Jeong, R.Albert, and A.-L. Barabási. Internet: Diameter of the world-wide web. Nature, 401:130-131, 1999.

[2] A.-L. Barabási and R.Albert. Emergence of scaling in random networks. Science, 286:509-512, 1999.

[3] V. Pareto. Cours d'Économie Politique: Nouvelle édition par G.- H. Bousquet et G. Busino, Librairie Droz, Geneva, 299–345, 1964.

[4] A.-L. Barabási. Linked: The New Science of Networks. (Plume, New York). ISBN 0-452-28439-2, 2002.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. Proceedings of SIGCOMM. Comput. Commun. Rev. 29: 251-262, 1999.

[6] R. Pastor-Satorras and A.Vespignani. Evolution and Structure of the Internet: A Statistical Physics Approach. (Cambridge University Press, Cambridge), 2004.

[7] D. J. De Solla Price. Networks of Scientific Papers, Science 149: 510-515, 1965.

[8] S. Redner. How Popular is Your Paper? An Empirical Study of the Citation Distribution, Eur. Phys. J. B 4, 131, 1998.

[9] A. L. Barabási, R.Albert, and H. Jeong. Mean-field theory of scale- free random networks.Physica A 272:173-187, 1999.

[10] R. Kumar, P. Raghavan, S. Rajalopagan, and A.Tomkins. Extracting Large-Scale Knowledge Bases from the Web. Proceedings of the 25thVLDB-Conference, Edinburgh,Scotland,pp.639-650,1999.

[11] H. Jeong, B.Tombor, R.Albert, Z. N. Oltvai, and A.-L. Barabási. The

large-scale organization of metabolic networks, Nature 407, 651-654, 2000.

[12] W. Aiello, F. Chung, and L.A. Lu. Random graph model for massive graphs, Proc. 32nd ACM Symp.Theor. Comp, 2000.

[13] H. Jeong, B.Tombor, S. P. Mason,A.-L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks, Nature 411, 41-42, 2001.

[14] A.-L. Barabási, H. Jeong, E. Ravasz, Z. Néda, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations, Physica A 311: 590-614, 2002.

[15] M. E. J. Newman. The structure of scientific collaboration networks, Proc. Natl.Acad. Sci. 98(2), 404-409, 2001.

[16] F.Liljeros,C.R.Edling,L.A.N.Amaral,H.E.Stanley,andY.Aberg. The Web of Human Sexual Contacts, Nature 411, 907-908, 2001.

[17]! R. Ferrer i Cancho and R.V. Solé.The small world of human language, Proc. R. Soc. Lond. B 268: 2261-2265, 2001.

[18] S.Yook, H. Jeong, and A.-L. Barabási. (unpublished), (2001).

[19] R. Ferrer i Cancho, C. Janssen, and R.V. Solé.Topology of technology graphs: Small world patterns in electronic circuits, Phys. Rev. E 64, 046119, 2001.

[20] S.Valverde and R.V. Sole. Hierarchical Small Worlds in Software Architecture, arXiv:cond-mat/0307278, 2003.

[21] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e- mail networks, Phys. Rev. E 66, 035103(R), 2002.

[22] J. P. K. Doye. Network Topology of a Potential Energy Landscape: A Static Scale-Free Network, Phys. Rev. Lett. 88, 238701, 2002.

[23] H. Kwak, C. Lee, H. Park, S. Moon.What is Twitter, a social network or a news media?, Proceedings of the 19th international conference on World wide web, 591-600, 2010.

[24] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi. Measuring user influence in Twitter:The million follower fallacy, Proceedings of international AAAI Conference on Weblogs and Social, 2010.

[25] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow.The Anatomy of the Facebook Social Graph, arXiv:1111.4503, 2011.

[26] B. Bollobás and O. Riordan.The Diameter of a Scale-Free Random Graph. Combinatorica, 24: 5-34, 2004.

[27] R. Cohen and S. Havlin. Scale free networks are ultrasmall, Phys.

Rev. Lett. 90, 058701, 2003.

[28] R. Cohen and S. Havlin. Complex Networks - Structure, Robustness and Function. (Cambridge University Press, Cambridge),2010.

[29] K.-I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. Phys. Rev. Lett. 87, 278701, 2001.

[30] F. Karinthy. Láncszemek, in Minden másképpen van (Budapest: Atheneum Irodai es Nyomdai R.-T. Kiadása, 1929), 85–90. English translation is available in: M.E. J. Newman, A.-L. Barabási, and D. J. Watts, The Structure and Dynamics of Networks (Princeton University Press, Princeton), 2006.

[31] P. S. Dodds, R. Muhamad and D.J.Watts.An experimental study to search in global social networks, Science 301, 827-829, 2003.

[32] P. Erd s and T. Gallai. Graphs with given degrees of vertices. Matematikai Lapok, 11:264-274, 1960.

[33] C.I. Del Genio, H. Kim, Z.Toroczkai, and K.E. Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. PLoS ONE, 5(4):e10012, 04 2010.

[34]V.Havel.A remark on the existence of finite graphs.Casopis Pest. Mat., 80:477-480, 1955.

[35] S. Hakimi. On the realizability of a set of integers as degrees of the vertices of a graph. SIAM J.Appl. Math., 10:496-506, 1962.

[36] I. Charo Del Genio, G.Thilo, and Kevin E. Bassler.All scale-free networks are sparse. Phys. Rev. Lett., 107:178701, 10 2011.

[37] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of la- belled regular graphs, European J. Combin. 1: 311– 316, 1980.

[38] M. Molloy and B. A. Reed. Critical Point for Random Graphs with a Given Degree Sequence. Random Structures and Algorithms, 6: 161-180, 1995.

[39] M. Newman. Networks: An Introduction. (Oxford University, Oxford), 2010.

[40] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. Science, 296:910-913, 2002.

[41] G. Caldarelli, I. A. Capocci, P. De Los Rios, and M.A. Muñoz. Scale-Free Networks from Varying Vertex Intrinsic Fitness, Phys. Rev. Lett. 89, 258702, 2002.

[42] B. Söderberg. General formalism for inhomogeneous random

graphs, Phys. Rev. E 66, 066121, 2002.

[43] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables, Phys. Rev. E 68, 036112, 2003.

[44]A.Clauset,C.R.Shalizi,and M.E.J.Newman.Power-law distributions in empirical data. SIAM Review S1: 661-703. ArXiv e- prints, 2009.

[45] S. Redner. Citation statistics from 110 years of physical review. Physics Today, 58:49, 2005.

CHAPTER 5

THE BARABÁSI-ALBERT MODEL

Introduction

Growth and preferential attachment 1

The Barabási-Albert Model 2

Degree dynamics 3

Degree distribution 4

The absence of growth or preferential attachment 5

Non-linear preferential attachment 6

The origins of preferential attachment 7

Diameter and clustering coefficient 8

Homework

Summary

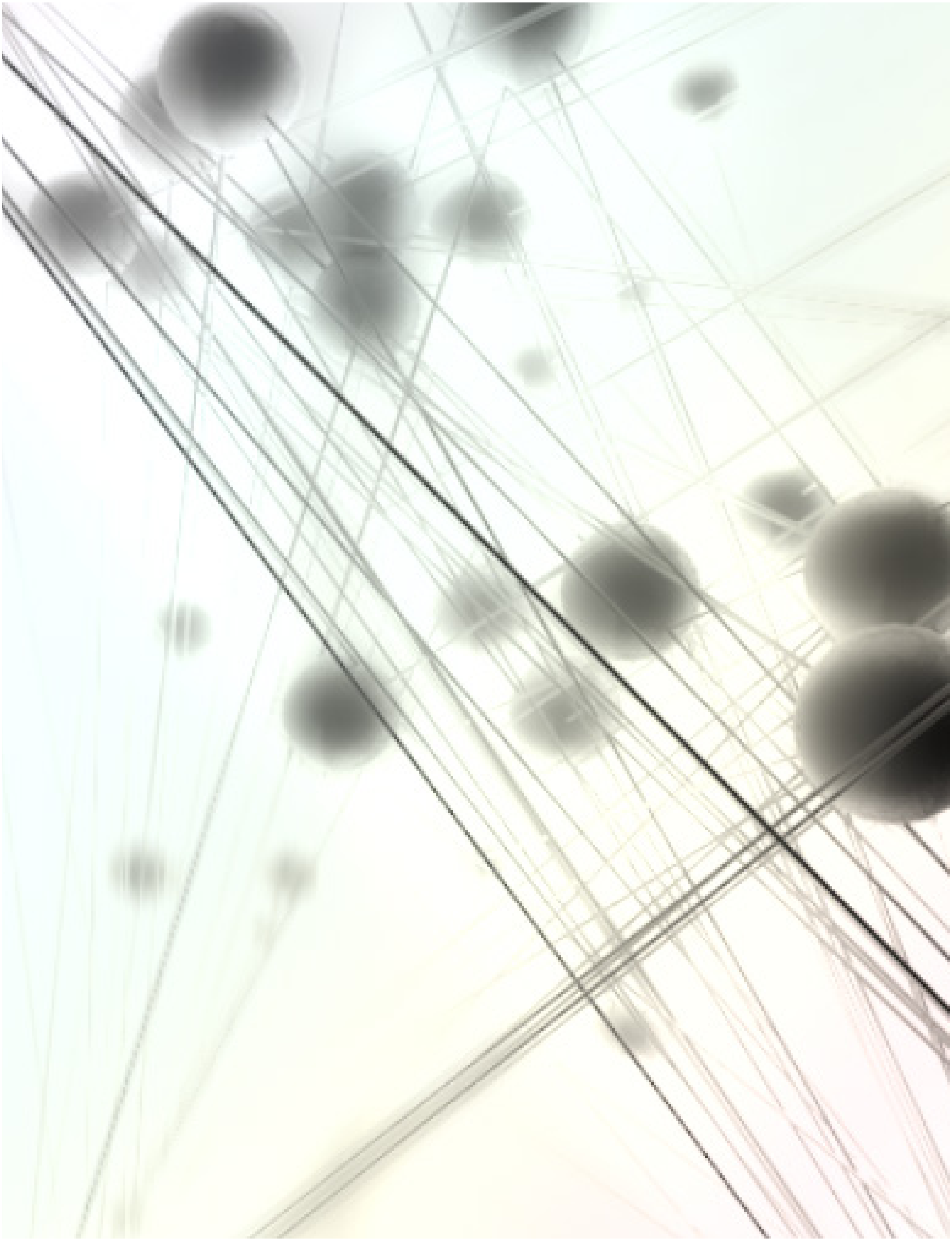
ADVANCED TOPICS 5.A
Deriving the degree distribution

ADVANCED TOPICS 5.B
Nonlinear preferential attachment

ADVANCED TOPICS 5.C
The clustering coefficient

Bibliography

Figure 5.0 (front cover)
Universality: G. Musella



SECTION 5.0

INTRODUCTION

Hubs represent the most striking difference between a random and a scale-free network. Their very existence raises several fundamental questions:

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

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

The last 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 proteins or metabolites, while the nodes of the WWW are documents, representing information without a physical manifestation.

- The links within the cell are binding interactions and chemical reactions, 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 a few decades old.

- The purpose of the metabolic network is to chemically build the basic chemical components the cells need for life, while the purpose of the WWW is information access and delivery.

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 major differences between the systems that display the scale-free property, the explanation must be simple and fundamental. The answers will change the way we view and model networks, forcing us to move from describing a network's topology to modeling the evolution of complex systems.

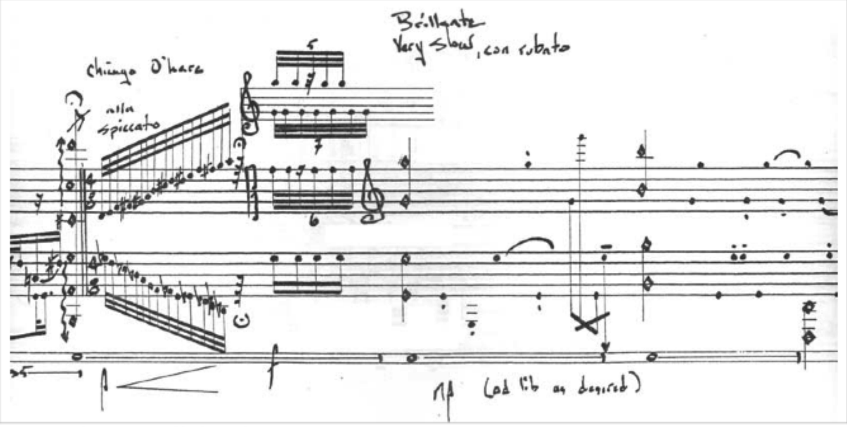


Figure 5.1
Scale-free sonata

Composed by Michael Edward Edgerton in 2003, *1 sonata for piano* was inspired by scale-free networks. The music obeys the principles of a growing network, incorporating growth and preferential attachment. The interplay between the music and networks is explained by the composer:

"6 hubs of different length and procedure were distributed over the 2nd and 3rd movements. Musically, the notion of an airport was utilized by diverting all traffic into a limited landing space, while the density of procedure and duration were varied considerably between the 6 differing occurrences."

The Image shows the beginning of what Edgerton calls Hub #5.

GROWTH AND PREFERENTIAL ATTACHMENT

Our journey towards understanding the origin of the scale-free property by asking: why are hubs and power laws absent from the model? The answer emerged in 1999, highlighting two hidden assumptions of the Erdős-Rényi model, each of which are violated in real networks [1]. Next we discuss these two assumptions separately.

NETWORKS EXPAND THROUGH THE ADDITION OF NEW NODES

The random network model assumes that we have a *fixed* number of nodes, *N*. The role of the modeler is to connect these nodes, while keeping *N* unchanged. Yet, in most real networks the number of nodes is not fixed, but continually *grows* thanks to the addition of new nodes. Let us consider a few examples:

- In 2001 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 **Fig. 5.2a**.

- The collaboration and citation networks continually expand through the publication of new research papers **Fig. 5.2b**.

- The Hollywood actor network continues to expand through the release of new movies **Fig. 5.2c**.

- At first the protein interaction network within our cells 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 today in a human cell over four billion years.

Consequently, if we wish to model these networks, we cannot resort to a static model. Rather our approach must 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. In most real networks, however, new nodes prefer to link to the more connected nodes, a process called *preferential attachment*. 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, but 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 connected nodes, we are more likely to link to a high-degree node than to a node with only few links.

- With more than a million scientific papers published each year, no scientist can attempt to read them all. The more cited is a paper, the more likely that we will notice it. Therefore, 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:

GROWTH

While the random network model assumes that the number of nodes, *N*, is fixed (time invariant), real networks are the result of a growth process that continuously increases *N*.

PREFERENTIAL ATTACHMENT

While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.

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, growth and preferential attachment have a particularly important role shaping a network's degree distribution.

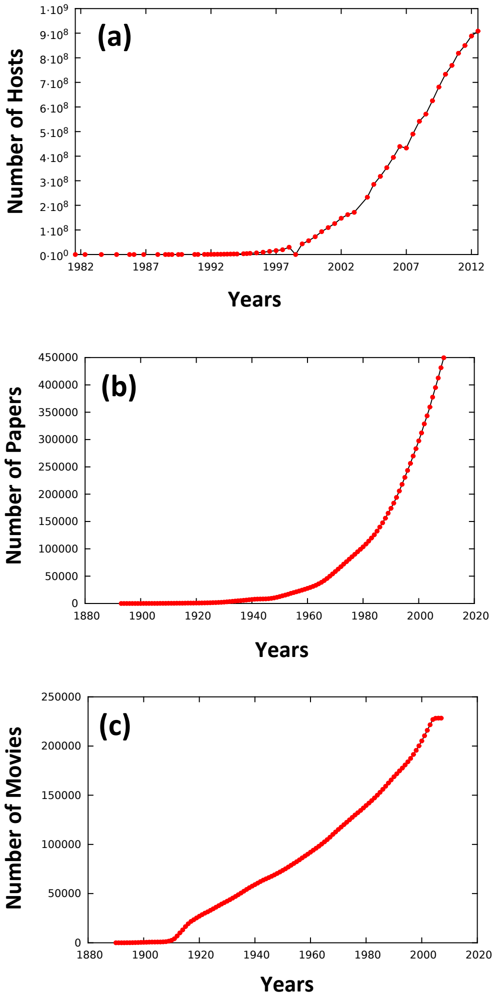


Figure 5.2
The growth of networks

(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* journals since the journal's funding in 1893. The observed growth drives the growth of both the science collaboration network as well as the citation network. Over the century the *Physical Review* portfolio has split several times, responding to the exponential growth of the number of research papers and to specialization. Today the corpus features *Physical Review Letters*, *Physical Review A, B, C, D, E, X* and *Reviews of Modern Physics*.

(c) Number of movies listed in IMDB.com, reflecting the growth of the Hollywood movie enterprise, and with that the growth of the actor network.

BOX 5.1

PREFERENTIAL ATTACHMENT: A BRIEF HISTORY

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

- It made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya (1887-1985) [2], proposed to explain the nature of certain distributions. Hence, in mathematics preferential attachment is often called a *Pólya process.*

- George Udmy Yule (1871-1951) in 1925 used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a *Yule process*.

- Rober Gibrat (1904-1980) in 1931 proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called *proportional growth*, this is a form of preferential attachment.

- George Kinsley Zipf (1902-1950) in 1941 used preferential attachment to explain the fat tailed distribution of wealth in the society [5].

- Modern analytical treatments of preferential attachment use of the master equation approach pioneered by the economist Herbert Alexander Simon (1916-2001). Simon used preferential attachment in 1955 to explain the fat-tailed nature of the distributions describing city sizes, word frequencies in a text, or the number of papers published by scientists [6].

- Building on Simon's work, Derek de Solla Price (1922-1983) used preferential attachment to explain the citation statistics of scientific publications, referring to it as *cumulative advantage* [7].

- In sociology preferential attachment is often called the *Matthew effect*, named by Robert Merton (1910-2003) [8] after a passage in the Gospel of Matthew: "For everyone who has will be given more, and he will have an abundance. Whoever does not have, even what he has will be taken from him."

- The term *preferential attachment* was introduced in the 1999 paper by Barabási and Albert [1] to explain the ubiquity of power laws in networks.

Note that the distributions characterized from Pólya to Merton describe scalar quantities, like the number of individuals with the same income or the size of cities. In contrast the Barabási-Albert model aims to describe networks. Networks have a wide array of topological characteristics that are absent from scalar distributions, but which are deeply affected by the power-law nature of the degree distribution.

SECTION 5.2

THE BARABÁSI-ALBERT MODEL

The recognition that growth and preferential attachment coexist in real networks has lead to the introduction of a minimal model capable of generating networks with power-law degree distribution [1]. The model is defined as follows:

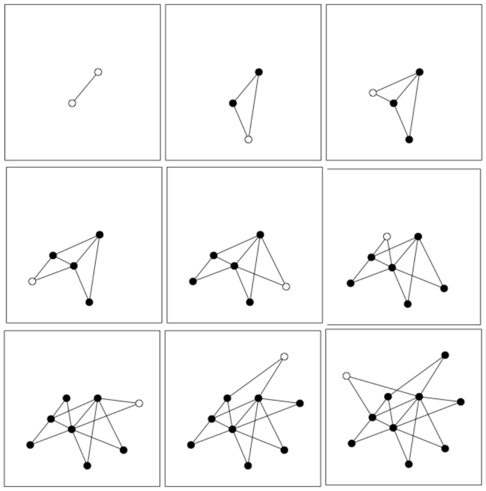


Figure 5.3
Time evolution of the Barabási-Albert model

The sequence of images shows the gradual emergence of a few highly connected nodes, or hubs, through growth and preferential attachment. White circles mark the newly added node to the network, which decides where to connect its two links (m=2) through preferential attachment **Eq. 5.1**. After [9].

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 **Fig. 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 π(k) that one of the links of the new node connects to node i depends on the degree k_i of node i as

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

Preferential attachment is a probabilistic rule: a new node is free to connect to any node in the network, whether it is a hub or has a single link. **Eq. 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. The model defined by steps (A) and (B) is called the *Barabási-Albert* model after the authors of the paper that introduced it in 1999 [1]. One may also encounter it in the literature as the BA model or the *scale-free* model. After t timesteps the Barabási-Albert model generates a network with $N = t + m_0$ nodes and $m_0 + mt$ links. As **Fig. 5.4** shows, the network generated by the model has a power-law degree distribution, a with a degree exponent γ=3.

As **Fig. 5.3** indicates, while most nodes in the network have only a few links, a few gradually turn into hubs. The hubs are the result of a *rich-gets-*

richer phenomenon: due to preferential attachment new nodes are more likely to connect to the more connected nodes than to the smaller degree nodes. Hence, the more connected nodes will acquire links at the expense of the less connected nodes, eventually turning into hubs.

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

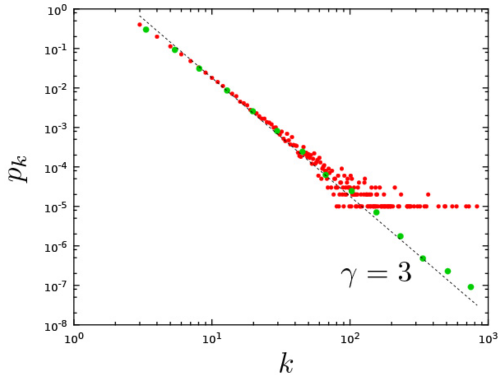


Figure 5.4
The degree distribution

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

To: ~~ALB~~ REKA ALBERT

from A.L. Barabási

Réka: Próbáld meg beprogramozni a következő két modellt; ami esetleg megadja a várt hatványfüggvényt:

MODEL 1: $t=0$: m vertices, no ~~d~~ edges.

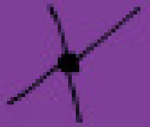
at time t: always add a new vertex, which has m edges (non-directed) coming out of it. The ends of the edges will be randomly connected to the existing vertices in the system.

thus: $t=1$: we have $(m+1)$ vertices, and m edges, coming out of the new vertex.

$t=2$: $(m+2)$ vertices; $2m$ edges

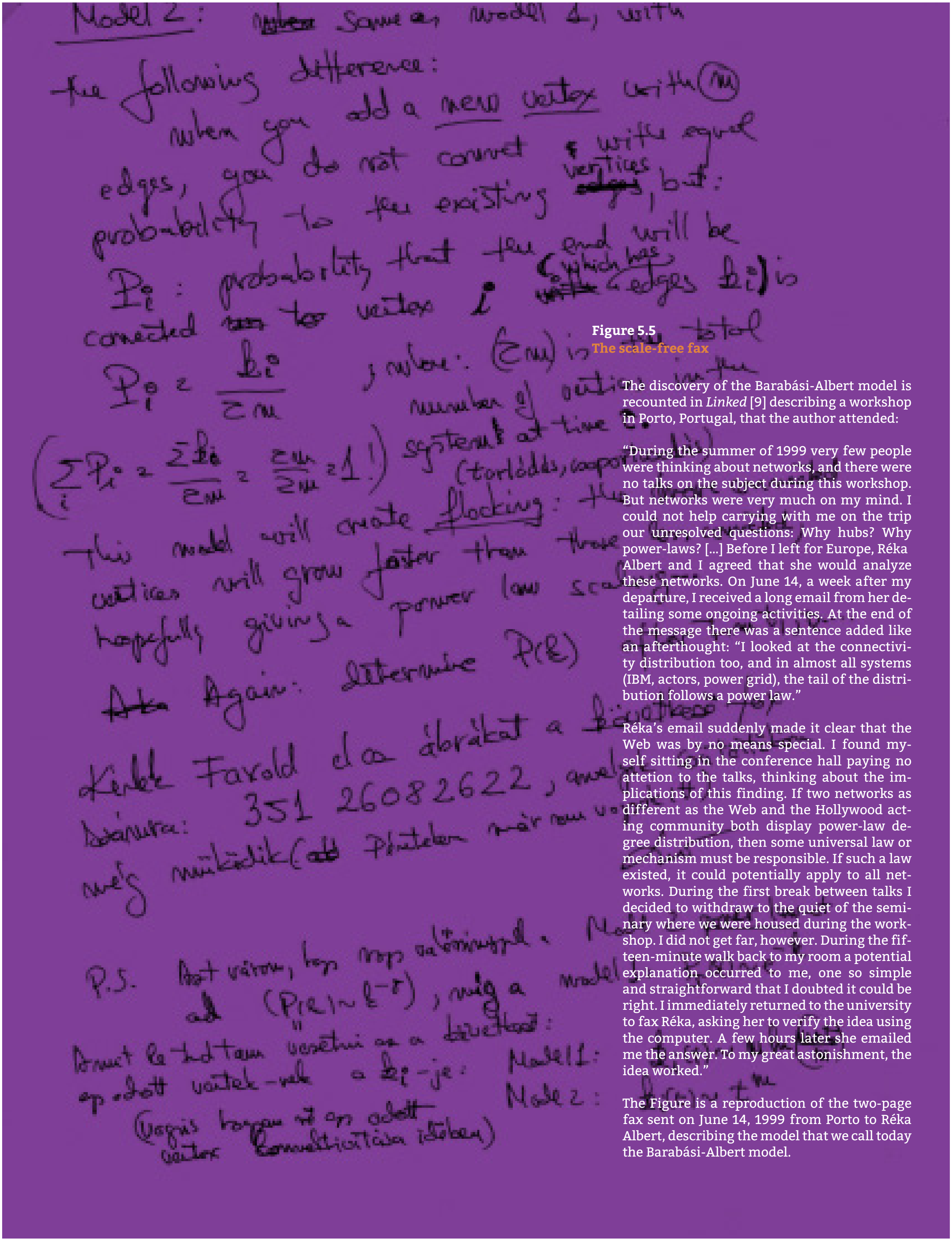
$\vdots$

$t=\tau$ $(m+\tau)$ vertices; τm edges

thus a new vertex always has m edges  whose ends are connected randomly to the already present vertices in the system.

The average connectivity after time τ: $\langle k\rangle = \frac{\tau m}{\tau + m} \to \underline{m}$

Determine $P(k)$ at different times $t_1, t_2 \ldots, t_n$ (where $t_1 \ldots$ are large), and see if it is exponential or power law!



Model 2: ~~when~~ same as model 1, with the following difference:
when you add a new vertex with (m) edges, you do not connect ~~&~~ with equal probability to the existing ~~edges~~ vertices, but:
P_i: probability that the end will be connected ~~to~~ to vertex i (which has ~~with~~ k_i edges) is

$P_i = \frac{k_i}{2m}$, where: $(2m)$ is total number of ~~out~~ in the system at time t.

$\left(\sum_i P_i = \frac{\sum k_i}{2m} = \frac{2m}{2m} = 1!\right)$

(torlódás, csoport...)

this model will create flocking: the ... those vertices will grow faster than ... hopefully giving a power law sca...

~~Ata~~ Again: determine $P(k)$

Kérlek Faxold el a ábrákat a következő számra: 351 2608 2622, amely még működik (~~ab~~ Pénteken már nem v...

P.S. Azt várom, hogy nagy valószínűséggel Mod... ad $(P(k) \sim k^{-\gamma})$, míg a model ...

Amit le tudtam vezetni az a következő: az adott vertek-nek a k_i-je: Model1: ...
Model 2: ...
(vagyis hogyan nő az adott vertex konnektivitása időben)

Figure 5.5
The scale-free fax

The discovery of the Barabási-Albert model is recounted in *Linked* [9] describing a workshop in Porto, Portugal, that the author attended:

"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? [...] Before I left for Europe, Réka Albert and I agreed that she would analyze these networks. On June 14, a week after my departure, I received a long email from her detailing some ongoing activities. At the end of the message there was a sentence added like an afterthought: "I looked at the connectivity distribution too, and in almost all systems (IBM, actors, power grid), the tail of the distribution follows a power law."

Réka's email suddenly made it clear that the Web was by no means special. I found myself sitting in the conference hall paying no attetion 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 housed during the workshop. 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 emailed me the answer. To my great astonishment, the idea worked."

The Figure is a reproduction of the two-page fax sent on June 14, 1999 from Porto to Réka Albert, describing the model that we call today the Barabási-Albert model.

BOX 5.2

THE MATHEMATICAL DEFINITION OF THE BARABÁSI-ALBERT MODEL

The definition of the Barabási-Albert model provided in this chapter leaves many mathematical details of the model unspecified:

- 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, independent of each other, or simultaneously. These problems were recognized by Riordan and Bollobás [10], who offered a definition that addresses these shortcomings. In contrast with the original model, Riordan and Bollobás allows for multiple edges and loops, showing later that their number will be negligible. The resulting model, called the Linearized Chord Diagram (LCD), is defined as follows:

Consider a fixed sequence of nodes v_1, v_2, ..., where the degree of the node v_i is k_i. We build a graph $(G_1^{(t)})_{t \geq 0}$ so that $G^{(t)}$ is a graph on v_i , $1 \leq i \leq t$ as follows: start with $G_1^{(0)}$ corresponding to an empty graph with no nodes, or with $G_1^{(1)}$ graph with one node and one loop. Given $G_1^{(t-1)}$ generate $G_1^{(t)}$ by adding the node v_t together with a single link between v_t and v_i, where i is chosen with probability

$$p_r(i=s)=\begin{cases} k_{{G^{(t-1)}}_1}\dfrac{k(v_s)}{(2t-1)}, & if\; 1 \leq s \leq t-1 \\ \dfrac{1}{(2t-1)}, & if\; s=t \end{cases} \tag{5.2}$$

That is, we place a link from node v_t to node $v_{i,}$ where the probability that node i is chosen as the target of this new link is proportional to its degree k_i at the time, the new link already contributing to the degree of v_t. Hence, the new node v_t can also link to itself with probability 1/(2t - 1). For $m > 1$, we add m links from v_t one by one, counting the previous links together with the outward half of the newly added link as contributing to the degrees.

DEGREE DYNAMICS

To understand the time evolution of the Barabási-Albert model, we first focus on the time-dependent degree of a node [11]. In the model a node has a chance to increase its degree each time a new node enters the network. When a new node joins the network, it will link to m of the $N(t)$ nodes present in the system. The probability that it chooses node i is given by Eq. 5.1. Assuming that k_i is a time-dependent continuous real variable, the rate at which node i acquires links follows the equation

$$\frac{\partial k_i}{\partial t} = m\Pi(k_i) = m\frac{k_i}{\sum\limits_{j=1}^{N-1} k_j}. \tag{5.3}$$

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 Eq. 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 Eq. 5.4 becomes

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

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

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

By integrating Eq. 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

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

The exponent β is the network's dynamical exponent and has the value $\beta = \frac{1}{2}$. Eq. 5.7 offers a number of predictions:

- The degree of each node increases following a power-law with the same dynamical exponent β =1/2 Fig. 5.6, implying that all nodes follow the same growth law.

- The growth in the degrees is sublinear (i.e. $\beta < 1$). In contrast in the Erdős-Rényi model $\langle k \rangle$ increases as $k_i \sim t$ if we add links one by one to the network. The sublinear nature of Eq. 5.7 is a consequence of the growing nature of the Barabási-Albert model: each new node has more nodes to link to than the previous nodes. Hence, with time each node competes for links with an increasing pool of nodes.

- The earlier node i was added, the higher is its degree $k_i(t)$. Hence, hubs are large not because they grow faster, but because they arrived earlier, a phenomenon called *first-mover advantage* in marketing and business.

- The growth rate of a node (i.e. the rate at which the node i acquires new links) is given by the derivative of Eq. 5.7

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

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

Taken together, the Barabási-Albert model offers a dynamical description of a network's evolution, capturing the fact that in real networks nodes arrive one after the other, connecting to the earlier nodes. This sets up a competition for links during which the older nodes have an advantage over the younger nodes, eventually turning into hubs.

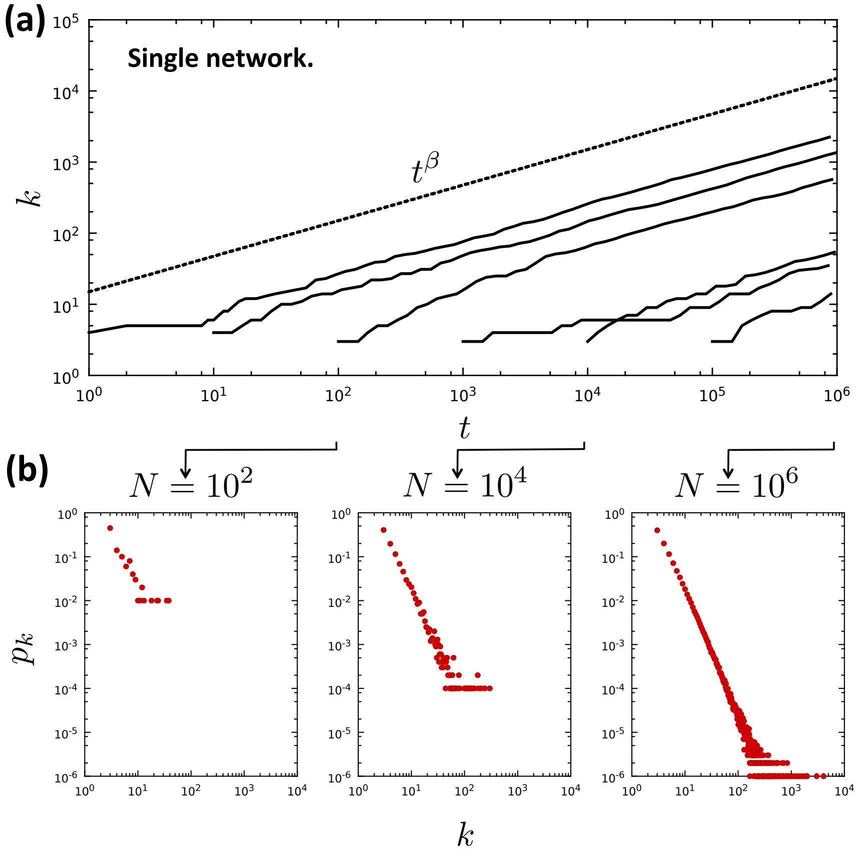


Figure 5.6
Degree dynamics

(a) Time dependence 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). One can see that each node increases its degree following the same law **Eq. 5.7**. Also, at any moment the older nodes have higher degrees. The dotted line corresponds to the analytical prediction **Eq. 5.7** with $\beta = 1/2$.

(b) Degree distribution of the network after the addition of $N = 10^2$, 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 p_k is plotted using linear binning, to better show the gradual emergence of the scale-free state.

BOX 5.3

THE MATHEMATICAL DEFINITION OF THE BARABÁSI-ALBERT MODEL

As we compare the predictions of the various network models with real data, we often have to decide how to measure time in networks. Real networks have evolved over rather different time scales: the first webpage was created in 1991, giving the WWW a history of a few decades at most. Given its trillion documents, this means that on average the WWW added more than a thousand nodes each second. In contrast the human cell is the result of 4 billion years of evolution; hence with roughly 20,000 genes, the cellular network added a node every 200,000 years. Given these enormous time-scale differences it seems impossible to use real time to compare the dynamics of these networks. Therefore, in network theory we use event time, that is, we advance time each time there is a change in the network topology. For example, in the Barabási-Albert model the addition of each new node corresponds to a new time step. Consequently in the model t=N. In more complicated models a distinct time step is assigned to each event—like the addition of a new node, the arrival of a new link, or the deletion of a node, any attempt to change the network topology. Obviously, if needed, we can establish a direct mapping between event time and the physical time.

DEGREE DISTRIBUTION

The distinguishing feature of networks generated by the Barabási-Albert model is their power-law degree distribution Fig. 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 model's degree distribution. The simplest is the continuum theory that we started developing in the previous section [1, 11]. It predicts that the degree distribution follows BOX 5.4,

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

with

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

Eq. 5.9 tells us that the degree distribution follows a power law with exponent γ=3, in agreement with the numerical results shown in Fig. 5.4 and Fig. 5.7. In turn Eq. 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. While the continuum theory predicts the correct degree exponent (γ=3), it fails to accurately predict the pre-factors of Eq. 5.9. This is why we use a proportional sign in Eq. 5.9, rather than equality. The exact degree distribution, with 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. As we show in ADVANCED TOPICS 5.A, the exact form of the degree distribution of the Barabási-Albert model is

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

Eq. 5.11 has several notable implications:

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

- The degree exponent γ is independent of m, a prediction that agrees with the numerical results Fig. 5.7a.

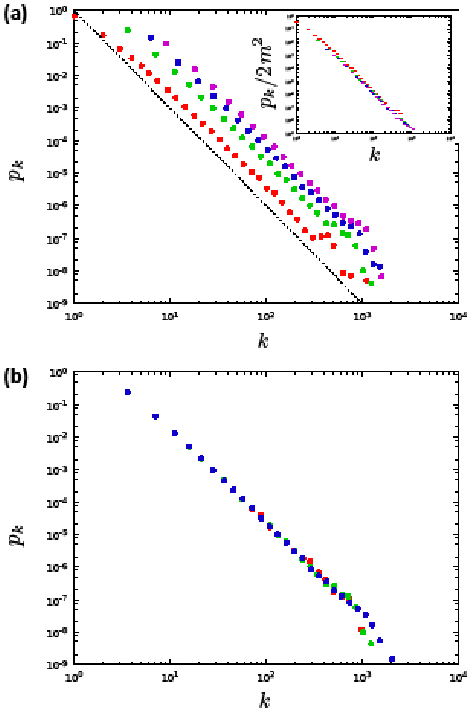


Figure 5.7
Probing the analytical predictions

(a) To show that p_k is independent of the parameters m and m_0, we generated networks with N=100,000 and m_0=m=1 (red), 3 (green), 5 (blue), and 7 (purple). The fact that the curves are parallel to each other indicates that γ is independent of m and m_0. The slope of the dashed line is -3. Inset: Eq. 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 all curves 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 (red), 100,000 (green), and 200,000 (blue), with m_0=m=3. The obtained p_k are practically indistinguishable, indicating that the degree distribution is time invariant.

- The power-law degree distribution observed in real networks describes systems of rather different age and size. Hence, a proper model should lead to a time-independent degree distribution. Indeed, Eq. 5.11 predicts that the degree distribution of the Barabási-Albert model is time independent, resulting in 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 Fig. 5.7b.

- Eq. 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 Fig. 5.7, inset).

In summary, the analytical calculations confirm that the Barabási-Albert model generates a power-law degree distribution, predicting the value of the degree exponent as γ=3. The exponent is independent of the parameters m and m_0. The calculations predict that the degree distribution is stationary (i.e. time invariant), explaining why networks with different history, size and age develop a similar degree distribution.

BOX 5.4

CONTINUUM THEORY

To calculate the degree distribution of the Barabási-Albert model we first determine the probability that the degree $k_i(t)$ of node i is smaller than a value k, i.e. $P(k_i(t) < k)$. Using Eq. 5.7, we can write

$$P(k_i(t) < k) = P\left(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}\right). \quad (5.12)$$

In the model we add the nodes at equal time intervals BOX 5.3. To capture this temporal uniformity we write the probablity that a node arrives at time ti as a random variable with a constant probability density

$$P(t_i) = \frac{1}{m_0 + t}. \quad (5.13)$$

Substituting Eq. 5.13 into Eq. 5.12 we obtain the cumulative distribution

$$P(k) = P\left(t_i \le \frac{m^{1/\beta}t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta}t}{k^{1/\beta}(t + m_0)} \quad (5.14)$$

We obtain the degree distribution $p(k)$ by taking the derivative of the cumulative function, i.e.

$$p(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^{1/\beta}t}{m_0 + t}\frac{1}{k^{1/\beta+1}}, \quad (5.15)$$

which for $t \gg m_o$ reduces to Eq. 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:

(1) Growth

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

(2) Random attachment

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

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

That is, $\pi(k_i)$ is independent of k_i, indicating that the new nodes choose randomly the nodes they link to. The continuum theory predicts that for Model A $k_i(t)$ increases logarithmically with time, i.e.

$$k(t) = m \ln\left(e \frac{m_0 + t + 1}{m_0 + t_i + 1}\right) \tag{5.17}$$

a much slower increase than the power law Eq. 5.7 derived earlier. Consequently the degree distribution becomes exponential Fig. 5.8a

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

As an exponential function decays much faster than a power law, it does not support hubs. Therefore the lack of preferential attachment eliminates the network's scale-free character and the hubs.

MODEL B

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

Preferential Attachment:

At each time step a node is selected randomly and connects to a node *i* with degree k_i already present in the network, where *i* is chosen with probability **Eq. 5.1**.

In Model B, the number of nodes remains constant during the network's evolution, but the number of links increases linearly with time. As a result the degree of each node also increases linearly with time **Fig. 5.8b**, inset

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

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. $L \ll N$), 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 **Fig. 5.8b**. Yet, p_k is not stationary, as after roughly $T \simeq N^2$ time steps the network converges to a complete graph. Consequently, after a transient period ($t \ll N$) the node degrees start to converge to the average degree **Eq. 5.19** and the degree distribution becomes peaked **Fig. 5.8b**. For $t \to N(N\text{-}1)/2$ the degree distribution becomes $p_k = \delta\ (N\text{-}1)$, i.e. the network turns into a complete graph in which all nodes have degree k_{max}=N-1. Therefore, in the absence of growth the network is not stationary, becoming a complete graph with time.

In summary, the 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.

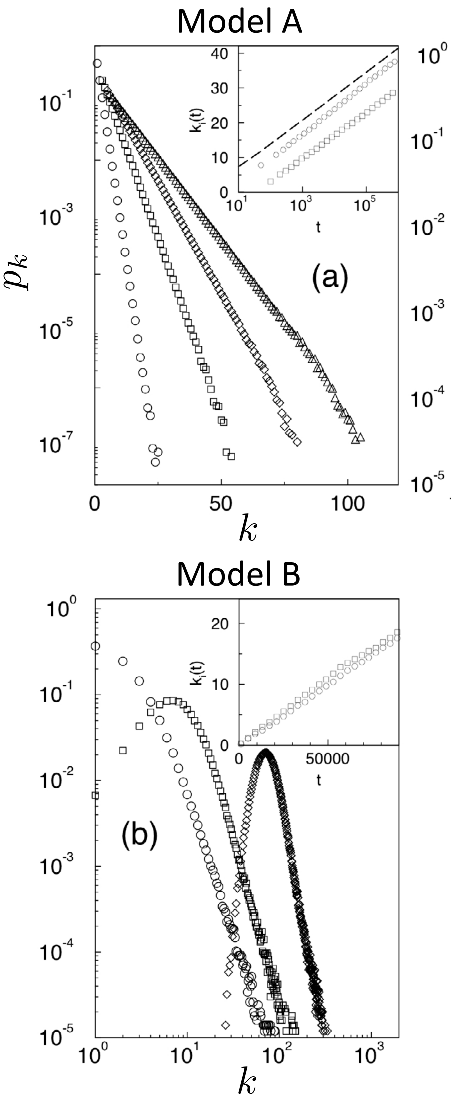


Figure 5.8
Model A and Model B

Numerical simulations of Model A and B, probing the role of growth andpreferential attachment.

(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.

Inset: Time evolution of the degree of two vertices added at t^1=7 and t^2=97 for m_0=m=3. The dashed line follows $k_i(t)$=m $ln(m_0$+t-1) as predicted by **Eq. 5.16** for large t.

(b) Degree distribution for Model B, that lacks growth but incorporates preferential attachment, shown for N=10,000 and t=N (circles), t=5N (squares), and t=40N (diamonds).

Inset: Time dependence of the degrees of two vertices for system size N=10,000, indicating that $k_i(t)$ grows linearly, as predicted by **Eq. 5.19**. After [14].

MEASURING PREFERENTIAL ATTACHMENT IN REAL NETWORKS

In the previous sections we showed that growth and preferential attachment are responsible for the scale-free property. The presence of growth in real systems is obvious: all large networks arrived to their current size by continuously 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. We start by noting that preferential attachment incorporates two hypotheses:

HYPOTHESIS 1

The likelihood to connect to a node depends on the 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 $\pi(k)$. To be specific, we can determine $\pi(k)$ for networks for which we [14, 15] 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.

Consider a network for which we have two different maps, the first taken at time t and the other at time $t + \Delta t$ Fig. 5.9. During the Δt time frame some nodes did not change their degree, so for these $k(t+\Delta t) = k(t)$. For nodes that did alter their degree we measure the change $\Delta k_i = k_i\,(t+\Delta t\,) - k_i(t)$. According to Eq. 5.1, the relative change $\Delta k_i/\Delta t$ should follow

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

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

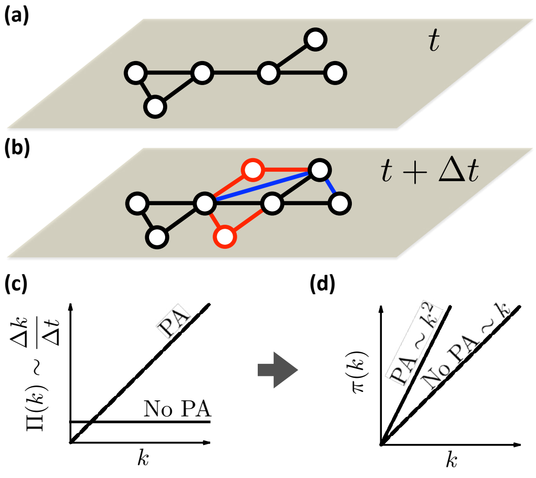


Figure 5.9
Detecting preferential attachment

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 that governs preferential attachment. Specifically we look at nodes that have gained new links thanks to the arrival of new nodes, like the two new red nodes at $t+\Delta t$. The blue lines correspond to links that connect previously disconnected nodes, called internal links. Their role is discussed in CHAPTER 6.

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

(d) The scaling of the cumulative preferential attachment function helps us detect the presence or absence of preferential attachment.

In practice the obtained $\Delta k_i/\Delta t$ curve is typically noisy, particularly for small networks. To reduce this noise we often measure the cumulative function

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

In the absence of preferential attachment we expect $\pi(k_i)$=constant, hence, $\pi(k) \sim$ k according to Eq. 5.21. If preferential attachment is present, i.e. $\pi(k_i)=k_i$, we expect $\pi(k) \sim k^2$. Fig. 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. Fig. 5.10 also suggests that $\pi(k)$ can be approximated with

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

For the Internet and citation networks we have $\alpha \simeq 1$, indicating that $\pi(k)$ depends linearly on k, as assumed in Eq. 5.2. This is in line with Hypotheses 1 and 2. For the co-authorship and the actor network the best fit provides $\alpha=0.9\pm0.1$ indicating the potential presence of a sublinear preferential attachment.

In summary, Eq. 5.20 helps us detect the presence (or absence) of preferential attachment in real networks. The measurements show that the attachment probability depends on the node degree, in line with Hypothesis 1. Yet, we also find that while in some systems preferential attachment is linear, in others it can be sublinear, hence, Hypothesis 2 is occasionally violated. The implications of this non-linearity is discussed in the next section.

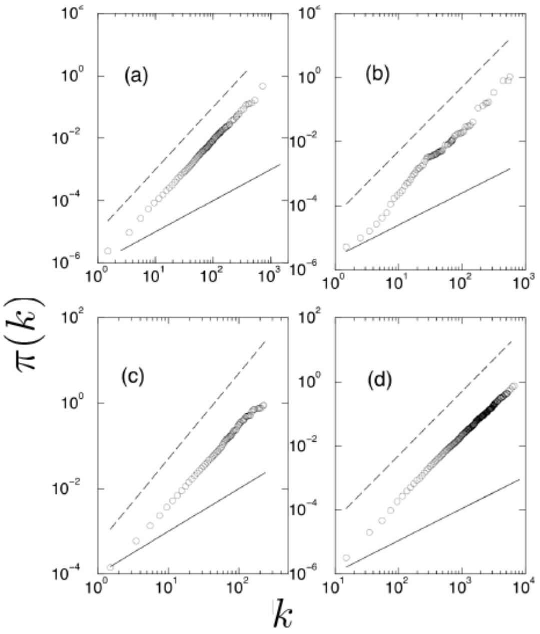


Figure 5.10
Evidence of preferential attachment

The figure shows the cumulative preferential attachment function $\pi(k)$, defined in Eq. 5.21, for several real systems:

(a) A citation network
(b) The Internet
(c) Neuroscience scientific collaboration network
(d) Actor network

In each panel we have two lines to guide the eye: the dashed line corresponds to linear preferential attachment ($\pi(k) \sim k^2$) and the continuous line indicates the absence of preferential attachment ($\pi(k) \sim k$). In line with Hypothesis 1 we detect a k-dependence in each dataset. Yet, in (c) and (d) $\pi(k)$ grows slower than k^2, indicating that for these two systems preferential attachment is sublinear, violating Hypothesis 2. Note that these measurements only consider links added through the arrival of new nodes, ignoring the addition of internal links. After [14].

SECTION 5.7

NON-LINEAR PREFERENTIAL ATTACHMENT

The observation of sublinear preferential attachment in Fig. 5.9 raises an important question: what is the impact of this nonlinearity on the network topology? To answer this we replace the linear preferential attachment Eq. 5.1 with Eq. 5.21 and re-calculate the degree distribution of the Barabási-Albert model. The behavior for α=0 is clear: in the absence of preferential attachment, the model reduces to Model A discussed in SECTION 5.4. Consequently the degree distribution will follow the simple exponential function Eq. 5.17. For α = 1 we recover the Barabási-Albert model, obtaining a scale-free network with degree distribution Eq. 5.14. We next focus on the case when α ≠ 0 and α ≠ 1. The calculation, providing p_k for an arbitrary α, is presented in ADVANCED TOPICS 5.B, predicting several scaling regimes [13]:

SUBLINEAR PREFERENTIAL ATTACHMENT (0 < α < 1)

For any α > 0 new nodes favor the more connected nodes over the less connected nodes. Yet, for α < 1 the bias is not sufficient to generate a scale-free degree distribution. Instead, in this regime the degrees follow the stretched exponential distribution SECT. 4.10

$$p_k \sim k^{-\alpha}\exp\left(\frac{2\mu(\alpha)}{\langle k\rangle(1-\alpha)}k^{1-\alpha}\right) \tag{5.23}$$

where *μ(α)* is a function that depends only weakly on α. For α → 1 Eq. 5.22 reduces to the degree distribution of the BA model. Indeed for *α*=1 we have *μ*=2, and $\lim_{\alpha\to 1}\frac{k^{1-\alpha}}{1-\alpha}=\ln k$. Therefore $p_k \sim k^{-1}\exp(-2\mathrm{ln}k)=k^{-3}$. The exponential cutoff in Eq. 5.22 implies that sublinear preferential attachment limits the size and the number of the hubs.

Sublinear preferential attachment also affects the size of the largest degree, k_{max}. In CHAPTER 4 we showed that for a scale-free network the degree of the largest node scales polynomially with time Eq. 4.14. For sublinear preferential attachment we have

$$k_{\max} \sim (\ln t)^{1/(1-\alpha)} \tag{5.24}$$

a logarithmic dependence that predicts a much slower growth of the maximum degee than the polynomial. This slower growth is the reason why the hubs are smaller for α < 1 Fig. 5.10.

SUPERLINEAR PREFERENTIAL ATTACHMENT (α > 1)

For α > 1 the tendency to link to highly connected nodes is enhanced, accelerating the *rich-gets-richer process*. The consequence of this is most obvious for α > 2, when the model predicts a *winner-takes-all* phenomenon: almost all nodes connect to a single or a few super-hubs. Hence, we observe the emergence of a hub-and-spoke network, in which most nodes link directly to a few central nodes. The situation for 1 < α < 2 is less extreme, but similar. This winner-takes-all process impacts the time dependence of the largest hub as well, finding that Fig. 5.12.

Hence for α > 1 the largest hub links to a finite fraction of nodes in the system.

$$k_{\max} \sim t \tag{5.25}$$

In summary, nonlinear preferential attachment introduces deviations from the power law degree distribution, either limiting the size of the hubs (α < 1), or leading to super-hubs (α > 1, Fig. 5.12). Hence, $\pi(k)$ needs to depend strictly linearly on the degrees for the resulting network to have a pure power law p_k. While in many systems we do observe such a linear behavior, in others, like the scientific collaboration network and the actor network, preferential attachment is sublinear, limiting the size of the hubs. This sublinear form of $\pi(k)$ could be responsible for the systematic deviations from a pure power-law degree distribution observed in the previous chapter. Hence for these systems a stretched exponential Eq. 5.22 should offer a better fit to the degree distribution.

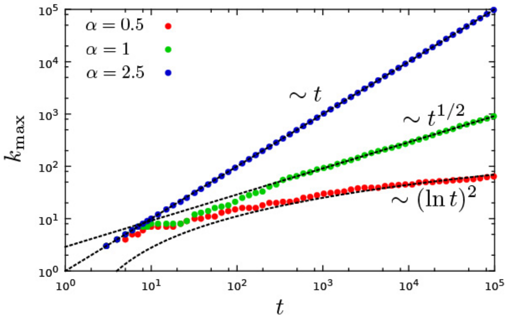


Figure 5.11
The growth of the hubs

The nature of preferential attachment affects the degree of the largest node. While in a scale-free network (α=1) the biggest hub grows as $t^{1/2}$ (green curve) Eq. 4.14, for sublinear preferential (α<1) attachment this dependence becomes logarithmic (red curve, Eq. 5.24. For superlinear preferential attachment the biggest hub grows linearly with time, always grabbing a finite fraction of all links (blue curve), Eq. 5.25. The symbols are provided by a numerical simulation; the dotted lines represent the analytical predictions.

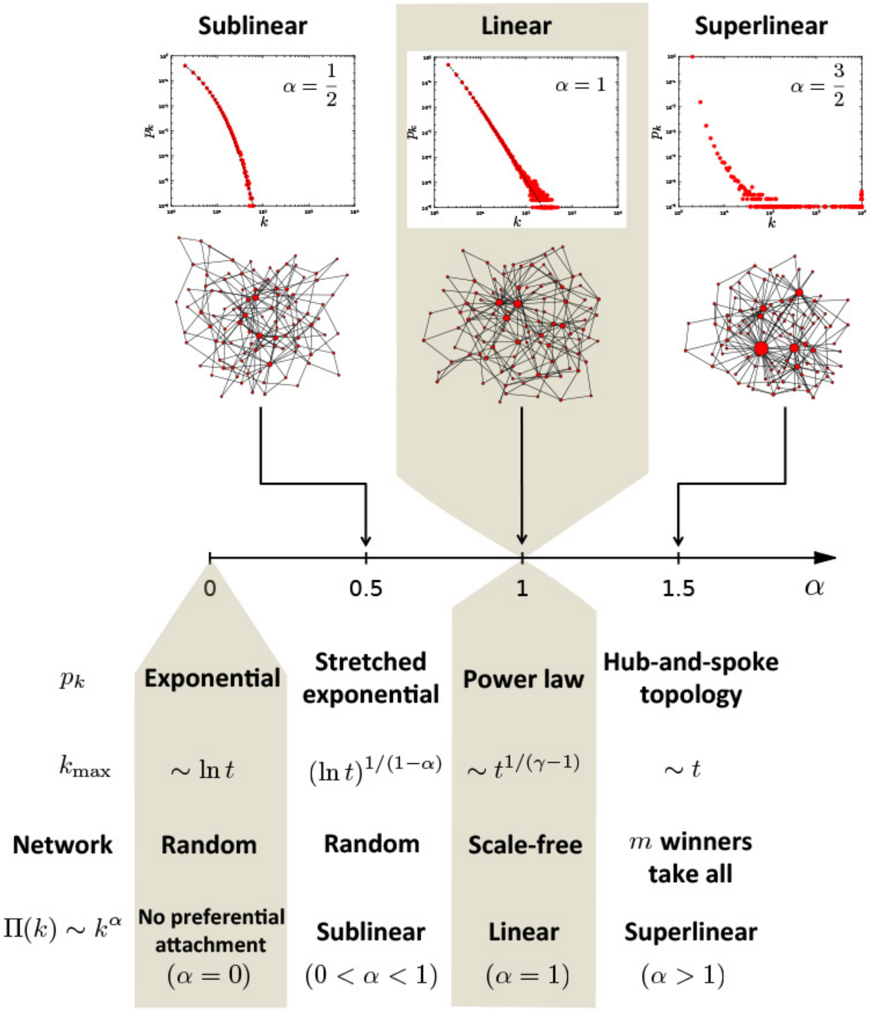


Figure 5.12

Nonlinear preferential attachment

The scaling regimes characterizing networks drivenby non-linear preferential attachment. The three toppanels show p_k for different α values (N=10^4). The network maps under them show the obtained topologies for N=100 nodes. The theoretical results indicate the existence of four scaling regimes:

No preferential attachment (α=0)
The network has a simple exponential degree distribution following **Eq. 5.18**. Hubs are absent and the resulting network behaves like a random network.

Sublinear preferential attachment (0<α<1)
The degree distribution follows the stretched exponential **Eq. 5.23**, resulting in fewer and smaller hubs than in a scale-free network. As α → 1 the cutoff length increases, hence p_k follows a power law over an increasing range of degrees.

Linear preferential attachment (α=1)
This corresponds to the Barabási-Albert model, hence the degree distribution follows a power law.

Superlinear preferential attachment (α>1)
The high-degree nodes are disproportionately attractive and the network follows a winner-takes-all dynamics, developing a hub-and-spoke topology. In this configuration the earliest nodes become super hubs and all subsequent nodes link to them. Hence p_k, shown for α=1.5 (top panel), indicates the coexistence of many small nodes with a few *super hubs* in the vicinity of k=10^4.

SECTION 5.8

THE ORIGINS OF PREFERENTIAL ATTACHMENT

Given the key the role preferential attachment plays in the evolution of real networks, we must ask, Where does preferential attachment come from? The question can be broken down 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 approaches to these questions. In the first class belong models that view preferential attachment as a result of an interplay between random events and some structural property of a network. These mechanisms do not require global knowledge of the network and rely on random actions, hence we will call them *local* or *random* mechanisms. A second class of models assume that each new node or link is preceeded by a cost-benefit analysis, balancing various needs with the available resources. These models assume familiarity with the whole network and rely on optimization principles, prompting us to call them *global* or *optimized* mechanisms. The purpose of this section is to discuss these two approaches.

LOCAL MECHANISMS

The link selection model offers perhaps the simplest example of a local or random mechanism capable of generating preferential attachment [16]. It is defined as follows:

- **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 one of the two nodes at the two ends of the selected link. This procedure is inherently local and random, as one does not need to know anything about the overall network topology to connect the new node. To show that this simple mechanism generates linear preferential attachment, we write the probability q_k that the node at the end of a randomly chosen

link has degree k as

$$q_k = Ckp_k\,. \tag{5.26}$$

Eq. 5.26 captures two effects:
(i) The higher the degree of a node, the higher the chance that it will be located at the end of the chosen link.

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

In Eq. 5.26 the value of C can be calculated using the normalization condition Σq_k = 1, obtaining C=1/$\langle k \rangle$. Hence the probability that we find a degree-k node at the end of a randomly chosen link is

$$q_k = \frac{kp_k}{\langle k \rangle}\,. \tag{5.27}$$

a quantity called excess degree. Eq. 5.27 also represents the probability that a new node connects to a node with degree k in the link selection model, hence it plays the role of the preferential attachment $\pi(k)$. Therefore Eq. 5.26 indicates that random link selection generates preferential attachment that is linear in then degree. While link selection is perhaps the simplest mechanism for preferential attachment, it is neither the first nor the most popular in the class of models relying on local mechanisms. That distinction goes to is the *copying model*, described in Fig. 5.13.

OPTIMIZATION

A longstanding assumption of economics is that humans make rational decisions, balancing cost against benefits. Innother 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 discuss below, such rational decisions can lead to preferential attachment as well [22, 23, 24].

Consider the Internet, whose nodes are routers or autonomous systems (AS), connected to each other via cables. Establishing a new Internet connection between two routers requires laying down a cable between them. As this can be costly, each new link is preceded by careful cost-benefit analysis. Each new router must agree with the nodes it links to that they will transmit the data packets leaving from or arriving to the new node (peering relationship). Therefore each new node will choose its link to balance access to good network performance (like 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 does not always offer the best network performance. For simplicity let us assume that the nodes are all located on a unit square. At each time step we add a new node by randomly choosing a point within the square. When deciding where to connect the new node i to the existing nodes, we calculate the cost function [22]

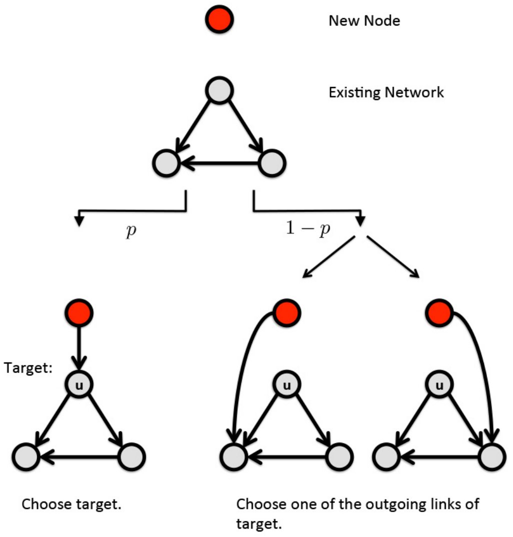


Figure 5.13
Copying model

When building a new webpage, authors tend to borrow links from webpages covering similar topics, a process captured by the copying model [17, 18]. In the model, in each time step a new node with a single link is added to the network. To choose the target node we randomly select a node u and follow a two-step procedure:

(a) **Random Connection**: With probability p the new node links to u.

(b) **Copying**: With probability 1-p we randomly choose an outgoing link of node u and link the new node to the link's target. Hence, the new node *copies* one of the links of an earlier node. For step (a) the probability of selecting a particular node is $1/N$. Step (b) is equivalent with selecting a node linked to a randomly selected link. The probability of selecting a degree-k node through the copying process of step (b) is $k/2L$ for undirected networks. That is, the likelihood that the new node will connect to a degree-k node follows preferential attachment $\pi(k) = p\ / N+(1- p)k/(2L)$, which is linear in k. The popularity of the copying model lies in its adaptability 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. Without friends, it is difficult to make new friends.

- **Citation Networks**: No scientist can be familiar with all papers published on a certain topic. If we assume that authors decide what to cite by randomly selecting references from the papers they have already read, then papers with more citations are more likely to be cited again.

- **Protein interaction networks**: Gene duplication, a common mechanism leading to next genes in a cell, can be mapped into the copying model, explaining the scale-free nature of protein interactions networks [19, 20].

$$C_i = \min_j[\delta d_{ij} + h_j] \tag{5.28}$$

for each node j already in the network, where d_{ij} is the Euclidean distance between i and j, and h_j is the network-based distance of node j to the first node in the network, designated as the "center " of the network Fig. 5.14. Hence h_j captures the "resources" offered by node j, in the form of 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 Eq. 5.28 and Fig. 5.15:

STAR NETWORK $\delta < (1/2)^{1/2}$

For δ = 0 the Euclidean distances are irrelevant, hence each node will simply link to the central node, turning the network into a star. This star configuration persist for any δ < $(1/2)^{1/2}$, guaranteeing that the h_{ij} term dominates over the δd_{ij} term in Eq. 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 Eq. 5.27. In this case each new node will connect to the node closest to it. The resulting graph is a dynamic version of the Euclidean minimum spanning tree. The resulting network will have a bounded degree distribution, like a random network Fig. 5.15.

SCALE-FREE NETWORK $4 \leq \delta \leq N^{1/2}$

Numerical simulations and analytical calculations [22] indicate that for intermediate δ values the network develops a scale-free topology.

The origin of the power law 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 Fig. 5.14. The size of each basin correlate with h_j of node j at its center, which in turn correlates with the node's degree k_j Fig 5.14f.

(ii) Randomness: We choose randomly the position of the new node, ending in one of the N basins of attraction. The node with the largest degree largest basin of attraction, will gain the most new nodes and links. This leads to preferential attachment, documented in Fig. 5.15d.

In summary, the microscopic mechanisms responsible for preferential attachment can have two fundamentally different origins BOX 5.5: it can be rooted in random processes, like link selection or copying, or in optimization process, when new nodes balance conflicting criteria as they decide where to connect. These results help us understand why preferential attachment is present in so different systems as the cell or the Internet. The diversity of the mechanisms discussed in this section suggest that preferential attachment is so widespread 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.

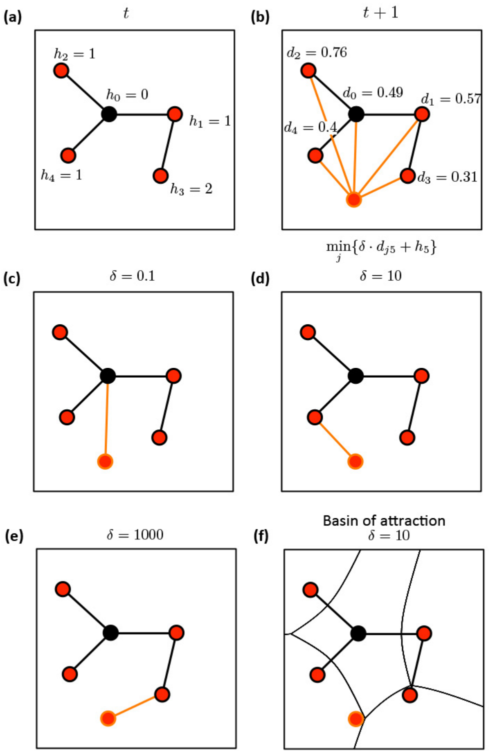


Figure 5.14
Optimization Model

(a) A small network configuration, where the h_j term in the cost function of Eq. 5.28 is shown for each node. Here h_j represents the network-based distance of node j from node i=0. Hence h_0=0 and h_3=2.

(b) A new node, shown in orange, will choose the node j to connect to by minimizing c_j of Eq. 5.28 . If δ =0 or δ is small the new node will connect to the central node with h_j =0.

(c)-(e) As we increase δ, the balance in Eq. 5.28 changes, forcing the new node to connect to different nodes. The panels (c)-(e) show the choice of the new node (orange) for a different values of δ for the given network configuration.

(f) The basin of attraction for each node for δ=10. A new node arriving inside a particular 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 hj, the larger can be the distancento the new node while still minimizing the cost Eq. 5.28. Yet, the higher the degree of node j, the smaller is its expected distance to the central node h_j.

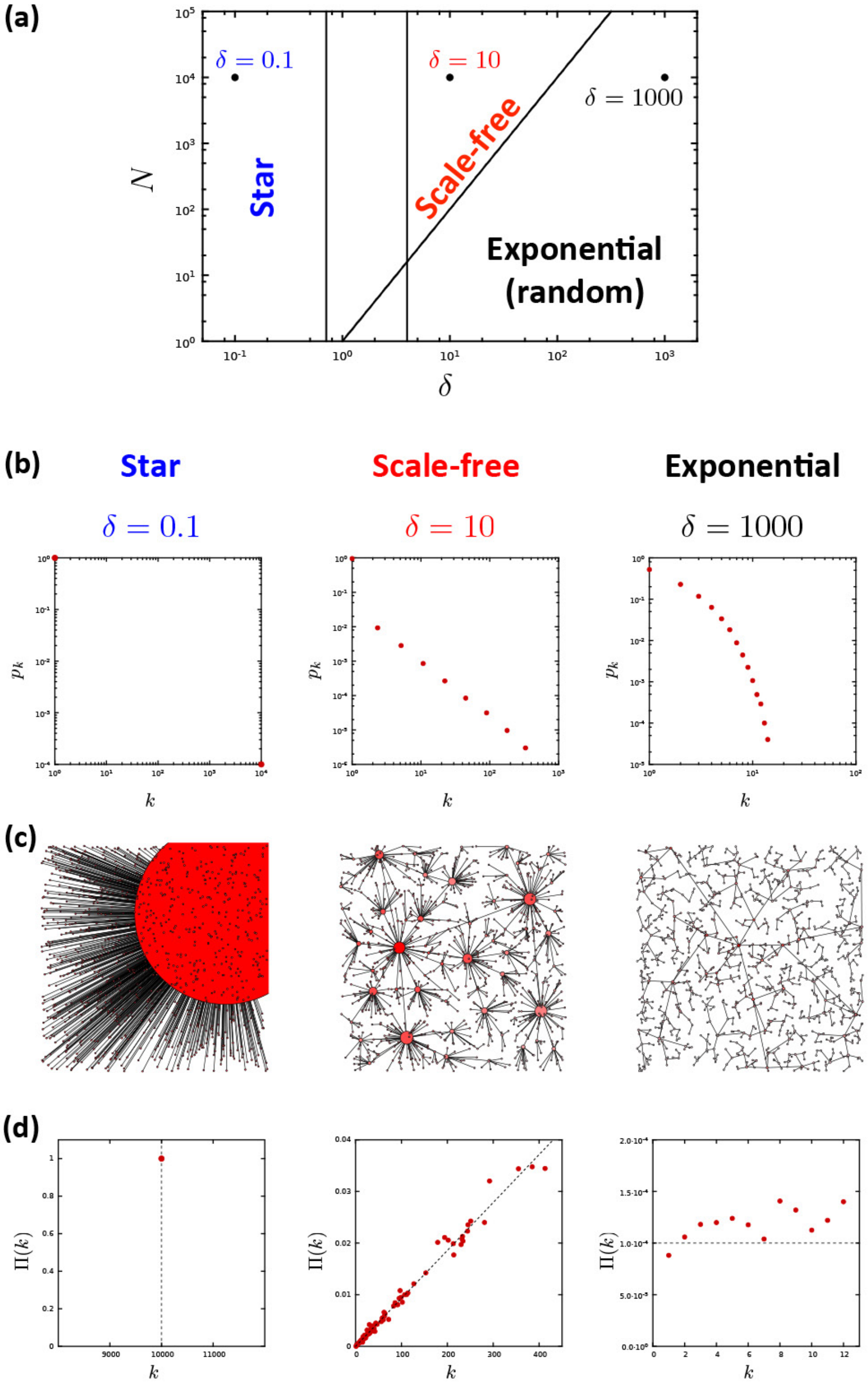


Figure 5.15

Scaling in the optimization model

(a) A schematic diagram, describing the three main classes of networks generated by the optimization model: star topology, scale-free topology and exponential networks. The structure of the network in the unmarked area is unknown. The boundary of the star configuration is given by s=(1/2)$^{1/2}$. Indeed, the maximum distance between two nodes on a square lattice with unit length, over which the model is defined, is the diagonal 2$^{1/2}$. Therefore if $\delta <$ 1/2$^{1/2}$, for any new node $\delta\ d_i$<1. In this case the cost of connecting to the central node is ci = δd_{ij}+0 , which is 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 network, see (c)). The oblique line making the 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}~$N^{-1/2}$ 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 ~log N, is scale-free networks networks h_j~lnln$N$). Therefore C_i is dominated by the $\delta\ d_{ij}$ term and the smallest C_i is achieved by minimizing the distance-dependent term. Note that strictly speaking the transition only occurs in the $N \rightarrow \infty$.

(b) Degree distribution for networks emerging in the three phases discussed above for N=10^4.

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

(d) We used the method described in SECT. 5.7 to measure the preferential attachment function $\pi(k)$. 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 indicate the presence of a linear preferential attachment in the scale-free phase, but its absence in the star and the exponential phases.

BOX 5.5

LUCK OR REASON: AN ANCIENT DEBATE

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 randomness is responsible for the power-law nature of word frequencies. Mandelbrot, however, fiercely defended an optimization-based framework.

The debate spanned seven papers and several years and is one of the most vicious scientific disagreement on record. It is hard to know what set it off—it may have been Simon's brief note in his 1955 paper [26], dismissing Mandelbrot's explanation that power laws observed in linguistics are routed in an optimization process [27]. Mandelbrot responded with a comment [28] stating that 'Simon's model is analytically circular.' The essence of Simon's lengthy reply a year later is well summarized in its abstract: 'Dr. Mandelbrot's principal and mathematical objections to the model are shown to be unfounded' [29]. This prompted a 19 page response by Mandelbrot entitled 'Final Note [...]', stating that 'most of Simon's (1960) reply was irrelevant' [30] and, ensuring that this will not be the final note. Sure enough, Simon's subsequent reply states that 'this present "Reply" refutes the almost entirely new arguments introduced by Dr. Mandelbrot' [31].

That inspired a paper creatively entitled a "Post Scriptum to "Final Note,"" by Mandlebrot [32], stating that 'My criticism has not changed since I first had the privilege of commenting upon a draft of Simon,' Simon's final note ends but does not resolve the debate: "Dr. Mandelbrot has proposed a new set of objections to my 1955 models of Yule distributions. Like earlier objections, these are invalid." [33].

In the context of networks the argument titled in Simon's favor the power laws observed in complex networks appear to be driven by randomness and preferential attachment. Yet, as we seek to explain the origins of preferential attachment, the optimization-based ideas proposed by Mandelbrot play an important role.

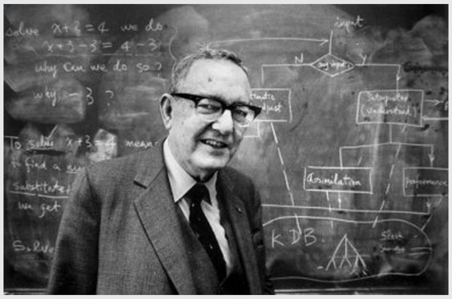


Figure 5.15a
Herbert Simon

Figure 5.15b
Benoit Mandelbrot

DIAMETER AND CLUSTERING COEFFICIENT

To complete the characterization of the Barabási-Albert model we need to discuss the behavior of two additional measures: the network diameter and the clustering coefficient. Both quantities play an important role in comparing the model predictions to the properties of real systems.

DIAMETER

The network diameter, representing the maximum distance in the Barabási-Albert network, is predicted to follow

$$D \sim \frac{\log N}{\log \log N} \cdot \qquad (5.29)$$

a result obtained independently by Cohen and Havlin [34] and Bollobás and Riordan [35], the latter also offering an exact proof. Eq. 5.29 tells us that the network diameter grows slower than log N, hence the distances in the Barabási-Albert model are smaller than the distances observed in a random graph of similar size. The difference is particularly relevant for large N. Note that while Eq. 5.29 is derived for the diameter, we expect that the average distance $\langle d \rangle$ scales in a similar fashion. The impact of the log log N the log N term captures the scaling of $\langle d \rangle$ with N, but for large $N(\geq 10^4)$ the impact of the logarithmic correction becomes noticeable.

CLUSTERING COEFFICIENT

The clustering coefficient of the Barabási-Albert model follows ADVANCED TOPICS 5.C.

$$C = \frac{m-1}{8}\frac{(\ln N)^2}{N}, \qquad (5.30)$$

a result obtained by Klemm and Eguiluz [36], and proved by Bollobás [37]. The prediction Eq. 5.30 is quite different from the 1/N dependence obtained for the random network model Fig. 3.20. The difference comes in the (ln$N$)^2 term, that increases the clustering coefficient for large N. Consequently the clustering coefficient of the Barabási-Albert model decays slower than expected for a random network, indicating that the obtained network is locally more clustered.

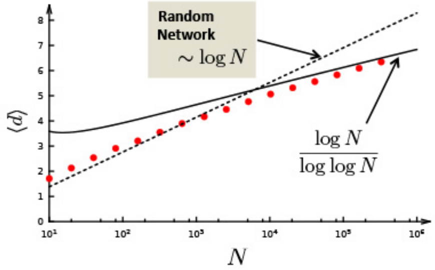


Figure 5.16
Average distance

The dependence of the average distance on the system size in the Barabási-Albert model. The continuous line corresponds to the exact result Eq. 5.29, while the dotted line corresponds to the prediction obtained in CH. 3 for a random network. Note that 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 $m = 2$.

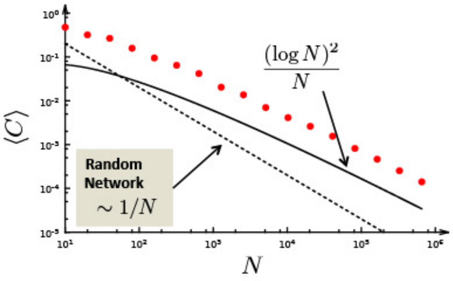


Figure 5.17
Clustering coefficient

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 Eq. 5.30, while the dotted line corresponds to the prediction for a random network, for which $\langle C \rangle$ ~1/N. The results were averaged for ten independent runs for $m = 2$. The dashed and continuous curves are only drawn to indicate the N dependent trends. Hence, they do not represent a precise fit.

SUMMARY

The most important message of the Barabási-Albert model is that network structure and evolution are inseparable. Indeed, in the Erdős-Rényi, configuration or the hidden parameter models the role of the modeler is to 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 feasible to the original brush strokes. consequently, the modeling philosophy of the Barabási-Albert model is simple: *to understand the topology of a complex system, we first need to describe how it came into being*.

Dynamics and network assembly take the driving role and the structural characteristics of the network, like the degree distribution, is a byproduct of this modeling philosophy. Random graphs, the configuration and the hidden parameter models will continue to play an important role as we try to understand how certain network properties 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. In its current form the Barabási-Albert model cannot describe the wide range of network characteristics observed in real systems. This is illustrated by the model's notable limitations:

- If predicts γ=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 known to occur in networks, from linking already existing nodes to the disappearance of links and nodes, are absent from the model.

- The model does not allow 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 is not designed to capture the details of any particular real network systems. 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. Finally, the results discussed in this chapter allow us to formulate the next law:

The Third Law of Networks: Growth and Preferential Attachment.
Hubs and power laws are a joint consequence of growth and preferential attachment.

Let us revisit the three criteria we used earlier to establish the validity of a network law:

(a) Quantitative formulation of the third law is provided by the Barabási-Albert model, together with its documented ability to generate scale-free networks based on growth and preferential attachment.

(b) Universality: SECTION 5.7 offers direct empirical evidence that real networks that exhibit the scale-free property are characterized by preferential attachment; SECTION 5.2 offers evidence of growth.

(c) Non-random origin: Preferential attachment is obviously absent from random networks, which is the main reason why random networks do not develop hubs and power laws.

BOX 5.6

AT A GLANCE: BARABÁSI-ALBERT MODEL

Number of nodes

$N = t$

Number of links

$N = mt$

Average Degree

$\langle k \rangle = 2m$

Degree dynamics

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

Dynamical exponent

$\beta = 1/2$

Degree distribution

$p_k \sim k^{-\gamma}$

Degree exponent

$\gamma = 3$

Average distance

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

Clustering coefficient

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

SECTION 5.11

HOMEWORK

1. Calculate the degree distribution of the directed Barabási-Albert model. That is, in each time a new node arrives, that connects with a directed link to a node chosen with preferential attachment Eq. 5.1, where $\pi(k_{in})$ depends only the node's in-degree. Discuss both the in and out-degree distribution of the resulting network.

2. Use the rate equation approach described above that the directed copying model leads to a scale-free network with the incoming degree exponent $\gamma_{in} = (2 - p) / (1- p)$, hence the degree exponent varies between $\gamma_{in} = 2$ for $p \to 0$ and $\gamma_{in} = \infty$ for $p \to 1$.

ADVANCED TOPICS 5.A DERIVING THE DEGREE DISTRIBUTION

A number of analytical techniques are available to calculate the exact form of the degree exponent provided in Eq. 5.11. Next we derive it using the rate equation approach [12, 13]. The method is rather general, allowing us to explore the properties of a wide range of growing networks. Consequently, the calculations described here will be of direct relevance for many systems, from the WWW to protein interaction networks. Let us denote with $N(k, t)$ the number of nodes with degree k at time t. The degree distribution $p_k(t)$ relates to this quantity via $p_k(t) = N(k,t)/N$.

Since at each time-step we add a new node to the network, we have $N = t$. That is, at any moment the total number of nodes equals the number of timesteps BOX 5.3. We write preferential attachment as

$$\Pi(k) = \frac{k}{\sum_j k_j} = \frac{k}{2mt} \cdot \tag{5.31}$$

where the $2m$ term captures the fact that that in an undirected network each link contributes to the degree of two nodes. Our goal is to calculate the changes in the number of nodes with degree k after a new node is added to the network. For this we inspect the two events that alter $N(k, t)$ (and hence $p_k(t)$) following the arrival of a new node:

(i) A new node can link to a degree-k node, turning it into a degree (k+1 node, hence decreasing $N(k, t)$.

(ii) A new node can link to a degree (k-1) node, turning it into a degree k node, hence increasing $N(k, t)$.

The number of links that are expected to connect to degree k nodes after the arrival of a new node is

$$\frac{k}{2mt} \times Np_k(t) \times m = \frac{k}{2} p_k(t), \tag{5.32}$$

where the first term captures the probability that the new node will link to a degree-k node (preferential attachment); the second term provides the total number of nodes with degree *k*, as the more nodes are in this category, the more likely that a new node will attach to one of them; the third term is simply the degree of the incoming node, as the higher *m*, the higher the chance that the new node will link to a degree-*k* node. We next apply **Eq. 5.32** to cases (i) and (ii) above:

(i') The number of degree *k* nodes that acquire a new link becoming (*k*+1) degree nodes, is

$$\frac{k}{2} p_k(t) \tag{5.33}$$

(ii') The number of degree (*k*-1) nodes that acquire a new link, increasing their degree to *k* is

$$\frac{k-1}{2} p_{k-1}(t). \tag{5.34}$$

Combining **Eq. 5.32** and **Eq. 5.33** we obtain the expected number of degree-*k* nodes after the addition of a new node

$$(N+1)p_k(t+1) = Np_k(t) + \frac{k-1}{2} p_{k-1}(t) - \frac{k}{2} p_k(t). \tag{5.35}$$

This equation applies to all nodes with degree *k*>*m*. As we lack nodes with degree *k*=0,1, ... , *m*-1 in the network (each new node arrives with degree *m*) we need a separate equation for degree *m* modes. Following the arguments we used to derive **Eq. 5.35**, we obtain

$$(N+1)p_m(t+1) = Np_m(t) + 1 - \frac{m}{2} p_m(t)\ . \tag{5.36}$$

Eq. 5.35 and **5.36** are the starting point of the recursive process that provides p_k. Let us use the fact that we are looking for a stationary degree distribution, supported by numerical simulations **Fig. 5.6**. This means that in the $N = t \rightarrow \infty$ limit, $p_k(\infty) = p_k$. Using this we can write the l.h.s. of **Eq. 5.35** and **5.36** as $(N+1)p_k\,(t+1) - Np_k\,(t) \rightarrow Np_k\,(\infty) + p_k\,(\infty) - Np_k\,(\infty) = p_k\,(\infty) = p_k\,(N+1)\ p_{mk}\,(t+1) - Np_m(t) \rightarrow p_m$. Therefore the rate equations **Eq. 5.35** and **5.36** take the form:

$$p_k = \frac{k-1}{k+2} p_{k-1} \quad k > m \tag{5.37}$$

$$p_m = \frac{2}{m+2} \tag{5.38}$$

Note that **Fig. 5.37** can be rewritten as

$$p_{k+1} = \frac{k}{k+3} p_k \tag{5.39}$$

via a *k*→*k*+1 variable change. To obtain the degree distribution, we use a recursive approach. That is, we write the degree distribution for the smallest degree, *k*=*m*, using **Eq.. 5.38** and then use **Eq. 5.39** to calculate p_k for the higher degrees:

$$\begin{aligned} p_{m+1} &= \frac{m}{m+3} p_m = \frac{2m}{(m+2)(m+3)} \\ p_{m+2} &= \frac{m+1}{m+4} p_{m+1} = \frac{2m(m+1)}{(m+2)(m+3)(m+4)} \\ p_{m+3} &= \frac{m+2}{m+5} p_{m+2} = \frac{2m(m+1)}{(m+3)(m+4)(m+5)} \end{aligned} \tag{5.40}$$

At this point we notice a simple recursive pattern: by replacing m+3 with k we obtain the probability to observe a node with degree k

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}, \tag{5.41}$$

which represents the exact form of the degree distribution for the Barabási-Albert model. Note that:

- For large k this becomes p_k~ k^{-3}, in agreement with the numerical result.

- The prefactor of Eq. 5.11 or Eq. 5.41 is different from the prefactor derived in Eq. 5.9.

This form was derived independently in [12] and [13], and the mathematical proof of its validity is provided in [10]. Note that the rate equation formalism offers an elegant continuum equation satisfied by the degree distribution of the Barabási-Albert model [16]. Starting from the equation

$$p_k = \frac{k-1}{2}p_{k-1} - \frac{k}{2}p_\infty \tag{5.42}$$

we can write

$$2p_k = (k-1)p_{k-1} - kp(k) = -p_{k-1} - k[p_k - p_{k-1}], \tag{5.43}$$

$$2p_k = -p_{k-1} - k\frac{p_k - p_{k-1}}{k-(k-1)} = -p_{k-1} - k\frac{\partial p_k}{\partial k} \tag{5.44}$$

obtaining

$$p_k = \frac{1}{2}\frac{\partial[kp_k]}{\partial k} \tag{5.45}$$

One can check that the solution of Eq. 5.45 is

$$p_k \sim k^{-3}. \tag{5.46}$$

ADVANCED TOPICS 5.B
NONLINEAR PREFERENTIAL ATTACHMENT

The purpose of this section is to derive the degree distribution of an evolving networks governed by a nonlinear preferential attachment. We follow Krapivsky et al. [13]. As the results of Ref. [13] were derived for undirected networks, here we adjusted the calculation to cover undirected networks.

Strictly speaking the degree distribution only exists for α ≤ 1. For α > 1 a few nodes attract a finite fraction of links, as explained in SECT. 5.7, and we do not have a stationary p_k. Therefore, we limit ourself to the α ≤ 1 case. We start with the Barabási-Albert model, in which at each time step a new node is added with m new links. We connect each new link to an existing node with probability

$$\Pi(k_i) = \frac{k_i^\alpha}{\mu(\alpha)} \, . \tag{5.47}$$

where k_i is the degree of node i, 0 < α ≤ 1 and

$$\mu(\alpha,t) = \sum_k k^\alpha p_k(t) \, . \tag{5.48}$$

is the normalization factor. Note that $\mu(0,t) = \sum p_k(t) = 1$ and $\mu(1,t) = \sum_k kp_k(t) = \langle k \rangle = 2mt/N$ is the average degree. Since 0 < α ≤ 1,

$$\mu(0,t) \le \mu(\alpha,t) \le \mu(1,t) \, . \tag{5.49}$$

Therefore in the long time limit

$$\mu(\alpha, t \to \infty) = \text{constant} \, . \tag{5.50}$$

whose precise value will be calculated later. For simplicity, we adopt the notation $\mu \equiv \mu(\alpha, t \to \infty)$.

Following the rate equation approach introduced in Advanced TOPICS 5.A, we write the rate equation for the network's degree distribution as

$$p_k(t+1) = \frac{m}{\mu(\alpha,t)}[(k-1)^\alpha p_{k-1}(t) - k^\alpha p_k(t)] + \delta_{k,m}. \qquad (5.51)$$

The first term on the r.h.s. describes the rate at which nodes with degree (*k*-1) gain new links; the second term describes the loss of degree-*k* nodes when they gain new links, turning into (*k*+1) degree nodes; the last term represents the newly added nodes with degree *m*. Asymptotically, in the *t*→∞ limit we can write p_k=p_k(*t* + 1)=p_k(*t*). Substituting *k*=*m* in Eq. 5.51 we obtain:

$$p_m = -\frac{m}{\mu} - m^\alpha p_m + 1$$

$$p_m = -\frac{\mu/m}{\mu/m + m^\alpha}. \qquad (5.52)$$

For *k* > *m*

$$p_k = \frac{m}{\mu}[(k-1)^\alpha p_{k-1} - k^\alpha p_k] \qquad (5.53)$$

$$p_k = \frac{(k-1)^\alpha}{\mu/m + k^\alpha} p_{k-1} \qquad (5.54)$$

Solving Eq. 5.53 recursively we obtain

$$p_m = \frac{\mu/m}{\mu/m + m^\alpha} \qquad (5.55)$$

$$p_{m+1} = \frac{\mu/m \cdot m^\alpha}{\mu/m + (m+1)} \frac{\mu/m}{\mu/m + m^\alpha} \qquad (5.56)$$

$$p_k = \frac{\mu/m}{k^\alpha} \prod_{k}^{j=m} \left(1 + \frac{\mu/m}{j^\alpha}\right)^{-1} \qquad (5.57)$$

To determine the large *k* behavior of p_k we take the logarithm of (52):

$$\ln p_k = \ln(\mu/m) - \alpha \ln k - \sum_{k}^{j=m} \left(1 + \frac{\mu/m}{j^\alpha}\right) \qquad (5.58)$$

Using the series expansion $\ln(1+x) = \sum_{n=1}^{\infty} (-1)_i^{n+1} / n \cdot x^n$ we obtain

$$\ln p_k = \ln(\mu/m) - \alpha \ln k - \sum_{i=m}^{k} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (\mu/m)^n j^{-\alpha \cdot n} \qquad (5.59)$$

We approximate the sum over *j* with the integral

$$\sum_{k}^{j=m} j_X^{-n\alpha} \approx \int_k^m x^{-n\alpha} dx = \frac{1}{1-n\alpha}(k^{1-n\alpha} - m^{1-n\alpha}) \qquad (5.60)$$

which in the special case of *nα* =1 becomes

$$\sum_{j=m}^{k} j^{-1} \approx \int_m^k x^{-1} dx = \ln k - \ln m. \qquad (5.61)$$

Hence we obtain

$$\ln p_k = \ln(\mu / m) - \alpha \ln k - \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \frac{(\mu / m)^n}{1 - n\alpha} (k^{1-n.\alpha} - m^{1-n\alpha}) \tag{5.62}$$

Consequently the degree distribution has the form

$$p_k = C_\alpha k^{-\alpha} e^{-\sum_{\infty}^{n=1} \frac{(-1)^{n+1}}{n} \frac{(\mu/m)^n}{1-n\alpha} k^{1-n\alpha}}, \tag{5.63}$$

where

$$C_\alpha = \frac{\mu}{m} e^{\sum_{\infty}^{n=1} \frac{(-1)^{n+1}}{n} \frac{(\mu/m)^n}{1-n\alpha} m^{1-n\alpha}} \tag{5.64}$$

The vanishing terms in the exponential do not influence the k → ∞ asymptotic behavior, being relevant only if $1-n\alpha \geq 1$. Consequently the precise form of p_k depends on α as:

$$p_k \sim \{ \begin{matrix} k^{-\alpha} e^{\frac{-\mu/m}{1-\alpha} k^{1-\alpha}} , 1/2 < \alpha < 1 \\ k^{-\frac{1}{2} + \frac{1}{2}(\frac{\mu}{m})^2} e^{-\frac{1}{2}\frac{\mu}{m} k^{-2}} , \alpha = 1/2 \\ k^{-\alpha} e^{-\frac{\mu/m}{1-\alpha} k^{1-\alpha} + \frac{1}{2} \frac{(\mu/m)^2}{1-2\alpha} k^{1-2\alpha}} , 1/3 < \alpha < 1/2 \\ \vdots \end{matrix} \tag{5.65}$$

That is, for 1/2 < α < 1 the degree distribution follows a stretched exponential. As we lower α, new corrections start contributing each time α becomes smaller than *1/n*, where n is an integer. For α→1 the degree distribution scales as k^{-3}, as expected for the Barabási-Albert model. Indeed for α = 1 we have μ=2, and

$$\lim_{a \to 1} \frac{k^{1-a}}{1-a} = \ln k. \tag{5.66}$$

Therefore $p_k \sim k^{-1} \exp(-2\ln k) = k^{-3}$.

Finally we need to calculate $\mu(\alpha) = \sum^{j} j^\alpha p_j$. For this we sum Eq. 5.58:

$$\sum_{k=m}^{\infty} k^\alpha p_k = \sum_{k=m}^{\infty} \frac{\mu(\alpha)}{m} \prod_{k}^{j=m} \left(1 + \frac{\mu(\alpha)/m}{j^\alpha}\right)^{-1} \tag{5.67}$$

$$1 = \frac{1}{m} \sum_{k=m}^{\infty} \prod_{i=m}^{k} \left(1 + \frac{\mu(\alpha)/m}{j^\alpha}\right)^{-1} \tag{5.68}$$

We obtain μ(α) by solving Eq. 5.52 numerically.

ADVANCED TOPICS 5.C
THE CLUSTERING COEFFICIENT

The purpose of this section is to derive the average clustering coefficient, Eq. 5.30, for the Barabási-Albert model. The derivation follows the an argument proposed by Klemm and Eguiluz [36], that was supported by the exact calculation of Bollobás [37]. We aim to calculate the number of triangles expected in the model, as the number of triangles can be linked to the clustering coefficient SECT. 2.10. We denote the probability to have a link between node *i* and *j* with *P(i, j)*. Therefore, the probability that three nodes *i, j, l* form a triangle is *P(i, j) P(i, l) P(j, l)*. The expected number of triangles in which node *l* with degree k_l participates is thus given by the sum of the probabilities that node *l* participates in triangles with an arbitrary chosen node *i* and *j* in the network. This can be written as

$$Nr_l(\triangleleft) = \int_{i=1}^{N} di \int_{j=1}^{N} dj P(i,j)P(i,l)P(j,l) \qquad (5.69)$$

To proceed we need to calculate *P(i,j)*, which requires us to consider how the Barabási-Albert model evolves. Let us denote the time when node *j* arrived with $t_j = j$, which we can do as in each time step we added only one new node. Hence the probability that at its arrival node *j* links to node *i* with degree k_i is given by preferential attachment:

$$P(i,j) = m\Pi(k_i(j)) = m\frac{k_i(j)}{\sum\limits_{l=1}^{j} k_l} = m\frac{k_i(j)}{2mj} \cdot \qquad (5.70)$$

Using Eq. 5.7, we can write

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}} = m\left(\frac{j}{i}\right)^{\frac{1}{2}}, \qquad (5.71)$$

where we used the fact that the arrival time of node *j* is $t_j = j$ and the arrival time of node is $t_i = i$. Hence Eq. 5.70 now becomes

$$P(i,j) = \frac{m}{2}(ij)^{-\frac{1}{2}}. \qquad (5.72)$$

Using this result we can calculate the number of triangles in Eq. 5.62, writing

$$Nr_l(\lhd) = \int_{i=1}^{N} di \int_{j=1}^{N} dj P(i,j)P(i,l)P(j,l) \tag{5.73}$$

$$= \frac{m^3}{8} \int_{i=1}^{N} di \int_{j=1}^{N} dj (ij)^{-\frac{1}{2}} (il)^{-\frac{1}{2}} (jl)^{-\frac{1}{2}}$$

$$= \frac{m^3}{8l} \int_{i=1}^{N} \frac{di}{i} \int_{j=1}^{N} \frac{dj}{j} = \frac{m^3}{8l} (\ln N)^2 \tag{5.74}$$

The clustering coefficient can be written as $C = \frac{2Nr_l(\lhd)}{k_l(k_l-1)}$, hence we obtain

$$C = \frac{\frac{m^3}{4l}(\ln N)^2}{k_l(k_l-1)} \tag{5.75}$$

To simplify Eq. 5.74, we note that according to Eq. 5.7 we have

$$k_l(t) = m\left(\frac{N}{l}\right)^{\frac{1}{2}} \tag{5.76}$$

which is the degree of node *l* at time *t* = *N*. Hence, for large k_l we have

$$k_l(k_l-1) \approx k_l{}^2 = m^2 \frac{N}{l} \tag{5.77}$$

allowing us to write the clustering coefficient of the Barabási-Albert model as

$$C = \frac{2Nr_l(\lhd)}{k_l(k_l-1)} \tag{5.78}$$

Eq. 5.78, apart from a factor 2, is the result Eq. 5.30.

BIBLIOGRAPHY

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. Science, 286:509-512, 1999.

[2] F. Eggenberger and G. Pólya. Über die Statistik Verketteter Vorgänge. ZAMM - Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 3:279-289, 1923.

[3] G. Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, f.r.s. Philosophical Transactions of the Royal Society of London. Series B, 213:21-87, 1925.

[4] Gibrat R. "Les Inégalités économiques", Paris, France, 1931.

[5] G. K. Zipf. Human behavior and the principle of least reort. Addison-Wesley Press, Oxford, England, 1949.

[6] H. A. Simon. On a class of skew distribution functions. Biometrika, 42:425-440, 1955.

[7] D. De Solla Price. A general theory of bibliometric and othercumulative advantage processes. Journal of the American Society for Information Science, 27:292-306, 1976.

[8] R. K. Merton. The Matthew effect in science. Science, 159(3810):56-63, 1968.

[9] A.-L. Barabási. Linked: The new science of networks. Perseus, New York, 2002.

[10] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. Random Structures and Algorithms, 18:279-290, 2001.

[11] A.-L. Barabási, H. Jeong, R. Albert. Mean-field theory for scale free

random networks. Physica A, 272:173-187, 1999.

[12] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. Phys. Rev. Lett., 85:4633-4636, 2000.

[13] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. Phys. Rev. Lett., 85:4629-4632, 2000.

[14] H. Jeong, Z. Néda. A.-L. Barabási. Measuring preferential attachment in evolving networks. Europhysics Letters, 61:567-572, 2003.

[15] M. E. J. Newman. Clustering and preferential attachment in growing networks, Phys. Rev. E 64, 025102, 2001.

[16] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks. Oxford Clarendon Press, 2002.

[17] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: measurements, models and methods. Proceedings of the International Conference on Combinatorics and Computing, 1999.

[18] R. Kumar, P. Raghavan, S. Rajalopagan, D. Divakumar, A. S. Tomkins, and E. Upfal. The Web as a graph. Proceedings of the 19th Symposium on principles of database systems, 2000.

[19] Pastor-Satorras, R., Smith, E. & Sole, R. Evolving proteinminteraction networks through gene duplication. J. Theor. Biol. 222:m199–210, 2003.

[20] Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of protein interaction networks. ComPlexUs 1:38–44, 2003.

[21] G.S. Becker, The economic approach to Human Behavior. Chicago, 1976.

[22] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. Heuristically optimized trade-offs: a new paradigm for power laws in the internet. In Proceedings of the 29th International Colloquium on Automata, Languages, and Programming (ICALP), pages 110-122, Malaga, Spain, July 2002.

[23] RM. D'Souza, C. Borgs, J. T. Chayes, N. Berger, and R. D. Kleinberg, Emergence of tempered preferential attachment from optimization, PNAS 104, 6112-6117, 2007.

[24] F. Papadopoulos, M. Kitsak, M. Angeles Serrano, M. Boguna, and D. Krioukov, Popularity versus similarity in growing networks, Nature, 489: 537, 2012.

[25] A.-L. Barabási Network science: luck or reason, Nature 489: 1-2, 2012.

[26] H. A. Simon. On a class of skew distribution functions. Biometrika 42:3, 425-440, 1955.

[27] B. Mandelbrot. An Informational Theory of the Statistical Structure of Languages. In Communication Theory, edited by W. Jackson, pp. 486-502. Woburn, MA: Butterworth, 1953.

[28] B. Mandelbrot. A note on a class of skew distribution function: analysis and critique of a Paper by H.A. Simon. Information and control 2: 90-99, 1959.

[29] H. A. Simon. Some Further Notes on a class of skew distribution functions. Information and Control 3: 80-88, 1960.

[30] B. Mandelbrot. Final Note on a Class of Skew Distribution Functions: Analysis and Critique of a Model due to H.A. Simon. Information and Control 4: 198-216., 1961.

[31] H. A. Simon. Reply to final note. Information and Control 4:, 217-223, 1961.

[32] B. Mandelbrot. Post scriptum to final note. Information and Control 4: 300-304 1961.

[33] H. A. Simon. Reply to Dr. Mandelbrot's Post Scriptum. Information and Control 4: 305-308, 1961.

[34] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. Phys. Rev. Lett., 90:058701, 2003.

[35] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. Combinatorica, 24:5-34, 2004.

[36] K. Klemm and V. M. Eguluz. Growing scale-free networks with small-world behavior. Phys. Rev. E, 65:057102, 2002.

[37] B. Bollobás, O.M. Riordan. Mathematical results on scale-free random graphs, in the Handbook of Graphs and Networks, edited by S. Bormholdt and A. G. Schuster, Wiley, 2003.

CHAPTER 6

EVOLVING NETWORKS

Introduction

The Bianconi-Barabási model 1

Measuring fitness 2

Bose-Einstein condensation 3

Evolving Networks 4

Summary

Homework

ADVANCED TOPICS 6.A
Solving the fitness model

Bibliography

Figure 6.0 (front cover)
Network representation by Mauro Martino

INTRODUCTION

Founded six years after birth of the World Wide Web, Google was a latecomer to search. By the late 1990s Alta Vista and Inktomi, two search engines with an early start, have been dominating the search market. Yet Google, the third mover, soon not only became the leading search engine, but acquired links at such an incredible rate that by 2000 became the most connected node of the Web as well [1]. But its status didn't last: in 2011 Facebook, with an even later start, took over as the Web's biggest hub.

This competition for the top spot is by no means unique to the online world: the history of business is full of companies whose consumers were hijacked by a more successful latecomer. Take Apple, whose ingenious Newton handheld, introduced in 1987, was wiped off the market by Palm. A decade later Apple engineered a dramatic comeback, creating the iPad, that changed the concept of a handheld computer. If we view the market as a bipartite network whose nodes are products and whose links are purchasing decisions, we can say that Apple's links in the 1990s were rewired to Palm, only to be re-captured by Apple again a decade later. This competitive landscape highlights an important limitation of our current modeling framework: the network models we encountered so far cannot account for it. Indeed, in the Erdős-Rényi model the identity of the biggest node is driven entirely by chance. The Barabási-Albert model offers a more realistic picture, predicting that each node increases its degree following $k(t) \sim t^{1/2}$ Eq. 5.6. This means that the oldest node always has the most links, a phenomena called the "first mover's advantage" in the business literature. It also means that late nodes can never turn into the largest hubs.

In reality a node's growth does not depend on the node's age only. Instead webpages, companies, or actors have intrinsic qualities that influence the rate at which they acquire links. Some show up late and nevertheless grab most links within a short timeframe. Others rise early yet never quite make it. The goal of this chapter is to understand how the differences in the node's ability to acquire links, and other processes not captured by the Barabási-Albert model, like node and link deletion or aging, affect the network topology.

SECTION 6.1

THE BIANCONI-BARABÁSI MODEL

Some people have a knack for turning each random encounter into a lasting social link; some companies turn each consumer into a loyal partner; some webpages turn visitors into addicts. A common feature of these successful nodes is some intrinsic property that propels them ahead of the other nodes. We will call this property fitness. Fitness is an individual's skill to turn a random encounter into a lasting friendship; it is a company's competence in acquiring consumers relative to its competition; a webpage's ability to bring us back on a daily basis despite the many other pages that compete for our attention. Fitness may have genetic roots in people, it may be related to management quality and innovativeness in companies and may depend on the content offered by a website. In the Barabási-Albert model we assumed that a node's growth rate is determined solely by its degree. To incorporate the role of fitness we assume that preferential attachment is driven by the product of a node's fitness, η, and its degree k.

Movie 6.1

The evolution of the Bianconi-Barabási model

The movie shows a growing network in which each new node acquires a randomly chosen fitness parameter at birth, represented by the color of the node. Each new node chooses the nodes it links to following generalized preferential attachment, making each node's growth rate proportional to its fitness. The node size is shown proportionally to its degree, illustrating that with time the nodes with the highest fitness turn into the largest hubs.

→ 🎞

Video courtesy of D. Wang.

The resulting model consists of the following two steps [2, 3]:

- **Growth:** In each timestep a new node j with m links and fitness η_j is added to the system, where η_j is a random number chosen from a distribution $\rho(\eta)$. Once assigned, a node's fitness does not change.

- **Preferential Attachment:** The probability that a link of a new node connects to a pre-existing node i is proportional to the product of node i's degree k_i and its fitness η_i

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j}. \qquad (6.1)$$

In Eq. 6.1 the dependence of Π_i on k_i captures the fact that higher-degree nodes are easier to encounter, hence we are more likely to link to them. The dependence of Π_i on η_i implies that between two nodes with the same degree, the one with higher fitness is selected with a higher probability. Hence, Eq. 6.1 assures that even a relatively young node with initially only a few links can acquire links rapidly if it has larger fitness than the rest of

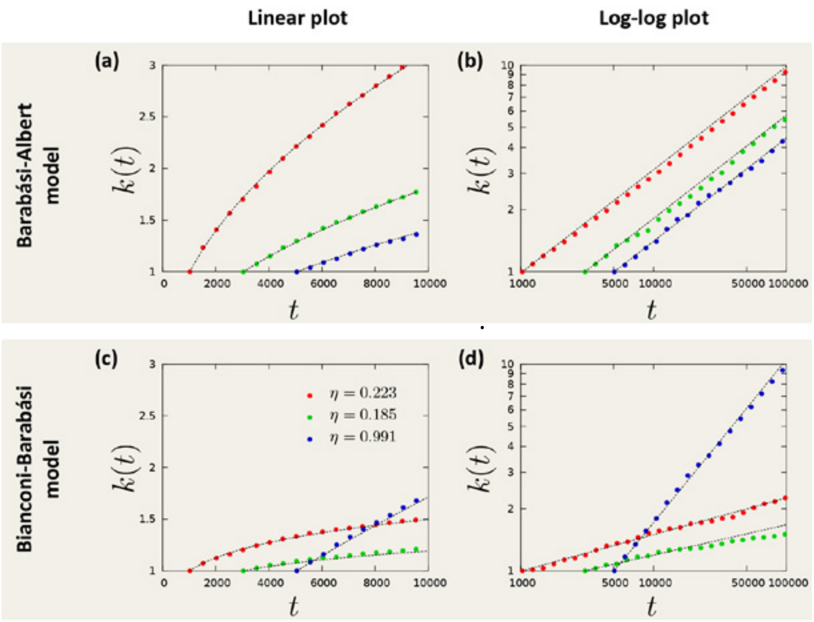


Figure 6.1
Competition in the Bianconi-Barabási model

(a) In the Barabási-Albert model all nodes increase their degree at the same rate, hence the earlier a node joins the network, the larger will be its degree. The figure shows the time dependence of the degree for nodes that arrived at different times, indicating that the later nodes are unable to pass the earlier nodes.

(b) Same as in (a) but in a log-log plot, demonstrating that each node follows the same growth law with identical dynamical exponents $\beta = 1/2$.

(c) In the Bianconi-Barabási model nodes increase their degree at a rate that is determined by their individual fitness. Hence a latecomer node (blue symbols) can overcome the earlier nodes.

(d) Same as in (c) but on a log-log plot, demonstrating that each node follows a growth curve with its own fitness-dependent dynamical exponent β, as predicted by **Eq. 6.3** and **Eq. 6.4**.

In (a)-(d) each curve corresponds to average over several independent runs using the same fitness sequence.

the nodes. We will call the model introduced above the *Bianconi-Barabási* model after the authors of the paper that introduced it [2, 3]. In the literature one may also account it as the *fitness model.*

DEGREE DYNAMICS

We can use the continuum theory to predict a node's temporal evolution in the model defined above. According to Eq. 6.1, the degree of node *i* changes at the rate

$$\frac{\partial k_i}{\partial t} = m \frac{\eta_i k_i}{\sum_k \eta_j k_j} \qquad (6.2)$$

Let us assume that the time evolution of k_i follows a power law with a fitness-dependent exponent $\beta(\eta_i)$ Fig. 6.1,

$$k_{\eta_i}(t, t_i) = m\left(\frac{t}{t_i}\right)^{\beta(\eta_i)}. \qquad (6.3)$$

Inserting Eq. 6.3 into Eq. 6.2 we find that the dynamic exponent satisfies ADVANCED TOPICS 6.A

$$\beta(\eta) = \frac{\eta}{C} \qquad (6.4)$$

with

$$C = \int \rho(\eta) \frac{\eta}{1-\beta(\eta)} d\eta. \qquad (6.5)$$

In the Barabási-Albert model we have $\beta = 1/2$, indicating that the degree of each node increases as a square root of time. In contrast, according to Eq. 6.4, in the Bianconi-Barabási model the dynamic exponent is proportional to the node's fitness, η, hence each node has its own dynamic exponent. Consequently, a node with a higher fitness will increase its degree faster. Given sufficient time, the fitter node will leave behind each node that has a smaller fitness BOX 6.1. Facebook is a poster child of this phenomenon: a latecomer with an addictive product, it acquired links faster than its competitors, eventually becoming the Web's biggest hub.

DEGREE DISTRIBUTION

The degree distribution of the network generated by the Bianconi-Barabási model can be calculated using the continuum theory ADVANCED TOPICS 6.A, predicting that

$$p_k \sim C\int d\eta \frac{\rho(\eta)}{\eta}\left(\frac{m}{k}\right)^{\frac{C}{\eta}+1}. \qquad (6.6)$$

Eq. 6.6 is a weighted sum of multiple power-laws, indicating that p_k depends on the precise form of the fitness distribution, *ρ(η)*. To illustrate the properties of the model we apply Eq. 6.4 and Eq. 6.6 to calculate *β(η)* and p_k for two different fitness distributions:

- Equal fitnesses

  When all fitnesses are equal, the Bianconi-Barabási model should reduce to the Barabási-Albert model. Indeed, let us use $\rho(\eta) = \delta(\eta - 1)$, capturing the fact that each node has the same fitness $\eta = 1$. In this case, Eq. 6.5 predicts $C = 2$. Using Eq. 6.4 we obtain $\beta = 1/2$ and Eq. 6.6 predicts $p_k \sim k^{-3}$, the known scaling of the degree distribution in the Barabási-Albert model.

- Uniform fitness distribution

  The model's behavior is more interesting when nodes have different fitnesses. Let us choose *η* to be uniformly distributed in the [0,1] interval. In this case *C* is the solution of the transcendental equation Eq. 6.5

$$\exp(-2/C) = 1 - 1/C \qquad (6.7)$$

whose numerical solution is $C^* = 1.255$. According to Eq. 6.4, each node *i* has a different dynamic exponent, $\beta(\eta_i) = \eta_i / C^*$. Using Eq. 6.6 we obtain

$$p_k \sim \int_1^0 d\eta \frac{C^*}{\eta}\frac{1}{k^{1+C^*/\eta}} \sim \frac{k^{-(1+C^*)}}{\ln k}, \qquad (6.8)$$

predicting that the degree distribution follows a power law with degree exponent γ = 2.255, affected by an inverse logarithmic correction 1/ln*k*.

Numerical support for these predictions is provided in Fig. 6.1 and Fig. 6.2. The simulations confirm that $k_i(t)$ follows a power law for each *η* and that the dynamical exponent *β(η)* increases with the fitness *η*. As Fig. 6.2 a indicates, the measured dynamical exponents are in excellent agreement with the prediction of Eq. 6.4. Fig. 6.2b also documents an agreement between Eq. 6.8 and the numerically obtained degree distribution.

In summary, the Bianconi-Barabási model can account for the different rate at which nodes with different internal characteristics acquire links. It predicts that a node's growth rate is directly determined by its fitness *η* and allows us to calculate the dependence of the degree distribution on the fitness distribution *ρ(η)*.

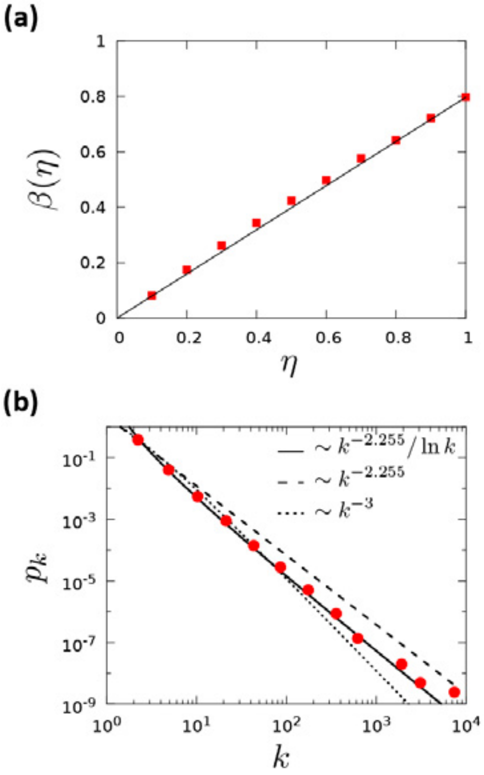


Figure 6.2
Characterizing the Bianconi-Barabási model

(a) The measured dynamic exponent *β(η)* shown in function of *η* in the case of a uniform *ρ(η)* distribution. The squares were obtained from numerical simulations while the solid line corresponds to the analytical prediction $\beta(\eta) = \eta/1.255$.

(b) Degree distribution of the fitness model obtained numerically for a network with *m*= and $N = 10^6$ and for fitnesses chosen uniformly from the $\eta \in [0, 1]$ interval. The solid line corresponds to the theoretical prediction Eq. 6.8 with $\gamma = 2.255$. The dashed line corresponds to a simple fit $p_k \sim k^{-2.255}$ without the logarithmic correction, while the long-dashed curve correspond to $p_k \sim k^{-3}$, expected if all fitness are equal. Note that the best fit is provided by Eq. 6.8.

SECTION 6.2

MEASURING FITNESS

Measuring the fitness of a node could help us identify web sites that are poised to grow in visibility, research papers that will become influential, or actors on their way to stardom. Yet, our ability to determine the utility of a webpage is prone to errors: while a small segment of the population might find a webpage on sumo wrestling fascinating, most individuals are indifferent to it and some might even find it repulsive. Hence, different individuals will inevitably assign different fitnesses to the same node. Yet, according to Eq. 6.1 fitness reflects the network's collective perception of a node's importance relative to the other nodes. Thus, we can determine a node's fitness by comparing its time evolution to the time evolution of other nodes in the network. In this section we show that if we have dynamical information about the evolution of the individual nodes, the conceptual framework of the Bianconi-Barabási model allows us to determine the fitness of each node.

To relate a node's growth rate to its fitness we take the logarithm of Eq. 6.3,

$$\log k_{n_i}(t,t_i) = \beta(n_i)\log t + \beta_i\ . \qquad (6.9)$$

where $B_i = \log(m/t_i^{\beta(\eta_i)})$ is a time-independent parameter. Hence, the slope of $\log k_{\eta_i}(t, t_i)$ is a linear function of the dynamical exponent $\beta(\eta_i)$, which depends linearly on η_i according to Eq.6.4. Therefore, if we can track the time evolution of the degree for a large number of nodes, the distribution of the obtained growth exponent $\beta(\eta_i)$ will be identical with the fitness distribution $\rho(\eta)$. Such measurement were first carried out in the context of the WWW, relying on a dataset that crawled the links of about 22 million web documents per month for 13 months [9]. While most nodes (documents) did not change their degree during this time frame, 6.5% of nodes showed sufficient changes to allow the determination of their growth exponent via Eq. 6.9. The obtained fitness distribution $\rho(\eta)$ has an exponential form Fig. 6.3, indicating that high fitness nodes are exponentially rare. This is somewhat unexpected, as one would be tempted to assume that on the web fitness varies widely: Google is probably significantly more interesting to Web

BOX 6.1

THE GENETIC ORIGINS OF FITNESS

Could fitness, an ability to acquire friends in a social network, have genetic origins? To answer this researchers examined the social network characteristics of 1,110 school-age twins [6, 7], using a technique previously developed to identify the heritability of a variety of traits and behaviors. The measurements indicate that:

- Genetic factors account for 46% of the variation in a student's in-degree (i.e. the number of students that name a given student as a friend).

- Generic factors are not significant for out-degrees (i.e. the number of students a given student names as friends).

This suggests that an individual's ability to acquire links, i.e. its fitness, is heritable. Hence, fitness may have genetic origins. This conclusion is also supported by research that associated a particular genetic variation with variation in popularity [8].

surfers than my personal webpage. Yet the exponential form of $\rho(\eta)$ indicates that most Web documents have comparable fitness. Consequently, the observed large differences in the degree of various web documents is the result of the system's dynamics: growth and preferential attachment amplifies the small fitness differences, turning nodes with slightly higher fitness into much bigger nodes. To illustrate this amplification, consider two nodes that arrived at the same time, but have different fitnesses $\eta_2 > \eta_1$. According to Eq. 6.3 and Eq. 6.4, the relative difference between their degrees grows with time as

$$\frac{k_2 - k_1}{k_1} \sim t^{\frac{\eta_2 - \eta_1}{C}} \,. \tag{6.10}$$

while the difference between η_2 and η_1 may be small, far into the future (large t) the relative difference between their degrees can become quite significant.

CASE STUDY: MEASURING THE FITNESS OF A SCIENTIFIC PUBLICATION

Eq. 6.9 assumes that Eq. 6.3 fully captures a Web document's temporal evolution. In some systems nodes follow a more complex dynamics, that we must account for when we try to measure their fitness. We illustrate this by determining the fitness of a research publication, allowing us to predict its future impact [11]. While most research papers acquire only a few citations, a small number of publications collect thousands and even tens of thousands of citations. These differences capture the considerable impact disparity characterizing the scientific enterprise BOX 6.3. These impact differences mirror differences in the novelty and the content of various publications. In general, we can write the probability that paper i is cited at time t after publication as [11]

$$\Pi_i \sim \eta_i c_i^t P_i(t)\,, \tag{6.11}$$

where η_i is the paper's fitness, accounting for the perceived noveltyand importance of the reported discovery; c_t^i is the cumulative number of citations acquired by paper i at time t after publication, accounting for the fact that well-cited papers are more likely to be cited again than less-cited contributions. The last term in Eq. 6.11 captures the fact that new ideas are integrated in subsequent work, hence the novelty of each paper fades with time [11, 12]. The measurements indicate that this decay has the log-normal form

$$P_i(t) = \frac{1}{\sqrt{2\pi}\sigma_i t} e^{-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}} \tag{6.12}$$

By solving the master equation behind Eq. 6.11, we obtain

$$c_i^t = m\left(e^{\frac{\beta\eta_i}{A}\Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)}\right), \tag{6.13}$$

where

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \tag{6.14}$$

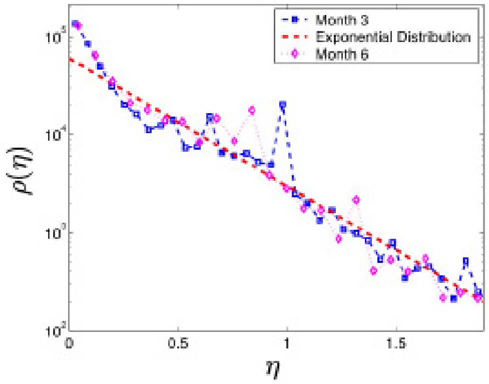


Figure 6.3

The fitness distribution of the WWW

The fitness distribution obtained from Eq. 6.9 by comparing the degree evolution of a large number of Web documents. The measurements indicate that each node's degree has a power law time dependence, as predicted by Eq. 6.3. The slope of each curve is $\beta(\eta_j)$, which corresponds to the node's fitness η_i up to a multiplicative constant according to Eq. 6.4. The plot shows the result of two measurements based on datasets recorded three months apart, demonstrating that the fitness distribution is time independent. The dashed line indicates that the fitness distribution follows an exponential form. After [9].

BOX 6.2

ULTIMATE IMPACT

Citation counts offer only the momentary impact of a research paper. Therefore, they represent an inherently weak measure of long-term impact, as we do not know if a paper that acquired a hundred citations in two years has already had its run, or will continue to grow in impact, acquiring thousands more. Ideally we would like to predict how many citations will a paper acquire during its lifetime, or its ultimate impact. The citation model Eq. 6.11 and Eq. 6.14 allows us to determine the ultimate impact by taking the t → ∞ limit in Eq. 6.13, finding [11]

$$c_i^\infty = m(e^{\eta'_i} - 1)\,. \tag{6.15}$$

is the cumulative normal distribution and m, β, and A are global parameters. Eq. 6.13 and Eq. 6.14 predict that the citation history of paper i is characterized by three fundamental parameters: the relative fitness $\eta_i' \equiv \eta_i \beta/A$, measuring a paper's importance relative to other papers; the immediacy μ_i, governing the time for a paper to reach its citation peak and the longevity σ_i, capturing the decay rate.

We fit Eq. 6.13 to the citation history of individual papers published by a given journal to obtain the journal's fitness distribution Fig. 6.4. We find that *Cell* has a fitness distribution shifted to the right, indicating that *Cell* papers tend to have high fitness. By comparison the fitness of papers published in *Physical Review* are shifted to the left, indicating that the journal publishes fewer high impact papers.

In summary, the framework offered by the Bianconi-Barabási model allows us to experimentally determine the fitness of individual nodes and the shape of the fitness distribution $\rho(\eta)$. The fitness distribution is typically bounded, meaning that differences in fitness between different nodes are small. With time these differences are magnified however, resulting in an unbounded (power law) degree distribution in incoming links in the case of the WWW or broad citation distribution in citation networks.

Eq. 6.15 predicts that despite the myriad of factors that contribute to the success and the citation history of a research paper, its ultimate impact is determined only by its fitness η'_i. As fitness can be determined by fitting Eq. 6.13 to a paper's existing citation history, we can use Eq. 6.15 to predict the ultimate impact of a publication.

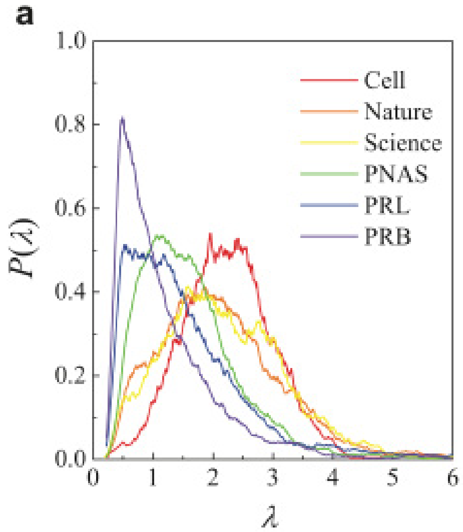


Figure 6.4
Fitness distribution of research papers

The fitness distribution of papers published in six journals in 1990. Each paper's fitness was obtained by fitting Eq. 6.13 to the paper's citation history for a decade long time interval following 1990. Two journals are from physics (Physical Review B and Physical Review Letters), one from biology (Cell) and three are interdisciplinary, meaning that they publish papers from different areas of science (Nature, Science, and PNAS).

The obtained fitness distributions are shifted relative to each other, indicating that Cell publishes papers with the highest fitness, followed by Nature and Science, PNAS, Physical Reviews Letters and Physical Review B. After [11].

BOX 6.3

THE TOP ONE PERCENT

The “one percent” phrase has dominated the discourse during the 2012 US presidential election, reminding everyone that one percent of the population earns a disproportional 17.42% of the total US income. To those familiar with power laws this is hardly surprising: it is a consequence of the fat-tailed nature of the income distribution. Therefore, the “one percent” phenomenon is present in any quantity that follows a power law, from the links of the WWW to scientific impact [10].

The “one percent” debate is not as much about the magnitude of the income disparity, but its trends: income disparity dropped between 1940 and 1970, only to skyrocket again in the past decades (see black line). As models explaining the distribution of income predict a time-invariant distribution, the observed changes offer evidence of endogenous shifts in the share of the top one percent. As the red line indicates, the impact disparity in physical sciences has also been rising steadily over the past century. Indeed, while in 1930 a year after publication the top 1% of papers got only about 5% of the citations, today the magnitude of this impact disparity is comparable to the income disparity.

This shift of the bulk of the citations to a few of publications may reflect the fact that while the number of research papers exploded, the time we devote to reading them has not. Hence, we increasingly rely on crowdsourcing to discover relevant work, a process that favors the highly cited publications.

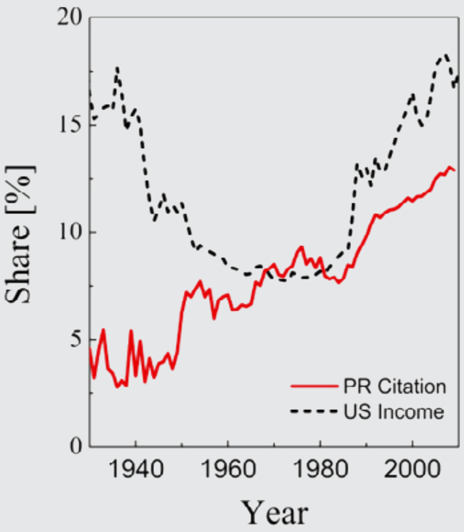


Figure 6.5
The 1% of Science

The share of citations in a given year received by the top 1% of all papers published during the previous year in Physical Review. The data captures the citation history of 463,348 papers published between 1893 and 2009 (red line). Also shown is the fraction of income earned by the top 1% of the population in US (black dashed line).

SECTION 6.3

BOSE-EINSTEIN CONDENSATION

In the previous section we found that the Web's fitness distribution follows a simple exponential Fig. 6.3, while the fitness of research papers follows a peaked distribution Fig. 6.4. The diversity of the observed fitness distributions raises an important question: how does the network topology depend on the precise shape of $\rho(\eta)$? Technically, the answer is provided by Eq. 6.6 that links p_k to $\rho(\eta)$. Yet, the true impact of the fitness distribution was realized only after the discovery that some networks can undergo a Bose-Einstein condensation BOX 6.5, with significant consequences on the network topology [13]. We start by establishing a formal link between the Bianconi-Barabási model and a Bose gas, whose properties have been extensively studied in physics Fig. 6.5:

- **Fitness → Energy**: to each node with fitness η_i we assign an energy ε_i using

$$\varepsilon_i = \frac{1}{\beta_T}\log\eta_i\,. \tag{6.16}$$

In physical systems β_T plays the role of the inverse temperature. Hence, we use the subscript T to distinguish β_T from the dynamic exponent β. According to Eq. 6.16, each node in a network corresponds to an energy level in a Bose gas. The larger the node's fitness, the lower is its energy.

- **Links → Particles:** for each link between nodes i and j we add a particle at the energy levels ε_i and ε_j, respectively.

- **Nodes → Energy levels:** the arrival of a new node with m links corresponds to adding a new energy level ε_j and $2m$ new particles to the Bose gas. Half of these particles land on level ε_j, corresponding to the links that start from the node j; the remaining m particles are distributed between the energy levels that correspond to the nodes to which the new node links to.

If we follow the mathematical consequences of this mapping, we find that in the resulting gas the number of particles on each energy level fol-

Movie 6.2
Bose-Einstein condensation in networks

The movie shows the time evolution of a growing network in which one node (purple) has a much higher fitness than the rest of the nodes. Consequently this high fitness node attracts most links, forcing the system to undergo a Bose-Einstein condensation.

→ 🎞

Video courtesy of D. Wang.

lows a Bose statistics, a formula derived by Satyendra Nath Bose in 1924, representing a fundamental result in quantum statistics BOX 6.6. Consequently, the links of the fitness model behave like subatomic particles in a quantum gas. This mapping to a Bose gas is exact and predicts the existence of two distinct phases [13, 14]:

SCALE-FREE PHASE

For most fitness distributions the network displays a fit-gets-rich dynamics, meaning that the degree of each node is ultimately determined by its fitness. While the fittest node will inevitably become the largest hub, in the scale-free phase the fittest node is not significantly bigger than the next fittest node.

BOX 6.4

BOSE-EINSTEIN CONDENSATION

In classical physics atoms can be distinguished and individually numbered, like the numbered balls used to pick the winning number in lottery. In the subatomic world particles differ in our ability to distinguish them: Fermi particles, like electrons, can be distinguished; in contrast Bose particles, like photons, are indistinguishable. Distinguishability impacts the energy of a particle. In classical physics the kinetic energy of a moving particle, $E= mv^2 /2$, can have any value between zero (at rest) and an arbitrarily large E, when it moves very fast. In quantum mechanics energy is quantized, which means that it can only take up discrete (quantized) values. This is where distinguishability matters: the distinguishable Fermi particles are forbidden to have the same energy. Hence, only one electron can occupy a given energy level Fig. 6.7a. As Bose particles cannot be distinguished, many can crowd on the same energy level Fig. 6.7b.

At high temperatures, when thermal agitation forces the particles to take up different energies, the difference between a Fermi and a Bose gas is negligible Fig. 6.7a, b. The difference becomes significant at low temperatures when all particles are forced to take up their lowest allowed energy. In a Fermi gas at low temperatures the particles fill the energy levels from bottom up, just like pouring water fills up a vase Fig. 6.7c. However, as any number of Bose particles can share the same energy level, they can all crowd at the lowest energy Fig. 6.7d. Hence, no matter how much "Bose water" we pour into the vase, it will stay at the bottom of the vessel, never filling it up. This phenomenon is called a Bose-Einstein condensation and it was first proposed by Einstein in 1924. Experimental evidence for Bose-Einstein condensation emerged only in 1995 and was recognized with the 2001 Nobel prize in physics.

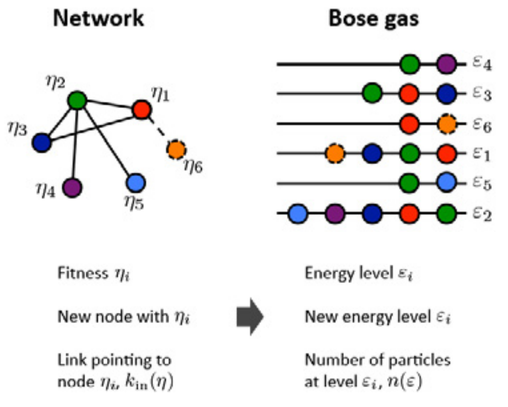


Figure 6.6

Mapping networks to a Bose gas

Left: A network of six nodes, each node characterized by a unique fitness, η_i, indicated by the color of the node. The individual fitnesses are chosen from the distribution $\rho(\eta)$.

Right: The mapping assigns an energy level ε to each fitness η, resulting in a Bose gas with random energy levels. A link from node i to node j corresponds to two particles, one at level ε_i and the other at level ε_j.

Growth: The network grows by adding a new node, like the node with fitness η_6, at each time step. The new node connects to m=1 other nodes (dashed link), chosen randomly following Eq. 6.1. In the Bose gas this results in the addition of a new energy level ε_6 (dashed line), populated by two particles, and the deposition of another particle at ε_1, the energy level to which η_6 connects to.

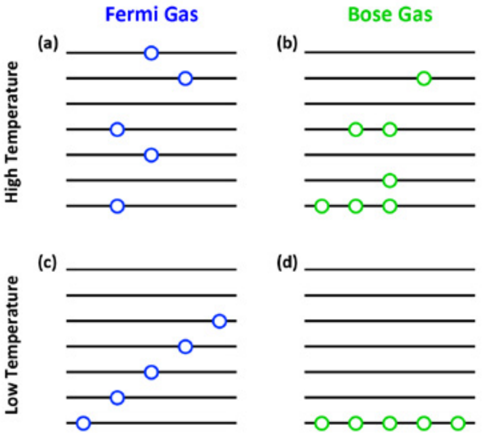


Figure 6.7

Bose and Fermi statistics

In a Fermi gas (a, c) only one particle is allowed on each energy level, while in a Bose gas (b,d) there is no such a restriction. At high temperatures it is hard, to notice the difference between the two gases. At low temperatures, however, particles want to occupy the lowest possible energy and the difference between the two gasses becomes significant.

Indeed, at any moment the degree distribution follows a power law, indicating that the largest hub is closely followed by a few slightly smaller hubs, with almost as many links as the fittest node Fig. 6.8a. The uniform fitness distribution discussed in the previous section results in a scale-free network.

BOSE-EINSTEIN CONDENSATION

The unexpected outcome of the mapping to a Bose gas is the possibility of a Bose-Einstein condensation for some fitness distributions *ρ(η)* BOX 6.7. In a Bose-Einstein condensate all particles crowd to the lowest energy level, leaving the rest of the energy levels unpopulated BOX 6.5. In a network this means that the fittest node grabs a finite fraction of the links, turning into a super-hub Fig. 6.8b, and the network develops a hub-and-spoke topology. In these networks the rich-gets-richer process is so dominant that becomes a winner takes-all phenomenon. Consequently, the network will loose its scale-free nature.

In summary, the precise shape of the fitness distribution, *ρ(η)*, plays an important role in shaping the topology of a growing network. While most fitness distributions (like the uniform distribution) lead to a power law degree distribution, some *ρ(η)* allow for Bose-Einstein condensation. If a network undergoes a Bose-Einstein condensation, then one or a few nodes grab most of the links. Hence, the rich-gets-richer process that generates the scale-free state, turns into a winner-takes-all phenomenon. The Bose-Einstein condensation has such an obvious impact on a network's structure that, if present, it is hard to miss: it destroys the hierarchy of hubs characterizing a scale-free network, turning it into a star-like topology BOX 6.8.

BOX 6.5

FROM FITNESS TO A BOSE GAS

In the context of the Bose gas of Fig. 6.6 the probability that a particle lands on level *i* is given by

$$\Pi_i = \frac{e^{-\beta_T \varepsilon_i} k_i}{\sum_j e^{-\beta_T \varepsilon_j} k_i} \,. \qquad (6.17)$$

Hence, the rate at which the energy level ε_i accumulates particles is [13]

$$\frac{\partial k_i(\varepsilon_i, t, t_i)}{\partial t} = m \frac{e^{-\beta_T \varepsilon_i} k_i(\varepsilon_i, t, t_i)}{Z_t} \qquad (6.18)$$

where $k_i(\varepsilon_i, t, t_j)$ is the occupation number of level *i* Fig. 6.6 and

$$Z_t \equiv \sum_{j=1}^{t} t e^{-\beta_T \varepsilon_j} k_j(\varepsilon_i, t, t_j) \qquad (6.18a)$$

is the partition function. The solution of Eq. 6.18 is

$$k_i(\varepsilon_i, t, t_i) = m \left(\frac{t}{t_i} \right)^{f(\varepsilon_i)} \qquad (6.19)$$

where $f(\varepsilon) = e^{-\beta_T(\varepsilon - \mu)}$ and μ is the chemical potential satisfying

$$\int \deg(\varepsilon) \frac{1}{e^{\beta_T(\varepsilon - \mu)} - 1} = 1 \,. \qquad (6.20)$$

Here, deg(ε) is the degeneracy of the energy level ε. Eq. 6.20 suggests that in the limit t → ∞ the occupation number, representing the number of particles with energy ε, follows the well-known Bose statistics

$$n(\varepsilon) \frac{1}{e^{\beta_T(\varepsilon - \mu)} - 1} \,. \qquad (6.21)$$

This concludes the mapping of the fitness model to a Bose gas, indicating that the node degrees in the fitness model follow Bose statistics.

BOX 6.6

FITNESSES DISTRIBUTION LEADING TO BOSE-EINSTEIN CONDENSATION

In physical systems Bose-Einstein condensation is induced by lowering the temperature of the Bose gas below some critical temperature. In networks, the temperature β_T in Eq. 6.16 is a dummy variable, disappearing from all topologically relevant quantities, like the degree distribution p_k. Hence, the presence or absence of Bose-Einstein condensation depends only on the form of the fitness distribution $\rho(\eta)$. In order for a network to undergo Bose-Einstein condensation, the fitness distribution needs to satisfy the following conditions:

(a) $\rho(\eta)$ must have a maximum $\eta_{\max}$.This means that η needs to have a clear upper bound.

(b) $\rho(\eta_{max})$=0, i.e. the system requires an infinite time to reach $\eta_{\max}$.

The uniform distribution, $\eta \in [0, 1]$ satisfies (a), as it is bounded, having $\eta_{\max}$=1. It fails, however, the criteria (b), as it can reach $\eta_{\max}$= 1 with a finite probability. Consequently, we cannot observe a Bose-Einstein condensation in this case. A fitness distribution that can lead to a Bose-Einstein condensation is

$$\rho(\eta) = (1-\eta)^{\zeta} \qquad (6.22)$$

satisfying both (a) and (b). Indeed, $\eta_{\max} = 1$ and $\eta(1) = 0$, which is the reason why, upon varying ζ, we can observe Bose-Einstein condensation Fig. 6.8. Indeed, the existence of the solution of Eq. 6.20 depends on the functional form of the energy distribution, $g(\varepsilon)$, determined by the $\rho(\eta)$ fitness distribution. Specifically, if Eq. 6.22 has no non-negative solution for a given $g(\varepsilon)$, we observe a Bose-Einstein condensation, indicating that a finite fraction of the particles agglomerate at the lowest energy level.

BOX 6.7

MICROSOFT AND BOSE-EINSTEIN CONDENSATION

Think of the operating systems (OS) that run on each computer as nodes that compete for links in terms of users or computers. Each time a user installs Windows on his or her computer, a link is added to Microsoft. If a fit-gets-rich behavior of scale-free networks prevails in the marketplace, there should be a hierarchy of operating systems, such that the most popular node is followed closely by several less popular nodes.

In the OS market, however, such hierarchy is absent. True, Windows is not the only available operating system. All Apple products run Mac OS; DOS, the precursor of Windows, is still installed on some PCs; Linux, a free operating system, continues to gain market share and UNIX runs on many computers devoted to number crunching.

But all these operating systems are dwarfed by Windows, as in 2010 its different versions were humming on a whopping 86 percent of all personal computers. The second most popular operating system had only a 5 percent market share. Hence the OS market carries the signatures of a network that has undergone Bose-Einstein condensation [1].

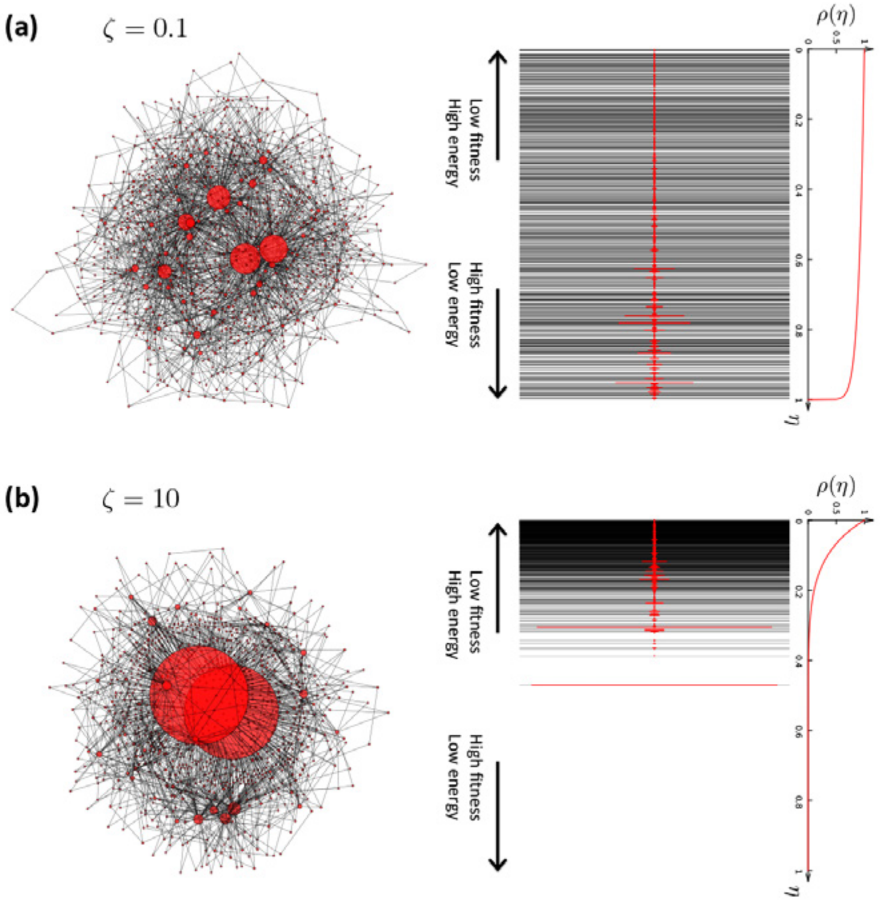


Figure 6.8

Bose Einstein Condensation in Networks

Left panels: (a) A scale-free network and (b) a network that has undergone a Bose-Einstein condensation generated by the fitness model with *ρ(η)* following **Eq. 6.22**.

Middle panels: the energy levels (black lines) and the deposited particles (red dots) for a network with *m*=2 and *N*=1,000. Each energy level corresponds to the fitness of a node on the network shown in the left. Each link connected to a node is represented by a particle on the corresponding energy level.

Right panels: the fitness distribution *ρ(η)*, given by **Eq. 6.22**, illustrating the difference in the shape of the two *ρ(η)* functions. The difference is determined by the parameter *ζ*.

SECTION 6.4

EVOLVING NETWORKS

The Barabási-Albert model is a minimal model, its main purpose being to capture the core mechanisms responsible for the emergence of the scale-free property. Consequently, it has several well-known limitations:

(i) It predicts $\gamma = 3$ while the experimentally observed degree exponents vary between 2 and 4 Table 4.1.

(ii) It predicts a pure power-law degree distribution, while real systems are characterized by various deviations from a power law, like small-degree saturation or high-degree cutoff BOX 4.18.

(iii) It ignores a number of elementary processes that are obviously present in many real networks, like the addition of internal links and node or link removal.

These limitations have inspired considerable research in the network science community, clarifying how various elementary processes influence the network topology. The purpose of this section is to systematically extend the Barabási-Albert model to capture the wide range of phenomena shaping the structure of real networks.

INITIAL ATTRACTIVENESS

In the Barabási-Albert model an isolated node cannot acquire links, as according to preferential attachment Eq. 4.1 the likelihood that a new node attaches to a k=0 node is strictly zero. In real networks, however, even isolated nodes acquire links. Indeed, each new research paper has a finite probability of being cited or a person that moves to a new city will quickly acquire acquaintances. In growing networks zero-degree nodes can acquire links if we add a constant to the preferential attachment function Eq. 4.1, obtaining

$$\Pi(k) \sim A + k \,. \qquad (6.23)$$

In Eq. 6.23 the parameter A is called initial attractiveness. As $\Pi(0) \sim A$, initial attractiveness represents the probability that a node will acquire its first link. We can detect the presence of initial attractiveness in real networks by measuring *Π(k)* Fig 6.9. Once present, initial attractiveness has two immediate consequences:

- **Increases the degree exponent:** If in the Barabási-Albert model we place Eq. 4.1 with Eq. 6.23, the degree exponent becomes [15, 16]

$$\gamma = 3 + \frac{A}{m} \cdot \tag{6.24}$$

  By increasing *γ*, initial attractiveness makes a network more homogeneous and reduces the size of the hubs. Indeed, initial attractiveness adds a random component to the probability of attaching to a node. This random component favors the numerous small-degree nodes, weakening the role of preferential attachment. For high-degree nodes the *A* term in Eq. 6.23 is negligible.

- **Generates a small-degree cutoff**: The solution of the continuum equation indicates that the degree distribution of a network governed by Eq. 6.23 does not follow a pure power-law, but has the form

$$p_k = C(k + A\ )^{-\gamma}. \tag{6.25}$$

Therefore, initial attractiveness induces a small-degree saturation at *k<A*. This saturation is again rooted in the fact that initial attractiveness enhances the probability that new nodes link to the small-degree nodes, which decreases the number of nodes with small *k*. For high degrees ($k \gg A$), where initial attractiveness loses its relevance, the degree distribution continues to follow a power law.

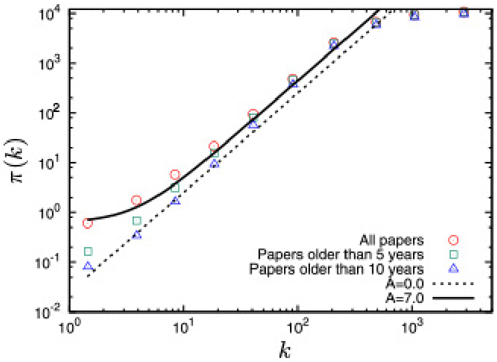


Figure 6.9
Initial Attractiveness

Cumulative preferential attachment function

$$\pi(k) = \sum_{k' \le k} \pi(k')$$

for the citation network, capturing the citation patterns of research papers published from 2007 to 2008. The *π(k)* curve was measured using the methodology described in SECTION 5.7. The continuous line corresponds to *C(k+A)−γ* with initial attractiveness *A* ~ 7.0. The dashed line corresponds to *A* = 0, i.e. the case without attractiveness. After [17].

INTERNAL LINKS

In many networks most new links are added between pre-existing nodes. For example, the vast majority of new links on the WWW are internal links, corresponding to newly added URLs between existing web documents. Similarly, virtually all new social/friendship links form between individuals that already have other acquaintances and friends. Measurements show that in collaboration networks internal links follow double preferential attachment, i.e. the probability for a new internal link to connect two nodes with degree *k* and *k'* is [18]

$$\Pi(k, k') \sim (A + Bk)(A' + B'k'). \tag{6.26}$$

We explore several limiting cases to understand the impact of internal links:

- **Double preferential attachment (A=A′=0):** In this case both ends of a new link are chosen proportional to the degree of the nodes they connect. Consider an extension of the Barabási-Albert model, where in each time step we add a new node with m links, followed by n internal links, each selected with probability Eq. 6.26 with *A*=*A′*=0. The degree exponent of the resulting network is [19, 20]

$$\gamma = 2 + \frac{m}{m+2n}, \qquad (6.27)$$

indicating that γ varies between 2 and 3. This means that double preferential attachment lowers the degree exponent from 3 to 2, hence increasing the network's heterogeneity. Indeed, by preferentially connecting the hubs to each other, it simultaneously helps both hubs to grow faster than they do in the Barabási-Albert model.

- **Random attachment (B=B′=0):** In this case the internal links are blind to the degree of the nodes they connect, implying that they are added between randomly chosen node pairs. Let us again consider the Barabási-Albert model, where after each new node we add n randomly selected links. In this case the degree exponent becomes [20]

$$\gamma = 3 + \frac{2n}{m}. \qquad (6.28)$$

  Hence, $\gamma \geq 3$ for any n, indicating that the resulting network will be more homogenous than the network generated by the Barabási-Albert model. Indeed, randomly added internal links mimic the process observed in random networks, making the node degrees more similar to each other.

NODE DELETION

In many real systems nodes and links systematically disappear, leading to node or link deletion. For example, nodes are deleted from an organizational network when employees leave the company or from the WWW when web documents are removed. At the same time in some networks node removal is virtually impossible Fig. 6.10.

To explore the impact of node removal, let us start again from the Barabási-Albert model. In each time step we add a new node with m links and with probability r we remove a node. The observed topologies depend on the value of r [23, 24, 25, 26, 27, 28]:

- **Scale-free phase:** For $r < 1$ the number of removed nodes is smaller than the number of new nodes, hence the network continues to grow. In this case the degree exponent has the value

$$\gamma = 3 + \frac{2r}{1-r}. \qquad (6.29)$$

  Hence, random node removal increases γ, homogenizing the network.

- **Exponential phase:** For r=1 the network has fixed size, as nodes arrive and are removed at the same rate (i.e. N=constant). In this case the network will loose its scale-free nature. Indeed, for $r \to 1$ we have $\gamma \to \infty$ in Eq. 6.29.

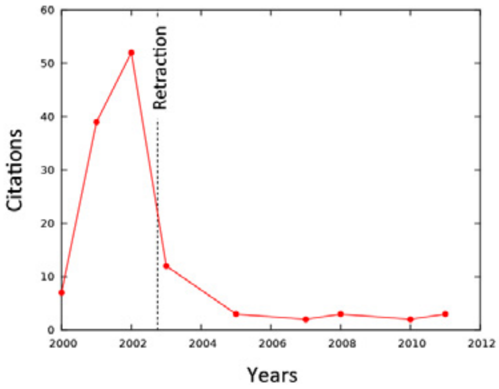


Figure 6.10
The impossibility of node removal

The citation history of a research paper by Jan Hendrik Schön published in *Science* [21]. Schön rose to prominence after a series of apparent breakthroughs in the area of semico ductors. Schön's findings were published by prominent scientific journals, like *Science* and *Nature*. His productivity was phenomenal: in 2001 he has coauthored one research paper every eight days. However, research groups around the world had difficulty reproducing some of his results.

Soon after Schön published a paper reporting a groundbreaking discovery on single-molecule semiconductors, researchers noticed that two experiments carried out at very different temperatures had identical noise [22].

The ensuing questions prompted Lucent Technologies, which ran the storied Bell Labs where Schön worked, to start a formal investigation. Eventually, Schön admitted falsifying some data to show more convincing evidence for the behavior that he observed. Several dozens of his papers, like the one whose citation pattern is shown in this figure, were retracted and the *University of Konstanz* revoked his PhD degree for "dishonorable conduct."

While the papers' retraction lead to a dramatic drop in citations, his papers continue to be cited even after their official "removal" from the literature, as shown in the figure above. This indicates that in the citation network it is virtually impossible to remove a node.

- **Declining networks:** For $r > 1$ the number of removed nodes exceeds the number of new nodes, hence the network declines BOX 6.10. Declining networks are important in several areas: Alzheimer's research focuses on the progressive loss of neurons with age and ecology focuses on the role of gradual habitat loss [29, 30, 31]. A classical example of a declining network is the telegraph, that dominated long distance communication in the second part of the 19th century and early 20th century. It was once a growing network: in the United States the length of the telegraph lines grew from 40 miles in 1846 to 23,000 in 1852. Yet, following the second World War, the telegraph gradually disappeared.

Note that node removal is not always random but can depend on the removed node's degree BOX 6.9. Furthermore, the behavior of a network can be rather complex if additional elementary processes are considered, inducing phase transitions between scale-free and exponential networks **Box 6.10**.

In summary, in most networks some nodes can disappear. Yet as long as the network continues to grow, its scale-free nature can persist. The degree exponent depends, however, on the detail governing the node removal process.

ACCELERATED GROWTH

In the models discussed so far the number of links increases linearly with the number of nodes. In other words, we assumed that $L=\langle k\rangle N$, where $\langle k\rangle$ is independent of time. This is a reasonable assumption for many real networks. Yet, some real networks experience accelerated growth, meaning that the number of links grows faster than N, hence $\langle k\rangle$ increases. For example the average degree of the Internet increased from $\langle k\rangle$=3.42 in November 1997 to 3.96 by December 1998 [32]; the WWW increased its average degree from 7.22 to 7.86 during a five month interval [33, 34]; in metabolic networks the average degree of the metabolites grows approximately linearly with the number of metabolites [35]. To explore the consequence of such accelerated growth let us assume that in a growing network the number of links arriving with each new node follows [36, 37, 38, 39]

$$m(t) = m_0 t^{\theta}. \tag{6.30}$$

For θ=0 each node has the same number of links; for θ>0, however, the network follows accelerated growth. The degree exponent of the Barabási-Albert model with accelerated growth Eq. 6.30 is

$$\gamma = 3 + \frac{2\theta}{1-\theta} \cdot \tag{6.31}$$

Hence, accelerated growth increases the degree exponent beyond γ=3, making the network more homogenous. For θ=1 the degree exponent diverges, leading to hyper-accelerating growth [37]. In this case $\langle k \rangle$ grows linearly with time and the network looses its scale-free nature.

AGING

In many real systems nodes have a limited lifetime. For example, actors have a finite professional life span, capturing the period when they still act in movies. So do scientists, whose professional lifespan corresponds to the time frame they continue to publish scientific papers. These nodes do not disappear abruptly, but fade away through a slow aging process, gradually reducing the rate at which they acquire new links [40, 41, 42, 43]. Capacity limitations can induce a similar phenomena: if nodes have finite resources to handle links, once they approach their limit, they will stop accepting new links [41].

To understand the impact of aging let us assume that the probability that a new node connects to node i is $\Pi(k, t-t_i)$, where ti is the time node i was added to the network. Hence, $t-t_i$ is the node's age. In analytical calculations slow aging is often modeled by choosing [40]

$$\Pi(k,t-t_i) \sim k(t-t_i)^{-\nu}, \qquad (6.32)$$

where ν is a tunable parameter governing the dependence of the attachment probability on the node's age. Depending on the value of ν we can distinguish three scaling regimes:

- **For negative ν** the older is node i, the more likely that a new node will link to it. Hence, $\nu < 0$ enhances the role of preferential attachment. In the extreme case $\nu \rightarrow -\infty$, each new node will only connect to the oldest node, resulting in a hub-and-spoke topology Fig. 6.11a. The calculations show that the scale-free state persists in this regime, but the degree exponent drops under 3. Hence, $\nu < 0$ makes the network more heterogeneous

- **A positive ν** will encourage the new nodes to attach to younger nodes. In the extreme case $\nu \rightarrow \infty$ each node will connect to its immediate predecessor Fig. 6.11a. We do not need a very large ν to experience the impact on aging: the degree exponent diverges as we approach ν=1. Hence gradual aging homogenizes the network by shadowing the older hubs.

- **For $\nu > 1$** the aging effect overcomes the role of preferential attachment, leading to the loss of the scale-free property.

In summary, the results discussed in this section indicate that a wide range of elementary processes can affect the structure of a growing network Table 6.1. These results highlight the true power of the evolving network paradigm: it allows us to address, using a mathematically predictive framework, the impact of a wide range of elementary processes on the network topology and evolution.

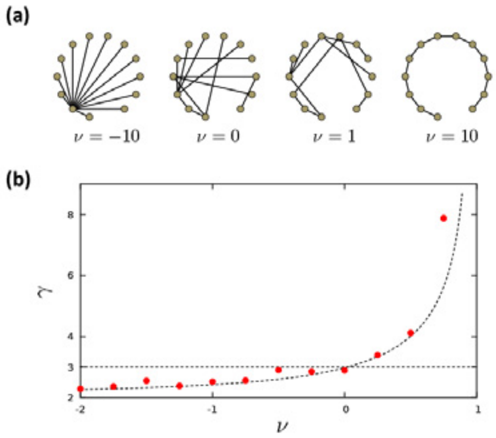


Figure 6.11
The impact of aging

(a) A schematic illustration of the expected network topologies for various aging exponents ν. In the context of a growing network model we assume that the probability to attach to a node is proportional to $k\tau^{-\nu}$, where τ is the age of the node. For negative ν nodes prefer the oldest nodes, turning the network into a hub-and-spoke topology. For positive ν the most recent nodes are the most attractive. Hence for large ν the network turns into a chain, as the last (hence the youngest) node is the most attractive for the new node. The network is shown for m=1 for clarity but note that the degree exponent is independent of m.

(b) The degree exponent γ vs the aging exponent ν, as predicted by the analytical solution of the rate equation. The red symbols are the result of simulations, each representing a single network of N=10,000 and m=1. The degree exponent is estimated using the method described in CHAPTER 4. Redrawn after Ref. [40].

BOX 6.8

DECLINING NETWORKS AND FASHION

The properties of declining networks is well illustrated by the New York City garment industry, whose nodes are designers and contractors that are connected to each other by the annual coproduction of lines of clothing. As the industry decayed, the network has persistently shrunk. This is illustrated by the fate of the largest connected component, that collapsed from 3,249 nodes in 1985 to 190 nodes in 2003. Interestingly, the network's degree distribution remained relatively unchanged during this period. The analysis of the network's evolution allowed researchers to uncover several interesting properties of declining networks [23]:

- **Preferential Attachment:** While overall the network was shrinking with time, new nodes continued to arrive. The measurements indicate that the attachment probability of these new nodes follows $\Pi(k)$ ~k^{α} with α=1.20 ± 0.06 **PANEL A**, offering evidence of superlinear preferential attachment.

- **Link deletion:** The measurements also show that the probability that a firm looses a link decreased proportionally with the firms' degree, as *k(t)–η* with η=0.41±0.04. This documents a weak-gets-weaker phenomenon where the less connected firms are more likely to loose their links.

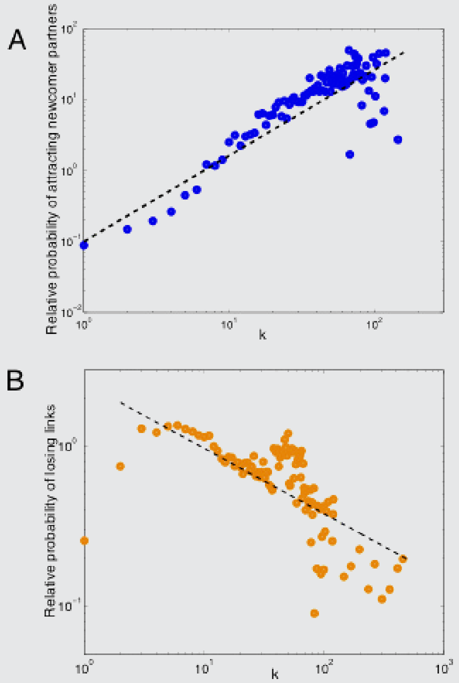


Figure 6.12
The decline of the garment industry

(a) Preferential attachment. The relative probability $\Pi(k)$ that a newcomer firm added at time t connects to an incumbent firm with k links. The dashed line has slope α=1.2.

(b) Link deletion. The relative probability, $R_{k(t)}$ of deleting a link from a degree node, compared with random link removal. The dashed line has slope η=0.41.

If link addition and removal were to be random, we would expect $\Pi(k)$ ~ 1 and $R_{k(t)}$ ~ 1 for all k. After [23].

Figure 6.13
Garment district

The Garment District is a Manhattan neighborhood located between Fifth Avenue and Ninth Avenue, from 34th to 42nd Street. Since the early 20th century it has been the center for fashion manufacturing and design in the United States. The Needle threading a button sculpture and a sculpture of a tailor, located in the heart of the district, pay tribute to the neighborhood's past and present.

PROCESS	PROCESS	γ	OBSERVATIONS
Preferential attachment	$\Pi(k) \sim k$	3	
Initial attractiveness	$\Pi(k) = A + k$	$3 + \frac{A}{m}$	Small degree cutoff $p_k \sim (k + A)^{-\gamma}$
Internal links	$\Pi(k, k') =$ $(A + Bk)(A' + B'k')$	$2 + \frac{m}{m + 2n}$	Double preferential attachment $A = A' = 0 , B \quad 0, B' \quad 0$
		$3 + \frac{2n}{m}$	Random internal attachment $B = B' = 0$
Node (link) deletionN	ode removal rate r	$3 + \frac{2r}{1 - r}$	Scale-free for $r < r^*$ Stretched exponential for $r = r^*$ Exponential for $r > r^*$
Accelerated growth	$m(t) = t^{\Theta}$	$\frac{3 - \theta}{1 - \theta}$	For $\theta = 1$ we have hyper-accelerated growth and the scale-free state disappears.
Aging	$\Pi(k) \sim (t - t_i)^{-\nu}$	See Figure 6.11	For $\nu > 1$ the network looses its scale-free topology.

Table 6.1 Elementary processes

A summary of the various elementary processes discussed in this section and their impact on the degree distribution.

BOX 6.9

NODE REMOVAL INDUCED PHASE TRANSITIONS

The coexistence of node removal with other elementary processes can lead to interesting topological phase transitions. This is illustrated by a simple model in which the network's growth is governed by Eq. 6.23, i.e. preferential attachment with initial attractiveness, and we also remove nodes with rate *r*. The network displays three distinct phases, captured by the phase diagram shown below:

Subcritical node removal ($r < r^*(A)$): If the rate of node removal is under a critical value $r^*(A)$, the network will be scale-free.

Critical node removal ($r = r^*(A)$): Once r reaches a critical value $r^*(A)$, the degree distribution turns into a stretched exponential SECT. 4.A.

Exponential networks ($r > r^*(A)$): In this regime the network looses its scale-free nature, developing an exponential degree distribution.

Therefore, the coexistence of multiple elementary processes in a network can lead to discontinuous changes in the network topology. To be specific, a continuous increase in the node removal rate leads to a transition from a scale-free to an exponential network.

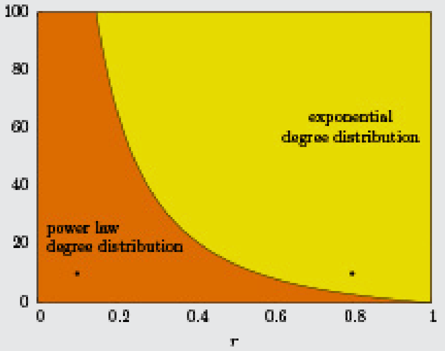


Figure 6.14
Scaling under node deletion

The degree distribution of a network whose growth is driven by preferential attachment with initial attractiveness A and node removal rate *r*. After [28].

SUMMARY

As we illustrated in this chapter, rather diverse processes, from node fitness to internal links or aging, can influence the topology of real networks. By exploring these processes, we came to see how to use the evolving network modeling framework to accurately predict the impact of various frequently encountered elementary events on a network's topology and evolution. The most important conclusion of the examples discussed in this chapter is that *if we want to understand the structure of a network we must first get its dynamics right. The topology is the bonus of this approach.*

In CHAPTER 4 we documented the difficulties we encounter when we attempt to fit a pure power law to the degree distribution of real networks. The roots of this problem were revealed in this chapter: if we account for the detailed dynamical processes that contribute to the evolution of real networks, we expect systematic deviations from a pure power law. Indeed, the previous sections predicted several analytical forms for the degree distribution:

- **Power law:** A pure power law emerges if a growing network is governed by preferential attachment only, lacking nonlinearities or initial attractiveness. In its pure form a power law is observed only in a few systems. Yet, it is the starting point for understanding the degree distribution of most real networks.

- **Stretched exponential:** If preferential attachment is sublinear, the degree distribution follows a stretched exponential SECTION 5.7. A stretched exponential degree-distribution can also appear under node removal at the critical point SECTION 6.5.

- **Fitness-induced corrections:** In the presence of fitnesses the precise form of p_k depends on the fitness distribution $\rho(\eta)$, which determines p_k via Eq. 6.6. For example, a uniform fitness distribution induces a logarithmic correction in p_k Eq. 6.8. Other forms of $\rho(\eta)$ can lead to rather exotic p_k.

- **Small-degree cutoffs:** Initial attractiveness adds a random component to preferential attachment. Consequently, the degree distribution develops a small-degree saturation.

- **Exponential cutoffs:** Node and link removal, present in many real systems, can also induce exponential cutoffs in the degree distribution. Furthermore, random node-removal can deplete the small-degree nodes, inducing a peak in p_k.

In most networks several of the elementary processes discussed in this chapter appear together. For example, in scientific collaboration networks we have sublinear preferential attachment with initial attractiveness and links can be both external and internal. As researchers have different creativity, fitness also plays a role, requiring us to know the appropriate fitness distribution. Therefore, the degree distribution is expected to display small degree saturation (thanks to initial attractiveness), stretched exponential cutoff at high degrees (thanks to sublinear preferential attachment), and some unknown corrections due to the particular form of the fitness distribution $\rho(\eta)$. These findings indicate that if our goal is to obtain an accurate fit to the degree distribution, we first need to build a generative model that analytically predicts the expected functional form of p_k. Yet, in many systems developing an accurate theory for p_k may be an overkill. Hence, it is often sufficient to establish if we are dealing with a broad or a bounded degree distribution SECTION 4.9, as the system's properties will be primarily driven by this distinction.

The results of this chapter also allow us to reflect on the role of the various network models. We can categorize these models into three main classes Table 6.2:

Static Models: The random network model of Erdős and Rényi CHAPTER 3 and the small world network model of Watts and Strogatz Fig. 3.15 have a fixed number of nodes, prompting us to call them static. They both assume that the role of the network modeler is to cleverly place the links between the nodes. Both models predict a bounded degree distribution.

Generative Models: The configuration and the hidden parameter models discussed in SECTION 4.8 generate networks with some predefined degree distribution. Hence, these models are not mechanistic, in the sense that they do not tell us why a network develops a particular degree distribution. Rather, they help us understand how various network properties, from clustering to path lengths, depend on the degree distribution.

Evolving Network Models: These models aim to capture the mechanisms that govern the time evolution of a network. The most studied example in the Barabási-Albert model, but equally important are the extenions discussed in this chapter, from the Bianconi-Barabási model to models involving internal links, aging, node and link deletion, or accelerated growth. These models are motivated by the hypothesis that if we correctly capture all microscopic processes that contribute to a network's evolution, then the network's large-scale characteristics follow from that. There is an important role in network theory for each three modeling frameworks. If our interest is limited to the role of the network environment on some phenomena, like spreading processes or network robustness, the generative models offer an excellent starting point. If, however, we want to understand the origin of a certain network property, we must resort to evolving network models, that capture the processes that have built the network in the first place.

Finally, the results of this chapter allow us to formulate our next network law:

The fifth law, the role of diversity.

With time the fittest nodes turn into the largest hubs.

A. Quantitative formulation

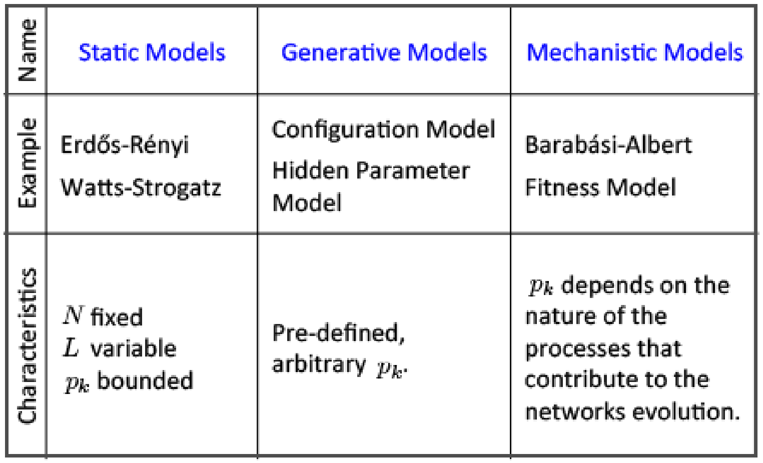
Eq. 6.4 offers the quantitative formulation of the fifth law, predicting that the dynamical exponent, capturing the rate at which a node acquires links, is proportional to the node's fitness. Hence the higher a node's fitness, the higher the rate it acquires links. Consequently, with time the nodes with the highest fitness will turn into the largest hubs.

B. Universality

In most networks nodes with different qualities and capabilities compete for links. Hence node fitness, capturing a node's ability to attract links, is present in most real networks.

C. Non-random origin

The dynamics of fitness-driven networks is quite different from the dynamics of the random network model, in which nodes acquire links at comparable rate. Hence, the properties of these networks cannot be explained within the random network framework.



Name	Static Models	Generative Models	Mechanistic Models
Example	Erdős-Rényi Watts-Strogatz	Configuration Model Hidden Parameter Model	Barabási-Albert Fitness Model
Characteristics	N fixed L variable p_k bounded	Pre-defined, arbitrary p_k.	p_k depends on the nature of the processes that contribute to the networks evolution.

Table 6.2
Models of network science

The table shows the three main modeling frameworks we encountered so far, together with their main distinguishing features.

HOMEWORK

1. Calculate the degree exponent and the dynamical exponent for a growing network with two distinct fitnesses. To be specific, let us assume that the fitnesses follow the double delta distribution

$$\rho(\eta)=\delta(\eta-a)+\delta(\eta-1) \quad \text{with } 0\leq a\leq 1. \qquad (6.33)$$

Discuss how the degree exponent depends on the parameter *a*.

2. Calculate the degree exponent of the directed Barabási-Albert model with accelerated growth, i.e. when $m(t)=t^{\Theta}$.

3. Assume that a network is driven by a preferential attachment with additive fitness, $\pi(k_i) \sim A_i+k_i$, where A_i is chosen from a $\rho(A_i)$ distribution [44]. Calculate and discuss the degree distribution of the resulting network.

ADVANCED TOPICS 6.A
SOLVING THE FITNESS MODEL

The purpose of this section is to derive the degree distribution of the fitness model [2, 13, 14]. We start by calculating the mean of the sum $\sum_j \eta_j k_j$ over all possible realizations of the quenched fitnesses *η*. Since each node is born at a different time t_o, we can write the sum over *j* as an integral over t_o

$$\left\langle \sum_j \eta_j k_j \right\rangle = \int d\eta \rho(\eta)\eta \int_1^t dt_0 k_\eta(t,t_0) \qquad (6.34)$$

By replacing $k_\eta(t, t_0)$ with Eq. 6.3 and performing the integral over t_0, we obtain

$$\left\langle \sum_j \eta_j k_j \right\rangle = \int d\eta \rho(\eta)\eta m \frac{t - t^{\beta(\eta)}}{1-\beta(\eta)} \ . \qquad (6.35)$$

The dynamic exponent *β(η)* is bounded, i.e. $0<\beta(\eta)<1$ because a node can only increase its degree with time $(\beta(\eta)>0$ and $k_i(t)$ cannot increase faster than $t(\beta(\eta)<1)$. Therefore in the limit t→∞ in Eq. 6.35 the term $t^{\beta(n)}$ can be neglected compared to *t*, obtaining

$$\left\langle \sum_j \eta_j k_j \right\rangle \overset{t\to\infty}{=} Cmt(1-O(t^{-\varepsilon})) \ , \qquad (6.36)$$

where $\varepsilon = (1 - \max_\eta \beta(\eta)) > 0$ and

$$C = \int d\eta \rho(\eta) \frac{\eta}{1-\beta(\eta)} \ . \qquad (6.37)$$

Using Eq. 6.36, and the notation $k_\eta = k_{\eta i}(t, t_0)$, the dynamic equation Eq. 6.2 can be written as

$$\frac{\partial k_\eta}{\partial t} = \frac{\eta k_\eta}{Ct} \ , \qquad (6.38)$$

which has a solution of the form Eq. 6.3, given that

$$\beta(\eta) = \frac{\eta}{C} \ , \qquad (6.39)$$

confirming the self-consistent nature of the assumption Eq. 6.3. To complete the calculation we need to determine C from Eq. 6.37. After substituting $\beta(n)$ with η/C , we obtain

$$1 = \int_0^{\eta_{\max}} d\eta\rho(\eta)\frac{1}{\frac{C}{\eta}-1} \ , \tag{6.40}$$

where η_{max} is the maximum possible fitness in the system. The integral Eq. 6.40 is singular. However, since $\beta(\eta)=\eta/c < 1$ for any η, we have $C > \eta_{max}$, thus the integration limit never reaches the singularity. Note also that, since

$$\sum_j \eta_j k_j \leq \eta_{\max}\sum_j k_j = 2mt\eta_{\max} \tag{6.41}$$

we have $C \leq \eta_{max}$.

If there is a single dynamic exponent β, the degree distribution should follow the power law $p_k \sim k{-}\gamma$, where the degree exponent is given by $\gamma=1/\beta+1$. However, in the fitness model we have a spectrum of dynamic exponents $\beta(\eta)$, thus p_k is given by a weighted sum over different power-laws. To find p_k we need to calculate the cumulative probability that a randomly chosen node's degree satisfies $k_\eta(t)>k$. This cumulative probability is given by

$$P(k_\eta(t) > k) = P\left(t_0 < t\left(\frac{m}{k}\right)^{C/\eta}\right) = t\left(\frac{m}{k}\right)^{C/\eta}. \tag{6.42}$$

Thus, the degree distribution is given by the integral

$$P_k = \int_{\eta_{max}}^{0} d\eta \frac{\partial P(k_\eta(t) > k)}{\partial t} \propto \int d\eta\rho(\eta)\frac{C}{\eta}\left(\frac{m}{k}\right)^{\frac{C}{\eta}+1}. \tag{6.43}$$

BIBLIOGRAPHY

[1] A.L. Barabási, Linked: The New Science of Networks. (Perseus, Boston) 2001.

[2] G. Bianconi and A.-L. Barabási, Competition and multiscaling in evolving networks, Europhysics Letters 54: 436-442, 2001.

[3] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the world wide web, Science 287: 2115, 2000.

[4] P. L. Krapivsky and S. Redner. Statistics of changes in lead node in connectivity-driven networks, Phys. Rev. Lett. 89:258703, 2002.

[5] C. Godreche and J. M. Luck. On leaders and condensates in a growing network, J. Stat. Mech. P07031, 2010.

[6] J. H. Fowler, C. T. Dawes, and N. A. Christakis. Model of Genetic Variation in Human Social Networks, PNAS 106: 1720-1724, 2009.

[7] M. O. Jackson. Genetic influences on social network characteristics, PNAS 106:1687–1688, 2009.

[8] S. A. Burt. Genes and popularity: Evidence of an evocative gene environment correlation, Psychol. Sci. 19:112–113, 2008.

[9] J. S. Kong, N. Sarshar, and V. P. Roychowdhury. Experience versus talent shapes the structure of the Web, PNAS 105:13724-9, 2008.

[10] A.-L. Barabási, C. Song, and D. Wang. Handful of papers dominates citation, Nature 491:40, 2012.

[11] D. Wang, C. Song, and A.-L. Barabási. Quantifying Long term scientific impact, preprint, 2013.

[12] M. Medo, G. Cimini, and S. Gualdi. Temporal effects in the growth of

networks, Phys. Rev. Lett., 107:238701, 2011.

[13] G. Bianconi and A.-L. Barabási. Bose-Einstein condensation in complex networks, Phys. Rev. Lett. 86: 5632–5635, 2001.

[14] C. Borgs, J. Chayes, C. Daskalakis, and S. Roch. First to market is not everything: analysis of preferential attachment with fitness, STOC'07, San Diego, California, 2007.

[15] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking, Phys. Rev. Lett. 85: 4633, 2000.

[16] C. Godreche, H. Grandclaude, and J. M. Luck. Finite-time fluctuations in the degree statistics of growing networks, J. of Stat. Phys. 137:1117-1146, 2009.

[17] Y.-H. Eom and S. Fortunato. Characterizing and Modeling Citation Dynamics, PLoS ONE 6(9): e24926, 2011.

[18] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311: 590-614, 2002.

[19] R. Albert, and A.-L. Barabási. Topology of evolving networks: local events and universality, Phys. Rev. Lett. 85:5234-5237, 2000.

[20] G. Goshal, L. Ping, and A.-L Barabási, preprint, 2013.

[21] J. H. Schön, Ch. Kloc, R. C. Haddon, and B. Batlogg. A superconducting field-effect switch, Science 288: 656–8. 2000.

[22] D. Agin. Junk Science: An Overdue Indictment of Government, Industry, and Faith Groups That Twist Science for Their Own Gain (Macmillan; New York), 2007.

[23] S. Saavedra, F. Reed-Tsochas, and B. Uzzi. Asymmetric disassembly and robustness in declining networks, PNAS105:16466–16471, 2008.

[24] F. Chung and L. Lu. Coupling on-line and off-line analyses for random power-law graphs, Int. Math. 1: 409-461, 2004.

[25] C. Cooper, A. Frieze, and J. Vera. Random deletion in a scalefree random graph process, Int. Math. 1, 463-483, 2004.

[26] S. N. Dorogovtsev and J. Mendes. Scaling behavior of developing and decaying networks, Europhys. Lett. 52: 33-39, 2000.

[27] C. Moore, G. Ghoshal, and M. E. J. Newman. Exact solutions for models of evolving networks with addition and deletion of nodes, Phys. Rev. E 74: 036121, 2006.

[28] H. Bauke, C. Moore, J. Rouquier, and D. Sherrington, Topological phase transition in a network model with preferential attachment and node removal, The European Physical Journal B: 83: 519-524, 2011.

[29] M. Pascual and J. Dunne, (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs (Oxford Univ Press, Oxford), 2005.

[30] R. Sole and J. Bascompte. Self-Organization in Complex Ecosystems (Princeton Univ Press, Princeton), 2006.

[31] U. T. Srinivasan, J. A. Dunne, J. Harte, and N. D. Martinez. Response of complex food webs to realistic extinction sequencesm, Ecology 88:671–682, 2007.

[32] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology, ACM SIGCOMM Computer Communication Review 29: 251-262, 1999.

[33] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, and A. Tomkins. Graph structure in the web, Computer networks 33: 309-320, 2000.

[34] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM TKDD07, ACM Transactions on Knowledge Discovery from Data (2007), 1(1).

[35] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, The large-scale organization of metabolic networks, Nature 407: 651–655, 2000.

[36] S. Dorogovtsev and J. Mendes. Effect of the accelerating growth of communications networks on their structure, Phys. Rev. E 63: 025101(R), 2001.

[37] M. J. Gagen and J. S. Mattick. Accelerating, hyperaccelerating, and decelerating networks, Phys. Rev. E 72: 016123, 2005.

[38] C. Cooper and P. Prałat. Scale-free graphs of increasing degree, Random Structures & Algorithms 38: 396–421, 2011.

[39] N. Deo and A. Cami. Preferential deletion in dynamic models of web-like networks, Inf. Proc. Lett. 102: 156-162, 2007.

[40] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites, Phys. Rev. E, 62:1842, 2000.

[41] A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Classes of small-world networks. Proc. National Academy of Sciences USA 97: 11149, 2000.

[42] K. Klemm and V. M. Eguiluz. Highly clustered scale free networks,

Phys. Rev. E 65: 036123, 2002.

[43] X. Zhu, R. Wang, and J.-Y. Zhu. The effect of aging on network structure, Phys. Rev. E 68: 056121, 2003.

[44] G. Ergun and G. J. Rodgers, Growing random networks with fitness, Physica A 303: 261-272, 2002.

CHAPTER 7

DEGREE CORRELATIONS

Introduction

Assortativity and disassortativity 1

Measuring degree correlations 2

Structural cutoffs 3

Degree correlations in real networks 4

Generating correlated networks

The impact of degree correlations

Summary

ADVANCED TOPICS 7.A
Degree correlation coefficient

ADVANCED TOPICS 7.B
Structural Cutoffs

Bibliography

Figure 6.0 (front cover)
Network representation by Mauro Martino

SECTION 7.0

INTRODUCTION

Angelina Jolie and Brad Pitt, Ben Affleck and Jennifer Garner, Harrison Ford and Calista Flockhart, Michael Douglas and Catherine Zeta-Jones, Tom Cruise and Katie Holmes, Richard Gere and Cindy Crawford. An odd list, yet instantly recognizable to those immersed in the headline-driven world of celebrity couples. They are Hollywood stars that are or were married in the past. Their weddings (and breakups) sold millions of magazines and drawn countless hours of media coverage. Thanks to them we take for granted that celebrities marry each other. We rarely pause to ask: is this normal? What is the expected chance that a celebrity marries another celebrity?

Assuming that a celebrity could date anyone from a pool of about a billion (10^9) eligible individuals, the chances that their mate would be another celebrity from a generous list of 1,000 other celebrities is only 10^{-6}. Therefore, if dating is driven by random encounters, celebrities would never marry each other. Yet, they do, with some puzzling implications.

Even if you do not care about the dating habits of celebrities, we must pause and explore what this phenomenon tells us about the structure of the social network. Hollywood celebrities, political leaders, and CEOs of major corporations tend to know an exceptionally large number of individuals and are known by even more. They are hubs. Hence celebrity dating is a manifestation of an interesting property of social network: hubs tend to have ties to other hubs.

As obvious this may sound, this property is not present in all networks. Consider for example the protein-interaction network of yeast, shown in Fig. 7.2. Each node corresponds to a protein and a link between two proteins indicates a binding interaction. A quick inspection of the network reveals its scale-free nature: numerous one- and two-degree proteins coexist with a few highly connected hubs. These hubs, however, tend avoid linking to *each other.* They link instead to many small-degree nodes, generating a hub-and-spoke pattern. This is particularly obvious for the two hubs highlighted in Fig. 7.2: they almost exclusively interact with small-degree proteins while avoiding linking to each other.

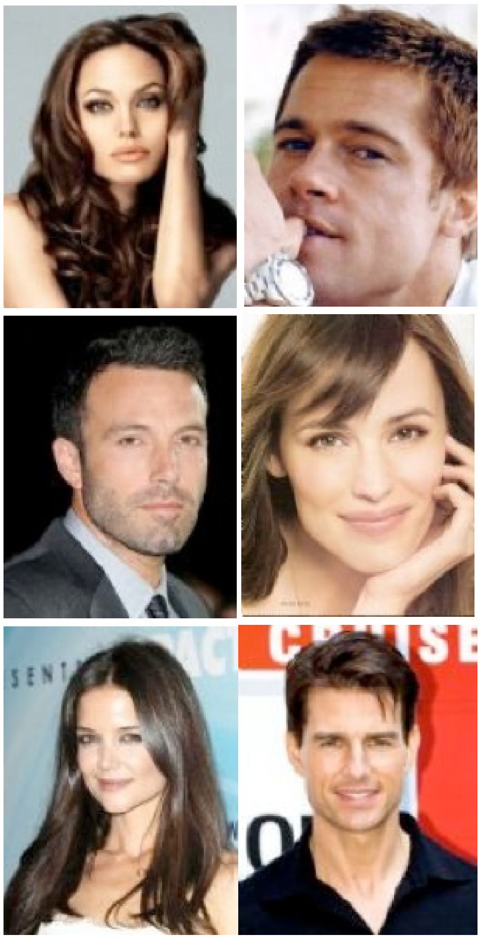


Figure 7.1
Hubs Dating Hubs

Celebrity couples, offering a vivid demonstration that in social networks hubs tend to know, date and marry each other (Images from http://www.whosdatedwho.com).

A brief calculation illustrates how unusual this pattern is. Let us assume that each node chooses randomly the nodes it connects to. Therefore the probability that two nodes with degree k and k' link to each other is

$$p_{k,k'} = \frac{kk'}{2L}. \qquad (7.1)$$

Eq. 7.1 tells us that hubs, by the virtue of the many links they have, are much more likely to connect to each other than to small degree nodes. Yet, the hubs highlighted in Fig. 7.2 almost exclusively connect to degree one nodes. By itself this is not unexpected: Fig. 7.1 also predicts that a hub with $k = 56$ connections should link to $N_1 P_{1,56} \approx 12$ nodes with degree 1. The problem is that this hub connects to 46 degree one neighbors, i.e. four times the expected number.

Furthermore, the likelihood that two largest hubs with degrees k=56 and $k' = 13$ have a direct link between them Fig. 7.2, is $p_{k_1\,k'} = 0.15$, which is 400 times larger than $p_{1,2} = 0.0004$, the likelihood that a degree-two node links to a degree-one node. Yet, there are no direct links between the hubs in Fig. 7.2, but we observe numerous direct links between small degree nodes.

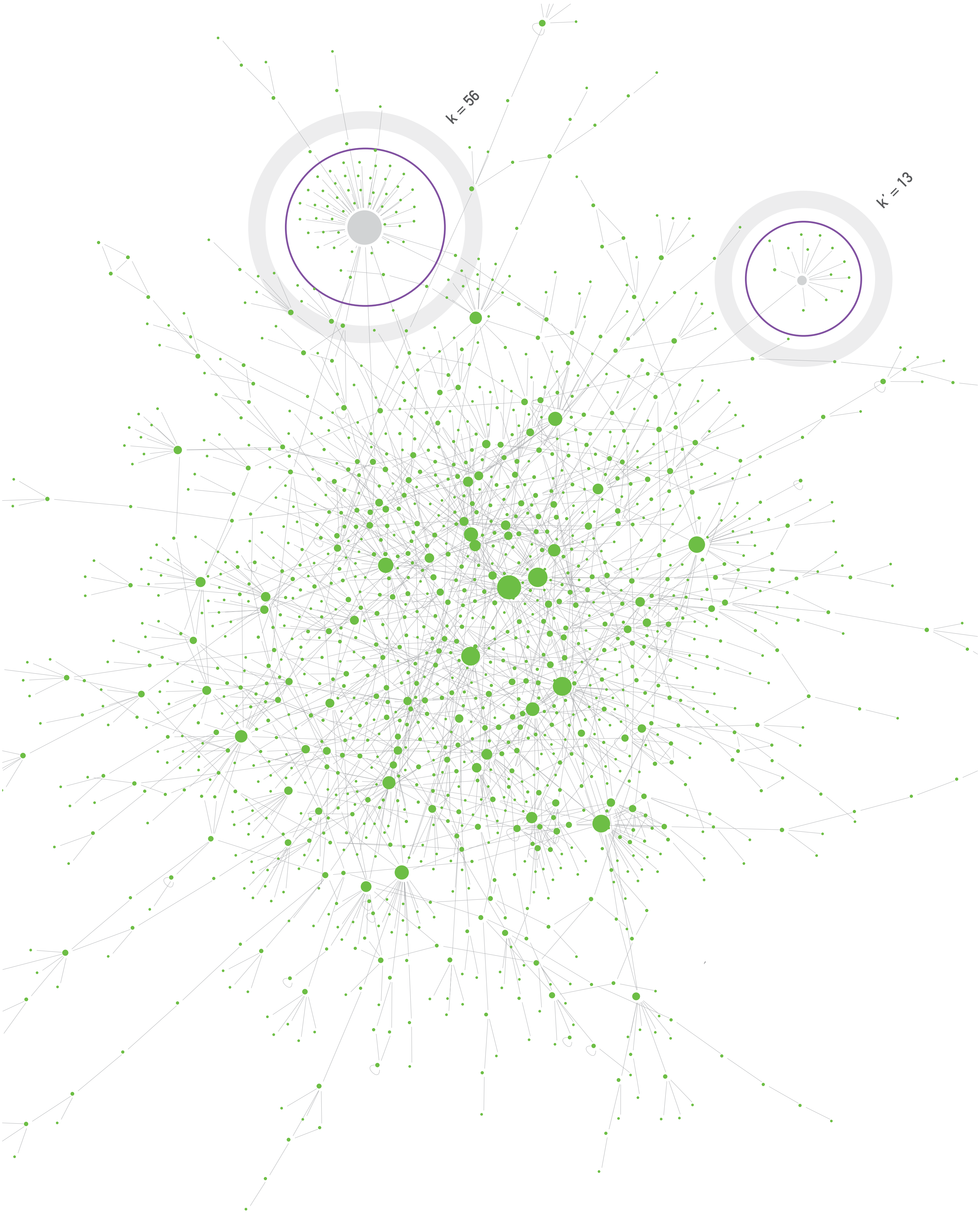
In summary, while in social networks hubs tend to “date” each other, in the protein interaction network the opposite is true: the hubs avoid linking to other hubs. While it is dangerous to extrapolate generic principles from two examples, the purpose of this chapter is to show that these patterns are manifestations of a general property of real networks: they exhibit a phenomena called *degree correlations*. We discuss how to measure such degree correlations and explore their impact on the network topology.

Figure 7.2 (folowing page)

Hubs Avoiding Hubs

The protein interaction map of yeast. Each node corresponds to a protein and two proteins are linked if there is experimental evidence that they can bind to each other in the cell. The two largest hubs, with degrees $k = 56$ (left) and $k' = 13$ (right) are highlighted in the figure. They both connect to many small degree nodes and avoid linking to each other.

The network has N = 1,870 proteins connected by L = 2,277 links, representing one of the earliest protein interaction maps [1, 2]. Only the largest component is shown. Note that the protein interaction network of yeast, discussed in TABLE 4.1, represents a later, more detailed map. Hence, it contains more nodes and links than the network shown in this figure. Redrawn after [3].


k = 56
k´ = 13

SECTION 7.1

ASSORTATIVITY AND DISASSORTATIVITY

Just by the virtue of the many links they have, hubs are expected to link to each other. Yet, as we have seen in the previous section, in some networks they do, in others they don't. This is illustrated in Fig. 7.3, that shows three networks with identical degree sequence but different topology:

- **Neutral Network**
  Fig. 7.3b shows a network whose wiring is truly a random. The network of Fig. 7.3b is *neutral*, meaning that the number of links between the hubs coincides with what we expect by chance, as predicted by Eq. 7.1. For clarity we highlighted in red the five largest nodes and the direct links between them, observing a few red links as the likelihood that two nodes link to each other increases with their degree.

- **Assortative Network**
  The network of Fig. 7.3a has precisely the same degree sequence as the one in Fig. 7.3b. Yet, there is a noticeable difference between the two networks: the hubs in Fig. 7.3a tend to link to each other, while avoiding linking to small-degree nodes. At the same time the small-degree nodes tend to connect to other small-degree nodes. Networks displaying such trends are *assortative*. An extreme manifestation of this pattern is a perfectly assortative network, in which degree-k nodes connect only to other degree-k nodes Fig. 7.4.

- **Disassortative Network**
  We observe the opposite trend in Fig. 7.3c, where the hubs completely avoid each other, linking mainly to small-degree nodes. Consequently the network displays a hub and-spoke character, making it *disassortative*.

In general a network displays degree correlations if the number of links between the high and low-degree nodes is systematically different from what is expected by chance. In other words, in correlated networks the number of links between nodes of degrees k and k' deviates from Eq. 7.1.

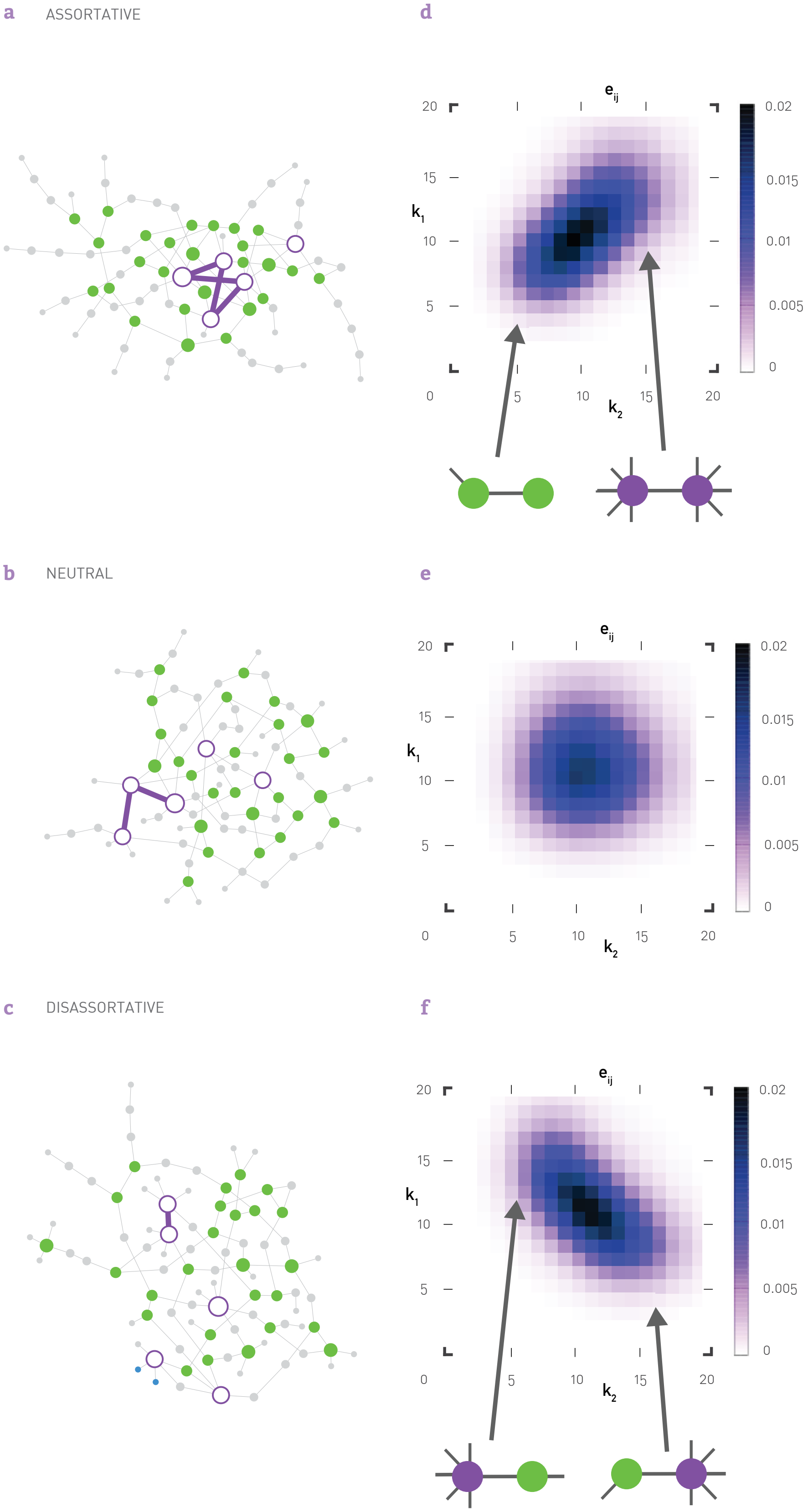


Figure 7.3

Degree correlation matrix

a, b, c Three networks that have precisely the same degree distribution (Poisson p_k), but display different degree correlations. We show only the largest component and we highlight the five nodes with the highest degree in red, together with the direct links between them.

d, e, f The degree correlation matrix e_{ij} for (d) an assortative, (e) a neutral and (f) a disassortative network with Poisson degree distribution and N=1,000, and $\langle k \rangle$=10. The colors correspond to the probability that there is a link between nodes with degrees k_1 and k_2.

a, d For assortative networks e_{ij} takes higher values along the main diagonal. This indicates that nodes of similar degree tend to link to each other: small-degree nodes to small-degree nodes and hubs to hubs. The network in (a) illustrates this by having numerous links between its hubs.

b, e In neutral networks nodes link to other nodes randomly. Hence, the density of links is symmetric around the average degree, indicating the lack of correlations in the linking pattern.

c, f In disassortative networks e_{ij} is higher along the secondary diagonal, indicating that hubs tend to connect to small-degree nodes, and small-degree nodes to hubs. This is illustrated by the hub and spoke character of the network in (c).

The complete information about potential degree correlations is contained in the degree correlation matrix, e_{ij}, which is the probability of finding *a* node with degrees *i* and *j* at the two ends of a randomly selected link. As e_{ij} is a probability, it obeys the normalization condition

$$\sum_{i,j} e_{ij} = 1. \tag{7.2}$$

In SECT. 5.8 we derived the probability q_k that there is a degree-*k* node at the end of the randomly selected link Eq. 5.29.

$$q_k = \frac{kp_k}{\langle k \rangle} \tag{7.3}$$

We can connect q_k to $e_{i,j}$ via

$$\sum_{j} e_{ij} = q_i. \tag{7.4}$$

In neutral networks, we expect

$$e_{ij} = q_i q_j. \tag{7.5}$$

A network displays degree correlations if eij deviates from the random expectation captured by Eq. 7.5, Eqs. 7.2 - 7.5 are valid for networks with an arbitrary degree distribution, hence they apply to both random and scale-free networks. Given that e_{ij} contains the complete information about potential degree correlations, we start with its visual inspection. Figs. 7.3 d, e, f shows e_{ij} for an assortative, a neutral and a disassortative network. In a neutral network small and high-degree nodes connect to each other randomly, hence e_{ij} lacks any trend Fig. 7.3e. In contrast, assortative networks show high correlations along the main diagonal, indicating that nodes predominantly connect to other nodes with comparable degree. Therefore low-degree nodes tend to link to other low-degree nodes and hubs to hubs Fig. 7.3d. In disassortative networks e_{ij} displays the opposite trend: it has high correlations along the secondary diagonal, indicating that high-degree nodes tend to connect to low-degree nodes Fig. 7.3f.

In summary information about degree correlations is carried by the degree correlation matrix e_{ij}. Yet, the study of degree correlations through the inspection of e_{ij} has numerous disadvantages:

- It is difficult to extract information from the visual inspection of a matrix.

- Unable to infer the magnitude of the correlations, it is difficult to compare networks with different correlations.

- e_{jk} contains approximately k^2_{max} independent variables, representing a huge amount of information that is difficult to model in analytical calculations and simulations.

We therefore need to develop a more compact way to detect the presence and the magnitude of degree correlations.

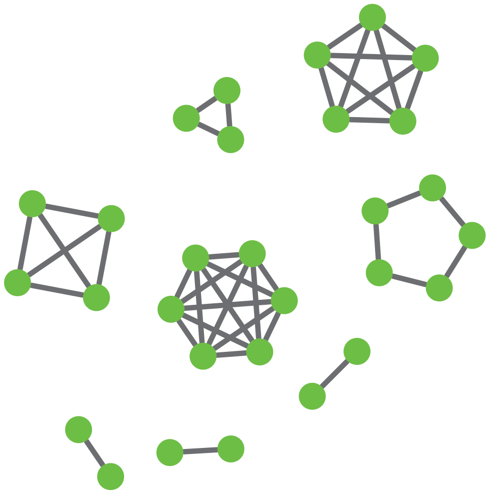


Figure 7.4

A perfectly associative network

Maximal assortativity is obtained when each degree-*k* node links only to other degree-*k* nodes. For such a perfectly assortative network $e_{jk} = \delta_{jk}q_k$, where δ_{jk} is the Kronecker delta. In this case the non-diagonal elements of the e_{jk} matrix are zero. The figure shows such a perfectly assortative network, consisting of complete *k*-clusters.

MEASURING DEGREE CORRELATIONS

While e_{ij} contains the complete information about the potential degree correlations characterizing a network, it is difficult to interpret its content. The purpose of this section is to introduce the degree correlation function, which offers a simpler way to measure degree correlations.

Degree correlations capture the relationship between the degrees of nodes that link to each other. One way to quantify their magnitude is to measure for each node i the average degree of its neighbors Fig. 7.5.

$$k_{nn}(k_i) = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}k_j\,. \tag{7.6}$$

If we wish to calculate Eq. 7.6 for all nodes with the same degree k, we define the degree correlation function as [4, 5]

$$k_{nn}(k) \equiv \sum_{k'} k'P(k'\mid k) \tag{7.7}$$

where $P(k' \mid k)$ is the conditional probability that following a link of a k-degree node we reach a degree-k node. To quantify degree correlations we inspect the dependence of $k_{nn}(k)$ on k. For *neutral networks*, using Eqs. 7.3-7.5, we have

$$P(k'\mid k) = \frac{e_{kk'}}{\sum_{k'} e_{kk'}} = \frac{e_{kk'}}{q_k} = \frac{q_{k'}q_k}{q_k} = q_{k'}\,. \tag{7.8}$$

Hence $k_{nn}(k)$ can be expressed as

$$k_{nn}(k) = \sum_{k'} k'q_{k'} = \sum_{k'} k'\frac{k'p(k')}{\langle k\rangle} = \frac{\langle k^2\rangle}{\langle k\rangle}\,.\!. \tag{7.9}$$

Therefore, in a neutral network the average degree of a node's neighbors is independent of the node's degree k and depends only on $\langle k\rangle$ and $\langle k^2\rangle$. So plotting $k_{nn}(k)$ in function of k is expected to result in a horizontal line at $\langle k^2\rangle/\langle k\rangle$, as observed in the case of the power grid in Fig. 7.6b. Eq. 7.9 also reflects an intriguing property of real networks: that our friends are more popular than we are, a phenomenon called the *friendship paradox* BOX 7.1.

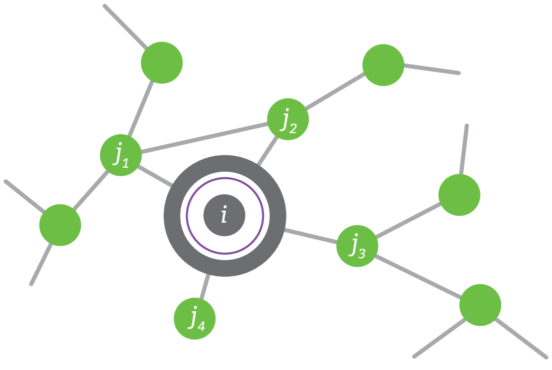


Figure 7.5

Nearest neighbor degree: K_{nn}

To determine $k_{nn}(k)$, we calculate the average degree of a node's neighbors. The figure illustrates the calculation of $k_{nn}(k)$ for node i shown in red. As the degree of the node i is $k_i = 4$, by averaging the degree of its neighbors j_1, j_2, j_3 and j_4, we obtain $k_{nn}(4) = (4 + 3 + 3 + 1)/4 = 2.75$.

- **Assortative Network**
  In this case hubs tend to connect to other hubs, hence the higher is the degree k of a node, the higher should be the average degree of its nearest neighbors. Consequently for assortative networks $k_{nn}(k)$ increases with k, as observed in collaboration networks in **Fig. 7.6a**.

- **Disassortative Network**
  In this case hubs prefer to link to low-degree nodes. Consequently $k_{nn}(k)$ decreases with k, as observed for the protein-protein interaction network **Fig. 7.6c**.

The scaling observed in **Fig. 7.6** prompts us to approximate the degree correlation function with [4]

$$k_{nn}(k) = ak^{\mu}. \tag{7.10}$$

If **Eq. 7.10** holds, then the nature of degree correlations characterizing a network is determined by the sign of the *correlation exponent* μ:

- **For assortative Networks μ > 0**
  Indeed, a fit to $k_{nn}(k)$ for the science collaboration network provides $\mu = 0.37 \pm 0.11$ **Fig. 7.6a**.

- **For neutral networks we have μ = 0**
  As according to **Eq. 7.9** $k_{nn}(k)$ is independent of k. For the power grid we obtain $\mu = 0.04 \pm 0.05$, which is indistinguishable from zero **Fig. 7.6b**.

- **For disassortative networks we expect μ < 0**
  Indeed, for the metabolic network we obtain $\mu = -0.76 \pm 0.04$ **Fig. 7.6c**.

In summary, the degree correlation function helps us capture the presence or absence of correlations in real networks. The $k_{nn}(k)$ function also plays an important role in analytical calculations, allowing us to calculate the impact of degree correlations on various network characteristics **SECT. 7.6**. Note that it is often convenient to extract a single number to capture the magnitude of correlations present in a network. This can be achieved either through the correlation exponent μ defined in **Eq. 7.10**, or using the degree correlation coefficient discussed in **BOX 7.2**.

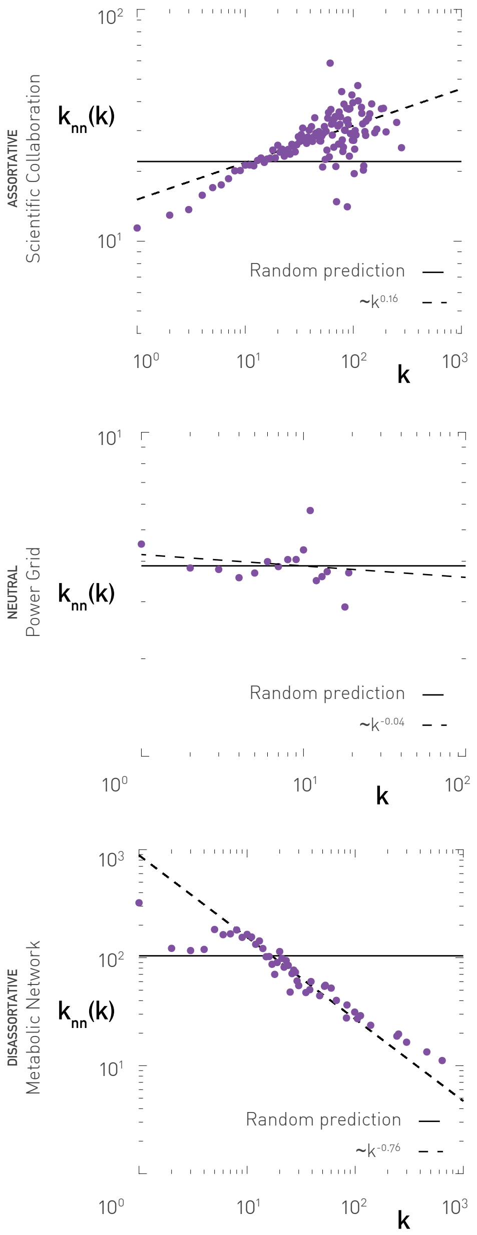


Figure 7.6
Degree correlation function

The degree correlation function $k_{nn}(k)$ for three real networks. The panels show $k_{nn}(k)$ on a log-log plot to test the validity of **Eq. 7.10**.

(a) Collaboration network of astrophysicists. The increasing $k_{nn}(k)$ with k indicates that the network is assortative.

(b) Power grid. The horizontal $k_{nn}(k)$ indicates the lack of degree correlations, as predicted by **Eq. 7.9** for neutral networks.

(c) Metabolic network. The decreasing $k_{nn}(k)$ documents the network's disassortative nature.

On each panel the horizontal dotted line corresponds to the prediction **Eq. 7.9** and the oblique dashed line is a fit to **Eq. 7.10**. The slope in (a) is $\mu = 0.37$, in (b) is $\mu = 0.04$ while the slope in (c) is $\mu = -0.76$.

BOX 7.1

FRIENDSHIP PARADOX

While most people believe that they have more friends than their friends [7], the friendship paradox, discovered by sociologist Scott L. Feld, states the opposite: on average your friends are more popular than you are [6].

The roots of the friendship paradox is Eq. 7.9, telling us that the average degree of a node's neighbors is not simply $\langle k \rangle$, but depends on $\langle k^2 \rangle$ as well. Consider for example a random (Erdős-Rényi) network, for which $\langle k^2 \rangle = \langle k \rangle(1 + \langle k \rangle)$. According to Eq. 7.9

$$k_{nn}(k) = 1 + \langle k \rangle. \tag{7.11}$$

Therefore the average degree of a node's neighbors is always higher than the average degree of the network $\langle k \rangle$. The gap between $\langle k \rangle$ and our friends' degree can be particularly large in scale-free networks, for which $\langle k^2 \rangle / \langle k \rangle$ is significantly larger than $\langle k \rangle$ Fig. 4.7. Consider for example the email network, for which $\langle k^2 \rangle / \langle k \rangle = 390.45$, or the actor network, for which $\langle k^2 \rangle / \langle k \rangle = 565.70$. Hence in these networks the average degree of the friends of a randomly selected node can be hundreds of times higher than the expected degree of the node itself, which is $\langle k \rangle$.

To understand the origin of the friendship paradox, we must realize that for a randomly chosen node, the degree distribution of the nodes at the other end of each link do not follow p_k, but are biased towards higher-degree nodes, as indicated by Eq. 7.3. In other words, we are more likely to be friends with hubs than with small-degree nodes, simply because hubs have more friends than the small-nodes. Hence our friends do not reflect the whole population - they are biased towards the hubs.

SECTION 7.3

STRUCTURAL CUTOFFS

Throughout this book we assumed that the networks we explore are simple, meaning that there is at most one link between any two nodes CHAPTER 2. For example, in the email network we place a single link between two individuals that are in email contact, despite the fact that they may have exchanged multiple messages; in the actor network we connect two actors with a single link if they acted together, independent of the number of movies they jointly made. All datasets discussed in TABLE 4.1 are simple networks. In simple networks there is a puzzling conflict between the scale-free property and degree correlations [10, 11]. Consider for example the scale-free network of Fig. 7.7a, whose two largest hubs have degrees $k = 55$ and $k' = 46$, connected by a link. In a network with degree correlations $e_{kk'}$ the expected number of links between k and k' is

$$E_{kk'} = e_{kk'} \langle k \rangle N \qquad (7.14)$$

For a neutral network $e_{kk'}$ is given by Eq. 7.5, which, using Eq. 7.3, predicts

$$E_{kk'} = \frac{k_k p_k k' p_{k'}}{\langle k \rangle} N = \frac{\frac{55}{300}\frac{46}{300}}{3} 300 = 2.8\ . \qquad (7.15)$$

Therefore, given the size of these two hubs, they should be connected to each other by two to three links to comply with the network's neutral nature. Yet, in a simple network we are allowed only one link between them, raising a conflict between degree correlations and the scale-free property. Such conflict emerges in a simple network each time the degrees violate the $E_{kk'} \leq 1$ condition. The goal of this section is to understand the origin and the consequences of this conflict.

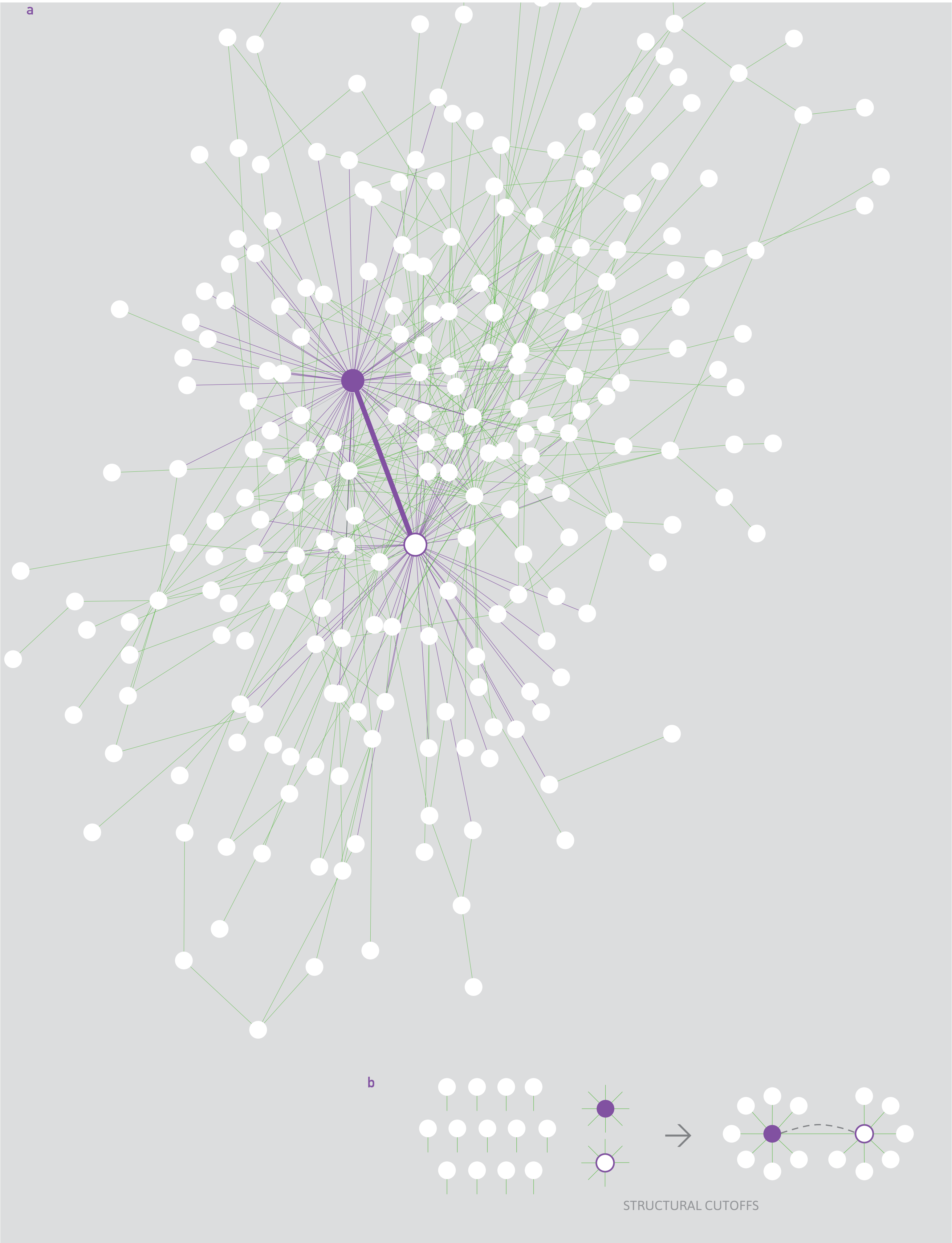
For small k and k Eq. 7.15 predicts that $E_{kk'}$ is also small, i.e. we expect less than one link between the two nodes. Only for nodes whose degree exceeds some threshold k_s will Eq. 7.15 predict multiple links. As we show in ADVANCED TOPICS 7.B, this k_s, that we

Figure 7.7 (following page)

Structural disassortativity

(a) A scale-free network with N=300, L=450, according Eq. 7.15, and γ=2.2, generated by the configuration model, while forbidding self-loops and multiple links between two nodes, making the network simple. The blue and the red nodes are the two largest nodes in the network and are connected by the red link. As Eq. 7.15 predicts,to maintain the network's neutral nature, we would need two to three links between these two nodes. The fact that we do not allow multiple links (simple network representation) makes the network disassortative, a phenomena we call structural disassortativity.

(b) To illustrate the origins of structural correlations, we start from a fixed degree sequence, shown as stubs on the left, and we randomly connect the stubs (configuration model). In this case, the expected number of links between the nodes with degree 8 and 7 is 8 x 7/28 ≈ 2. Yet, if we do not allow multilinks, there can only be one link, making the network structurally dissasortative.


a
b
STRUCTURAL CUTOFFS

$$k_s(N) \sim (\langle k \rangle N)^{1/2}. \qquad (7.16)$$

In other words, nodes whose degree exceeds Eq. 7.16 are expected to have $E_{kk'} > 1$, a conflict that as we show below gives rise to degree correlations.

To fully understand the consequences of the described conflict, we must first ask if a network has nodes whose degrees exceeds Eq. 7.16. For this we compare the structural cutoff, k_s, with the natural cutoff, k_{max}, which is the expected largest degree in a network with degree distribution p_k. According to Eq. 7.14, for a scale-free network $k_{max} \sim N^{\frac{1}{\gamma-1}}$. The relative magnitude of k_{max}, vs. k_s, gives raise to two regimes:

- For scale-free networks with $\gamma \geq 3$ and random networks, ks is always larger than k_{max}, hence we lack nodes for which $E_{kk'} > 1$.

- For scale-fee networks with $\gamma < 3$, k_s is smaller than k_{max}, hence all nodes between k_s and $k_{\max}$ violate $E_{kk'} > 1$. Consequently, the network has fewer links between its hubs than expected based on Eq. 7.15. As a result, these networks will be disassortative, a phenomenon we call *structural disassortativity*. This is illustrated in Figs. 7.8a, b that show a simple scale-free network generated by the configuration model. The network shows disassortative tendencies, despite the fact that we did not impose degree correlations.

We have two avenues to generate networks that are free of structural disassortativity:

(i) We relax the simple network requirement, allowing multiple links between the nodes. The conflict disappears and the network will be neutral Figs. 7.8c, d.

(ii) If we insist of having a simple scale-free network that is neutral or associative, we must remove all hubs with degrees larger than k_s. This is illustrated in Fig 7.8 e, f: the obtained network, missing nodes with $k \geq 100$, is neutral.

How can we convince ourselves that the correlations observed in a particular network are a consequence of structural dissasortativity, or are generated by some unknown process? Degree-preserving randomization Fig. 4.14 helps us distinguish these two possibilities:

(i) Degree preserving randomization with simple links (R-S): We apply degree-preserving randomization to the original network, while making sure that we do not allow for more than one link between any pair of nodes. On the algorithmic side this means that each rewiring that results in multiple links between two nodes is discarded. If the real $k_{nn}(k)$ and the randomized $k_{nn}^{R-S}(k)$ are indistinguishable, then the correlations observed in a real system are all structural,

BOX 7.2

DEGREE CORRELATION COEFFICIENT

If we wish to characterize degree correlations using a single number, we can also use the degree correlation coefficient, introduced by Mark Newman and defined as [8,9]

$$r = \sum_{jk} \frac{jk(e_{jk} - q_j q_k)}{\sigma_r^2} \qquad (7.12)$$

with

$$\sigma_r^2 = \sum_k k^2 q_k - \left[\sum_k^{jk} k q_c \right]^2 \qquad (7.13)$$

Hence r is the Pearson correlation coefficient between the degrees at the two end of the same link. It varies between $-1 \leq r \leq 1$: for $v < 0$ the network is assortative, for $r = 0$ the network is neutral and for $v > 0$ the network is disassortative. For example, for the collaboration network we obtain $r = 0.13$, in line with its assortative nature; for the protein interaction network $r = -0.04$, supporting its disassortative nature and for the power grid $r = 0$. Note that the degree correlation coefficient r assumes that $k_{nn}(k)$ is a linear function of k with slope r. In contrast the correlation exponent μ assumes that $k_{nn}(k)$ follows the power law Eq. 7.10. Naturally, both scaling laws cannot be valid simultaneously. The analytical models of SECT. 7.7 offer some guidance, supporting the validity of Eq. 7.10. As we show in ADVANCED TOPICS 7.B, while r correlates with μ, we need to be cautious when we use it to measure degree correlations.

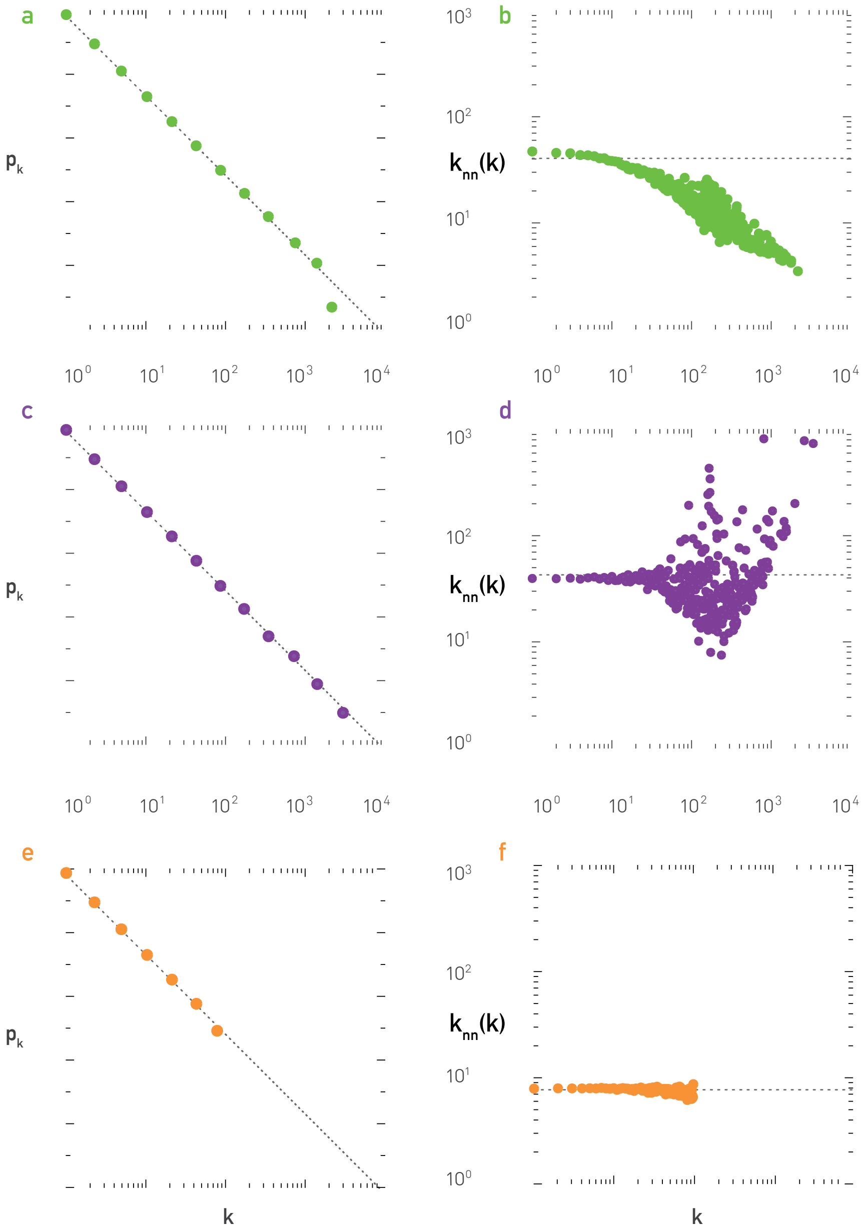


Figure 7.8

Natural and structural cutoffs

The figure illustrates the tension between the scale-free property and degree correlations. It shows the degree distribution (left panels) and the degree correlation function $k_{nn}(k)$ (right panels) of a scale-free network with $N = 10{,}000$ and $\gamma = 2.5$, generated by the configuration model.

(a, b) If we generate a scale-free network with the power-law degree distribution shown in (a), and we forbid self-loops and multi-links, the network displays structural disassortativity, as indicated by $k_{nn}(k)$ in (b). In this case, we lack a sufficient number of links between the high-degree nodes to maintain the neutral nature of the network, hence for high k the $k_{nn}(k)$ function decays.

(c, d) We can eliminate structural disassortativity by allowing multiple links, i.e. relaxing the simple network requirement. As shown in (c,d), in this case we obtain a neutral scale-free network.

(e, f) If we artificially impose an upper cutoff by removing all nodes with $k \geq k_s$ predicted by **Eq. 7.16**, the network becomes neutral, as seen in (f).

fully explained by the degree distribution. If the randomized $k_{nn}^{R-S}(k)$ does not show degree correlations while $k_{nn}(k)$ does, there is some unknown process that generates the observed degree correlations.

(ii) Degree preserving randomization with multiple links (*R-M*): For a self-consistency check it is useful to also perform degree-preserving randomization that allows for multiple links between the nodes. On the algorithmic side this means that we allow each random rewiring, even if they lead to multiple links. This process eliminates all degree correlations.

We have taken the three networks of in Fig. 7.6 and performed the randomizations discussed above. As Fig. 7.9a shows, the assortative nature of the scientific collaboration network disappears under both randomizations. This indicates that the observed assortative correlations are not linked to the scale-free nature of the underlying network. In contrast, for the metabolic network the observed disassortativity remains unchanged under *R-S* Fig. 7.9c. This indicates that the disassortativity of the metabolic network is structural, induced by its degree distribution.

In summary, the scale-free property can induce disassortativity in simple networks. To be specific, in neutral or assortative networks we expect multiple links between the hubs. If such multiple links are forbidden (simple graph), the network will display disassortative tendencies. This conflict vanishes for scale-free networks with $\gamma \geq 3$ and for random networks. It also vanishes if we allow for multiple links between the nodes.

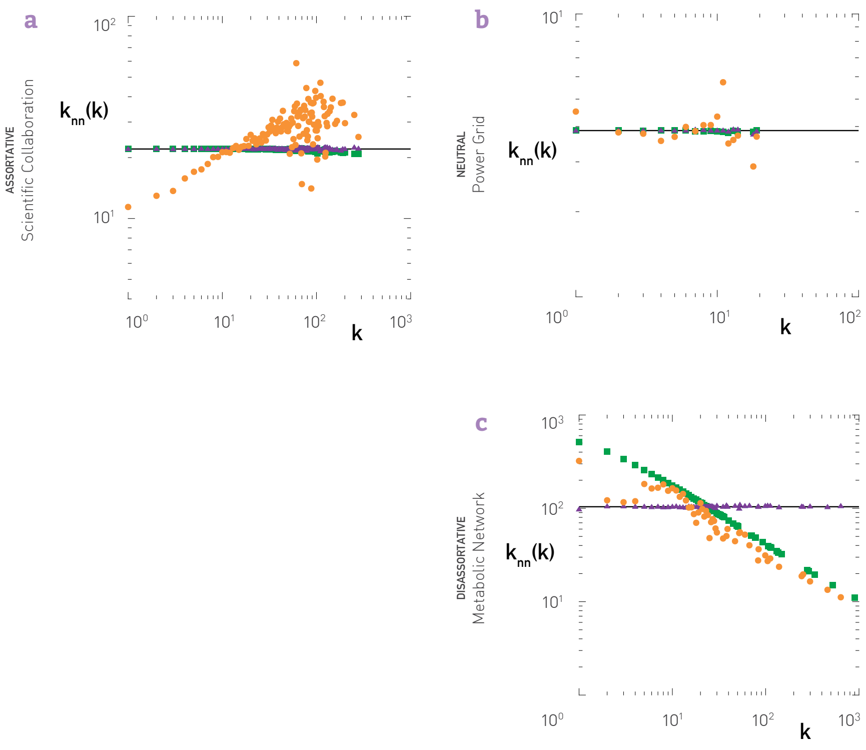


Figure 7.9

Randomization and degree correlations

To uncover the origin of the observed degree correlations, it is useful to compare $k_{nn}(k)$ with k_{nn}^{R-S} (k) and $^{R}k^{M}_{nn}$ (k) obtained after degree-preserving randomization. We perform two different randomizations for this purpose.

Green symbols
Degree-preserving randomization with simple links (*r-s*), in which case at each step of the randomization process we check that we do not have more than one link between any node pairs.

Blue symbols
Degree-preserving randomization with multiple links (*R-M*), in which case we allow multilinks during the randomization processes.

We performed these two randomizations for the networks of Fig. 7.6. The *R-M* procedure always generates a neutral network, consequently $k_{nn}^{R-M}(k)$ is always horizontal.The true insight is provided when we compare $k_{nn}(k)$ with $k_{nn}^{R-S}(k)$, allowing us to decide if the observed correlations are structural:

(a) Scientific collaboration network
The increasing $k_{nn}(k)$ differs from the horizontal $k_{nn}^{R-S}(k)$, indicating that the network's assortativity is not structural (i.e. it is not a consequence of the degree distribution), but it is generated by some process that governs the network's evolution. This is not unexpected: structural effects can generate only disassortativity, not assortativity.

(b) Power grid
The horizontal $k_{nn}(k)$, $k_{nn}^{R-S}(k)$ and $k_{nn}^{R-M}(k)$ all support the lack of degree correlations (neutral network).

(c) Metabolic network
As both $k_{nn}(k)$ and $k_{nn}^{R-S}(k)$ decrease, we conclude that the network's disassortative nature is induced by its scale-free property. Hence the observed degree correlations are structural.

DEGREE CORRELATIONS IN REAL NETWORKS

To truly understand the prevalence of degree correlations, we need to inspect the correlations characterizing various real networks. Therefore, in Fig. 7.10 we show the $k_{nn}(k)$ function for the ten reference networks of TABLE 4.1. Let us discuss the observed behavior:

- **Power grid**
  For the power grid $k_{nn}(k)$ is flat and indistinguishable from its randomized version, indicating a lack of degree correlations Fig. 7.10a. Hence the power grid is neutral.

- **Internet**
  For small degrees ($k \leq 30$) $k_{nn}(k)$ shows a clear assortative trend, an effect that levels off for high degrees Fig. 7.10b. The degree correlations vanish in the randomized networks. Hence the Internet is assortative, but structural cutoffs eliminate the effect for high *k*.

- **Social Networks**
  The three networks capturing social phenomena, like the mobile phone network, science collaboration networks and actor network, all have an increasing $k_{nn}(k)$, indicating that they are assortative Figs. 7.10c-e. Hence in these networks hubs tend to link to other hubs and low-degree nodes tend to link to low-degree nodes. For each of these networks the observed $k_{nn}(k)$, differs from the $k_{nn}^{R\text{-}S}(k)$, indicating that their assortative nature is not rooted in the degree distribution.

- **Email Network**
  While the email network is often used as an example of a social network, its $k_{nn}(k)$ decreases with *k*, documenting a clear disassortative behavior Fig. 7.10f. The randomized $k_{nn}^{R\text{-}S}(k)$ also decays, indicating that we are observing structural disassortativity, a consequence of the network's scale-free nature.

- **Biological Networks**
  The protein interaction and the metabolic network both have a nega-

tive μ, suggesting that these networks are disassortative Eq. 7.10. Yet, the scaling of $k_{min}^{R\text{-}S}$ *(k)* is indistinguishable from $k_{nn}(k)$, indicating that we are observing structural disassortativity, rooted in the scale-free nature of these networks Fig. 7.10g, h.

- **WWW**
  The decaying $k_{nn}(k)$ implies disassortative correlations Fig. 7.10i. The randomized $k_{min}^{R\text{-}S}$ *(k)* also decays, but not as rapidly as $k_{nn}(k)$. Hence the disassortative nature of the WWW is not fully explained by its degree distribution.

- **Citation network**
  This network displays a puzzling behavior: for $k \leq 20$, $k_{nn}(k)$ shows a clear assortative trend; for $k > 20$, however, we observe equally clear disassortative scaling Fig. 7.10j. Such mixed behavior can emerge in networks that display extreme assortativity SECT. 7.6. This suggests that the citation network is strongly assortative up to k_s, but its scale-free nature reverses the trend for $k \gg k_s$.

In summary, Fig. 7.10 indicates that to understand degree correlations, we must always compare $k_{nn}(k)$ to the degree randomized $k_{nn}^{R\text{-}S}(k)$. It also allows us to draw some interesting conclusions:

(i) Of the ten reference networks the power grid appears to be the only that is truly neutral. Hence most real networks display degree correlations.

(ii) All networks that display disassortative tendencies (email, protein, metabolic), do so thanks to their scale-free property. Hence, these are all structurally disassortative. Only the WWW shows disassortative correlations that are only partially explained by its degree distribution.

(iii) The degree correlations characterizing associative networks are not explained by their degree distribution. Most social networks (mobile phone calls, scientific collaboration, actor network) are in this class and so is the Internet and the citation network.

A number of proposals exist to explain the origin of the observed assortativity. For example, the tendency of individuals to form communities CHAPTER 9 has been shown to induce assortative scaling [12]. Similarly, the society has endless mechanisms, from professional committees to TV shows, to bring hubs together, enhancing the assortative nature of social and professional networks. Finally, homophily, a well documented social phenomena, [13], captures the fact that individuals have a tendency to associate with other individuals of similar background and characteristics. This tendency may also be responsible for the celebrity marriages discussed in SECT. 7.0.

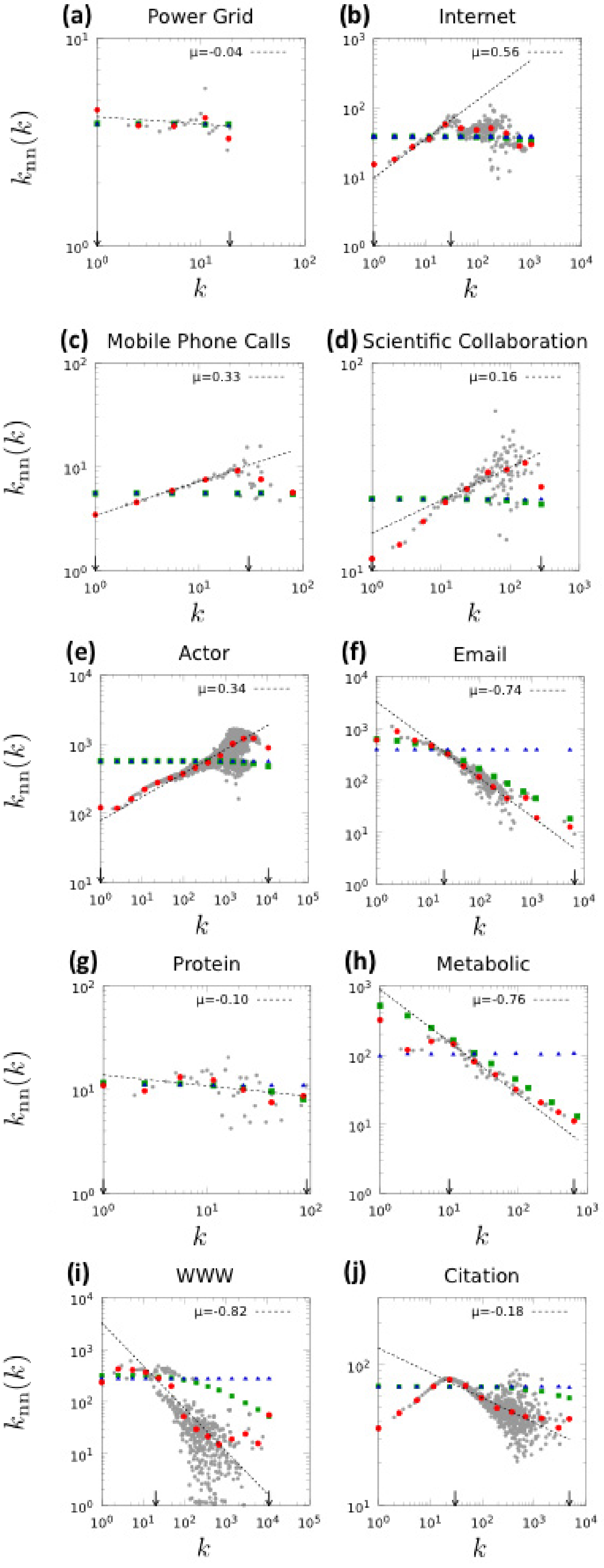


Figure 7.10

Randomization and degree correlations

The degree correlation function $k_{nn}(k)$ for the ten reference networks of **Table 4.1**. The grey symbols show the $k_{nn}(k)$ function under linear binning: red symbols represent the same data using log-binning **SECT. 4.10**. The dotted line corresponds to the best fit of the form **Eq. 7.10** and the small arrows at the bottom mark the fitting interval. Green squares represent $k_{nn}(k)$ obtained for 100 independent degree-preserving randomizations, making sure that we preserve the simple character of these networks; blue triangles correspond to k_{nn} (k), i.e. randomization that does allow self-loops and multiple links between two nodes. Note that we made directed networks undirected when we measured $k_{nn}(k)$. To fully characterize the correlations emerging in directed networks we must use the directed correlation function **BOX 7.3**.

BOX 7.3

CORRELATIONS IN DIRECTED NETWORKS

The degree correlation function $k_{nn}(k)$ in Eq. 7.7 is defined for undirected networks. To measure correlations in directed networks we must take into account that each node i is characterized by an incoming k_i^{in} and an outgoing k_i^{out} degree. Hence, we define four degree correlation functions, $k_{nn}^{\alpha,\beta}(k)$, where α and β refer to the in and out indices Figs. 7.11 a-d. In Fig. 7.11e we show $k_{nn}^{\alpha,\beta}(k)$ for citation networks, indicating a lack of in-out correlations, while a detectable assortative scaling for small k for the other three correlations.

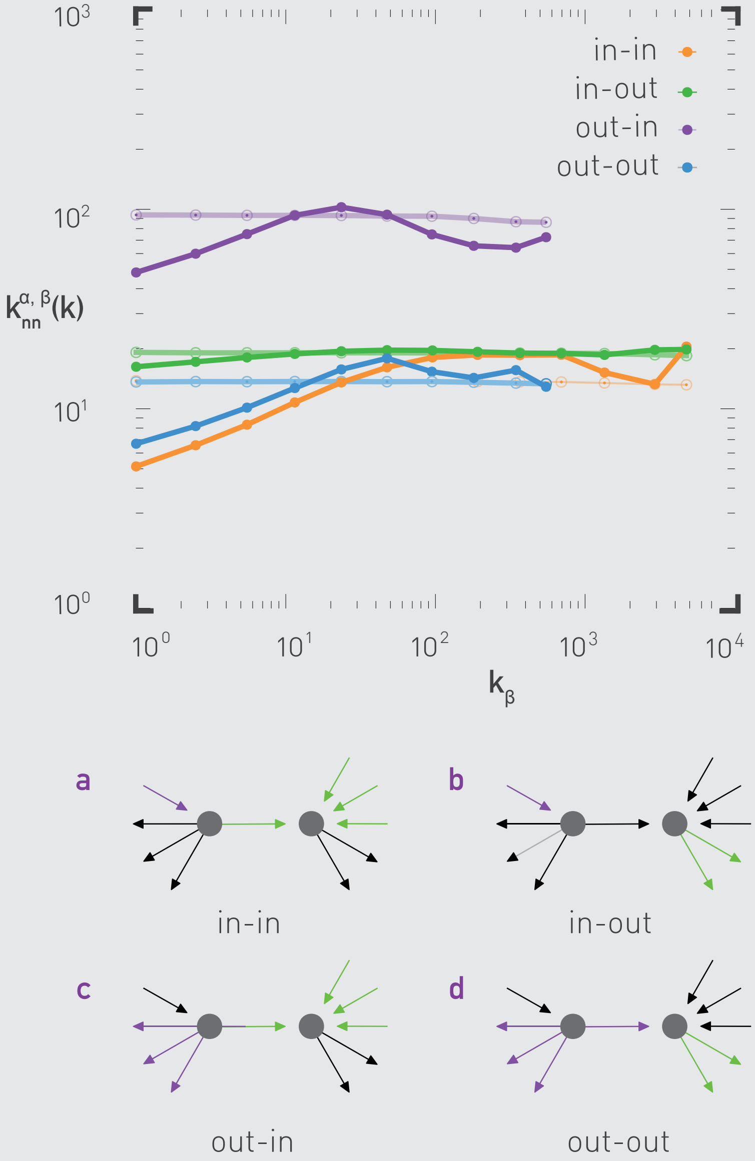


Figure 7.11

Correlation and directed network

Panels (a)-(d) illustrate the four possible correlations in directed networks. We show in red and green the (α, β) indices that define the appropriate correlation function [14]. For example, (a) describes the $k_{nn}^{in,in}(k)$ correlations between the in-degrees of two nodes connected by a link. (e) The $k_{nn}^{\alpha,\beta}(k)$ correlation function for citation networks, a directed network. For example $k_{nn}^{in,in}(k)$ is the average indegree of the in-neighbors of nodes with in-degree k_{in}. These functions show a clear assortative tendency for three of the four function up to degree $k \simeq 100$. The lighter symbols capture the degree randomized $k_{nn}^{\alpha,\beta}(k)$ for each correlation function.

SECTION 7.5

GENERATING CORRELATED NETWORKS

To study degree correlations and to explore their impact on various network characteristics, we need to build networks with tunable correlations. Given the conflicts between the scale-free property and degree correlations, this is not a trivial task. In this section we discuss the degree correlations characterizing some well-known network models, together with an algorithm capable of generating networks with tunable correlations.

DEGREE CORRELATIONS IN STATIC MODELS

Erdős-Rényi Model

The random network model is neutral by definition. As it lacks hubs, it does not develop structural correlations either. Hence for the Erdős-Rényi network $k_{nn}(k)$ is given by Eq. 7.9, predicting $\mu = 0$ for any $\langle k \rangle$ and N.

Configuration Model: The configuration model SECT. 4.7 is also neutral, independent of our choice of the degree distribution p_k. This is because the model allows for both multi links and self-loops. Consequently, any conflicts caused by the hubs are relieved by multiple links between them. If, however, we force the network to be simple, then the generated network will develop structural disassortativity Fig. 7.8.

Hidden Parameter Model

In the model e_{jk} is the product of the hidden variables η_j and η_k, which are chosen randomly, hence the network is technically uncorrelated SECT. 4.8. However, if we do not allow multiple links, for scale-free networks we again observe structural disassortativity. Analytical calculations indicate that in this case $k_{nn}(k) \sim k^{-1}$, i.e. we have $\mu = -1$ [10].

DEGREE CORRELATIONS IN EVOLVING NETWORKS

To understand the emergence (and absence) of degree correlations in growing networks, let us start with the initial attractiveness model discussed in SECT. 6.4. In the model preferential attachment follows $\Pi(k) \sim A + k$, where A is the initial attractiveness Eq. 6.23. The degree correlation function depends on A, the calculations predicting three scaling regimes [15]:

(i) If γ < 3 (i.e. $-m < A < 0$ according to Eq. 6.24, we have

$$k_{nn}(k) \simeq m\frac{(m+A)^{1-\frac{A}{m}}}{2m+A}\varsigma\left(\frac{2m}{2m+A}\right)N^{-\frac{A}{2m+A}}k^{\frac{A}{m}} \tag{7.17}$$

Hence the resulting network is disassortative, $k_{nn}(k)$ being characterized by the power-law decay [15, 16]

$$k_{nn}(k) \simeq k^{-\frac{|A|}{m}} \tag{7.18}$$

(ii) If $\gamma = 3$ ($A = 0$), the initial attractiveness model reduces to the Barabási-Albert model CHAPTER 5. In this case

$$k_{nn}(k) \simeq \frac{m}{2}\ln N, \tag{7.19}$$

that is, $k_{nn}(k)$ is independent of k, hence the network is neutral.

(iii) If $\gamma > 3$ ($A > 0$), the calculations predict

$$K_{nn}(k) \simeq (m+a)\ln\left(\frac{k}{m+a}\right). \tag{7.20}$$

As $k_{nn}(k)$ increases logarithmically with k, the resulting network displays a weak assortative tendency, but does not follow the scaling Eq. 7.10.

Bianconi-Barabási Model

With a uniform fitness distribution the Bianconi-Barabási model generates a disassortative network [5] Fig. 7.12. As the randomized version of the network is also disassortative, this is a structural disassortativity. Note, however, that the real $k_{nn}(k)$ and the randomized $k_{nn}^{R\text{-}S}(k)$ do not overlap, indicating that the Bianconi-Barabási model displays some disassortativity that is not fully explained by its scale-free nature.

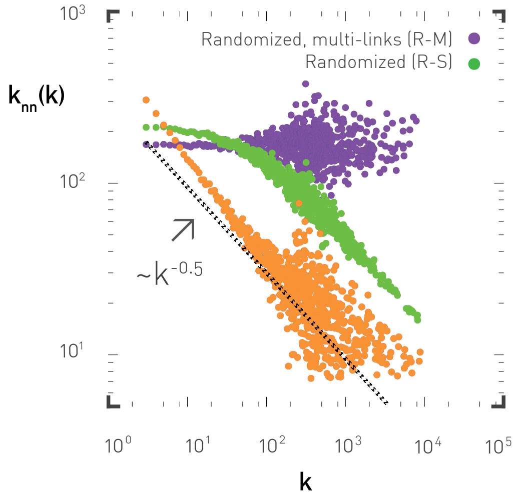


Figure 7.12
Correlations in the Bianconi-Barabási model

The degree correlation function of the Bianconi-Barabási model for $N = 10,000$, $m = 3$ and uniform fitness distribution SECT. 6.2. As the dotted line indicates, the network is disassortative, with $\mu = 0.5$. The green symbols show $k_{nn}^{R\text{-}S}(k)$, and the blue are for $k_{nn}^{R\text{-}M}(k)$. As $k_{nn}^{R\text{-}S}(k)$ also decreases, the bulk of the observed disassortativity is structural. But the difference between $k_{nn}^{R\text{-}S}(k)$ and correlations in the Bianconi-Barabási model suggests that structural effects cannot fully account for the observed degree correlation.

TUNING DEGREE CORRELATIONS

Several algorithms exist to generate networks with desired degree correlations [8, 17, 18]. Here we discuss a simplified version of the algorithm proposed by Xalvi-Brunet and Sokolov that generates maximally correlated networks with a predefined degree sequence [19, 20, 21]. It consists of the following steps Fig. 7.13a:

- **Step 1: Link selection**
  Choose at random two links. Label the four nodes at the end of these two links with *a, b, c,* and *d* such that their degrees k_a, k_b, k_c, and k_d are ordered as

$$k_a \geq k_b \geq k_c \geq k_d.$$

- **Step 2: Rewiring**
  Break the selected links and rewire them to form new pairs. Depending on the desired degree correlations the rewiring is done in two different ways:

- **Step 2A: Assortative**
  By pairing the two highest degrees (*a* with *b*) and the two lowest degrees (*c* with *d*), we are connecting nodes with comparable degrees, enhancing the network's assortative nature.

- **Step 2B, Disassortative**
  By pairing the highest and the lowest degree nodes (*a* with *d* and *b* with *c*), we tend to connect nodes with rather different degrees, enhancing the network's disassortative nature.

By iterating these steps we gradually enhance the network's assortative (2A) or disassortative (2B) features. If we aim to generate a simple network (free of multi-links), after *Step 2* we check whether the particular rewiring leads to multi-links. If it does, we reject it, returning to *Step 1*.

The correlations characterizing the networks generated by this algorithm converge to the maximal or minimal value one can reach for the given degree sequence **Fig. 7.13b**. We refer to these networks as maximally assortative or maximally disassortative. The model has no difficulty creating disassortative correlations **Figs. 7.13e, f**. In the assortative limit simple networks displays a mixed $k_{nn}(k)$: assortative for small *k* and disassortative for high *k* **Figs. 7.13b**. This is a consequence of structural cutoff: for scale-free networks the system is unable to sustain assortativity for high *k*. This behavior is reminiscent of the $k_{nn}(k)$ function observed for citation networks **Fig. 7.10j**.

The version of the Xalvi-Brunet & Sokolov algorithm discussed in **Fig. 7.13** generates maximally assortative or disassortative networks. We can tune the magnitude of the generated degree correlations if we use the original version of the proposed algorithm, discussed in **Fig. 7.14**.

In summary, static models, like the configuration or hidden parameter models, are neutral if we allow multi-links, and develop structural disassortativity if we force them to generate simple networks. To generate networks with tunable correlations, we can use for example the Xalve-Brunet & Sokolov algorithm. An important result of this section is **Eq. 7.17**, predicts the functional form of the degree correlation function for a growing network, offering analytical backing for the scaling hypothesis **Eq. 7.10**.

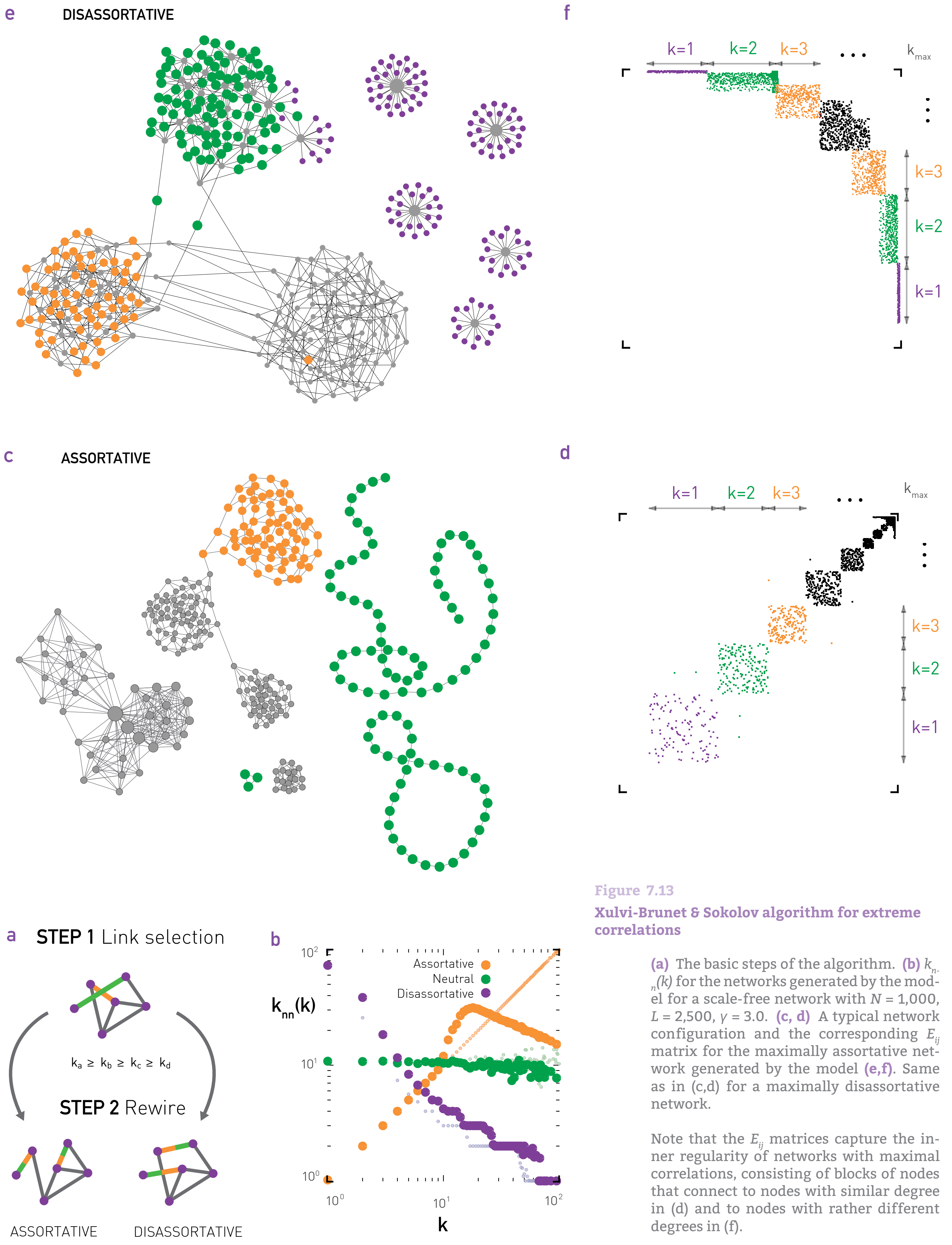


Figure 7.13

Xulvi-Brunet & Sokolov algorithm for extreme correlations

(a) The basic steps of the algorithm. **(b)** $k_{nn}(k)$ for the networks generated by the model for a scale-free network with $N = 1{,}000$, $L = 2{,}500$, $\gamma = 3.0$. **(c, d)** A typical network configuration and the corresponding E_{ij} matrix for the maximally assortative network generated by the model **(e,f)**. Same as in (c,d) for a maximally disassortative network.

Note that the E_{ij} matrices capture the inner regularity of networks with maximal correlations, consisting of blocks of nodes that connect to nodes with similar degree in (d) and to nodes with rather different degrees in (f).

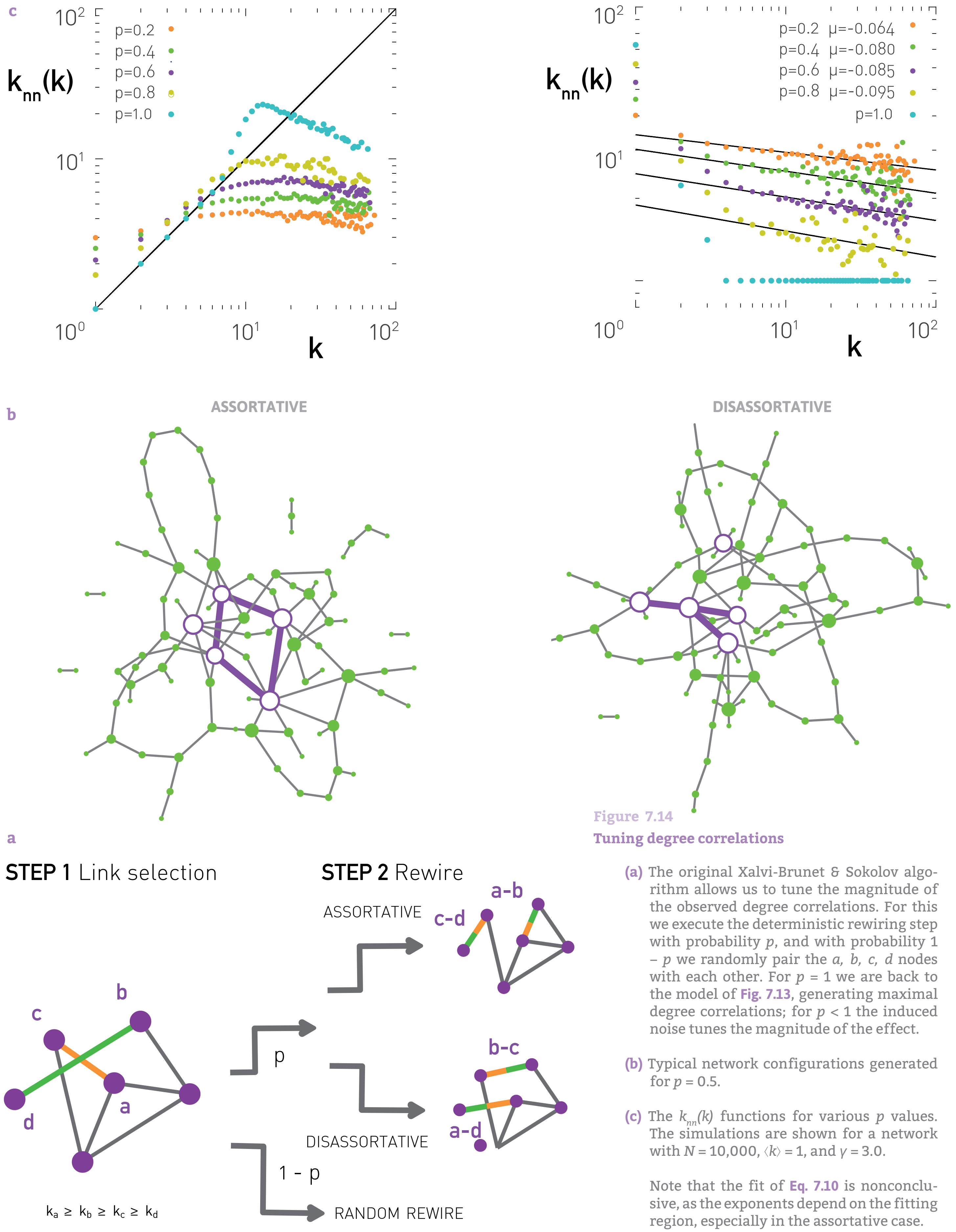


Figure 7.14

Tuning degree correlations

(a) The original Xalvi-Brunet & Sokolov algorithm allows us to tune the magnitude of the observed degree correlations. For this we execute the deterministic rewiring step with probability p, and with probability $1 - p$ we randomly pair the a, b, c, d nodes with each other. For $p = 1$ we are back to the model of **Fig. 7.13**, generating maximal degree correlations; for $p < 1$ the induced noise tunes the magnitude of the effect.

(b) Typical network configurations generated for $p = 0.5$.

(c) The $k_{nn}(k)$ functions for various p values. The simulations are shown for a network with $N = 10{,}000$, $\langle k \rangle = 1$, and $\gamma = 3.0$.

Note that the fit of **Eq. 7.10** is nonconclusive, as the exponents depend on the fitting region, especially in the assortative case.

SECTION 7.7

THE IMPACT OF DEGREE CORRELATIONS

As we have seen in SECT. 7.5, most real networks are characterized by some degree correlations. Social networks are assortative; biological networks display structural disassortativity. The presence of these correlations raise an important question: why do we care? In other words, do degree correlations alter the properties of a network? And which network properties do they influence? The purpose of this section is to briefly address these questions.

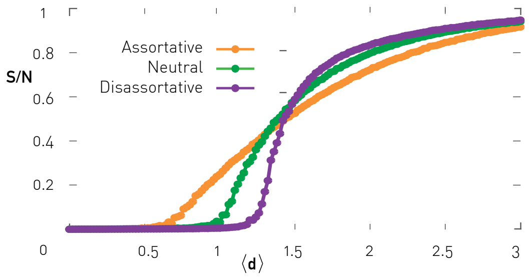


Figure 7.15
Degree correlations and the phase transition point

Relative size of the giant component for an Erdös-Rényi network of size N=10,000 (green curve), which is rewired using the Xalvi-Brunet & Sokolov algorithm with $p = 0.5$, to induce degree correlations (red and blue curve). Each point represents an average of 10 independent runs. The figure indicates that as we move from assortative to disassortative networks, the phase transition point is delayed and the size of the giant component increases for large $\langle k \rangle$.

As we have seen in SECT. 3.6, an important property of a random network is the emergence of a phase transition at $\langle k \rangle = 1$, marking the appearance of the giant component. Fig. 7.15 shows the relative size of the giant component for networks with different degree correlations, indicating that [8, 19, 20]:

- **For assortative networks**
  the phase transition point moves to a lower $\langle k \rangle$, hence a giant component emerges for $\langle k \rangle < 1$. The reason is that it is easier to create a giant component if the high-degree nodes tend to link to other high-degree ones.

- **For disassortative networks**
  the phase transition is delayed, as in these networks the hubs tend to connect to small degree nodes. Consequently, these networks have difficulty forming a giant component.

- For large $\langle k \rangle$ the giant component is smaller in assortative networks than in neutral or disassortative networks. Indeed, the high-degree nodes form a core group of high mean degree. As assortativity forces these hubs to mostly link to each other, they fail to attract to the giant component the numerous small degree nodes.

These changes in the size and the structure of the giant component have implications on the spread of diseases [22, 23, 24], a topic discussed in CHAP. 10. Indeed, as we have seen in SECT. 7.4, social networks tend to be assortative. The high degree nodes therefore form a giant component that acts as the “reservoir” for the disease, sustaining an epidemic even when on average the network is not sufficintly dense for the virus to persist.

The altered giant component has implications for network robustness as well [25]. As we discuss in CHAPTER 8, a network can be fragmented by the removal of its hubs. In assortative networks hub removal makes less damage because the hubs cluster together, forming a core group, hence many of them are redundant. The removal of the hubs is more damaging in disassortative networks, as in these the hubs connect to many small-degree nodes, which fall off the network once a hub is deleted.

Let us mention a few additional consequences of degree correlations:

- Fig. 7.16 shows the path-length distribution for a random network rewired to display different degree correlations. It indicates that in assortative networks the average path length is shorter than in neutral networks. Yet the most dramatic difference is in the network diameter d_{max}, which is significantly higher for assortative networks. Indeed, assortativity favors links between nodes with similar degree, hence it results in long chains of $k = 2$ nodes, enhancing d_{max} Fig. 7.13c.

- Degree correlations influence a system’s stability against stimuli and perturbations [26] as well as the synchronization of oscillators placed on a network [27, 28].

- Degree correlations have a fundamental impact on vertex cover problems [29], requiring us to find the minimal set of nodes such that each link is connected to at least one node in the vertex cover BOX 7.4.

- Finally, degree correlations have an impact on our ability to control a network, altering the number of input signals one needs to achieve full control [30].

In summary, degree correlations are not only of academic interest, but they alter numerous network characteristics and have a strong impact on various processes that take place on a network.

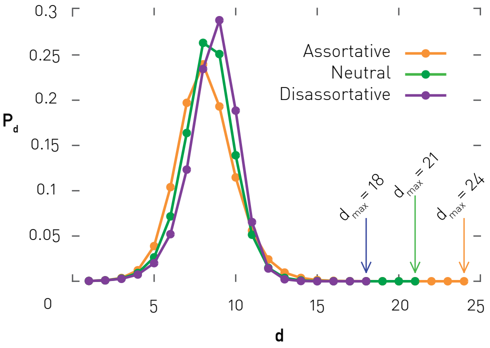


Figure 7.16
Degree correlations and path lengths

Shortest path distribution for a network with Poisson degree distribution of size N = 10, 000 and $\langle k \rangle$ = 3. Correlations are added using the Xalvi-Brunet & Sokolov algorithm with p = 0.5. Each curve presents is an average of 10 independent networks. The plots indicate that as we move from disassortative to assortative networks, the average path length decreases, but the diameter grows.

BOX 7.4

VERTEX COVER AND MUSEUM GUARDS

Imagine you are director of an open-air museum situated in a large park with numerous paths. You wish to place guards on crossroads to observe each path, but to save cost you want to use as few guards as possible. Let N be the number of crossroads and $m < N$ is the number of guards you can afford to hire. There are (Nm) ways of placing the m guards in the N positions, but most configurations will leave some paths unobserved [31].

The number of trials one needs to find a perfect solution grows exponentially with N. Indeed, this is one of the six basic NP-complete problems, called the vertex cover problem. By definition, the vertex cover of a network is a set of nodes such that each link is connected to at least one node of the set. The NP-completeness means that there is no known algorithm which can identify a vertex cover substantially faster than using as exhaustive search, i.e. checking each possible configuration individually. Obviously, the number of nodes needed to obtain a vertex cover depends on the network topology, being affected by the degree distribution and potential degree correlations [29].

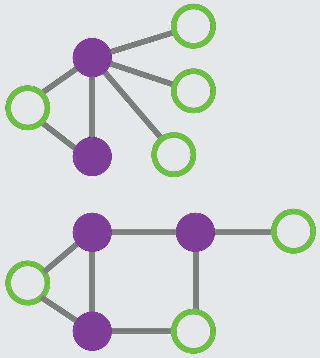


Figure 7.17
The minimum cover

Formally, a vertex cover of a network G is a set C of nodes such that each link of G connects to at least one node in C. A minimum vertex cover is a vertex cover of smallest possible size. The figure above shows examples of minimal vertex covers in two graphs, where the set C is shown in red. One can check that if we turn any of the red nodes into white nodes, we will have at least one link that does not connect to a red node.

SECTION 7.7

SUMMARY

There are at least three important reasons why we care about degree correlations:

- Degree correlations are present in most real networks SECT. 7.4.

- In the previous chapters we showed how much we can learn about a network by inspecting its degree distribution. Degree correlations force us to go beyond the degree distribution, demonstrating that there are quantifiable patters that govern the way nodes link to each other that are not captured by p_k alone.

- Once present, degree correlations change a network's behavior SECT. 7.6.

Despite the considerable effort devoted to characterizing degree correlations, our understanding of the phenomena is not yet complete. For example, while in SECT. 7.6 we showed how to tune degree correlations, the problem is far from being fully resolved. Indeed, the full degree correlations characterizing a network is contained in the e_{ij} matrix. Generating networks with an arbitrary e_{ij} remains a difficult task.

The results of this chapter allow us to formulate the next network law:

Structural Correlations

Simple scale-free networks are disassortative.

Let us inspect the validity of this law in the light of the three criteria established in CHAPTER 3:

A. Quantitative Formulation

The quantitative basis of this law is provided in SECT. 5.3 and ADVANCED TOPICS 7.B, where we derived the magnitude of the structural cutoff and the emergence of disassortative correlations beyond k_s.

B. Universality

In SECT. 7.4 we showed that many real networks, from biological to email networks, display structural disassortativity.

C. Non-random Character

As we showed in SECT. 5.3, structural disassortativity cannot appear in the random network model, as the degree of the largest node in a random network is smaller than the structural cutoff k_s.

BOX 7.5

DEGREE CORRELATIONS: BRIEF HISTORY

Degree correlations were first reported in 2001 in the context of the Internet in a classic paper by Romualdo Pastor-Satorras, Alexei Vazquez, and Alessandro Vespignani [4, 5]. This work introduced the degree correlation function $k_{nn}(k)$ and the scaling Eq. 7.10. A year later Kim Sneppen and Sergey Maslov used the full $p(k_i, k_j)$, rooted in the e_{ij} matrix, to discover the presence of degree correlations in protein-interaction networks [32]. In 2003 Mark Newman introduced the degree correlation coefficient [8, 9], allowing him to realize that two kinds of correlations can emerge in real systems. He also introduced the terminology “assortativity” and “disassortativity” to characterize this diversity. These terms have their roots in social sciences where they are used to capture mating preferences [33].

Assortative mating

reflects the tendency of individuals to date or marry individuals that are similar to them. For example, low-income individuals tend to marry low-income individuals, and college graduates marry college graduates. Network theory uses assortativity in the same spirit, capturing the degree-based similarities between nodes: in assortative networks hubs tend to connect to other hubs and small-degree nodes to small-degree nodes. In a network environment we can also encounter the traditional assortativity, when nodes of similar properties link to each other Fig. 7.18.

Disassortative mixing

when individuals link to individuals who are unlike them, is also observed in social systems. Sexual networks are perhaps the best example of this phenomena, as most sexual relationships are between individuals of different gender. Disassortative mixing is also common in economic settings. For example, trade typically takes place between individuals of different skills: the baker does not sell bread to other bakers, and the shoemaker rarely fixes other shoemaker’s shoes.

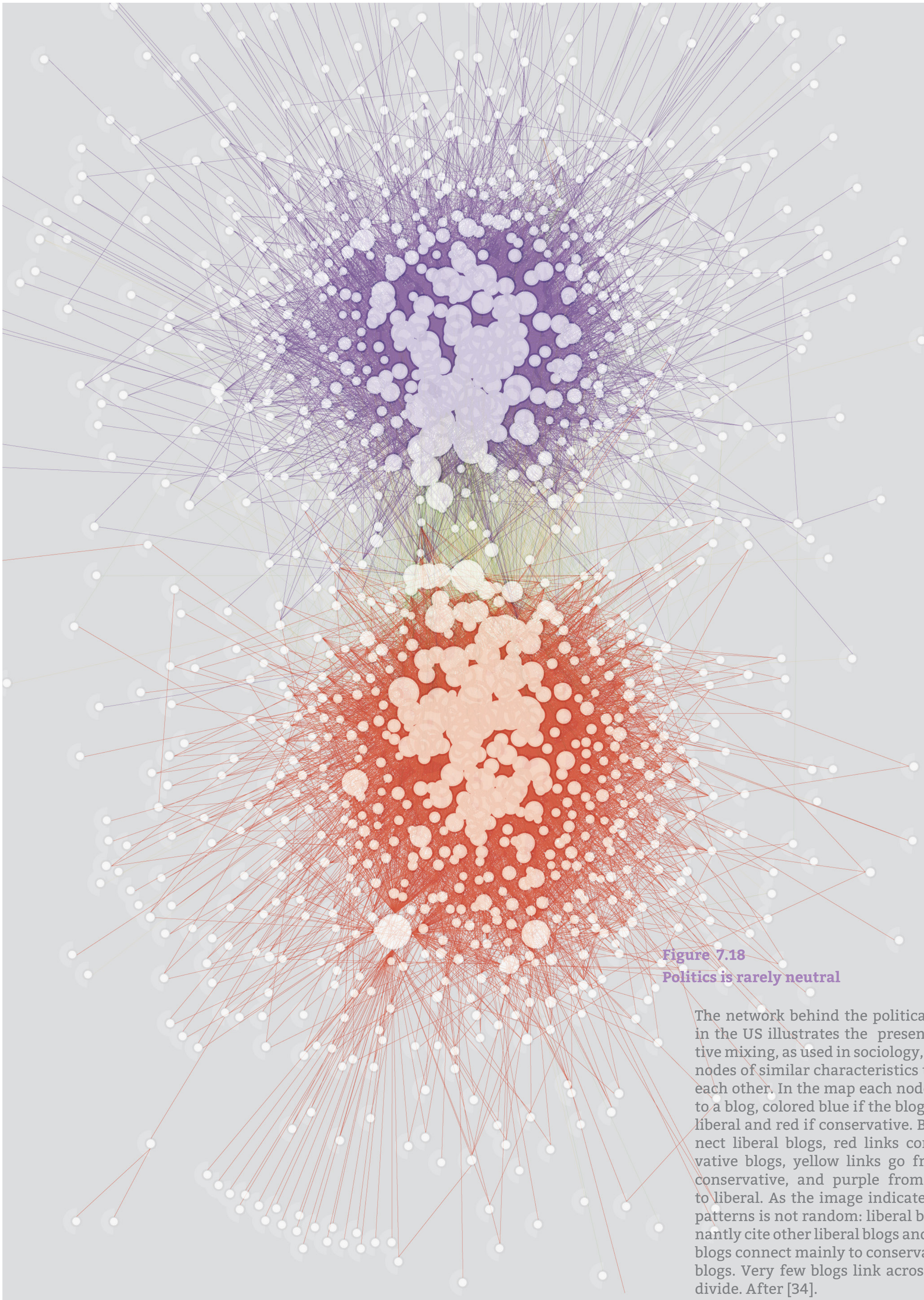


Figure 7.18
Politics is rarely neutral

The network behind the political blogosphere in the US illustrates the presence of assortative mixing, as used in sociology, meaning that nodes of similar characteristics tend to link to each other. In the map each node corresponds to a blog, colored blue if the blog is considered liberal and red if conservative. Blue links connect liberal blogs, red links connect conservative blogs, yellow links go from liberal to conservative, and purple from conservative to liberal. As the image indicates, the linkage patterns is not random: liberal blogs predominantly cite other liberal blogs and conservative blogs connect mainly to conservative-learning blogs. Very few blogs link across the political divide. After [34].

BOX 7.6

TWO-POINT, THREE-POINT CORRELATIONS

In their most general form, the degree correlations present in a network are determined by the conditional probability $P\ (k^{(1)}, k^{(2)}, ..., k^{(k)})$ that a node of degree k connects to nodes with degrees $k^{(1)}$, $k^{(2)}, ..., k^{(k)}$.

Two-point correlations

The simplest of these is the two-point degree correlation discussed in this chapter, being the conditional probability $P(k')$ that a node with degree k is connected to a node with degree k'. For uncorrelated networks this conditional probability is independent of k, hence $P(k') = k'P(k')/\langle k\rangle$ [18]. As the empirical evaluation of $P(k')$ in real networks is a cumbersome task, it is more practical to analyze the degree correlation function $k_{nn}(k)$ defined in Eq. 7.7.

Three-point correlations

In principle there is no reason to stop at two-point correlations. Correlations involving three nodes are determined by the probability $P\ (k^{(1)}, k^{(2)}\ |\ k)$ that a node with degree k is connected to nodes with degrees $k^{(1)}$ and $k^{(2)}$. This conditional probability determines the clustering coefficient Eq. 7.20. Indeed, the average clustering coefficient $C(k)$ of nodes with degree k [22, 23] can be formally written as the probability that a node of degree k is connected to nodes with degrees $k^{(1)}$ and $k^{(2)}$, and that those two are joined by a link, averaged over all the possible values of $k^{(1)}$ and $k^{(2)}$,

$$C(k) = \sum_{k^{(1)},k^{(2)}} P(k^{(1)},k^{(2)}\mid k)\,p^{k}_{k^{(1)},k^{(2)}}\ ,$$

where $p^k_{k^{(1)},\,k^{(2)}}$ is the probability that nodes $k^{(1)}$ and $k^{(2)}$ are connected, provided that they have a common neighbor with degree k [18]. For neutral networks the clustering coefficient is independent of k, following

$$C(k) = \frac{\left(\langle k^2\rangle - \langle k\rangle\right)^2}{\langle k\rangle^3 N}\ .$$

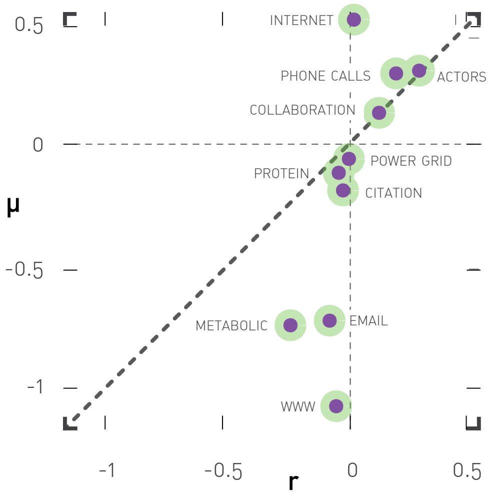


Figure 7.19
Correlation between r and N

To illustrate the relationship between r and μ, we estimated μ by fitting the knn function to Eq. 7.10, whether or not the power law scaling was statistically significant Fig. 7.8.

ADVANCED TOPICS 7.A
DEGREE CORRELATION COEFFICIENT

In BOX 7.2 we defined the degree correlation coefficient r as an alternative measure of degree correlations characterizing a network [8, 9]. The use of a single number to characterize degree correlations is extremely attractive, as it also offers an easy way to compare the correlations observed in networks of different nature and size. Yet, before we use r we must be aware of some of its limitations.

The hypothesis behind the correlation coefficient r is that the $k_{nn}(k)$ function can be approximated by the linear function

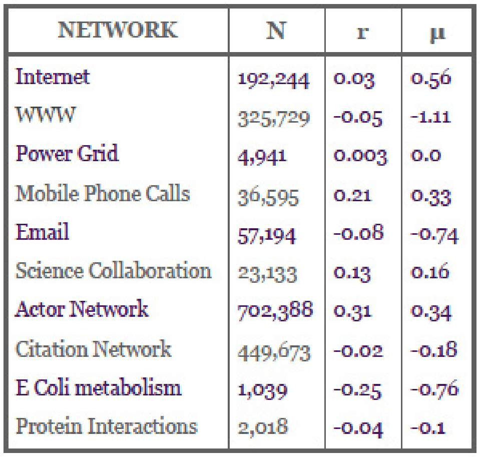
$$k_{nn}(k) \sim rk. \tag{7.21}$$

This is different from the scaling Eq. 7.10, which assumes a power law dependence on k. Eq. 7.21 raises several important issues:

- The linear dependence Eq. 7.21 is not supported by empirical data, numerical simulations, or analytical calculations. Indeed, analytical calculation of the initial attractiveness model predict a power law Eq. 7.18 or a logarithmic k-dependence Eq. 7.20 for the degree correlation function. Therefore, r forces a linear fit to an inherently nonlinear function. This discrepancy is illustrated in Fig. 7.20, which shows that for assortative and disassortative networks Eq. 7.21 offers a poor to the data.

- As we have seen in Fig. 7.10, the dependence of $k_{nn}(k)$ on k is rather complex, often changing trends for large k thanks to the structural cutoff. A linear fit ignores this inherent complexity. To illustrate the consequences of this phenomena, we calculated r and μ for the ten reference networks TABLE 7.1. The results are plotted in Fig. 7.19, indicating that while μ and r correlate for positive r, this correlation breaks down for negative r.

- As we discuss in BOX 7.8, the maximally correlated model has a vanishing r for large N, despite the fact that the network maintains its



NETWORK	N	r	μ
Internet	192,244	0.03	0.56
WWW	325,729	-0.05	-1.11
Power Grid	4,941	0.003	0.0
Mobile Phone Calls	36,595	0.21	0.33
Email	57,194	-0.08	-0.74
Science Collaboration	23,133	0.13	0.16
Actor Network	702,388	0.31	0.34
Citation Network	449,673	-0.02	-0.18
E Coli metabolism	1,039	-0.25	-0.76
Protein Interactions	2,018	-0.04	-0.1

Table 7.1

Degree correlations in reference networks

The table shows r and μ for the ten reference networks of TABLE 4.1. Directed networks were made undirected to measure r and μ. Alternatively, we can use the directed correlation coefficient to characterize such directed networks BOX 7.8.

degree correlations. This suggests that the degree correlation coefficient r has difficulty detecting correlations characterizing large networks.

RELATIONSHIP BETWEEN μ AND r

If $k_{nn}(k)$ follows the scaling Eq. 7.10, then the sign of the degree coefficient r should agree with the sign of μ. This is supported by Fig. 7.20 as well. To show the origin of this behavior, next we derive a direct relationship between μ and r. To be specific we assume the validity of Eq. 7.10 and determine the value of r for a network with a given correlation exponent μ.

We start by determining a from from Eq. 7.10. We can write the second moment of the degree distribution as

$$\langle k^2\rangle = \langle k_{nn}(k)k\rangle = \sum_{k'} ak^{\mu+1}p_k = a\langle k^{\mu+1}\rangle,$$

which leads to

$$a = \frac{\langle k^2\rangle}{\langle k^{\mu+1}\rangle}.$$

We now calculate r for a network with a given μ:

$$r = \frac{\sum_{k'} kak^{\mu}q_k - \frac{\langle k^2\rangle^2}{\langle k\rangle^2}}{\sigma_r^2} = \frac{\sum_{k'} a.k^{\mu+2}\frac{p_k}{\langle k\rangle} - \frac{\langle k^2\rangle^2}{\langle k\rangle^2}}{\sigma_r^2} = \frac{\frac{\langle k^2\rangle}{\langle k^{\mu+1}\rangle}\frac{\langle k^{\mu+2}\rangle}{\langle k\rangle} - \frac{\langle k^2\rangle^2}{\langle k\rangle^2}}{\sigma_r^2} = \tag{7.22}$$
$$= \frac{1}{\sigma_r^2}\frac{\langle k^2\rangle}{\langle k\rangle}\left(\frac{\langle k^{\mu+2}\rangle}{\langle k^{\mu+1}\rangle} - \frac{\langle k^2\rangle}{\langle k\rangle}\right).$$

For $\mu = 0$ the term in the last parenthesis vanishes, obtaining $r = 0$. Hence if $\mu = 0$ (neutral network), the network will be neutral based on r as well. For $k > 1$ Eq. 7.22 suggests that for $\mu > 0$ the parenthesis is positive, hence $r > 0$, and for $\mu < 0$ is negative, hence $r < 0$. Therefore, r and μ predict degree correlations of similar kind.

Therefore, if the degree correlation function follows Eq. 7.10, then the sign of the degree correlation exponent μ will determine the sign of the assortativity coefficient r:

$$\mu < 0 \rightarrow r < 0$$
$$\mu = 0 \rightarrow r = 0$$
$$\mu > 0 \rightarrow r > 0.$$

In summary, the degree correlation coefficient assumes that $k_{nn}(k)$ scales linearly with k, a hypothesis that lacks numerical and analytical support. Hence r forces a linear fit to $k_{nn}(k)$, giving occasionally rise to inconsistent results. While typically the sign of r and μ agree, overall r does not offer a natural characterization of the underlying degree correlations. An accurate characterization starts with e_{ij}, whose behavior is reasonably captured by $k_{nn}(k)$.

BOX 7.7

AT A GLANCE: DEGREE CORRELATIONS

Degree Correlation Matrix e_{ij}

probability of finding a node with degrees i and j at the two ends of a link.

Neutral networks:

$$e_{ij} = q_i q_i = \frac{k_i p_{k_i} k_j p_{k_j}}{\langle k\rangle^2}$$

Degree Correlation Function

$$k_{nn}(k) = \sum_{k'} k'\,p(k'|k)$$

Neutral networks:

$$k_{nn}(k) = \frac{\langle k^2\rangle}{\langle k\rangle}$$

Scaling Hypothesis

$$k_{nn}(k) \sim k^{\mu}$$

$\mu > 0$: *Assortative*
$\mu = 0$: *Neutral*
$\mu < 0$: *Dissasortative*

Degree Correlation Coefficient

$$r = \sum \frac{jk(e_{jk} - q_j q_k)}{\sigma_r^2}$$

$r > 0$: *Assortative*
$r = 0$: *Neutral*
$r < 0$: *Dissasortative*

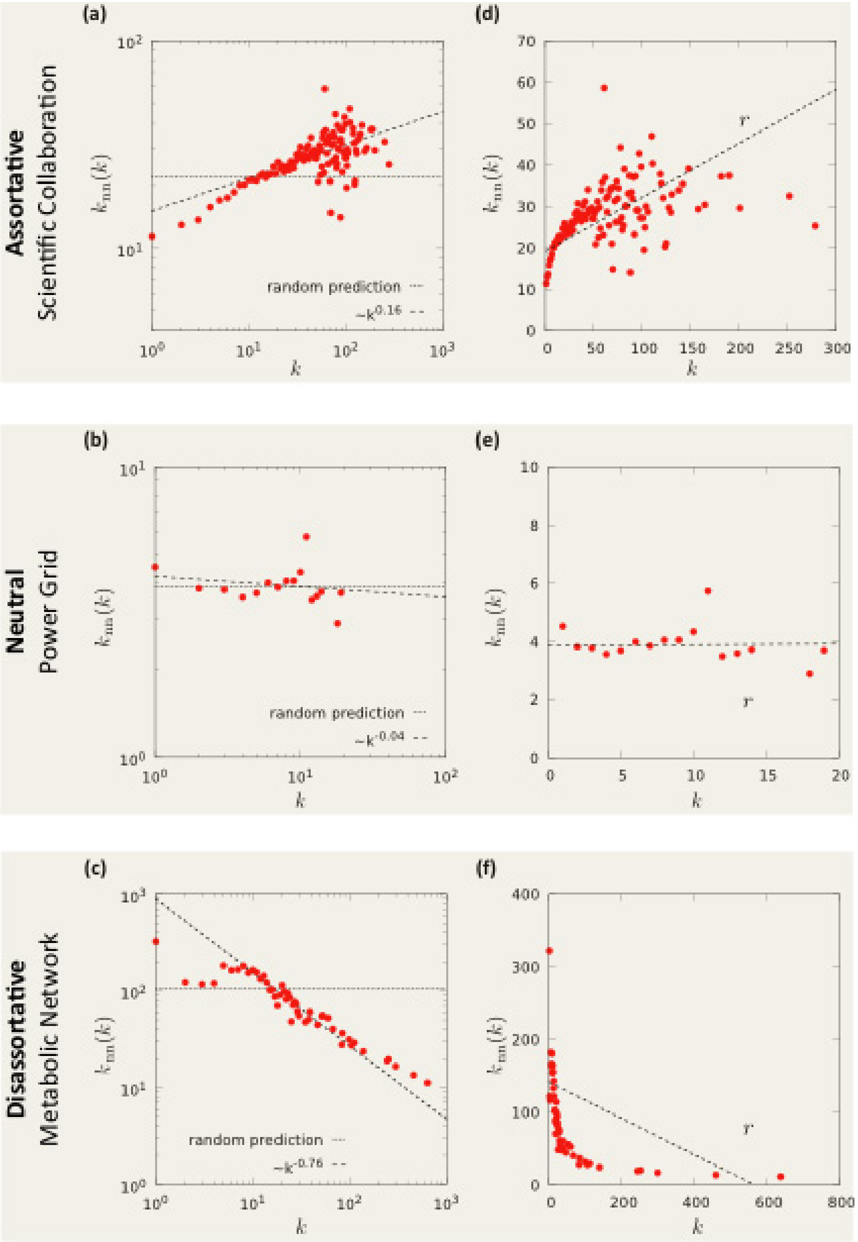


Figure 7.20

Degree correlation function

The degree correlation function knn for three real networks. The left panels show the cumulative function $k_{nn}(k)$ on a log-log plot to test the validity of Eq. 7.10. The right panels show $k_{nn}(k)$ on a *lin–lin* plot to test the validity of Eq. 7.21, i.e. the assumption that $k_{nn}(k)$ depends linearly on *k*, the hypothesis behind the correlation coefficient *r*. The slope of the dotted line corresponds to the correlation coefficient *r*. As the lin-lin plots illustrate, Eq. 7.21 offers a poor fit for assortative (d) and disassortative (f) networks.

BOX 7.9

CORRELATION COEFFICIENT FOR DIRECTED NETWORKS

To measure correlations in directed networks we must take into account that each node i is characterized by an incoming k_i^{in} and an outgoing k_i^{out} degree. Hence, we can define four degree correlation coefficients, $r_{in,\,in}$, $r_{in,out}$, $r_{out,in}$, $r_{out,\,out}$ capturing all possible combinations between the incoming and outgoing degrees of two nodes linked to each other Figs. 7.12 a-d. Formally we have [14].

$$r_{\alpha,\beta} = \frac{\sum_{jk} jk(e_{jk}^{\alpha,\beta} - q_j^{\alpha} q_k^{\beta})}{\sigma^{\alpha}\sigma^{\beta}}, \qquad (7.23)$$

where α and β refer to the in and out indices. To illustrate the use of Eq. 7.23, we show in Fig. 7.21e the four correlation coefficients for the five directed reference networks TABLE 7.1. For a complete characterization of degree correlations, it is desirable to measure the four $k_{nn}(k)$ functions as well BOX 7.2.

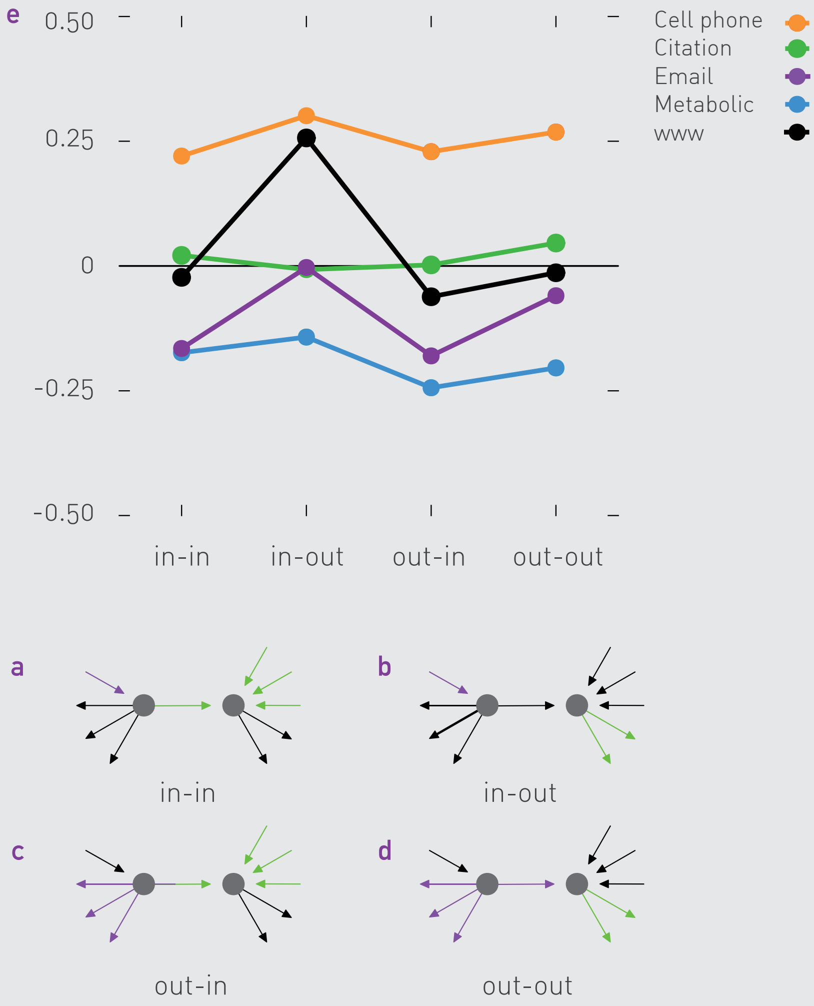


Figure 7.21

Panels (a)-(d) illustrate in red and green the (α, β) indices that define the appropriate correlation coefficient for directed ne tworks. (e) The correlation profile of the five directed reference networks, indicating, for example, that while citation networks have negligible correlations, all four correlation coefficients document strong assortative behavior for cell phone calls and strong disassortative behavior for metabolic networks. The case of the WWW is particularly interesting: while three of its correlation coefficients are close to zero, there is a strong assortative tendency for the (in, out) combinations.

SECTION 7.9

ADVANCED TOPICS 7.B
STRUCTURAL CUTOFFS

As discussed in SECT. 7.3, there is a fundamental conflict between the scale-free property and degree correlations, which leads to a structural cutoff in simple networks. In this section we derive Eq. 7.16, providing the system size dependence of the structural cutoff [11]. We start by defining

$$r_{kk'} = \frac{E_{kk'}}{m_{kk'}}, \tag{7.24}$$

where $E_{kk'}$ is the number of links between nodes of degrees k and k', and

$$m_{kk'} = \min\left\{kN_k, k'N_k, N_kN_{k'}\right\} \tag{7.25}$$

is the largest possible value of $E_{kk'}$. If multiple links are allowed, $m_{kk'}$ is simply $m_{kk'} = min\{k\ N_{k'}, k'N_{k'}, N_kN_{k'}\}$. The origin of Eq. 7.25 is explained in Fig. 7.22. Consequently, we can write the $r_{kk'}$ ratio as

$$r_{kk'} = \frac{E_{kk'}}{m_{kk'}} = \frac{\langle k\rangle P(k,k')}{\min\left\{kP(k), k'P(k'), NP(k)P(k')\right\}}. \tag{7.26}$$

As $m_{kk'}$ is the maximum of $E_{kk'}$, $r_{kk'}$ must be smaller than or equal to one for any k and k'. Yet, for some networks and for some k, k' pairs $r_{kk'}$ becomes larger than one. This is clearly non-physical and signals some conflict in the network configuration. Strictly speaking, in simple networks degree pairs for which $r_{kk'} > 1$ cannot exist. Hence, we define the structural cut off k_s as the solution of the equation

$$r_{k_sk_s} = 1. \tag{7.27}$$

Note that as soon as $k > NP(k')$ and $k' > NP(k)$, the effects of the restriction on the multiple links are already felt, turning the expression for $r_{kk'}$ into

$$r_{kk'} = \frac{\langle k\rangle P(k,k')}{Np_kp_{k'}}. \tag{7.28}$$

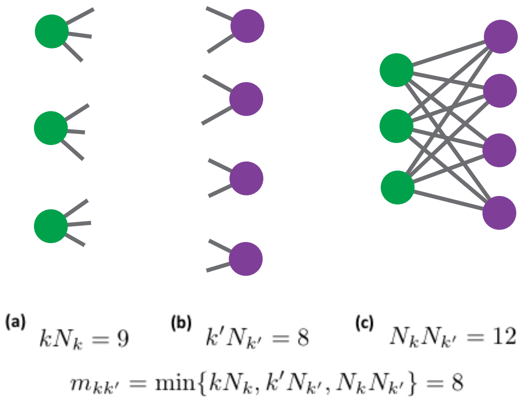


Figure 7.22

Correlation between *r* and *N*

Illustrating the maximum number of links one can have between two groups of nodes. The figure shows two groups of nodes, with degree k=3 and k'=3. The total number of links between these two groups must not exceed

(a) The total number of links available in k=3 group, which is kN_k=9;

(b) The total number of links available in k'=2 group, which is $k'N_{k'}$=8;

(c) The total number of links one can potentially have between the two groups, which is $N_kN_{k'}$.

In the example shown above the smallest of the three is $k'N_k$ = 8 of (b). The resulting configuration is shown on the top right. One can see that in this configuration, one link in the k=3 class remains unpaired.

For scale-free networks these conditions are fulfilled in the region $k, k' > (aN)^{1/(\gamma+1)}$, where a is a constant that depends on the function p_k. Note that this value is below the natural cut off. As a consequence, this scaling provides a lower bound for the structural cut off, in the sense that whenever the cut off of the degree distribution falls below this limit, the condition $r_{kk'} < 1$ is always satisfied.

For neutral networks the joint distribution factorizes as

$$P(k,k') = \frac{kk' p_k p_{k'}}{\langle k \rangle^2} . \qquad (7.29)$$

Hence, the ratio $r_{kk'}$ of Eq. 7.28 takes the form

$$r_{kk'} = \frac{kk'}{\langle k \rangle N} . \qquad (7.30)$$

Therefore, the structural cutoff needed to preserve the condition $r_{kk'} \leq 1$ has the form [35, 36, 37]

$$k_s(N) \sim (\langle k \rangle N)^{1/2} , \qquad (7.31)$$

which is Eq. 7.16. Note that Eq. 7.31 is independent of the degree distribution of the underlying network. Consequently, for a scale-free network $k_s(N)$ is independent of the degree exponent γ.

BIBLIOGRAPHY

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, RS Judson, JR Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623–627, 2000.

[2] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, D. Eisenberg DIP: the database of interacting proteins. Nucleic Acids Res. 28: 289–29, 2000.

[3] H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Nature 411: 41-42, 2001.

[4] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet. Phys. Rev. Lett. 87: 258701, 2001.

[5] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of Internet. Phys. Rev. E 65: 066130, 2002.

[6] S.L. Feld. Why your friends have more friends than you do. American Journal of Sociology 96: 1464–1477, 1991.

[7] E.W. Zuckerman and J.T. Jost. What makes you think you're so popular? Self evaluation maintenance and the subjective side of the "friendship paradox". Social Psychology Quarterly 64: 207–223, 2001.

[8] M. E. J. Newman. Assortative mixing in networks, Phys. Rev. Lett. 89: 208701, 2002.

[9] M. E. J. Newman. Mixing patterns in networks. Phys. Rev. E 67: 026126, 2003.

[10] S. Maslov, K. Sneppen, and A. Zaliznyak, Pattern detection in complex networks: Correlation profile of the Internet, e-print cond-mat/0205379, 2002.

[11] M. Boguna, R. Pastor-Satorras, A. Vespignani. Cut-offs and finite size effects in scale-free networks. Eur. Phys. J. B 38: 205, 2004.

[12] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks, arXiv:cond-mat/0305612v1.

[13] M. McPherson, L. Smith-Lovin, J. M. Cook. Birds of a feather: homophily in social networks. Annual Review of Sociology 27:415-444, 2001.

[14] J. G. Foster, D. V. Foster, P. Grassberger, M. Paczuski. Edge direction and the structure of networks. PNAS 107: 10815, 2010.

[15] A. Barrat and R. Pastor-Satorras. Rate equation approach for correlations in growing network models. Phys. Rev. E 71, 036127, 2005.

[16] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. Adv. Phys. 51: 1079, 2002.

[17] J. Berg and M. Lässig. Correlated random networks. Phys. Rev. Lett. 89: 228701, 2002

[18] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. Phys. Rev. E 68: 036112, 2003.

[19] R. Xulvi-Brunet and I. M. Sokolov. Reshuffling scale-free networks: From random to assortative. Phys. Rev. E 70: 066102, 2004.

[20] R. Xulvi-Brunet and I. M. Sokolov. Changing correlations in networks: assortativity and dissortativity. Acta Phys. Pol. B 36: 1431, 2005.

[21] J. Menche, A. Valleriani, and R. Lipowsky. Asymptotic properties of degree-correlated scale-free networks. Phys. Rev. E 81: 046103, 2010.

[22] V. M. Eguíluz and K. Klemm. Epidemic threshold in structured scale-free networks. Phys. Rev. Lett. 89:108701, 2002.

[23] M. Boguñá and R. Pastor-Satorras. Phys. Epidemic spreading in correlated complex networks. Rev. E 66: 047104, 2002.

[24] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. Phys. Rev. Lett. 90: 028701, 2003.

[25] A. Vázquez and Y. Moreno. Resilience to damage of graphs with degree correlations. Phys. Rev. E 67: 015101R, 2003.

[26] S.J.Wang,A.C.Wu,Z.X.Wu,X.J.Xu,andY.H.Wang. Response of degree-correlated scale-free networks to stimuli. Phys. Rev. E 75: 046113, 2007.

[27] F. Sorrentino, M. Di Bernardo, G. Cuellar, and S. Boccaletti. Synchronization in weighted scale-free networks with degree–degree correlation. Physica D 224: 123, 2006.

[28] M. Di Bernardo, F. Garofalo, and F. Sorrentino. Effects of degree correlation on the synchronization of networks of oscillators. Int. J. Bifurcation Chaos Appl. Sci. Eng. 17: 3499, 2007.

[29] A. Vazquez and M. Weigt. Computational complexity arising from degree correlations in networks, arXiv:cond-mat/0207035, 2002.

[30] M. Posfai, Y Y. Liu, J-J Slotine. A.-L. Barabási. Effect of correlations on network controllability, Scientific Reports 3: 1067,2013.

[31] M. Weigt and A. K. Hartmann. The number of guards needed by a museum: A phase transition in vertex covering of random graphs. Phys. Rev. Lett. 84: 6118, 2000.

[32] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. Science 296: 910–913, 2002.

[33] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: homophily in social networks. Annual Review of Sociology 27: 415–444, 2001.

[34] L. Adamic and N. Glance, The political blogosphere and the 2004 U.S. election: Divided they blog (2005).

[35] J. Park and M. E. J. Newman. The origin of degree correlations in the Internet and other networks. Phys. Rev. E 66: 026112, 2003.

[36] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. Annals of Combinatorics, 6: 125, 2002.

[37] Z. Burda and Z. Krzywicki. Uncorrelated random networks. Phys. Rev. E 67: 046118, 2003.

CHAPTER 8

NETWORK ROBUSTNESS

Introduction

Percolation theory 1

Robustness of scale-free networks 2

Attack tolerance 3

Cascading failures 4

Modeling cascading failures 5

Building Robustness 6

Summary

ADVANCED TOPICS 8.A
Random networks and percolation

ADVANCED TOPICS 8.B
Malloy-Reed Criteria

ADVANCED TOPICS 8.C
Critical threshold under random failures

ADVANCED TOPICS 8.D
Breakdown of a finite scale-free network

ADVANCED TOPICS 8.E
Threshold under attack

ADVANCED TOPICS 8.F
Modeling cascading failures

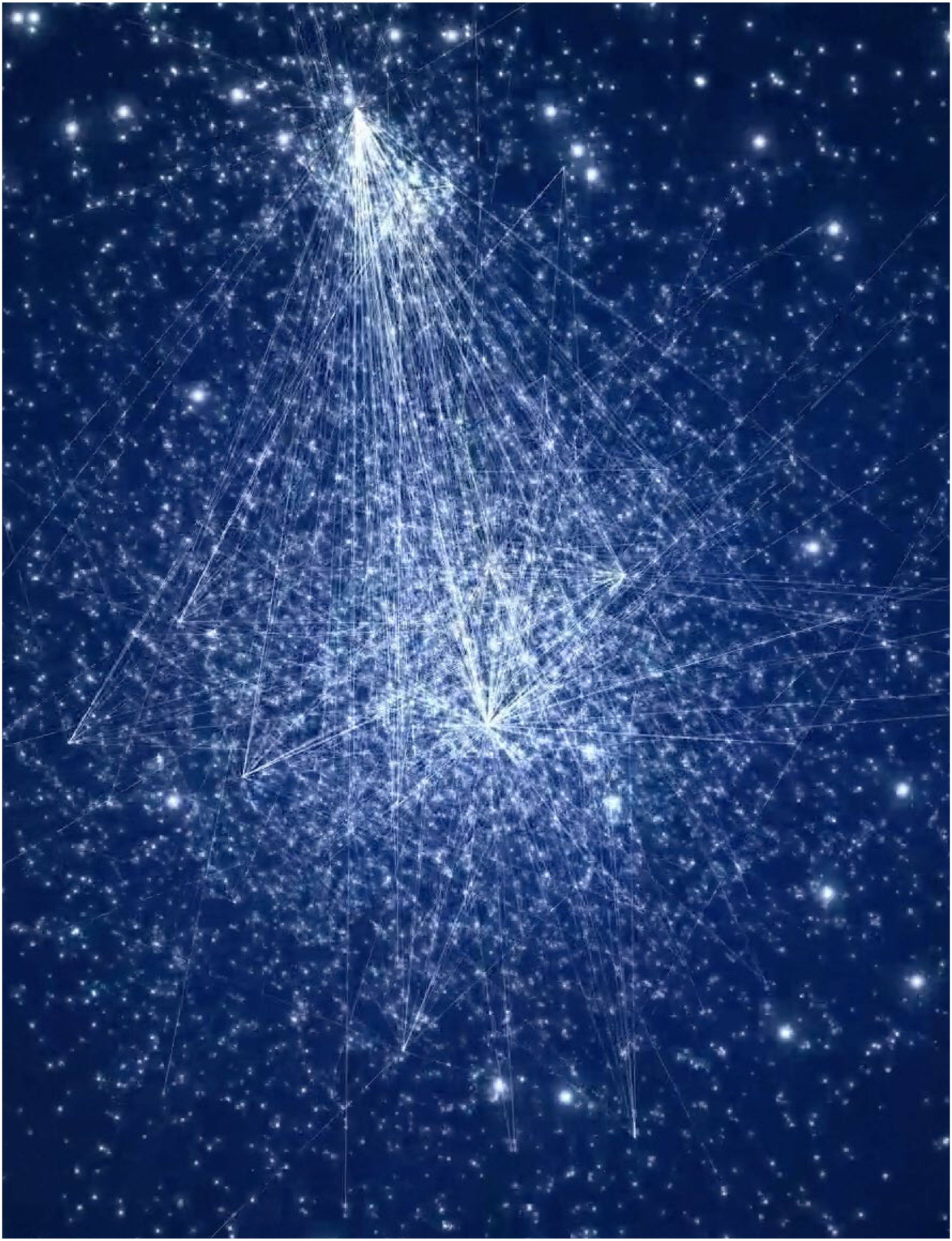
ADVANCED TOPICS 8.G
Attack and error tolerance of real networks

ADVANCED TOPICS 8.H
The optimal degree distribution

Homework

Bibliography

Figure 8.0 (front cover)
Network representation by Mauro Martino



SECTION 8.0

INTRODUCTION

Errors and failures can corrupt all human designs: the failure of a single component in your car's engine may force you to call for a tow truck or a wiring error in your computer chip can make your computer useless. Many natural and social systems, however, have a unique ability to sustain their basic functions even when several of their components fail. Indeed, there are countless protein misfolding errors or missed reactions in our cells, often without noticeable consequences; large organizations can function despite numerous missing employees. Understanding the origins of this robustness is important for many disciplines:

- Robustness is a central question in biology, which aims to understand how a cell or an organism functions under frequent internal errors and why some errors lead to diseases.

- It is of concern for social scientists and economists, who explore the stability of human societies and organizations in the face of famine, war, and changes in social and economic order.

- It is a key issue for ecologists and environmental scientists, who seek to estimate the chances that an ecosystem survives when faced with the disruptive effects of human activity.

- It is the ultimate goal in engineering, aiming to design communication systems, cars, or airplanes that can maintain a high readiness despite occasional component failures.

These biological, social and technological systems share a common feature: their functionality and robustness is guaranteed by densely interlinked networks. Indeed, cellular functions are encoded by intricate regulatory and metabolic networks; the society's resilience cannot be divorced from the interwoven social, professional, and communication web behind it; economic stability is guarded by a delicate network of financial and regulatory organizations; an ecosystem's survivability cannot be understood without a careful analysis of the food webs that sustain each species.

nature
Achilles' heel of the Internet
Obesity Mice that eat more but weigh less
Ocean anoxic events Not all at sea
Cell signalling Fringe sweetens Notch

Figure 8.1
Achilles' Heel of Complex Networks

The cover of the 27 July, 2000 issue of *Nature*, highlighting the paper entitled *Attack and error tolerance of complex networks* that sparked the interest in network robustness [1].

Whenever nature seeks robustness, it resorts to networks to achieve it.

The purpose of this chapter is to explore the role networks play in ensuring the robustness of a complex system. We show that understanding the structure of the underlying network is essential if we want to quantify a system's ability to survive random failures or deliberate attacks. We also explore the role of these networks in the emergence of cascading failures, a damaging phenomenon frequently encountered in real systems. Most importantly, we show that the laws governing the error and attack tolerance of complex networks and the emergence of cascading failures are universal.

Figure 8.2
Robust Robustness

"Robust" comes from the latin Quercus Robur, meaning oak, the symbol of strength and longevity in the ancient world.

PERCOLATION THEORY

Robustness requires us to understand the impact of node or link removal on the integrity of a network.The removal of a single node typically has only limited impact on a network's integrity Fig. 8.3a. The removal of multiple nodes, however, can break a network into isolated, non-communicating subgraphs Fig. 8.3c, d. Obviously, the more nodes we remove, the higher are the chances that we damage a network, prompting us to ask: How many nodes do we have to delete to fragment a network into isolated components? For example, what fraction of Internet routers must break down so that the Internet turns into isolated clusters of computers that are unable to communicate with each other? To answer these questions, we first introduce some concepts of percolation theory that offers the mathematical underpinnings of the network robustness problem.

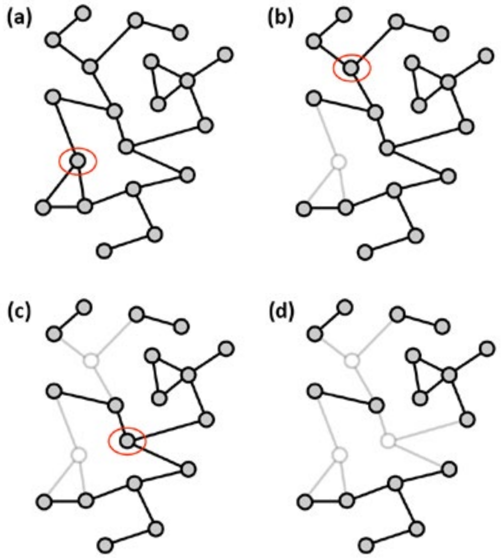


Figure 8.3
The impact of node removal

The gradual fragmentation of a small network following the breakdown of several nodes. In each panel we remove a new node (highlighted), together with its links. As the sequence of images indicates, while the removal of the first node has only limited impact on the network's integrity, the removal of the second node isolates two small clusters from the rest of the network and the removal of the third node fragments the network, breaking it into five non-communicating clusters of sizes s = 2, 2, 2, 5, 6.

PERCOLATION

Percolation theory is a highly developed subfield of statistical physics and mathematics [2, 3, 4]. A typical problem addressed by it is illustrated in Fig. 8.4, showing a square lattice with pebbles placed with probability p at each intersection. Pebbles next to each other are considered connected, forming clusters of size two or more. Given that the position of each pebble is decided by chance, we ask:

- What is the size of the largest cluster?
- What is the average cluster size?

Obviously, the higher is p, the larger are the individual clusters. Percolation theory predicts, however, that the cluster size does not change continuously with p. Rather, for a wide range of p values the lattice is populated with numerous tiny clusters Fig. 8.4a. If we increase p beyond a critical value p_c, these small clusters grow rapidly until a single large cluster emerges rather suddenly. We call this the percolating cluster as it percolates through the lattice by reaching its ends. In other words, at p_c we observe a phase transition from many small clusters to a percolating cluster that spans the whole lattice Fig. 8.4b.

To quantify the nature of this phase transition, we focus on several frequently measured quantities:

- The average cluster size, $\langle s \rangle$, represents the average size of all finite clusters observed for a given p. Percolation theory predicts that in the vicinity of p_c it follows

$$\langle s \rangle \sim |p - p_c|^{-\gamma_p} . \qquad (8.1)$$

In other words, the average cluster size diverges as we approach p_c Fig. 8.4c.

The *order parameter*, p_∞, represents the probability that a randomly chosen pebble belongs to the largest cluster. In the vicinity of the critical point p_∞ follows

$$P_\infty \sim (p - p_c)^\beta . \qquad (8.2)$$

Hence, as p decreases towards p_c the probability that a pebble belongs to the largest cluster drops zero Fig. 8.4d.

- The *correlation length*, ς, represents the mean distance between two pebbles that belong to the same cluster. In the vicinity of p_c it follows

$$\zeta \sim |p - p_c|^{-\nu} . \qquad (8.3)$$

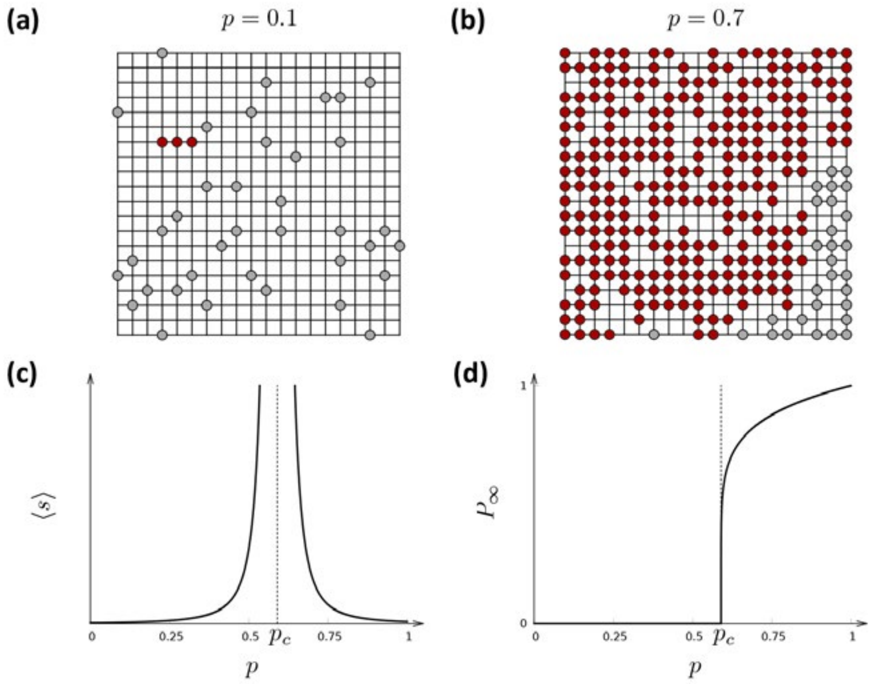


Figure 8.4
Percolation

A classical problem in percolation theory explores the random placement with probability p of pebbles on a square lattice.

(a) For small p most peebles are isolated. In this case the largest cluster has only three nodes, shown in red.

(b) For large p most (but not all) pebbles belong to a single giant component, colored red. This giant component is called the percolating cluster, as it spans the whole lattice (See also Fig. 8.6).

(c) The average cluster size, $\langle s \rangle$, in function of p. As we approach pc from below, the numerous small clusters coalesce and $\langle s \rangle$ diverges. The same divergence is observed above pc, where to measure $\langle s \rangle$ we remove the largest component from the average. The plot shows schematically the divergence of $\langle s \rangle$ as described by Fig. 8.1. The same exponent γ_p characterizes the divergence at both sides of the critical point.

(d) The p–dependence of the probability p_∞ that a pebble belongs to the largest connected component. For $p < p_c$ all components are small, so p_∞ is effectively zero. Once p reaches p_c a giant component emerges. Consequently beyond p_c there is a finite probability that a node belongs to the largest component, as predicted by Fig. 8.2.

While at $p < p_c$ the distance between the peebles in the same cluster is finite, at p_c the correlation length diverges. Therefore, at p_c, the linear size of the largest cluster becomes infinite, which is the reason it percolates the whole lattice.

The exponents γ_p, β, and ν are called critical exponents, as they characterize the system's behavior near the critical point p_c. Percolation theory predicts that these exponents are universal, meaning that they are independent of the nature of the lattice on which we place the peebles or the precise value of p_c. Therefore, if we place the peebles on a triangular or a hexagonal lattice, the behavior of $\langle s \rangle$, P_∞, and ζ is characterized by the same γ_p, β, and ν exponents. Consider the following examples to better understand this universality:

- The exponents depend only on the dimension of the lattice. In two dimensions, the case illustrated in Fig. 8.4, we have $\gamma_p = 43/18$, $\beta = 5/36$, $\nu = 4/3$, while in three dimensions $\gamma_p = 1.80$, $\gamma = 0.41$, $\gamma = 0.88$. For $d > 6$ we have $\gamma_p = 1$, $\gamma = 1$, $\gamma = 1/2$ [2], i.e. beyond $d = 6$ the exponents are independent of d.

- The value of p_c is not universal, as it depends on the lattice type. For example, for a two-dimensional square lattice Fig. 8.4 we have $p_c \approx 0.593$, while for two-dimensional triangular lattice we have $p_c = 1/2$ (site percolation).

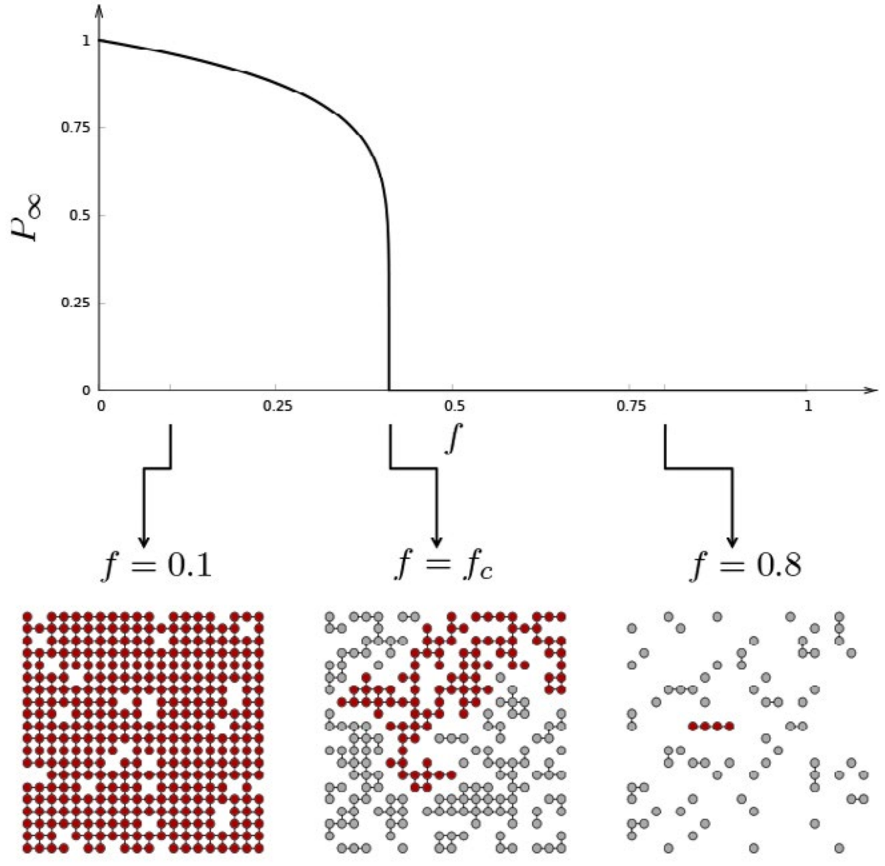
- The value of p_c also changes with the dimension: for a square lattice we have $p_c \approx 0.593$ (d = 2); for a simple cubic lattice ($d = 3$) we have $p_c \approx 0.3116$. Therefore, in $d = 3$, we need to cover a smaller fraction of the nodes with pebbles to reach the percolation transition.

ROBUSTNESS AS AN INVERSE PERCOLATION TRANSITION

We can use percolation theory to describe the impact of node failures in networks, the phenomena of primary interest in robustness. For this we view a square lattice as a network whose nodes are the intersections Fig. 8.5. Next, we randomly remove an f fraction of nodes, asking how their absence impacts the integrity of the lattice. If f is small, the few missing nodes do little damage to the network. Increasing f, however, can remove chunks of nodes from the giant component. Finally, for sufficiently large f the giant component breaks into tiny disconnected components.

Once again, the fragmentation process is not gradual, but it is characterized by a critical threshold f_c: for $f < f_c$ we continue to have a giant component, but once f exceeds f_c, the giant component vanishes. This is illustrated by the f-dependence of P_∞, representing the probability that a node is part of the giant component Fig. 8.5: P_∞ is finite under f_c, but it drops to zero as we approach f_c. The critical exponents characterizing this breakdown, γ_p, β, μ, are the same as those encountered in Eq. 8.1-8.3, as the two processes can be mapped into each other by choosing $f = 1 - p$. Furthermore, in ADVANCED TOPICS 8.A, we show that random networks under random node failures share the same scaling exponents as infinite-dimensional percolation. This equivalence predicts that the value of the critical exponents for a random network are $\gamma_p = 1$, $\beta = 1$ and $\nu = 1$, the $d > 6$ values encountered earlier.

In summary, the breakdown of a network under random node removal is not a gradual process. In general, removing a small fraction of nodes has limited impact on a network's integrity. Once the number of removed nodes reaches a critical threshold, the network abruptly breaks into disconnected components. In other words, random node failures induce a phase transition from a connected to a fragmented state. We can use the tools of percolation theory to characterize this transition in both regular and in random networks.



$0 < f < f_c$:

There is a giant component. The probability that a node belongs to the giant component:

$P_\infty \sim |f - f_c|^\beta$

$f = f_c$:

The giant component vanishes.

$f > f_c$:

The network is broken into many tiny clusters.

Figure 8.5

Network breakdown as inverse percolation

The consequences of node removal are accurately captured by the inverse of the percolation process discussed in **Fig. 8.4**. We start from a square lattice, viewed as a network whose nodes are the intersections. Next we randomly select and remove an f fraction of nodes, measuring the size of the largest component formed by the remaining nodes, captured by P_∞. The obtained networks are illustrated on the three bottom panels. Under each panel we list the characteristics of the corresponding phases.

BOX 8.1

PERCOLATION THEORY: A BRIEF HISTORY

Percolation theory began with a paper written by the mathematicians Simon Broadbent and John Hammersey in 1957, who proposed its name and formalized many of its mathematical concepts [5]. The theory rose to particular prominence in the 1960 and 70s with the development of critical phenomena in physics and the recognition that percolation offers an analytically treatable example of phase transitions. It also found important applications from oil exploration to transport phenomena in physics.

The spread of a fire in a forest is often used to illustrate the basic concepts of percolation theory. Let us assume that each pebble in Fig. 8.4 is a tree, hence the whole lattice describes a forest. If a tree catches fire, it ignites the neighboring trees; these, in turn ignite their neighbors. The fire continues to spread until no burning tree has a non-burning neighbor. The question we ask is the following: if we randomly ignite a tree, what fraction of the forest do we expect to burn down? And how long it takes the fire to burn out? The answer depends on the tree density, controlled by the parameter *p*. For small *p* the forest consists of many small islands of trees ($p = 0.55$ in Fig. 8.6), hence igniting any tree will only burn down the small cluster containing the igniting tree.

Consequently, the fire will die out quickly. For very large *p* most trees belong to a single large cluster, hence the fire will rapidly sweep through much of the dense forest (see $p = 0.62$ in Fig. 8.6). The simulations indicate that there is a critical p_c, for which it takes extremely long time for the fire to end. This p_c is the critical threshold of the percolation problem. Indeed, at $p = p_c$ the giant component just emerged through the union of many small clusters. Hence the fire has to follow a long and winding path to reach all clusters and all trees, a process that can be rather time consuming.

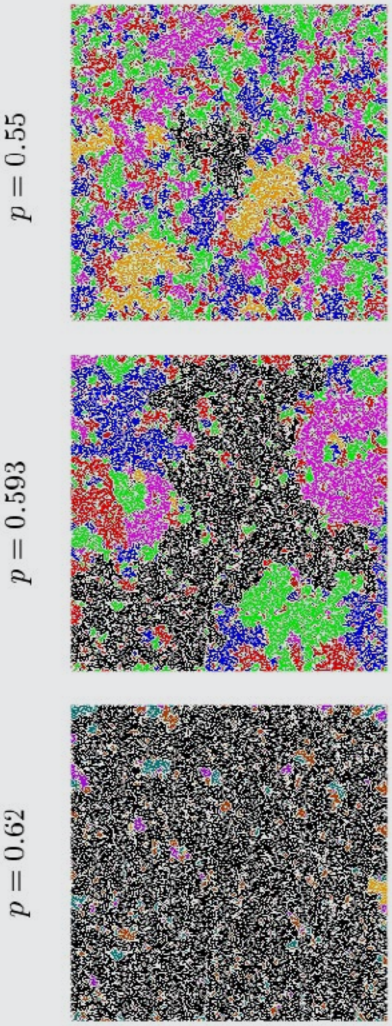


Figure 8.6
Forrest Fire

The emergence of the giant component on a square lattice as we change the occupation probability *p*. Each panel corresponds to a different p in the vicinity of p_c . The largest cluster is shown in black. For $p < p_c$ the largest cluster is tiny, as seen on the top panel. If we view this as a forest, where the pebbles are trees, a fire can at most consume only a small fraction of the trees, hence it burns out quickly. Once *p* reaches $p_c \simeq 0.593$, however, the largest cluster percolates the whole lattice and the fire can "percolate" through the forest. Increasing *p* beyond p_c "fattens" the largest cluster, connecting more pebbles (trees) to it, as seen for $p = 0.62$ on the bottom panel. Hence, the fire burns out quickly again.

SECTION 8.2

ROBUSTNESS OF SCALE-FREE NETWORKS

Percolation theory was developed either for regular lattices, whose nodes have identical degrees, or for random networks, whose nodes have comparable degrees. What happens, however, if the network is scale-free? Will the hubs affect the percolation transition? We can get a hint from a simulation testing the Internet's robustness to router failures [1]. We start from the router level map of the Internet and randomly select and remove nodes one-by-one. Percolation theory predicts that once the number of removed nodes reaches a critical value f_c, the Internet should fragment into many isolated subgraphs. The simulations indicate otherwise: the Internet refuses to break apart even under rather extensive random node removal. Instead, the size of the largest component decreases gradually, vanishing only in the vicinity of $f = 1$ Fig. 8.7a. Hence, the network behind the Internet shows an unusual robustness to random router failures: we must remove all nodes to destroy its giant component. This conclusion disagrees with percolation theory, which predicts that networks must fall apart after the removal of a finite fraction of nodes.

The behavior observed above is not unique to the Internet. Indeed,in Fig. 8.7b we show P_∞ for a scale-free network with degree exponent γ = 2.5, observing a similar pattern: under random node removal the giant component vanishes only in the vicinity of $f = 1$, rather than collapsing at some finite f_c. This indicates that the robustness observed for the Internet is a property of the scale-free topology. The goal of this section is to uncover and quantify the source and the characteristics of this remarkable robustness.

MALLOY-REED CRITERIA

To understand the origin of the anomalously high f_c for the Internet and for a scale-free network we must first calculate f_c for a network with an arbitrary degree distribution. To do so we rely on a simple observation: if a network has a giant component, then most nodes that belong to it must be connected to at least two other nodes Fig. 8.8. This leads to the Malloy-Reed criteria ADVANCED TOPICS 8.B, stating that the condition for the existence of a giant component is [6]

MOVIE 8.1

SCALE-FREE NETWORK UNDER NODE FAILURES

To illustrate the robustness of a scale-free network we start from the network we constructed in Movie 4.1 using the Barabási-Albert model. Next we randomly select and remove nodes one-by-one. As the movie illustrates, despite the fact that we remove a significant fraction of the nodes, the network refuses to break apart. The origin of this robustness to random failures is the topic of SECTION 8.2. Visualization by Dashun Wang.

→

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} > 2. \tag{8.4}$$

Consequently networks with κ < 2 must be fragmented into many disconnected components. The Malloy-Reed criteria links the network's integrity, as expressed by the presence or the absence of a giant component, to $\langle k \rangle$ and $\langle k^2 \rangle$, which depend only on the degree distribution p_k. To illustrate the predictive power of Fig. 8.4 we apply it to a random network, for which $\langle k^2 \rangle = \langle k \rangle(1 + \langle k \rangle)$. Hence, a random network has a giant component if

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\langle k \rangle(1+\langle k \rangle)}{\langle k \rangle} = 1 + \langle k \rangle > 2. \tag{8.5}$$

or

$$\langle k \rangle > 1. \tag{8.6}$$

The condition coincides with Eq. 3.10, the necessary condition for the existence of a giant component.

ROBUSTNESS OF SCALE-FREE NETWORKS

To understand the mathematical origin of the robustness observed in Fig. 8.8, we ask at what threshold f_c will a scale-free network loose its giant component. To answer this we apply the Malloy-Reed criteria to a network with an arbitrary degree distribution. We find that the critical threshold follows [7] (ADVANCED TOPICS 8.C)

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}. \tag{8.7}$$

The most remarkable prediction of Fig. 8.7 is that we can calculate the critical threshold f_c depends only from $\langle k \rangle$ and $\langle k^2 \rangle$, quantities that are uniquely determined by the degree distribution p_k. Let us illustrate the utility of Fig. 8.7 by calculating the breakdown threshold of a random network. Using $\langle k^2 \rangle = \langle k \rangle (\langle k \rangle + 1)$, we obtain

$$f_c^{ER} = 1 - \frac{1}{\langle k \rangle}. \tag{8.8}$$

Hence, the denser is a random network, the higher is f_c, i.e. the more nodes we need to remove to break it apart. Eq. 8.8 also predicts that a random network always has a finite f_c, consequently it always breaks apart after the removal of a finite fraction of nodes.

Most important, Fig. 8.7 helps us understand the roots of the enhanced robustness observed in Fig. 8.7. Indeed, for scale-free networks with γ < 3 in the N → ∞ limit the second moment diverges. If we insert $\langle k^2 \rangle \to \infty$ into Fig. 8.7, we find that f_c converges to $f_c = 1$. This means that to fragment a scale-free network we must remove all of its nodes. In other

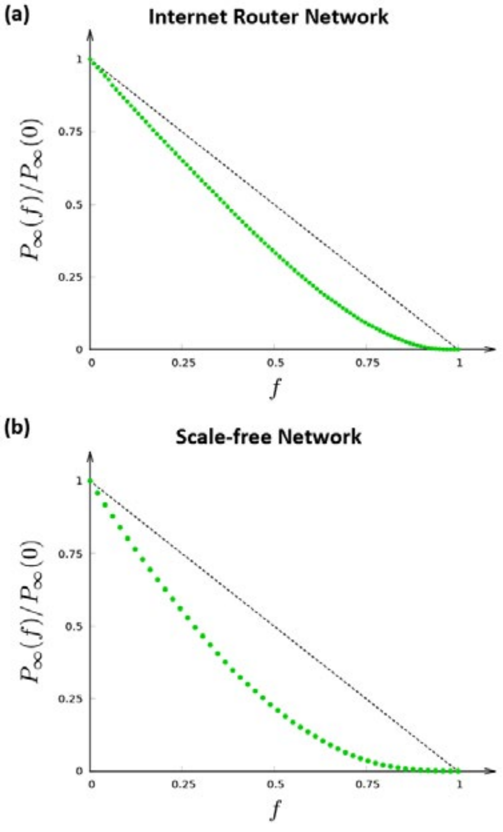


Figure 8.7
Robustness of scale-free networks

(a) The fraction of Internet routers that belong to the giant component after an f fraction of routers are randomly removed. The ratio $P_\infty(f)/P_\infty(0)$ provides the relative size of the giant component. The simulations use the router level Internet topology of Table 4.1.

(b) The fraction of nodes that belong to the giant component after the removal of an f fraction of nodes from a scalefree network with $\gamma = 2.5$, $N = 10,000$ and $k_{min} = 1$.

The plots indicate that the Internet and in general a scale-free network do not fall apart after the removal of a finite fraction of nodes. We need to remove almost all nodes (i.e. $f_c \simeq 1$) to fragment these networks.

words, the random removal of a finite fraction of its nodes does not break apart a large scale-free network. To further illustrate the roots of this anomaly we express ⟨k⟩ and ⟨k^2⟩ in terms of the parameters characterizing a scale-free network: the degree exponent γ and the minimal and maximal cutoffs, k_{min} and k_{max}, obtaining (ADVANCED TOPICS 5.D)

$$f_c = \begin{cases} 1 - \dfrac{1}{\dfrac{\gamma-2}{3-\gamma} k_{\min}{}^{\gamma-2} k_{\max}{}^{3-\gamma} - 1} & 2 < \gamma < 3 \\ 1 - \dfrac{1}{\dfrac{\gamma-2}{\gamma-3} k_{\min} - 1} & \gamma > 3 \end{cases} \qquad (8.9)$$

Eq. 8.9 predicts that Fig. 8.9:

- For γ > 3 the critical threshold f_c depends only on γ and k_{min}, hence f_c is independent of the network size N. In this regime a scale-free network behaves like a random network: it falls apart after the removal of a finite fraction of its nodes.

- For γ 3 the k_{max} diverges for large N (see Eq. 4.18). Therefore in the $N \rightarrow \infty$ limit Eq. 8.9 predicts $f_c \rightarrow 1$. Hence, to fragment an infinite scale-free network *we must remove all of its nodes*.

Eq. 8.6, 8.9 are the key results of this chapter, predicting that scale-free networks can withstand an arbitrary level of random failures without breaking apart. To understand the origin of this remarkable robustness we must inspect the role of the hubs. Random node failures by definition are blind to degree, affecting with the same probability a small or a large degree node. Yet, in a scale-free network we have far more small degree nodes than hubs. Therefore, random node removal will predominantly remove one of the numerous small nodes as the chances of removing one of the few large hubs is negligible. These small nodes contribute little to a network's integrity, hence their removal does not damage the network.

Returning to the airport analogy of Fig. 4.6, if we close a randomly selected airport, we will most likely be shutting down one of the numerous small airports. Its absence will be hardly noticed elsewhere in the world: you can still travel from New York to Tokyo, or from Los Angeles to Rio de Janeiro.

ROBUSTNESS OF FINITE NETWORKS

Eq. 8.9 predicts that for a scale-free network f_c converges to one only in the $k_{max} \rightarrow \infty$ (or $N \rightarrow \infty$) limit. While many networks of practical interest are very large, they are still finite, prompting us to ask if the observed anomaly is relevant for finite systems. We can address this by inserting Eq. 4.18 into Eq. 8.9, obtaining that f_c depends on the network size N as

Figure 8.8
Malloy-Reed criteria

To form a chain, each individual must hold the hand of two other individuals. Similarly, to have a giant component in a network, on average each of its nodes should have at least two neighbors. The Malloy-Reed criteria Fig. 8.4 exploits this property to help us calculate the critical point at which a network breaks apart.

$$f_c \simeq 1 - \frac{C}{N^{\frac{3-\gamma}{\gamma-1}}} \cdot \qquad (8.10)$$

where C collects all terms that do not depend on N. Eq. 8.10 indicates that the larger the size of a network, the closer will be its critical threshold to $f_c = 1$. To see how close fc can get in a real system to the theoretical $f_c = 1$, we calculate f_c for the Internet. The router level map of the Internet has $\langle k^2\rangle/\langle k\rangle = 37.94$ Table 4.1. Inserting this ratio into Eq. 8.7 we obtain $f_c = 0.972$. Therefore, we need to remove 97% of the routers to fragment the Internet into disconnected components. The probability that by chance 220,000 routers fail simultaneously, representing 97% of the N = 228, 263 routers on the Internet, is effectively zero. This is one of the reasons the topology of the Internet is so robust to random failures.

A network displays enhanced robustness if its breakdown threshold deviates from the random network prediction Eq. 8.8, i.e. if

$$f_c > f_c^{ER}. \qquad (8.11)$$

Enhanced robustness has several ramifications:

- The inequality Eq. 8.11 is satisfied for all networks for which $\langle k^2\rangle$ deviates from $\langle k\rangle$ ($\langle k\rangle$ + 1). According to Fig. 4.8, for virtually all networks in our reference network list $\langle k^2\rangle$ exceeds the random expectation. Hence the robustness predicted by Eq. 8.7 is not an isolated property of a few selected networks, but affects most networks of practical interest.

- Eq. 8.7 also predicts that the degree distribution of a network does not need to follow a strict power law to display enhanced robustness; all we need is a larger $\langle k^2\rangle$ than expected for a random network of similar size.

- Finally, enhanced robustness is not limited to node removal, but emerges under link removal as well Fig. 8.10.

In summary, in this section we encountered a fundamental property of real networks: their robustness to random failures. Eq. 8.7 predicts that the breakdown threshold of a network depends only on its degree distribution through $\langle k\rangle$ and $\langle k^2\rangle$. This predicts that random networks have a finite threshold, but for scale-free networks with $\gamma < 3$ the breakdown threshold converges to one. Therefore, we need to remove all nodes to break a scale-free network apart, indicating that these networks show enhanced robustness to random failures. The origin of this enhanced robustness is the large $\langle k^2\rangle$ term. As for most real networks $\langle k^2\rangle$ is larger than the random expectation, enhanced robustness is a generic property of many networks. This robustness is rooted in the fact that random failures affect mainly the numerous small nodes, which play a limited role in maintaining a network's integrity.

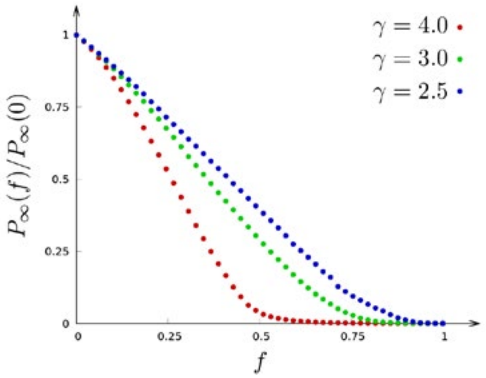


Figure 8.9
Robustness and degree exponent

The probability that a node belongs to the giant component after the removal of an f fraction of nodes from scale-free networks with different degree exponent γ. For γ = 4 we observe a finite critical point f_c, as predicted by Eq. 8.9. For $\gamma < 3$, however, we have $f_c \to 1$. The networks were generated with the configuration model using $k_{min} = 2$ and $N = 10,000$.

NETWORK	RANDOM FAILURES	RANDOM NETWORK	ATTACK
Internet	0.92	0.84	0.16
WWW	0.88	0.85	0.12
Power Grid	0.61	0.63	0.20
Mobile-Phone Call	0.78	0.68	0.20
Email	0.92	0.69	0.04
Science Collaboration	0.92	0.88	0.27
Actor Network	x	0.99	0.55
Citation Network	0.96	0.95	0.76
E. Coli Metabolism	0.96	0.90	0.49
Yeast Protein Interactions	0.88	0.66	0.06

Table 8.1

Breakdown thresholds under random failures and attacks

The table shows the estimated fc for random failures (second column) and attacks (fourth column) for ten reference networks. The procedure for determining fc is described in ADVANCED TOPICS 10.X. The third column (random network) offers fc for a network whose *N*, *L* coincides with the original values, but whose nodes are connected randomly to each other. Note that for most networks f_c for random failures exceeds f^{ER} for the corresponding random network, indicating that these networks display enhanced robustness, as defined in **Eq. 8.11**. Only the power grid lacks this property, a consequence of the fact that its degree distribution is exponential **Fig. 8.27e**.

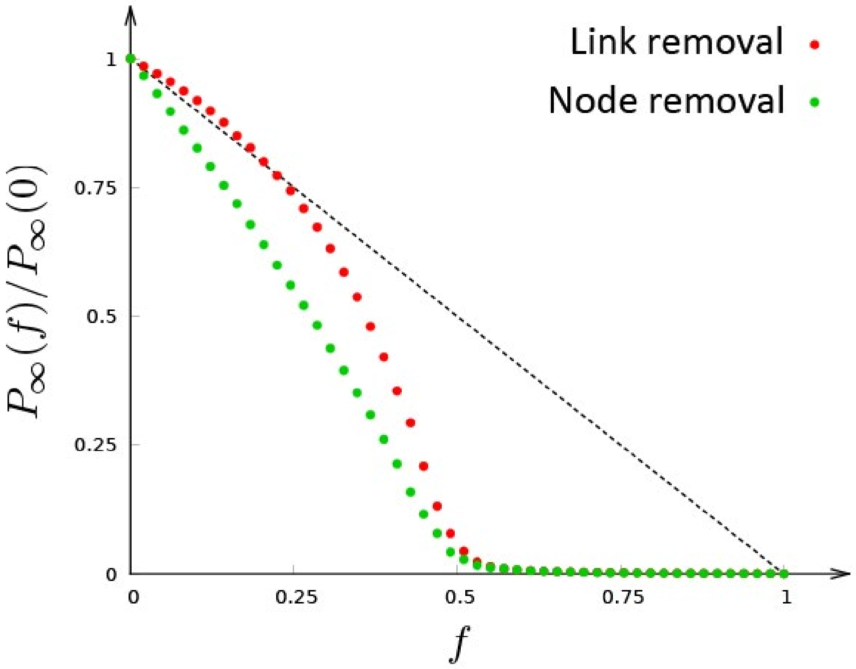


Figure 8.10

Robustness and link remova

Our focus on node removal prompts us to ask: what happens if we randomly remove the links rather than the nodes? That is, how robust are networks to link removal? The calculations predict that the critical threshold fc is the same for random link and node removal [7, 8]. To illustrate this,we compare the impact of random node and link removal on a random network with ⟨*k*⟩ = 2. The plot indicates that the network falls apart at the same critical threshold $f_c \simeq 0.5$. The difference in the shape of the two curves is rooted in the fact that the removal of an f fraction of nodes leaves us with a smaller giant component than the removal of an f fraction of links. This is not unexpected: on average each node removes ⟨*k*⟩ links, hence the removal of an *f* fraction of nodes is equivalent with removing an *f*⟨*k*⟩ fraction of the links, which clearly makes more damage than the removal of an *f* fraction of links.

ATTACK TOLERANCE

The important role the hubs play in holding together a scale-free network prompts our next question: what if we do not remove the nodes randomly, but go after the hubs? That is, we first remove the highest degree node, followed by the node with the next highest degree and so on. The likelihood that nodes would break in this particular order under normal conditions is essentially zero. Instead this process mimics an attack on the network, as it assumes a detailed knowledge of the network topology, an ability to target the hubs, and a desire to deliberately cripple the network [1]. The removal of a single hub is unlikely to fragment a network, as the remaining hubs can still hold the network together. After the removal of a few hubs, however, large chunks of nodes start falling off Movie 8.2. If the attack continues, it can rapidly break the network into tiny clusters.

The impact of hub removal is quite obvious in the case of the scale-free network shown in Fig. 8.11: the critical point, which is absent under random failures (green curve), reemerges under attacks (red curve). Not only reemerges, but it has a remarkably low value. This indicates that the removal of a small fraction of the nodes, namely the system's hubs, is sufficient to break a scale-free mnetwork into tiny clusters. The goal of this section is to identify the origin of this attack vulnerability and to quantify its magnitude.

CRITICAL THRESHOLD UNDER ATTACKS

As Fig. 8.11 indicates, an attack on a scale-free network has two consequences:

- The critical threshold, f_c, is smaller than $f_c = 1$, indicating that under attacks a scale-free network can be fragmented by the removal of a finite fraction of its hubs.
- The observed f_c is remarkably low, indicating that we need to remove only a tiny fraction of the hubs to cripple the network.

To quantify this process we need to analytically calculate f_c for a net-

MOVIE 8.2

SCALE-FREE NETWORK UNDER ATTACK

During an attack we aim to inflict maximum damage on a network. We can do this by removing first the highest degree node, followed by the next highest degree, and so on. As the movie illustrates, it is sufficient to remove only a few hubs to break a scale-free network into disconnected components. Compare this with the network's refusal to break apart under random node failures, shown in MOVIE 8.1. Visualization by Dashun Wang.

→

work under attack. To do this we rely on the fact that hub removal changes the underlying network in two different ways [9]:

- It changes the maximum degree of the network from k_{max} to k'_{max} as all nodes with degree larger than k'_{max} have been removed.

- The degree distribution of the network changes from p_k to $p'_{k'}$, as all nodes connected to the removed hubs will loose links, altering the degrees of the remaining nodes.

In ADVANCED TOPICS 8.E we combine these two changes and map the attack problem into the robustness problem discussed in the previous section. In other words, we can view an attack as random node removal from a network with adjusted k'_{max} and $p'_{k'}$. The calculations predict that the critical threshold f_c for attacks on a scale-free network with degree exponent γ is the solution of the equation [9, 10].

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} k_{min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right). \qquad (8.12)$$

Fig. 8.12 shows the numerical solution of Eq. 8.12 in function of γ, leading to several conclusions:

- While f_c for failures decreases monotonically with γ, f_c for attacks has a complex non-monotonic behavior.

- f_c for attacks is always smaller than f_c for random failures.

- For large γ a scale-free network behaves like a random network. As a random network lacks hubs, an attack on a random network will follow a scenario similar to random node removal. Numerical simulations support this expectation: Fig. 8.13 shows that a random network has a finite percolation threshold under both random failures and attack. The main difference is that f_c for attacks is lower than f_c for random failures.

- The failure and the attack thresholds converge to each other for large γ. Indeed, if $\gamma \to \infty$ then $p_k \to \delta(k - k_{min})$, meaning that all nodes have the same degree k_{min}. Therefore random failures and targeted attacks become indistinguishable in the $\gamma \to \infty$ limit, when $f_c \to 1 - 1/(k_{min} - 1)$.

The airport analogy helps us understand the fragility of scale-free networks to attacks: the closing of two hub airports, like Chicago's O'Hare Airport or the Atlanta International Airport for only a few hours would be headline news, altering travel throughout the US. Should some series of events lead to the simultaneous closure of the Atlanta, Chicago, Denver, and New York airports, the biggest hubs, air travel within the U.S. would come to a halt within hours.

In summary, while random node removal has difficulty fragmenting

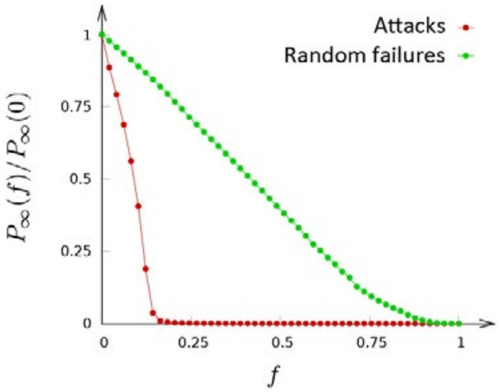


Figure 8.11

Scale-free networks under attack

The probability that a node belongs to the largest connected component in a scale-free network under attack (red) and under random failures (green). In the case of an attack the nodes are removed in a decreasing order of their degree: we first remove the biggest hub, followed by the next biggest and so on. In the case of failures, the order in which the nodes are chosen is random, independent of the node's degree. The plot illustrates the network's extreme fragility to attacks: f_c is rather small, implying that the removal of only a few hubs can disintegrate the network. The initial network has a degree exponent $\gamma = 2.5$, $k_{min} = 2$ and $N = 10,000$.

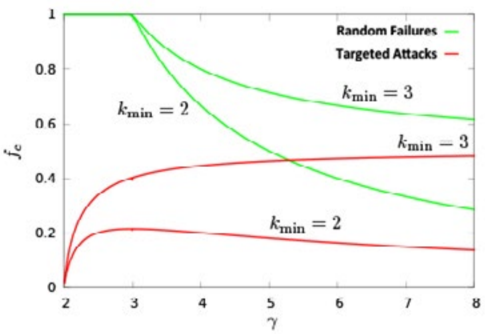


Figure 8.12

Critical threshold under attack

The dependence of the critical probability, f_c, on the degree exponent γ, for scale-free networks with $k_{min} = 2, 3$, as predicted by Eq. 8.12, for an attack (red curves) and by Eq. 8.7 for random failures (green curves). Note that Eq. 8.12 predicts that the attack threshold $f_c \to 0$ for $k_{min} = 2$ and $f_c \to 1/2$ for $k_{min} = 3$, in line with the behavior observed in the figure.

a scale-free network, an attack that targets the hubs can easily destroy a network. This fragility is bad news for the Internet, as it indicates that it is inherently vulnerable to deliberate attacks. It can be good news in medicine, as the vulnerability of bacteria to the removal of its hub proteins offers avenues to design drugs that target these hubs, potentially destroying the organism.

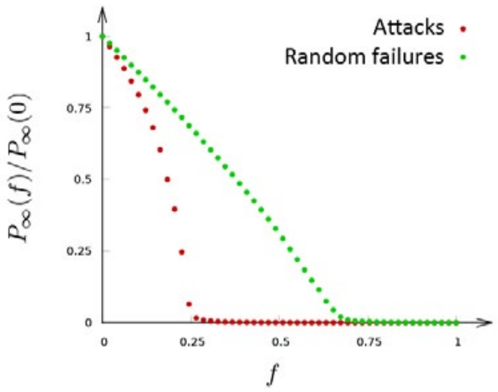


Figure 8.13

Attack and failures in random networks

The fraction of nodes that belong to the giant component in a random (i.e. Erdős-Rényi) network if an f fraction of nodes are removed randomly (random failure, green) and in decreasing order of their degree (attacks, red). Both curves indicate the existence of a finite threshold, in contrast with scale-free networks, for which $f_c \rightarrow 1$ under random failures. The simulations were performed for random networks with $N = 10,000$ and $\langle k \rangle = 3$.

BOX 8.2

PAUL BARAN AND THE INTERNET

In 1959 RAND, a Californian think-tank, has assigned Paul Baran, a young engineer at that time, to develop a communication system that can survive a Soviet nuclear attack. As a nuclear strike handicaps all equipment within the range of the detonation, Baran had to design a system whose users outside this range would not lose contact with one another. He described the communication system of his time as a "hierarchical structure of a set of stars connected in the form of a larger star," offering an early description of what we would call today a scale-free network. He concluded that this topology is too centralized to be viable under attack. He also discarded the hub-and-spoke topology shown in Fig. 8.14a noting that "The centralized network is obviously vulnerable as destruction of a single central node destroys communication between the end stations." Baran decided that the ideal survivable architecture was a distributed mesh-like network, shown in Fig. 8.14C, which is sufficiently redundant, so that even if some of its nodes break down, alternative paths can maintain the connection between the remaining nodes. Baran's ideas were ignored by the military, so when the Internet was born a decade later, it relied on distributed protocols that allowed each node to decide where to link. This decentralized approach paved the way to the emergence of a scale-free Internet, rather than the uniform mesh-like topology envisioned by Baran. Consequently the Internet today resembles more the decentralized structure B, the one that Baran wanted to avoid, than the distributed topology C he preferred.

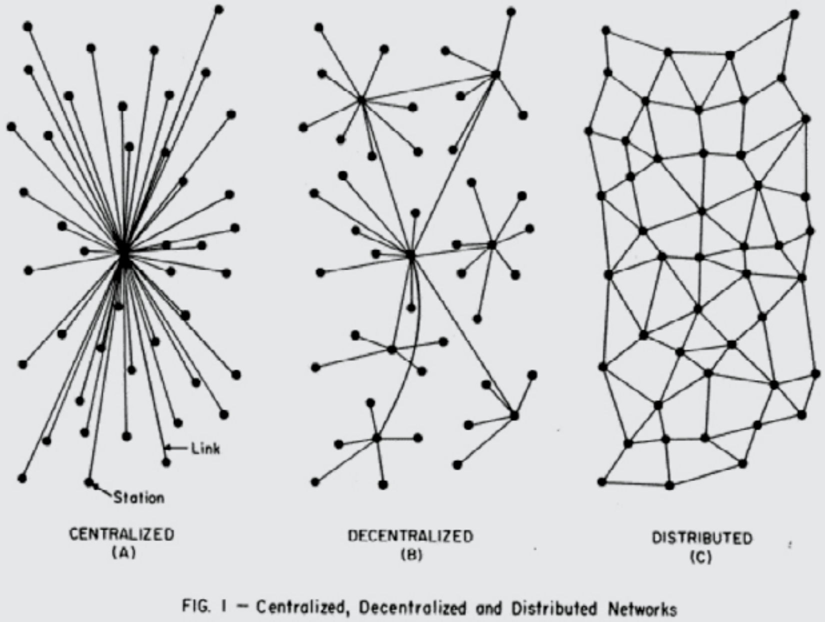

Link
Station
CENTRALIZED
(A)
DECENTRALIZED
(B)
DISTRIBUTED
(C)
FIG. I — Centralized, Decentralized and Distributed Networks

Figure 8.14
Baran's Network

Possible network configurations, described by Paul Baran in his 1959 report.

SECTION 8.4

CASCADING FAILURES

Throughout this chapter we assumed that the nodes in a network fail independently of each other. In reality, the activity of each node in a network depends on the activity of its neighboring nodes. Consequently the failure of one node can often induce the failure of the nodes connected to it. Let us consider a few examples:

- **Blackouts (power grid)**
  As electricity travels with the speed of light, after a node or link failure the electric currents are instantaneously reorganized on the rest of the power grid. For example, on August 10, 1996, a hot day in Oregon, a line carrying 1,300 megawatts sagged close to a tree and snapped. Because electricity cannot be stored, the current it carried was automatically shifted to two lower voltage lines with 115 and 230 kilovolt capacity. These were not designed to carry the excess current, so they also failed. Seconds later the excess current lead to the malfunction of thirteen generators, causing a blackout in eleven U.S. states and two Canadian provinces [11].

- **Denial of service attacks (Internet)**
  If a router malfunctions, responding too slowly or failing to transmit the packets received by it, the Internet protocols will alert the neighboring routers, which will re-route the packets, using alternative routes to avoid the troubled equipment. Consequently a failed router can place a significant burden on other routers, potentially inducing cascading failures in the form of a series of denial of service attacks distributed throughout the Internet [12].

- **Financial Crises**
  Cascading failures are common in economic systems as well. For example, the drop in the house prices in 2008 in the U.S. lead to a global financial meltdown that is considered the worst crisis since the 1930s Great Depression. In other words, the impact of the housing bubble has spread along the links of the financial network, inducing a cascade of failures throughout the economy, leading to failed banks,

Figure 8.15
Domino effect

In general we call the domino effect a sequence of events that is induced by a local change, yet it propagates through the whole system. The phenomena is similar to the fall of a series of dominos induced by the fall of the first domino. The domino effect represents perhaps the simplest illustration of cascading failures, the topic of this section.

BOX 8.3

NORTHEAST BLACKOUT OF 2003

One of the largest blackouts in North America took place on August 14, 2003, just before 4:10 p.m. Its cause was a software bug in the alarm system at a control room of the *FirstEnergy* Corporation in Ohio. Missing the alarm, the operators were unaware of the need to redistribute the power after an overloaded transmission line hit a tree. Consequently a normally manageable local failure stared a cascading failure that shut down more than 508 generating units at 265 power plants, leaving without electricity an estimated 10 million people in Ontario and 45 million people in eight U.S. states.

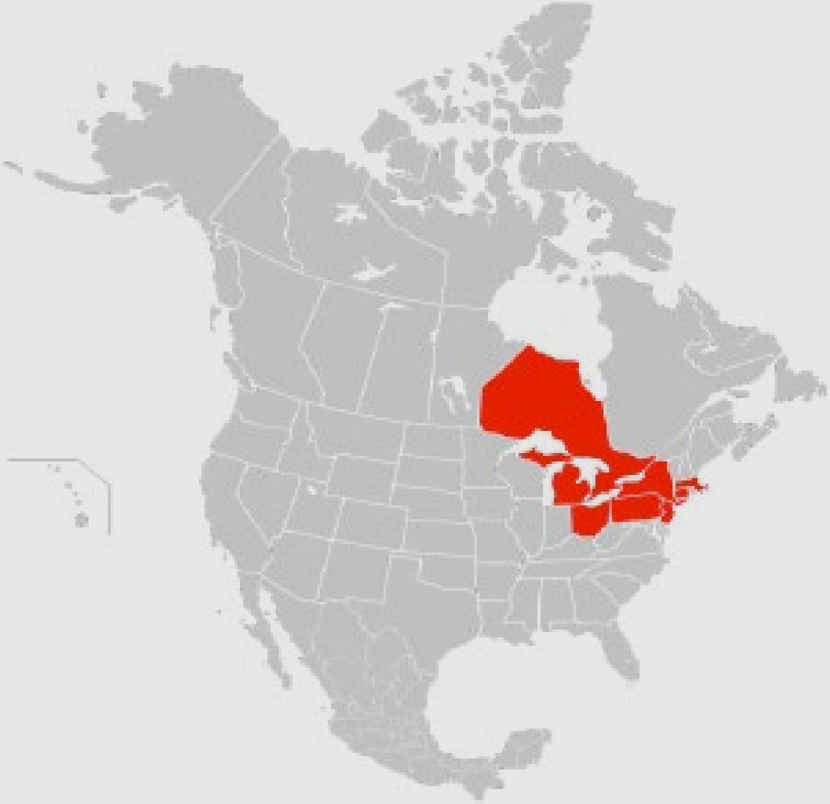


Figure 8.16
The 2003 Power Outage

Canadian and USA states affected by the August 14, 2003 blackout, illustrating how a local failure can turn into a major global event.

companies and even nations [13, 14, 15].

Despite covering different domains, these examples have several common characteristics. First, the initial failure had only limited impact on the network structure. Second, the initial failure did not stay localized, but it spread along the links of the network, inducing additional failures. Eventually, multiple nodes, one after the other, failed to carry out their normal functions. Each of these systems experienced cascading failures, a dangerous phenomena in many networks [16]. In this section we discuss the empirical patterns governing such cascading failures. The modeling of these events in the topic of the next section.

EMPIRICAL RESULTS

Cascading failures are well documented in the case of the power grid, information systems and tectonic motion, offering detailed statistics about their frequency and magnitude.

- **Blackouts**

  A blackout, also called a power outage or power failure, is a loss of the electric power in some area. It can be caused by failures at power stations, damage to electric transmission lines, substations, a short circuit, and so on. When the operating limits of a component is exceeded, it is typically automatically disconnected to protect it. In other words, a component can “fail” in the sense that it is not available to transmit power. Such failure redistributes the power previously carried by the failed component to other components, altering power flows, frequency, voltage and phase, and inducing changes in the operation of the control, monitoring and alarm systems. These changes can in turn disconnect other components as well, in some cases starting an avalanche of failures.

  A frequently recorded measure of blackout size is the energy unserved. Fig. 8.17a shows the probability distribution *p(s)* of energy unserved in all North American blackouts between 1984 and 1998. Electrical engineers approximate the obtained distribution with a power law [17],

$$p(s) \sim s^{-\alpha}, \tag{8.13}$$

  where the estimated value of the avalanche exponent α is listed in Table 8. 2 for several countries. The power law nature of this distribution indicates that most blackouts are rather small, affecting only a few consumers. These coexists, however, with occasional major blackouts, where millions of consumers lose power BOX 8.3.

- **Information cascades**

  Modern communication systems, from email to mobile phones, Facebook or Twitter, allow for the cascade-like spreading of information along the links of the social network. As the events pertaining to the spreading process often leave digital traces, these platforms al-

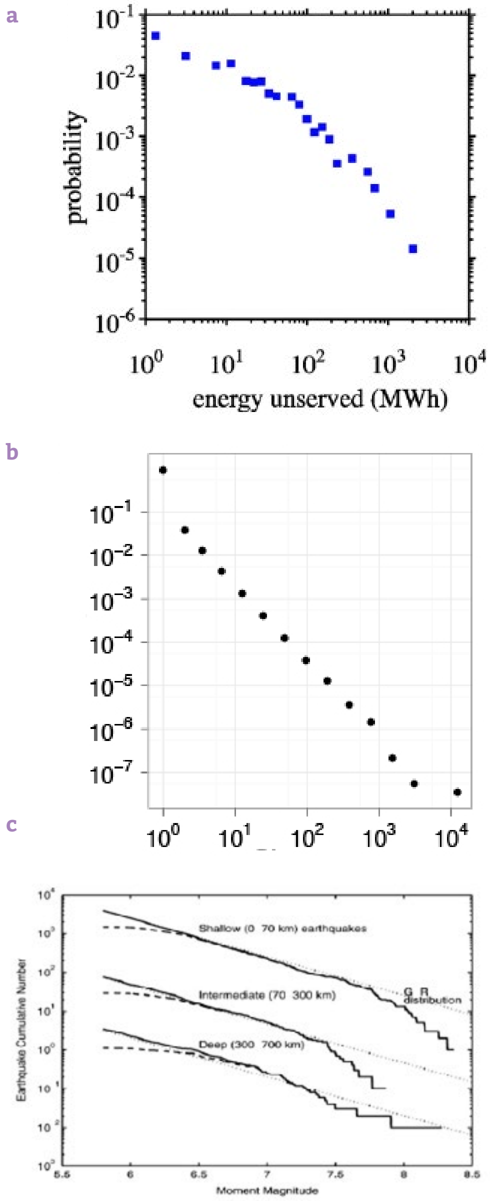


Figure 8.17
Cascade size distributions

(a) The distribution of energy loss for all North American blackouts between 1984 and 1998, as documented by the North American Electrical Reliability Council. The distribution is typically fitted to **Eq. 8.13**. The reported exponents for different countries are listed in **Table 8.2**. After [17].

(b) The distribution of cascade sizes on Twitter. While most tweets cause no reaction, bringing the average cascade size down to 1.14, a tiny fraction of tweets are shared thousands of times. Overall the retweet numbers are well approximated with **Eq. 8.13** with $\alpha \simeq 1.75$. After [18].

(c) The distribution of earthquake sizes recorded between 1977 and 2000. The dotted line indicates the power law fit **Eq. 8.13** used by seismologists to characterize the distribution. After [19].

low researchers to detect and explore the underlying cascades. The micro-blogging service Twitter has been particularly useful in this context. On Twitter the network of who follows whom can be reconstructed by crawling the follower graph behind the service. As users frequently share web-content using URL shorteners, the tracking of a spreading/sharing process is relatively straightforward. A study tracking 74 million such events over a two month interval in 2009 traced the diffusion of each URL from a particular seed node through its reposts until the end of a cascade Fig. 8.18.

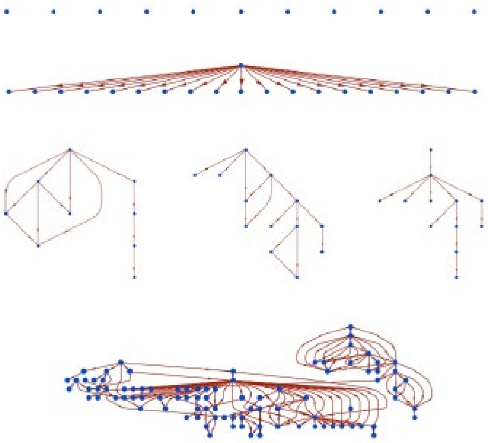


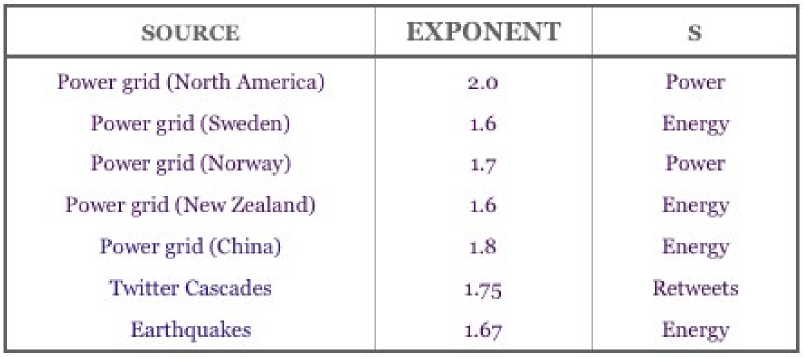
Figure 8.18
Information Cascades

Examples of information cascades on Twitter. Nodes denote Twitter users, the top node corresponding to the individual who first posted a certain shortened URL. The links correspond to those who retweeted it. The observed cascades capture the heterogeneity of information avalanches: most URLs are not retweeted at all (shown as individual nodes in the figure), but some are part of major retweet avalanches, like the one seen at the bottom panel. After [18].

As Fig. 8.17b indicates, the cascade size distribution follows the power-law Eq. 8.13 with an avalanche exponent $\alpha \simeq 1.75$. This indicates that the vast majority of posted URLs do not spread at all, a claim also supported by the fact that the average cascade size is only 1.14. Yet, a small fraction of URLs are reposted thousands of times.

- **Earthquakes**
  Most geological fault surfaces are irregular and sticky, not permitting a smooth slide against each other. Once a fault has locked, the continued relative motion of the plates increases the stress, accumulating an increasing amount if strain energy around the fault surface. This continues to accumulate until the stress is sufficient to break through the asperity, resulting in a sudden slide that releases the stored energy and causes an earthquake. Earthquakes can be also induced by the natural rupture of geological faults, by volcanic activity, landslides, mine blasts and even nuclear tests. Each year around 500,000 earthquakes are detected with instrumentation. Only about 100,000 of these are sufficiently strong to be felt by humans. Seismologists approximate the distribution of earthquake sizes with the power law Eq. 8.13 with $\alpha \approx 1.67$ Fig. 8.17c. Earthquakes are rarely considered a manifestly network phenomenon, given the difficulty of mapping out the precise interdependencies in the Earth's crust that causes them. Yet, the resulting cascading failures bear many similarities to network based cascading events, suggesting common mechanisms.

The power-law distribution Eq. 8.13 followed by blackouts, information cascades and earthquakes indicates that most cascading failures are relatively small.



SOURCE	EXPONENT	S
Power grid (North America)	2.0	Power
Power grid (Sweden)	1.6	Energy
Power grid (Norway)	1.7	Power
Power grid (New Zealand)	1.6	Energy
Power grid (China)	1.8	Energy
Twitter Cascades	1.75	Retweets
Earthquakes	1.67	Energy

Table 8.2
Avalanche exponents in real systems.

The reported avalanche exponents characterizing the power law distribution Eq. 8.13 of energy loss in various countries [17], twitter cascades [18] and earthquake sizes [19]. The third column indicates the nature of the measured cascade size s, corresponding to power or energy not served, the number of retweets generated by a typical tweet and earthquake magnitudes.

These capture the loss of electricity in a few houses, tweets of little interest to most users, or earthquakes so small that are detected only by sensitive instruments. Eq. 8.13 also predicts that these numerous small events coexist with a few exceptionally large events, like blackouts leaving millions without power, messages retweeted by hundreds or earthquakes causing major loss of human life. Examples of such major cascades include the 2003 power outage in North America BOX 8.3, the tweet *Iran Election Crisis: 10 Incredible YouTube Videos http://bit.ly/vPDLo* that was shared 1,399 times [20], or the January 2010 earthquake in Haiti, with over 200,000 victims Fig. 8.19. What is particularly intriguing as we compare these systems is that the avalanche exponents reported by electrical engineers, media researches and seismologists are so close to each other: they are all between 1.6 and 2 Table 8.2.

Figure 8.19
Earthquake Damage

Devastation caused in Port-au-Prince, Haiti, by a magnitude 7 earthquake that hit the island on January 12, 2010. After *http://en.wikipedia.org/wiki/File:Haiti_earthquake_damage.jpg*

Cascading failures are documented in many other environments:

- The consequences of bad weather or mechanical failures can cascade through airline schedules, delaying flights whose schedule is apparently unrelated to the original cause, in some cases stranding thousands of passengers BOX 8.5 [21].

- The extinction of a species can cascade through an ecosystem, inducing the extinction of numerous species and altering the habitat of others [22, 23, 24, 25].

- Cascading failures are common in international trade, when the shortage of a particular component cripples supply chains, affecting the availability of a numerous products. For example, the 2011 floods in Thailand have resulted in a chronic shortage of car components that disrupted the production chain of more than 1,000 automotive factories. Thanks to these cascading events the damage was not limited to the flooded factories, resulting in insurance claims reaching $20 billion [26].

In summary, cascading effects are observed in systems of rather different nature. Their size distribution is well approximated with a power law, implying that most cascades are too small to be noticed; a few, however, are huge, having global impact. The goal of the next section is to understand the origin of these phenomena.

BOX 8.4

CASCADING CONGESTION

In the U.S. flight delays have an economic impact of over $40 billions per year [27], caused by the need for enhanced operations, passenger loss of time, decreased productivity and missed business and leisure opportunities. A flight delay is defined as the time difference between the expected and actual departure/arrival times of a flight. Airline schedules include a buffer period between consecutive flights to allow for potential delays. When a delay exceeds this buffer, subsequent flights that rely on the same aircraft, flight crew or gate, are also delayed. Consequently the impact of a delay can propagate in a cascade-like fashion through the system.

A study found that while most flights in 2010 were on time, 37.5% arrived or departed late [21]. The delay distribution has a broad tail, similar to Eq. 8.13, implying that while most flights were delayed by just a few minutes, a few were hours behind schedule. These long delays induce correlated delay patterns, a signature of cascading congestions in the air transportation system Fig. 8.20.

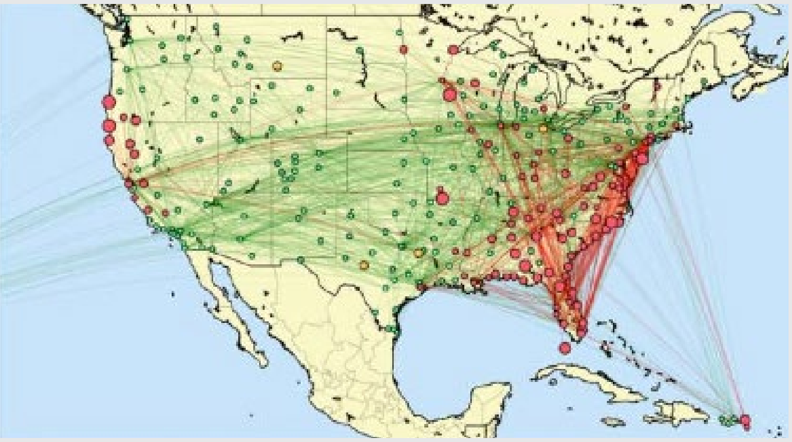


Figure 8.20
Clusters of congested airports

A U.S. aviation map showing the congested airports as red nodes, while those with normal traffic as green nodes. The lines correspond to the direct flights between them on March 12, 2010. Congested airports form a correlated cluster, a manifestation of cascading delays in air travel. After [21].

SECTION 8.5

MODELING CASCADING FAILURES

The unfolding of a cascading event depends on the structure of the network on which it propagates, the nature of the propagation process, and the breakdown criteria of each individual component. The empirical results discussed in the previous section indicate that despite the diversity of these factors, the statistics of cascading processes is universal, being independent of the particularities of the system. The purpose of this section is to understand the mechanisms governing cascading phenomena and to explain the power-law nature of the observed cascade size distributions.

Numerous models have been proposed to capture the dynamics of cascading events [17, 28, 29, 30, 31, 32, 33, 34]. While these models differ in the degree of fidelity they employ to capture specific phenomena, they indicate that systems that develop cascades share three key ingredients:

(i) Each system is characterized by some flow over a network, like the flow of electric current in the power grid or the transport of information in communication systems.

(ii) Each component has a local breakdown rule that determines when it contributes to a cascade, either by failing or choosing to pass on a piece of information.

(iii) Each system has a mechanism to redistribute the traffic or flow to other nodes upon the failure or the activation of a component.

Next, we discuss two models that offer an increasing level of abstraction and with that an increased ability to predict the characteristics of cascading failures.

FAILURE PROPAGATION MODEL

Introduced to model the spread of ideas and opinions, the failure propagation model [29] is used to describe both information cascades and cascading failures [34]. Consider a network with an arbitrary degree distribution p_k, where each node contains an agent. Each agent i can be in the state *0* (active or healthy) or 1 (inactive or failed), and is characterized by a breakdown threshold $\varphi_i \equiv \varphi$.

All agents are initially in state 0. At time $t = 0$ one agent is switched to state 1, capturing for example an initial failure or the birth of new information. In each subsequent time step, we randomly pick an agent and update its state following a simple threshold rule:

- If the selected agent *i* is in state 0, it inspects the state of its k_i neighbors. The agent *i* adopts state *1* (i.e. it also fails) if at least a φ fraction of its k_i neighbors are in state *1*, otherwise it retains its original state 0.

- If the selected agent *i* is in state *1*, it does not change its state.

Depending on the local network topology, an initial perturbation can die out immediately, failing to induce the failure of any other node. It can also lead to the failure of additional nodes, as illustrated in **Fig. 8.21a, b**. To characterize the dynamics of the model we focus on two quantities:

(i) The probability that a global cascade is triggered by a single node, a global cascade is defined as a sequence of failures that involves a finite fraction of all nodes **Fig. 8.21c**.

(ii) The expected size distribution of the observed cascades **Fig. 8.21d**.

Both quantities depend on the average degree $\langle k \rangle$ of the network and the threshold φ. The simulations document three regimes with distinct avalanche characteristics:

- **Subcritical Regime**
  If $\langle k \rangle$ is high, changing the state of a node is unlikely to move other nodes over their threshold, as the unflipped nodes have many neighbors that did not yet flip. In this regime cascades die out quickly and their sizes follow an exponential distribution. The system is subcritical, unable to support large global cascades (blue symbols in **Fig. 8.21c, d**).

- **Supercritical Regime**
  If $\langle k \rangle$ is very small, the flipping of a single node can put several of its neighbors over the threshold, triggering a global cascade. In this case virtually any perturbation induces a major breakdown, making the system supercritical (black symbols in **Fig. 8.21c, d**).

- **Critical Regime**
  At the boundary of the subcritical and supercritical regime the avalanches have widely different sizes. Numerical simulations indicate that in this regime the avalanche sizes s follow **Eq. 8.13** (green, red symbols in **Fig. 8.21d**), with $\alpha = 3/2$ if the underlying network is random.

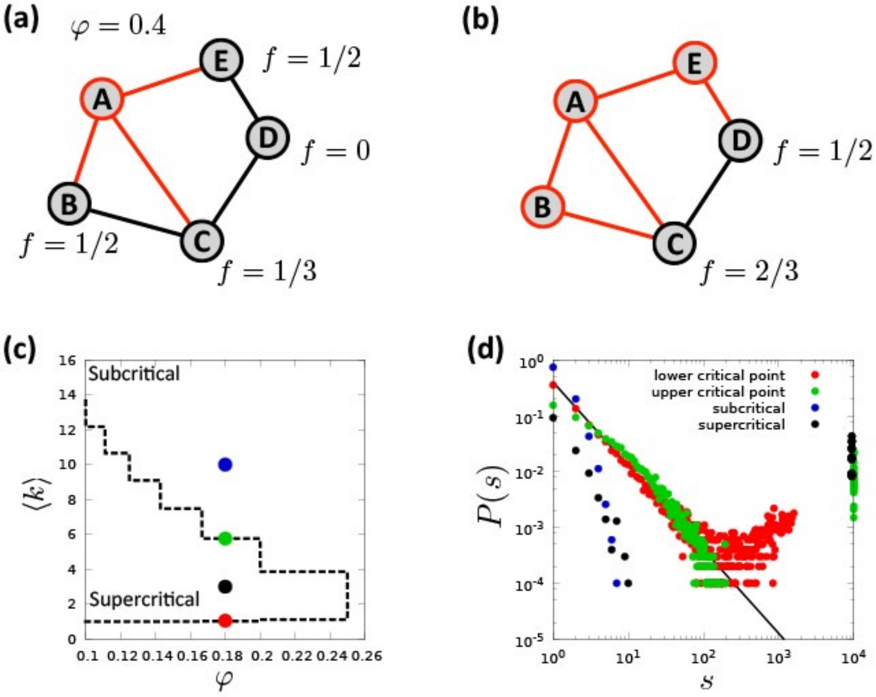


Figure 8.21
Failure propagation model

(a) The emergence of a cascade in a small network when each node has the same breakdown threshold $\varphi = 0.4$. Initially all nodes are in state 0, shown as black circles. After node A changes its state to 1 (red circle), its neighbors *B* and *E* will have a fraction $f = 1/2 > 0.4$ of their neighbors in state *1*. Consequently they also fail, changing their state to 1, as shown in (b).

(b) In the next time step *C* and *D* will also fail, having $f > 0.4$. Consequently at the end the cascade sweeps the whole network, reaching a size $s = 5$. One can check that if we initially flip node *B*, it will not induce an avalanche, as none of its neighbors pass the threshold φ.

(c) The phase diagram of the failure propagation model in terms of the threshold function φ and the network's average degree $\langle k \rangle$. The dashed line encloses the region of the $(\langle k \rangle, \varphi)$ plane in which the cascades can propagate in a random graph.

(d) Cascade size distributions for $N = 10{,}000$ and $\varphi = 0.18$, $\langle k \rangle = 1.05$ (red), $\langle k \rangle = 3.0$ (black), $\langle k \rangle = 5.76$ (green) and $\langle k \rangle = 10.0$ (blue). At the lower critical point we observe a power law *p(s)* with exponent $\alpha = 3/2$. In the supercritical regime we have only a few small avalanches, as most cascades are global. In the upper critical and subcritical regime we see only small avalanches. After [29].

In ADVANCED TOPICS 8.F we discuss two additional models describing cascading failures:

- The overload model is designed to capture power grid failures. Its distinguishing feature is a global flow process: as the electric current redistributes itself throughout the power grid, the model assumes that each failure instantaneously increases the load of all nodes. This is different from the local spread of failures in the failure propagation model.

- The self-organized critical model aims to model only the behavior of a system in the critical regime.

Despite there differences these two models predict the same avalanche exponent ($\alpha = 3/2$ for a random network) in the critical regime as the failure propagation model. The fact that the three models with rather different propagation dynamics and failure mechanisms predict similar scaling laws and avalanche exponents suggests that the underlying phenomena is universal, i.e. model independent.

BRANCHING MODEL

Given the complexity of the models discussed above, it is hard to analytically predict their scaling behavior. To understand the origin of the power-law nature of the observed *p(s)* and to calculate the avalanche exponent α, we turn to the branching model. This is perhaps the simplest model that still captures the basic features of cascading events.

The model builds on the observation that the history of a cascading failure (avalanche) can be described as a branching process. Let us designate the node whose failure triggers the avalanche as the root of a tree.

The branches of the tree are the nodes whose failure was directly triggered by this initial failure. For example, in Fig. 8.21 a, b, the breakdown of node *A* starts the avalanche, hence *A* is the root of the tree. The failure of *A* leads to the failure of *B* and *E*, which are the two branches of the tree. Subsequently *E* induces the failure of *D* and *B* leads to the failure of *C* Fig. 8.22a.

The branching model captures the essential features of this avalanche propagation process Fig. 8.22. The model starts with a single active node. In the next time step each active node produces *k* offsprings, where *k* is selected from a p_k distribution. If a node selects $k = 0$, that branch dies out permanently Fig. 8.22b. If it selects $k > 0$, it will have *k* new active sites. The size of an avalanche corresponds to the size of the tree when all active sites died out Fig. 8.22c.

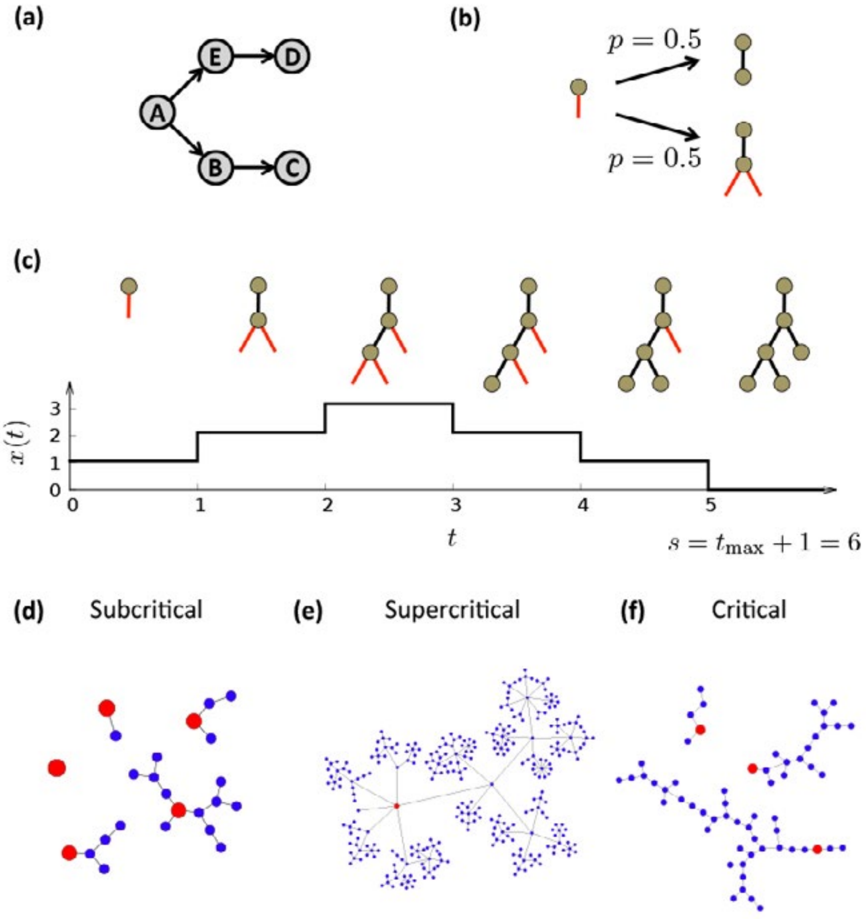


Figure 8.22
Branching process

(a) The branching process describing the propagation of the failure shown in **Fig. 8.20a,b**. The perturbation starts from node *A*, whose failure flips *B* and *E*, which in turn flip *C* and *D*, respectively.

(b) An elementary branching process. Each active link (red link) can become inactive with probability $p_0 = 1/2$ (top) or give birth to two new active links with probability $p_2 = 1/2$ (bottom).

(c) The number of active sites, *x(t)*, in function of time t. A nonzero *x(t)* means that the avalanche persists. When *x(t)* becomes zero, we loose all active sites and the avalanche ends. In the image this happens at $t = 5$, hence the size of the avalanche is $t_{max} + 1 = 6$. An exact mapping between the branching model and a one dimensional random walk helps us calculate the avalanche exponent. Consider a branching process starting from a stub with one active end. When the active site becomes inactive, it decreases the number of its active sites, i.e. $x \rightarrow x - 1$. When the active site branches, creates two active sites, i.e. $x \rightarrow x + 1$. This maps the avalanche size s to the time it takes for the walk that starts at $x = 1$ to reach $x = 0$ for the first time. This is a much studied process in random walk theory, predicting that the return time distribution follows a power law with exponent 3/2 [32]. For branching process corresponding to scale-free p_k, the avalanche exponent depends on γ, as predicted by **Fig. 8.15d, f**.

Typical avalanches generated by the branching model in the: subcritical *(d)*, supercritical *(e)* and critical regime *(f)*. The red node in each cascade marks the root of the tree, i.e. the first perturbation. In *d* and *f* we show multiple trees, while in *e* we show only one, as each tree grows indefinitely.

The branching model predicts the same three phases as those observed in the cascading failures model. These phases are determined by $\langle k \rangle$ of p_k:

- **Subcritical regime**
  If $\langle k \rangle < 1$, on average each branch has less then one offspring. Consequently each tree will terminate quickly Fig. 8.22d. In this regime the avalanche sizes follow an exponential distribution.

- **Supercritical regime**
  If $\langle k \rangle > 1$, on average each branch has more than one offspring. Consequently the tree will continue to grow indefinitely Fig. 8.22e. This captures the supercritical phase, when all avalanches are global.

- **Critical regime**
  If $\langle k \rangle = 1$, on average each branch has exactly one offspring. Consequently some trees are large; others die out shortly Fig. 8.21. Numerical simulations indicate that in this regime the avalanche size distribution follows a power law.

The branching model can be solved analytically, allowing us to predict the avalanche size distribution for an arbitrary p_k. If p_k is bounded, *e.g.* it follows a binomial or exponential form, the calculations predict $\alpha = 3/2$. If, however, p_k is scale-free, then the avalanche exponent depends on the power-law exponent γ as Fig. 8.23 [31, 32]

$$\alpha = \begin{cases} 3/2, & \gamma \geq 3 \\ \gamma/(\gamma-1), & 2 < \gamma < 3 \end{cases} . \qquad (8.15)$$

We can revisit Table 8.2 in the light of Eq. 8.15, to confirm that the empirically observed avalanche exponents are all between 1.5 and 2, as predicted by Eq. 8.15.

In summary, numerous models capture the dynamics of cascading failures. These models differ in their realism as well as the number and the nature of their tuning parameters. Yet, their predictions are consistent with each other:

- They predict the existence of a critical state, in which the avalanche sizes follow a power law. The value of the avalanche exponent α depends on the degree exponent of the network on which the avalanche propagates, as predicted by Eq. 8.15.

- They predict the existence of a subcritical regime, in which all perturbations die out immediately, and a supercritical regime, when most perturbations sweep the whole system.

Note that a detailed modeling of cascading failures should also account for the fact that nodes and links have different capacities to carry traffic [33]. Such models are best discussed in the context of weighted networks.

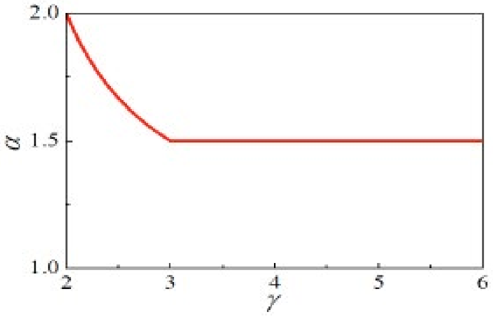


Figure 8.23
The avalanche exponent is universal

The dependence of the avalanche exponent α on the degree exponent γ of the network on which the avalanche propagates, according to Eq. 8.15. The plot indicates that between $2 < \gamma < 3$ the avalanche exponent depends on the exponent of p_k. Beyond $\gamma = 3$, however, the avalanches behave as they would be spreading on a random network.

BUILDING ROBUSTNESS

Can we enhance a network's robustness? In this section we take advantage of the insights we gained in the previous sections to design networks that are simultaneously robust to random failures and attacks. We also discuss mechanisms proposed to stop a cascading failure, allowing us to enhance a system's dynamical robustness. Finally, we apply the developed tools to the power grid, linking its robustness to its reliability.

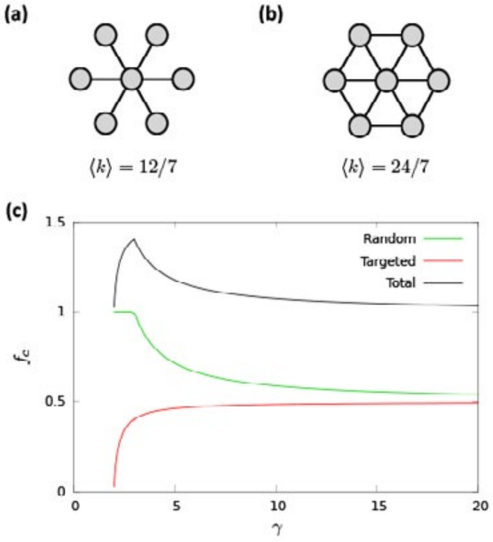


Figure 8.24
Enhancing Robustness

(a) A hub-and-spoke network is robust to random failures but has a low tolerance to an attack that removes its central hub.

(b) By connecting some of the small degree nodes, the reinforced network has a higher tolerance to targeted attacks. Yet, the cost, captured by the total number of links the network needs to maintain, i.e. $\langle k \rangle$, is higher in the reinforced network.

(c) Random, f_c^{rand}, targeted f_c^{targ} and total f_c^{tot} percolation thresholds for scale-free networks in function of the degree exponent γ. The plot is shown for $k_{min} = 3$.

DESIGNING ROBUST TOPOLOGIES

The coexistence of robustness to random failures and fragility to attacks of scale-free networks prompts us to ask: could we design networks that are simultaneously robust to attacks and random failures [35, 36, 37, 38]? This appears to be a conflicting desire. For example, the hub-and-spoke network of Fig. 8.24a is robust to random failures, as only the failure of its central node can break the network into isolated components. Therefore, the probability that a random failure will fragment the network is $1/(N-1)$, which is negligible for large N. At the same time this network is rather vulnerable to attacks, as the removal of a single node, its central hub, will break the network into isolated nodes.

We can enhance this network's robustness to both failures and attacks by connecting its peripheral nodes Fig. 8.24b. There is a price, however, for this enhanced robustness: it requires us to double the number of links. If we define the cost to build and maintain a network to be proportional to its average degree $\langle k \rangle$, the cost of the network of Fig. 8.24b is 24/7, which is double of the cost 12/7 of the network of Fig. 8.24a. The increased cost helps us refine our question: can we maximize the robustness of a network to both random failures and targeted attacks, without changing the cost, i.e. keeping $\langle k \rangle$ constant?

To enhance a network's robustness against random failures we can increase its percolation threshold f_c, which denotes the moment when the network falls apart. As f_c depends only on $\langle k \rangle$ and $\langle k^2 \rangle$ according to Eq. 8.7, the degree distribution which maximizes f_c needs to maximize $\langle k^2 \rangle$ for a fixed $\langle k \rangle$. This is achieved by a bimodal distribution, whose nodes have either degree k_{min} or k_{max}, the two extreme values allowed in the respective network.

In a similar spirit, if we wish to optimize the network topology against both random failures and attacks, we search for topologies that maximize the sum Fig. 8.24c

$$f_c^{tot} = f_c^{rand} + f_c^{t\arg}. \tag{8.16}$$

A combination of analytical arguments and numerical simulations indicate that this too is best achieved by the bimodal degree distribution [35, 36, 37, 38]

$$p_k \equiv (1-r)\delta(k-k_{\min}) + r\delta(k-k_{\max}) \tag{8.17}$$

describing a network in which an r fraction of nodes have degree k_{max} and the remaining (1 – r) fraction have degree k_{min}. As we show in ADVANCED TOPICS 8.G, the maximum of f_c^{tot} is obtained when $r = 1/N$, i.e. when there is a single node with degree k_{max} and the remaining nodes have degree k_{min}. The value of k_{max} depends on the system size as (ADVANCED TOPICS 8.G)

$$k_{\max} = AN^{2/3}. \tag{8.18}$$

In other words, a network that is robust to both random failures and attacks has a single hub with degree Eq. 8.18, while the rest of the nodes have the same degree k_{min}. This configuration is obviously robust against random failures as the chance of removing the central hub is rather small. The obtained network may appear to be vulnerable to an attack that removes its k_{max} hub, but it is not necessarily so. Indeed, the network's giant component is held together by both the central hub as well as by the many nodes with degree k_{min}, that for $k_{min} > 1$ form a giant component on their own. Hence while the removal of the k_{max} hub causes a major time loss, the remaining low degree nodes are robust against subsequent targeted removal Fig. 8.25.

HALTING CASCADING FAILURES

How can we avoid cascading failures? The first instinct is to reinforce the network through the addition of new links. If that is feasible, in some system may solve the problem. In others additional links could worsen the situation, offering more routes for the failure to spread. The true problem with reinforcement is that in most real systems the time frame needed to establish a new link is much larger than the timeframe of a cascading failure. For example thanks to regulatory, financial and legal barriers, building a new power line can take up to two decades. In contrast, a cascading failure can sweep the power grid in a few seconds. There is no way we can reinforce the network during this short time frame.

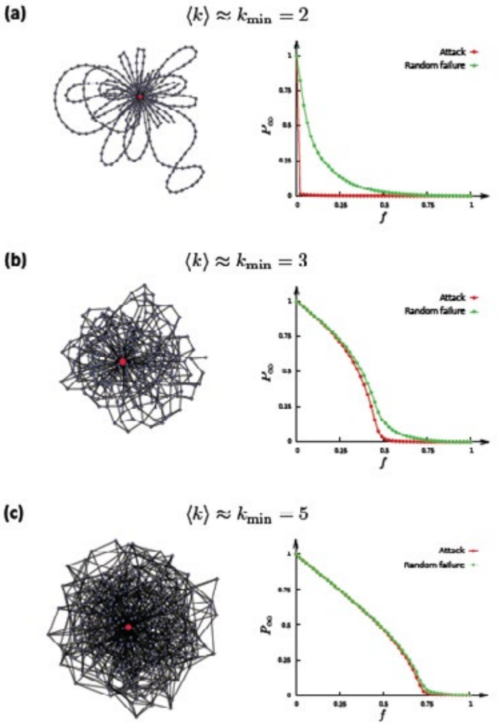


Figure 8.25
Optimizing attack and failure tolerance

The figure illustrates the optimal network topologies predicted by Eq. 8.16 and Eq. 8.17, consisting of a single hub of size Eq. 8.18 and the rest of the nodes have the same degree kmin determined by ⟨k⟩. The left panels illustrate the network topology for N = 300; the right panels show the failure/attack curves for N = 10000.

(a) For small ⟨k⟩ the hub plays a key role in holding the network together. Once we remove this central hub, given that ⟨k⟩ is small, the network breaks apart. Hence the attack and error curves are rather different.

(b) For larger ⟨k⟩ a giant component exists even without the central hub. Hence while the hub enhances the system's robustness to random failures, it is no longer essential for the network. In this case both the attack and error fc are large.

(c) For even larger ⟨k⟩ the error and the attack curves are indistinguishable, as the network is robust even without its central hub.

BOX 8.5

ROBUSTNESS, REDUNDANCY, RESILIENCE

Redundancy and resilience are concepts deeply linked to robustness. It is useful, therefore, to clarify the relationship between them.

Robustness

A system is robust if it can maintain its basic functions in the presence of internal and external errors. In a network context, robustness refers to the system's ability to carry out its basic functions, even when some of its nodes and links may be missing.

Resilience

A system is resilient when it can adapt to internal and external errors by changing its method of operations while continuing to function. Hence, resilience is a dynamical property that requires a shift in the system's core activities.

Redundancy

Redundancy implies the presence of parallel functions and components that, if needed, can replace a missing function or component. Networks show considerable redundancy in their ability to navigate information between two nodes, thanks to the multiple independent paths between most node pairs. For example, if you live in the United States, your local senator offers you a short path to the President. Yet, you may also reach to the president through many other, often equally short chains of acquaintances. A similar redundancy is built into the Internet: if a router fails, the packets normally handled by it are rerouted along alternative routes.

In a counterintuitive fashion, the impact of cascading failures can be lowered through selective node and link removal [39]. To do so we note that each cascading failure has two parts:

(i) Initial failure results in the breakdown of the first node or link, representing the source of the subsequent cascade.

(ii) Propagation, when the initial failure induces the breakdown of additional nodes and it starts cascading through the network.

In real networks the time interval between (i) and (ii) is much shorter than the time scale over which new nodes and links could be added to reinforce the network. Yet, simulations indicate that the size of a cascade can be reduced if we intentionally remove additional, well selected nodes, right after the initial failure, but before the failure could propagate. Even though the intentional removal of a node or a link increases the damage to the network, the removal of a well chosen component can suppress the cascade propagation. The mechanism is similar to the method used by firefighters, who set a controlled fire in the fire-line to consume the fuel in the path of a wildfire. The implementation of this procedure depends on the details of the spreading and failure mechanism, but simulations indicate that we can limit the size of the cascades if we remove nodes with small loads and links with large excess load in the vicinity of the initial failure.

A dramatic manifestation of the potentially positive effects of further damage is provided by the *Lazarus effect*, the ability to revive a bacteria that is unable to grow through the knockout of a few well selected genes [40] Fig. 8.26. Therefore, in a counterintuitive fashion, controlled damage can be beneficial to a network facing cascading failures.

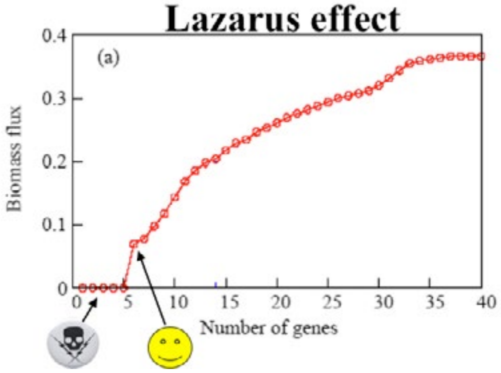


Figure 8.26
Lazarus effect in bacteria

A bacteria's growth is limited by its ability to generate biomass, the molecules the bacteria needs to build its cell wall, DNA and other cellular components. If some key genes are missing, preventing the bacteria from generating the necessary biomass, it cannot multiply and will likely die. Genes in whose absence the biomass flux is zero are called *essential*. The plot above shows the biomass flux for a mutant of E. *Coli*, a bacteria frequently studied by biologists. The mutant is missing an essential gene, hence its biomass flux is zero, as shown on the vertical axis. Consequently, it cannot multiply. Yet, the removal of five additional genes can turn on the biomass flux. Consequently, in a counterintuitive fashion, we can revive a dead organism, through the removal of further genes, a phenomena called *Lazarus effect* [40].

CASE STUDY: ESTIMATING ROBUSTNESS

The European power grid is an ensemble of more than twenty national power grids consisting of over 3,000 generators and substations (nodes) and 200,000 km of transmission lines Fig. 8.27a-d. The degree distribution of this network can be approximated with Fig. 8.27e [41, 42]

$$p_k = \frac{e^{-k/\langle k \rangle}}{\langle k \rangle} \tag{8.19}$$

indicating that its topology is characterized by a single parameter, $\langle k \rangle$. As we showed in SECTION 5.5, such p_k emerges in growing networks that lack preferential attachment. By determining $\langle k \rangle$ separately for each national power grid, we can predict the critical threshold f_c for attacks, using the tools SECTION 8.3. As shown in Fig. 8.27f, for national power grids with $\langle k \rangle > 1.5$ there is a reasonable agreement between the observed and the predicted f_c (Group 1). However, for national power grids with $\langle k \rangle <$ 1.5 (Group 2) the predicted f_c underestimates the real f_c, indicating that these national networks are more robust to attacks than expected based on their degree distribution. As we show next, this enhanced robustness correlates with the reliability of the respective national networks.

> To test the relationship between robustness and reliability, we use several quantities collected for each power failure: (1) energy not supplied; (2) total loss of power; (3) average interruption time, measured in minutes per year. The measurements indicate that Group 1 networks, for which the real and the theoretical f_c agree, represent two thirds of the full network size and share almost as much power and energy as the Group 2 networks. Yet, this group accumulates more than five times the average interruption time, more than two times the recorded power losses and almost four times the undelivered energy. Hence, Group 1 networks are more fragile than the Group 2 networks. This result offers direct evidence that networks that are topologically more robust are also more reliable. Note that this finding is rather counterintuitive: one would expect the denser networks to be more robust. We find, however, that the sparser power grids show enhanced robustness.

In summary, the results of this section indicate that a better understanding of the network topology is essential to develop strategies to improve robustness. We can improve robustness by either designing network topologies that are simultaneously robust to both random failures and attacks, or by designing interventions that limit the spread of cascading failures.

Our ability to design robust networks would suggest that we should redesign the topology of the Internet and the power grid to enhance their robustness [43]. Given the chance, this could indeed be achieved. Yet, these infrastructural networks were built incrementally over decades, following the self-organized growth process described in the previous chapters. Given the enormous cost of each link and node, it is unlikely that we would ever be given a chance to redesign them. In general the design principles of robust networks should be enforced only if robustness is an absolute criteria, like in the case of the wiring diagram of an airplane, whose failure could be fatal.

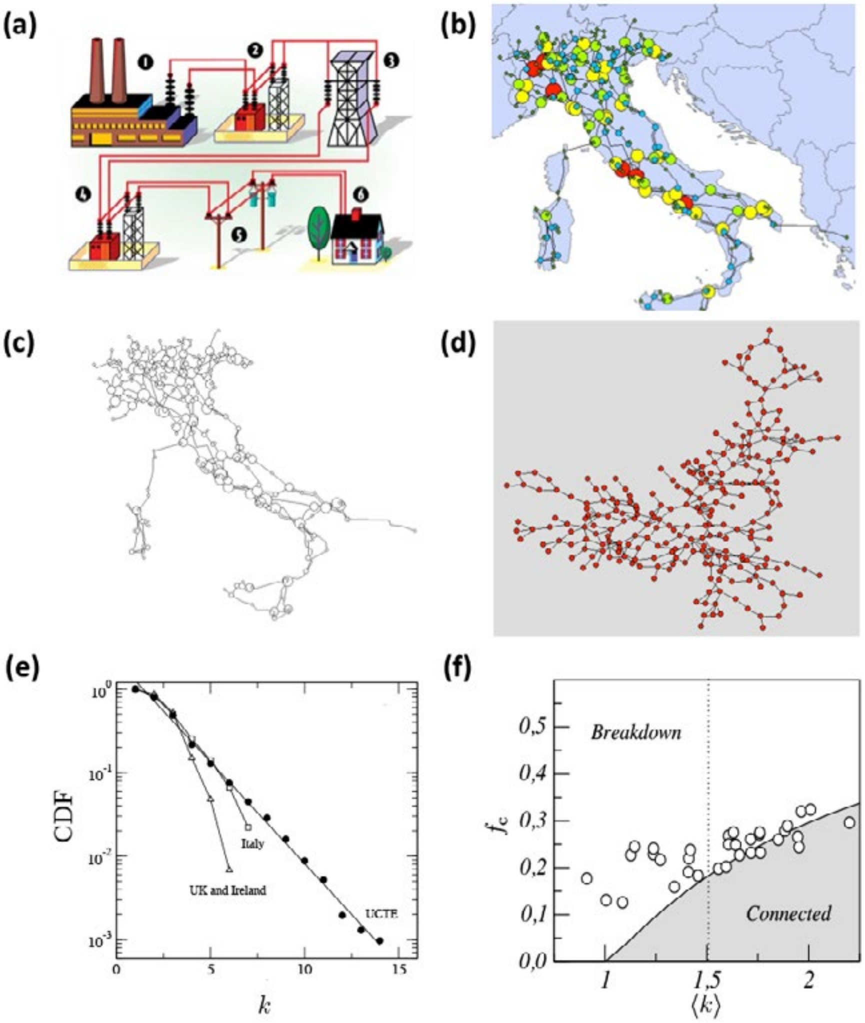


Figure 8.27
The power grid

(a) From a power engineer's perspective a power grid is a complex machinery co sisting of (1) power generators, (2) switching units, (3) the high voltage transmission grid, (4) transformers, (5) low voltage lines, (6) consumers, like households or busines es. When we study the network behind the power grid, many of these details are i nored. The necessary procedure to arrive to the network topologies amenable for study is illustrated in (b)-(d) for the Italian power grid.

(b) The power grid with the details of production and consumption. Once we strip these off, we obtain the spatial network shown in (c). Once the spatial information is also removed, we arrive to the network show in (d), which is the typical object of study at the network level.

(e) The cumulative degree distribution of the European power grid. The plot shows the data for the full network (UCTE) and separately for Italy, UK, and Ireland, indicating that the national grid's p_k also follows the exponential (8.19).

(f) Phase space $(f_c,\ k)$ for exponential uncorrelated networks under attack, where f_c is the fraction of hubs we must remove to fragment the network. The continuous curve corresponds to the critical boundary for attacks, below which the network has retains its giant component. The plot also shows the estimated $f_c(\langle k \rangle)$ for attacks from the thirty-three national power grids within EU, shown as circles. The plot allows us to distinguish two classes of power grids. For countries with $\langle k \rangle > 1.5$ (Group 1), the analytical prediction for f_c agrees with the numerically observed values. However, for countries with $\langle k \rangle > 1.5$ (Group 2) the analytical prediction underestimates the numerically observed values. Therefore, Group 2 national grids show enhanced robustness to attacks, which means that they are more robust than expected for a random network with the same degree sequence. Reliability measures indicate that the power grids in the robust Group 2 countries are more reliable. After [41].

SECTION 8.7

SUMMARY

The terrorist attacks of September 11, 2001 offered a vivid illustration of the important role hubs play in attacks. Indeed, the targets of the attack were not chosen at random: the World Trade Center in New York, the Pentagon, and the White House (an intended target) in Washington DC are the hubs of America's economic, military, and political power [44]. Yet, while causing a human tragedy far greater than any other event America has experienced since the Vietnam war, the attacks failed at their main goal: to topple the network. They did offer, however, an excuse to start new wars, like the Iraq and the Afghan wars, hence inducing a series of cascading events whose impact was far more devastating than the 9/11 terrorist attacks themselves. Yet, all networks, ranging from the economic to the military and the political web, survived. Hence, we can view 9/11 as a tale of robustness and network resilience. The roots of this robustness were uncovered in this chapter: real networks have a whole hierarchy of hubs. Taking out any one of them is not sufficient to topple the underlying network.

Network robustness represents good news for most complex systems. Indeed, there are uncountable errors in our cells, from misfolding proteins to the late arrival of a transcription factor. Yet, the robustness of the underlying cellular network helps our cells to carry on their normal functions. Network robustness also explains why we rarely notice the effect of router errors on the Internet or why the disappearance of a species does not lead to an immediate environmental catastrophe.

This topological robustness has its price, however: a fragility against attacks. As we showed in this chapter, the simultaneous removal of hubs will break any network. This is bad news for the Internet, as it allows crackers to design strategies that can harm this vital communication system. It is bad news for economic systems, as it indicates that hub removal can cripple the whole economy, vividly illustrated by the 2009 financial meltdown. Yet, it is good news for drug design, as it suggests that an accurate map of cellular networks can help us design drugs that can kill unwanted bacteria or cancer cells.

The main message of this chapter is simple: network topology, robustness, and fragility cannot be separated from one other. Rather, each complex system has its own Achilles' Heel: the networks behind them are robust to random failures but vulnerable to attacks. When considering robustness, we cannot ignore the fact that most systems have numerous controls and feedback loops that help them survive in the face of errors and failures. Internet protocols were designed to 'route around the trouble', guiding the traffic to avoid routers that malfunction; cells have numerous mechanisms to dismantle faulty proteins and to shut down malfunctioning genes. This chapter documented a new contribution to error tolerance: the structure of most complex systems preferred by nature offers them an enhanced error and failure tolerance. By the virtue of their topology only, real systems display a high degree of topological robustness.

The robustness of scale-free networks prompts us to ask: is this enhanced robustness the reason why many real networks are scalefree? Perhaps real systems have developed a scale-free architecture to satisfy their need for robustness. If this hypothesis is correct we should be able to set robustness as an optimization criteria and obtain a scale-free network. Yet, as we showed in SECTION 8.6, a network optimized for robustness has a hub-and-spoke topology. Its degree distribution is bimodal, rather than a power law. This suggests that robustness is not the force that drives the development of real networks. Rather, networks are scale-free thanks to growth and preferential attachment. It so happens that scale-free networks also have enhanced robustness. Yet, they are not the most robust networks we could design.

Finally, note that enhanced robustness does not require a network to be scale-free. Indeed, Eq. 8.7 links f_c to $\langle k^2\rangle$, hence any network whose $\langle k^2\rangle$ is larger than expected for a random network will display enhanced robustness. Of course, if a network is scale-free with $\gamma < 3$, yet automatically displays enhanced robustness.

The results of this chapter allow us to formulate our next law:

ACHILLES' HEEL

Scale-free networks are robust to random failures and fragile to attacks. Let us revisit the three criteria that prompts us to call this statement a network law:

A. Quantitative formulation

Eq. 8.7 indicates that the critical threshold of a scale-free network, capturing its response to random failures, converges to $f_c = 1$, implying an enhanced robustness to random node deletion. As we showed in SECTION 8.3, the finite threshold re-emergences under attacks.

B. Universality

Enhanced robustness is present in all networks whose $\langle k^2\rangle$ is higher than expected in a random network. According to Table 4.1, this is true for most real networks.

BOX 8.7

NETWORK ROBUSTNESS: BRIEF HISTORY

The systematic study of network robustness within network science started with a paper published in Nature Fig. 8.1 by Réka Albert, Hawoong Jeong and Albert-László Barabási [1], that discovered the robustness of scale-free networks to random failures and their fragility to attacks. Yet, when it comes to our understanding of network robustness, particularly important were the contributions of Shlomo Havlin and his collaborators, like Reuven Cohen, who showed that the percolation threshold of a scale-free network is determined by the first two moments of the degree distribution.

Figure 8.28
Shlomo Havlin

A statistical physicist at Bar Ilan University in Israel, Havlin has played an important role in the development of network science. He derived (8.7), demonstrating that the origin of the enhanced robustness is rooted in the convergence of the the percolation threshold to one. His contributions to the field are diverse, from discovering the selfsimilar nature of real networks [45] to introducing the study of layered interdependent networks [46].

C. Non-random origins

The phenomena discussed in this chapter is obviously absent from random networks, that have a finite threshold against both failures and attacks Fig. 8.13.

BOX 8.7

AT A GLANCE: NETWORK ROBUSTNESS

Malloy-Reed criteria:

A giant component exists if

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} > 2$$

Critical threshold for random failures:

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$$

Random Network: $f_c^{ER} = 1 - \frac{1}{\langle k \rangle}$

Enhanced robustness: $f_c > f_c^{ER}$

Critical threshold for attacks:

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} k_{\min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right)$$

Size distribution for cascading failures:

$$p(s) \sim s^{-\alpha}$$

$$\alpha = \begin{cases} 3/2 & \gamma > 3 \\ \frac{\gamma}{\gamma - 1} & 2 < \gamma < 3 \end{cases}$$

ADVANCED TOPICS 8.A RANDOM NETWORKS AND PERCOLATION

RANDOM NETWORKS AND INFINITE DIMENSIONAL LATTICE

The percolation transition observed when nodes are randomly removed from a random network is characterized by the same critical exponent as percolation in $d > 6$ dimensions. This equivalence is illustrated by the following two-step argument.

i. The Cayley tree, a regular branching tree shown in Fig. 8.29a, is a frequently used as a model of an infinite dimensional lattice. Indeed, in d-dimensions the volume of a hypersphere of radius r is proportional to r^d, whereas its surface is r^{d-1}. Hence, in general, we have *suface* $\propto$ *volume*$^{1-1/d}$. For example, in $d = 2$ the area (volume) of a circle of radius r is πr^2 and its circumference (surface) is $2\pi r$. In $d = 3$, the volume depends on r as r^3 whereas the area increases as r^2. In the $d \to \infty$ limit the surface of a d-dimensional hypersphere is proportional to its volume, as $1/d$ becomes negligible. This proportionality is valid for the Cayley tree: the number of nodes in the outer layer of a Cayley tree (surface) equals the number of nodes inside the tree (volume). Hence, we can view the Cayley tree as an infinite dimensional object Fig. 8.29a.

ii. At the same time the Cayley tree captures the local topology of a random network. Indeed, in a very large random network the probability of finding loops is negligible. Hence, locally the network is a tree. To see this consider a cluster of three nodes occupied with pebbles on a d-dimensional cubic lattice (red nodes in $d = 2$, Fig. 8.29b). If we add an additional pebble, for it to be part of the cluster, it has to be adjacent to at least one of the three previous pebbles. We can place the new pebble in $3(2d - 2)$ possible spots so that it does not form a loop (green sites); only one of the site closes the loop (red site). Therefore, the probability that the four pebbles form a loop decreases as,

$$\frac{1}{3(2d-2)+1} \tag{8.19b}$$

which is negligible in the d → ∞ limit [2]. Consequently, locally the nodes in a random network form a tree, well approximated by the Cayley tree of Fig. 8.29a.

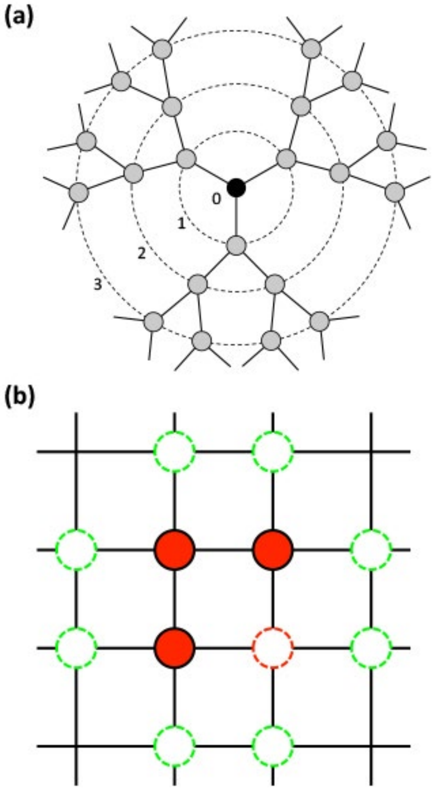


Figure 8.29
Infinite dimensional percolation

(a) The Cayley tree represents a model of an infinite dimensional lattice, as the number of nodes on its surface is proportional to the total number of nodes within the tree, corresponding to its volume. This is a property of an infinite dimensional lattice as well.

(b) If we form a four-node cluster in a square lattice, the likelihood that they form a loop is negligible. Indeed, we have positions for the new node to form a cluster (green dashed circles); one (red dashed) leads to a loop.

In summary, percolation in a random network is in the same universa ity class as percolation on a Cayley tree, which in turn is in the universality class of infinite dimensional percolation [2]. Therefore, $\gamma_p = 1$, $\beta = 1$ and $\gamma = 1$, the percolation critical exponents for $d = \infty$, characterize the behavior of the clusters near the critical point f_c when an fraction of nodes are randomly selected and removed from a random network.

PERCOLATION EXPONENTS FOR SCALE-FREE NETWORKS

To understand how a scale-free network breaks apart as we approach the threshold Eq. 8.7, we need to determine the critical exponents γ_p, β and ν. The calculations show that the scale-free property alters the value of these exponents, leading to systematic deviations from the exponents discussed in SECTION 8.1 that characterize random networks.

Let us start with the probability P_∞ that a randomly selected node belongs to the giant component. According to Eq. 8.2 this follows a power law near p_c (or f_c in the case of node removal). The calculations predict that for a scale-free network the exponent β depends on the degree exponent γ as [7, 47, 48, 49, 50]

$$\beta = \begin{cases} \dfrac{1}{3-\gamma} & 2 < \gamma < 3, \\ \dfrac{1}{\gamma - 3} & 3 < \gamma < 4, \\ 1 & \gamma > 4. \end{cases} \tag{8.20}$$

Hence, while for a random network (captured by the γ > 4 regime) we have β = 1, for most scale-free networks of practical interest β > 1. Therefore, the collapse of the giant component is steeper at the critical point in a scale-free network than in a random network.

The exponent describing the average component size near p_c follows [47]

$$\gamma_p = \begin{cases} 1 & \gamma > 3 \\ -1 & 2 < \gamma < 3. \end{cases} \tag{8.21}$$

The negative γ_p for γ < 3 may appear surprising. Note, however, that for γ < 3 we always have a giant component. Hence, the divergence Fig. 8.1 cannot be observed in this regime. For a random graph of arbitrary degree distribution the size distribution of the finite clusters follows [47, 49, 50]

$$n_s \sim s^{-\tau} e^{-s/s^*}. \tag{8.22}$$

Here, n_s is the number of clusters of size s and s^* is the crossover cluster size. At criticality

$$s^* \sim \left| q - q_c \right|^{-\sigma}. \tag{8.23}$$

The pertinent critical exponents are

$$\tau = \begin{cases} \frac{5}{2} & \gamma > 4 \\ \frac{2\gamma - 3}{\gamma - 2} & 2 < \gamma < 4, \end{cases}$$

$$\sigma = \begin{cases} \frac{3-\gamma}{\gamma - 2} & 2 < \gamma < 3 \\ \frac{\gamma - 3}{\gamma - 2} & 3 < \gamma < 4 \\ \frac{1}{2} & \gamma > 4. \end{cases} \tag{8.24}$$

Once again, the random network values $\tau = 5/2$ and $\sigma = 1/2$ are recovered for $\gamma > 4$. Finally, the exponent ν governs the average size of the finite components, obeying the scaling relation $\nu = (3 - \tau)/\sigma$. Hence,

$$\nu = \begin{cases} \frac{\gamma - 3}{\gamma - 2}, & 3 < \gamma < 4 \\ \frac{3-\gamma}{\gamma - 2}, & 2 < \gamma < 3. \end{cases} \tag{8.25}$$

In summary, the exponents describing the breakdown of a scale-free network depend on the degree exponent γ. This is true even in the range $3 < \gamma < 4$, where the percolation transition occurs at a finite threshold f_c. The mean-field behavior predicted for percolation in infinite dimensions, capturing the response of a random network to random failures, is recovered for $\gamma > 4$.

SECTION 8.9

ADVANCED TOPICS 8.B
MALLOY-REED CRITERIA

The purpose of this section is to derive the Malloy-Reed criteria BOX 8.7, which allows us to calculate the percolation threshold of an arbitrary network [6]. For a giant component to exist each node that belongs on average to it must be connected to at least two other nodes Fig. 8.8. Therefore, the average degree k_i of a randomly chosen node i that is part of the giant component should be at least 2. Denote with $p(k_i \mid i \leftrightarrow j)$ the joint probability that a node in a network with degree k_i is connected to an arbitrary node j that is part of the giant component. This conditional probability allows us to determine the expected degree of node i as

$$\langle k_i \mid i \leftrightarrow j \rangle = \sum_{k_i} k_i P(k_i \mid i \leftrightarrow j) = 2\ . \tag{8.26}$$

In other words, $\langle k_i \mid i \leftrightarrow j \rangle$ should be equal to two, the condition for node i to be part of the giant component. We can write the probability appearing in the sum Eq. 8.26 as

$$P(k_i \mid i \leftrightarrow j) = \frac{P(k_i, i \leftrightarrow j)}{P(i \leftrightarrow j)} = \frac{P(i \leftrightarrow j \mid k_i)p(k_i)}{P(i \leftrightarrow j)}\ \cdot \tag{8.27}$$

where we used Bayes' theorem in the last term. For a network with degree distribution p_k, in the absence of degree correlations, we can write

$$P(i \leftrightarrow j) = \frac{2L}{N(N-1)} = \frac{\langle k \rangle}{N-1} \qquad P(i \leftrightarrow j \mid k_i) = \frac{k_i}{N-1}\ \cdot \tag{8.28}$$

which expresses the fact that we can choose between $N - 1$ nodes to link to, each with probability $1/(N - 1)$ and that we can try this k_i times. We can now return to Eq. 8.26,

$$\sum_{k_i} k_i P(k_i \mid i \leftrightarrow j) = \sum_{k_i} k_i \frac{P(i \leftrightarrow j \mid k_i)p(k_i)}{P(i \leftrightarrow j)} = \sum_{k_i} k_i \frac{k_i p(k_i)}{\langle k \rangle} = \frac{\sum_{k_i} k_i^2 p(k_i)}{\langle k \rangle} \tag{8.29}$$

With that we arrive at the Malloy-Reed criteria Eq. 8.4, indicating that the condition to have a giant component is

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} > 2\ . \tag{8.30}$$

ADVANCED TOPICS 8.C
CRITICAL THRESHOLD UNDER RANDOM FAILURES

The purpose of this section is to derive Eq. 8.7, that provides the critical threshold for random node removal [7, 50]. The random removal of an f fraction of nodes has two consequences:

- It alters the degree of some nodes, as nodes that were previously connected to the removed nodes will lose some links [$k \rightarrow k' \leq k$].
- It changes the degree distribution, as the neighbors of the missing nodes will have an altered degree [$p_k \rightarrow p'_{k'}$].

To be specific, after we randomly remove an f fraction of nodes, a node with degree k turns into a node with degree k' with probability

$$\begin{pmatrix} k \\ k' \end{pmatrix} f^{k-k'}(1-f)^{k'} \quad k' \leq k \ . \tag{8.31}$$

The first f -dependent term in Eq. 8.31 accounts for the fact that the selected node lost $(k - k')$ links, each with probability f; the next term accounts for the fact that node removal leaves k' links untouched, each with probability $(1 - f)$.

The probability that we have a degree-k node in the original network is p_k; the probability that we have a new node with degree k' in the new network is

$$p'_{k'} \simeq \sum_{k=k'}^{k\equiv k'} p_k \begin{pmatrix} k \\ k' \end{pmatrix} f^{k-k'}(1-f)^{k'} . \tag{8.32}$$

Let us assume that we know $\langle k \rangle$ and $\langle k^2 \rangle$ for the original degree distribution p_k. Our goal is to calculate $\langle k' \rangle$, $\langle k'^2 \rangle$ for the new degree distribution $p'_{k'}$, obtained after we randomly removed an f fraction of the nodes. For this we write

$$\begin{aligned}\langle k'\rangle_f &= \sum_{k'=0}^{\infty} k' p'_{k'} \\ &= \sum_{k'=0}^{\infty} k' \sum_{k=k'}^{\infty} p_k \left(\frac{k!}{k'!(k-k')} \right) f^{k-k'} (1-f)^{k'} \\ &= \sum_{k'=0}^{\infty} k' \sum_{k'=k'}^{\infty} p_k \frac{k!}{k'!(k-k')!} f^{k-k'} (1-f)^{k'} .\end{aligned} \tag{8.33}$$

The sum above is performed over the triangle shown in Fig. 8.30. We can check that we are performing the same sum if we change the order of summation together with the limits of the sums as

$$= \sum_{k'=0}^{\infty} \sum_{k=k'}^{\infty} = \sum_{k=0}^{\infty} \sum_{k'=0}^{k} . \tag{8.34}$$

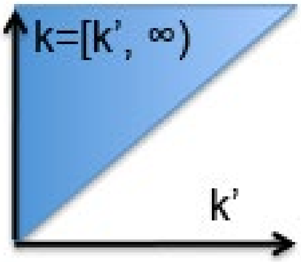


Figure 8.30
The integration domain

In Eq. 8.34 we changed the integration order, *i.e.* the order of the two sums. We could do so because both sums are defined over the triangle shown in blue in the figure.

Hence we obtain

$$\begin{aligned}\langle k'\rangle_f &= \sum_{k=0}^{\infty} k' \sum_{k'=0}^{k} p_k \frac{k!}{(k'-1)!(k-k')!} f^{k-k'} (1-f)^{k'-1} (1-f) \\ &= \sum_{k=0}^{\infty} (1-f) k p_k \sum_{k'=0}^{k} \frac{(k-1)!}{(k'-1)!(k-k')!} f^{k-k'} (1-f)^{k'-1} \\ &= \sum_{k=0}^{\infty} (1-f) k p_k \sum_{k'=0}^{k} \binom{k-1}{k'-1} f^{k-k'} (1-f)^{k'-1} \\ &= \sum_{k=0}^{\infty} (1-f) k p_k \\ &= (1-f)\langle k\rangle .\end{aligned} \tag{8.35}$$

This connects ⟨k'⟩ to the original ⟨k⟩ after the random removal of an f fraction of nodes. We perform a similar calculation for ⟨k'^2⟩:

$$\begin{aligned}\langle k'2\rangle_f &= \langle k'(k'-1)+k'\rangle_f \\ &= \langle k'(k'-1)\rangle_f + \langle k'\rangle_f \\ &= \sum_{k'=0}^{\infty} k'(k'-1) p'_{k'} + \langle k'\rangle_f .\end{aligned} \tag{8.36}$$

Again, we change the order of the sums Fig. 8.30

$$\begin{aligned}\langle k'(k'-1)\rangle_f &= \sum_{k'=0}^{\infty} k'(k'-1) p'_{k'} \\ &= \sum_{k'=0}^{\infty} k'(k'-1) \sum_{k=k'}^{\infty} p_k \binom{k}{k'} f^{k-k'} (1-f)^{k'} \\ &= \sum_{k'=0}^{\infty} k'(k'-1) \sum_{k'=0}^{k} p_k \frac{k!}{k'!(k-k')!} f^{k-k'} (1-f)^{k'} \\ &= \sum_{k'=0}^{\infty} \sum_{k'=0}^{k} p_k \frac{k!}{(k'-2)!(k-k')!} f^{k-k'} (1-f)^{k'-2} (1-f)^2 \\ &= \sum_{k=0}^{\infty} (1-f)^2 k(k-1) p_k \sum_{k'=0}^{k} \frac{(k-2)!}{(k'-2)!(k-k')!} f^{k-k'} (1-f)^{k'-2} \\ &= \sum_{k=0}^{\infty} (1-f)^2 k(k-1) p_k \sum_{k'=0}^{k} \binom{k-2}{k'-2} f^{k-k'} (1-f)^{k'-2} \\ &= \sum_{k=0}^{\infty} (1-f)^2 k(k-1) p_k \\ &= (1-f)^2 \langle k(k-1)\rangle .\end{aligned} \tag{8.37}$$

Hence we obtain

$$\begin{aligned}
\left\langle k'^2 \right\rangle_f &= \langle k'(k'-1)+k' \rangle_f \\
&= \langle k'(k'-1) \rangle_f + \langle k' \rangle_f \\
&= (1-f)^2 \langle k(k-1) \rangle + (1-f)\langle k \rangle \\
&= (1-f)^2 \left(\langle k^2 \rangle - \langle k \rangle \right) + (1-f)\langle k \rangle \qquad (8.38) \\
&= (1-f)^2 \langle k^2 \rangle - (1-f)^2 \langle k \rangle + (1-f)\langle k \rangle \\
&= (1-f)^2 \langle k^2 \rangle - \left(-f^2 + 2f - 1 + 1 - f \right)\langle k \rangle \\
&= (1-f)^2 \langle k^2 \rangle + f(1-f)\langle k \rangle .
\end{aligned}$$

which connects $\langle k'^2 \rangle$ to the original $\langle k^2 \rangle$ after the random removal of an f fraction of nodes. Let us put the results Eq. 8.35 and Eq. 8.38 together:

$$\langle k' \rangle_f = (1-f)\langle k \rangle \qquad (8.39)$$

$$\langle k' \rangle_f = (1-f)^2 \langle k^2 \rangle + f(1-f)\langle k \rangle \qquad (8.40)$$

According to the Malloy-Reed criteria the breakdown threshold is given by the equality

$$\kappa \equiv \frac{\langle k'^2 \rangle_f}{\langle k' \rangle_f} = 2 \qquad (8.41)$$

Inserting Eq. 8.38 and Eq. 8.40 into Eq. 8.41 we obtain our final result Eq. 8.7 in SECTION 8.3,

$$f_c = 1 - \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1} \qquad (8.42)$$

providing the breakdown threshold of networks with arbitrary p_k under random node removal.

SECTION 8.11

ADVANCED TOPICS 8.D BREAKDOWN OF A FINITE SCALE-FREE NETWORK

The goal of this section is to determine the dependence Eq. 8.10 of the breakdown threshold of a scale-free network on the network size N. We start by calculating the m^{th} moment of a power-law distribution

$$\langle k^m \rangle = (\gamma-1)k_{\min}^{\gamma-1}\int_{k_{min}}^{k_{\max}} k^{m-\gamma}dk \qquad = \frac{(\gamma-1)}{(m-\gamma+1)}k_{\min}^{\gamma-1}[k^{m-\gamma+1}]_{k_{min}}^{k_{max}} \tag{8.43}$$

As discussed in CHAPTER 4, we have

$$k_{max} = k_{min}N^{\frac{1}{\gamma-1}} \tag{8.44}$$

$$\langle k^m \rangle = \frac{(\gamma-1)}{(m-\gamma+1)}k_{min}^{\gamma-1}[k_{\max}^{m-\gamma+1} - k_{min}^{m-\gamma+1}] \tag{8.45}$$

To calculate f_c we determine the ratio

$$\kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{(2-\gamma)}{(3-\gamma)}\frac{k_{\max}^{3-\gamma} - k_{\min}^{3-\gamma}}{k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}}\ , \tag{8.46}$$

which in the $N \to \infty$ (and hence the $k_{max} \to \infty$) limit depends on γ as

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} = \left|\frac{2-\gamma}{3-\gamma}\right| \begin{cases} k_{min} & \gamma > 3 \\ k_{\max}^{3-\gamma}k_{\min}^{\gamma-2} & 3 > \gamma > 2 \\ k_{max} & 2 > \gamma > 1 \end{cases} \tag{8.47}$$

The breakdown threshold is given by

$$f_c = 1 - \frac{1}{\kappa - 1}\ , \tag{8.48}$$

where κ is given by Eq. 8.46. Inserting Eq. 8.43 into Eq. 8.42 and Eq. 8.47, we obtain

$$f_c \simeq 1 - \frac{C}{N^{\frac{3-\gamma}{\gamma-1}}}\ , \tag{8.49}$$

which is Eq. 8.10, providing the dependence of f_c on N.

SECTION 8.12

ADVANCED TOPICS 8.E
THRESHOLD UNDER ATTACK

The goal of this section is to explore how a scale-free network responds to attack, by deriving Eq. 8.12. In other words, we calculate f_c for an uncorrelated scale-free networks, generated by the configuration model with $p_k = c \cdot k^{-\gamma}$ where $k = k_{min}, ..., k_{max}$ and $c \approx (\gamma - 1)/(k_{min}^{-\gamma+1} - k_{max}^{-\gamma+1})$.

The removal of an *f* fraction of nodes in a decreasing order of their degree (hub removal) has two effects [9, 50]:

(i) The maximum degree of the network changes from k_{max} to k'_{max}.

(ii) The links connected to the removed hubs are also removed, changing the degree distribution of the remaining network.

The resulting network is still uncorrelated, therefore we can use the Molloy-Reed criteria to determine the existence of a giant component. We start by considering the impact of (i). The new upper cutoff, k'_{max}, is given by

$$f = \int_{k_{\max}}^{k'_{\max}} p_k dk = \frac{\gamma - 1}{\gamma - 1} \frac{k'^{\,-\gamma+1}_{\max} - k^{-\gamma+1}_{\max}}{k^{-\gamma+1}_{\min} - k^{-\gamma+1}_{\max}}, \qquad (8.50)$$

If we assume that $k_{max} \gg k'_{max}$ and $k_{max} \gg k_{min}$ (true for large scale-free networks with natural cutoff), we can ignore the k_{max} terms, obtaining

$$f = \left(\frac{k'_{\max}}{k_{\min}} \right)^{-\gamma+1}, \qquad (8.51)$$

which leads to

$$k'_{max} = k_{min} f^{\frac{1}{1-\gamma}}. \qquad (8.52)$$

Eq. 8.51 provides the new maximum degree of the network after we remove an *f* fraction of the hubs.

Next, we turn to (ii), accounting for the fact that hub removal also removes the links connected to these hubs, changing the degree distribution $p_k \rightarrow p'_k$. In the absence of degree correlations we can assume that the links of the removed hubs connect to randomly selected stubs. Consequently, we calculate the fraction of links removed 'randomly', $\tilde{f}$, as a consequence of removing an f fraction of the hubs:

$$\tilde{f} = \frac{\int_{k_{\max'}}^{k_{\max}} k p_k dk}{\langle k \rangle} = \frac{1}{\langle k \rangle} c \int_{k_{\max'}}^{k_{\max}} k^{-\gamma+1} dk \tag{8.53}$$

$$= \frac{1}{\langle k \rangle} \frac{1-\gamma}{2-\gamma} \frac{{k'_{\max}}^{-\gamma+2} - k_{\max}^{-\gamma+2}}{{k_{\min}}^{-\gamma+1} - k_{\max}^{-\gamma+2}}$$

Ignoring the k_{max} again and using $\langle k \rangle \approx \frac{\gamma-1}{\gamma-2} k_{\min}$ we obtain

$$\tilde{f} = \left(\frac{k'_{\max}}{k_{\min}} \right)^{-\gamma+2}. \tag{8.54}$$

Using Eq. 8.49 we obtain:

$$\tilde{f} = f^{\frac{2-\gamma}{1-\gamma}}. \tag{8.55}$$

For $\gamma \rightarrow 2$ we have $\tilde{f} \rightarrow 1$, which means that the removal of a tiny fraction of the hubs removes all links, potentially destroying the network. The reason is that for $\gamma = 2$ the hubs dominate the network. The degree distribution of the remaining network is

$$p'_{k'} = \sum_{k=k_{\min}}^{k'_{\max}} \begin{pmatrix} k \\ k' \end{pmatrix} \tilde{f}^{k-k'} (1-\tilde{f})^{k'} p_k. \tag{8.56}$$

Note that we obtained the same degree distribution as Eq. 8.27 in ADVANCED TOPICS 5.B. This means that now we can use the calculation method developed for random node removal. To be specific, we calculate κ for a scale-free network with k_{min} and k'_{max} using Eq. 8.45:

$$\kappa = \frac{2-\gamma}{3-\gamma} \frac{{k'}_{\max}^{3-\gamma} - k_{\min}^{3-\gamma}}{{k'}_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}}. \tag{8.57}$$

Substituting into this Eq. 8.57 we have

$$\kappa = \frac{2-\gamma}{3-\gamma} \frac{k_{\min}^{3-\gamma} f^{(3-\gamma)/(1-\gamma)} - k_{\min}^{3-\gamma}}{k_{\min}^{2-\gamma} f^{(2-\gamma)/(1-\gamma)} - k_{\min}^{2-\gamma}} = \frac{2-\gamma}{3-\gamma} k_{\min} \frac{f^{(3-\gamma)/(1-\gamma)} - 1}{f^{(2-\gamma)/(1-\gamma)} - 1}. \tag{8.58}$$

After simple transformations we obtain:

$$f_c^{\frac{2-\gamma}{1-\gamma}} = 2 + \frac{2-\gamma}{3-\gamma} k_{min} \left(f_c^{\frac{3-\gamma}{1-\gamma}} - 1 \right). \tag{8.59}$$

which is Eq. 8.12 in SECTION 8.3.

ADVANCED TOPICS 8.F
MODELING CASCADING FAILURES

In this section we discuss two additional cascading failure models, that together with the Failure Propagation Model and the Branching Model discussed in SECTION 8.5, help illustrate the universality of the mechanisms governing cascading failures.

OVERLOAD MODEL

The overload model was proposed to capture the emergence of large blackouts [17]. The model has N identical nodes (components), each node j assigned an initial load L_j, which is a random variable uniformly distributed between L_{min} and L_{max}. A node fails when its load exceeds a preassigned threshold, L_{fail} assumed to be the same for all nodes. When a node fails, a fixed amount of power P is transferred to all other nodes in the network. Hence the impact of each failure is global, affecting not only the neighbors of the failed node, but all other nodes.

This mimics the fact that after each node or link failure the electric currents rearrange themselves globally. Hence, the impact of a local failure is not limited to the failed node's or link's direct neighbors, but can alter the current flowing through all nodes and links. Consequently, the model's behavior is independent of the network topology: the system behaves as if it would be fully connected.

To begin a cascade, we assume an initial disturbance that adds to the load of each component an additional load P. Some nodes with high initial loads L_i may fail and each such failure distributes an additional load P to the remaining nodes, potentially causing further failures Fig. 8.31a, b. The model's behavior is captured by the phase diagram of Fig. 8.31c, predicting three regimes:

- **Subcritical regime**
  If the initial load L_{min} is under a global threshold, then most local perturbations die out, hence we do not observe global avalanches. In this regime the avalanche size distribution is bounded Fig. 8.31d.

- **Supercritical regime**
  If the initial load L_{min} is over the threshold, then the perturbations propagate, resulting in avalanches that involve most nodes. In this regime the avalanche size distribution is again bounded and bimodal, capturing the coexistence of only very large and very small avalanches.

- **Critical regime**
  At the boundary of the subcritical and the supercritical regime, the avalanche size distribution follows a power law. The observed avalanche exponent is $\alpha = 3/2$.

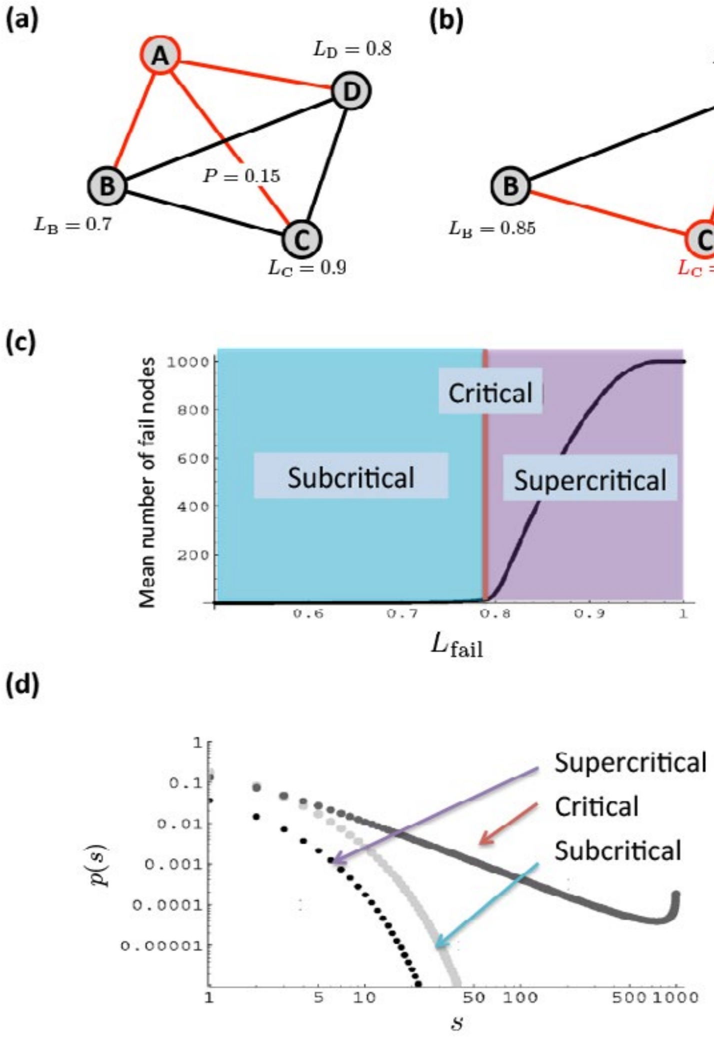


Figure 8.31

Avalanche sizes in the overload

(a) To illustrate the dynamics of the overload model, we start with four nodes, each assigned an initial load L_i. When node A is perturbed by increasing its load above the threshold $L_{fail} = 1$, it fails and redistributes a fixed amount of power $P = 0.15$ to all other nodes in the network. Hence, the remaining nodes will now have a load $L_i = L_i + P$.

(b) The extra power will cause node C to also fail, as now $L_c = 1.05 > L_{fail}$. That event adds another P power to the remaining nodes, prompting them to fail too, allowing the avalanche to sweep the full network.

(c) Phase Diagram: Mean number of failed components in function of L_{fail} in the overload model. At the critical loading $L_{fail} = 0.8$ the model changes its behavior, supporting large avalanches.

(d) Log-log plot showing the distribution of the number of components failed for three values of L_{fail}. Note the power-law scaling of $p(s)$ for the critical load $L = 0.8$.

After [17].

SANDPILE MODEL

A common feature of the cascade and the overload models is that the empirically documented power-law behavior appears only in the critical regime. Therefore, we need to tune the model parameters to reach this critical state. Outside this critical regime the avalanche sizes follow an exponential distribution. This raises an important question: how does nature drive these systems to criticality? Attempts to answer this question have lead to a family of models collectively called self-organized critical (SOC) models [51]. These models do not require external tuning, but self-organize to a critical state. Probably the best known of these is the sandpile model [52] that mimics the dropping of single grains on a plane. When the pile gets too high on a site, it topples by moving its grains to the neighboring sites. The model is typically defined on a lattice. With interest in cascading failures, its network version has been explored [30, 31]. The model starts from a network with an arbitrary wiring diagram and evolves following these steps:

1. Each node i is given a prescribed threshold $z_i(\le k_i)$. We denote with $[z_i]$ the smallest integer not smaller than z_i $(z_i \le k_i)$.

2. At each time step a grain is added at a randomly chosen node i, so that the height of the node i increases by one ($h_i \rightarrow h_i + 1$).

3. If the height of node *i* reaches or exceeds z_i, it becomes unstable and z_i grains on the node topple to z_i randomly chosen adjacent nodes among the k_i neighbors of node *i*. Therefore, $h_i \rightarrow h_i + 1 \text{-} z_i$ and $h_i \rightarrow h_i + 1$ for z_i neighboring nodes of *j*.

4. If this toppling causes any of the adjacent nodes to become unstable, z_i subsequent toppings follow until there is no unstable node left. This process defines an avalanche.

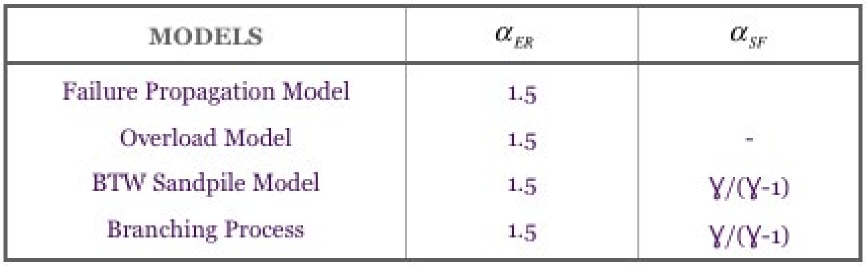
We repeat the steps 2–4 and determine the avalanche size *s* in each case, where *s* represents the number of toppling events in a given avalanche. The analytical calculations indicate that for a random network the avalanche size distribution follows [30, 31]

$$p(s) \sim s^{-3/2}. \tag{8.60}$$

For a scale-free network the distribution depends on γ as:

$$p(s) \sim \begin{cases} s^{-\gamma/(\gamma-1)} & 2 < \gamma < 3 \\ s^{-3/2(\ln s)^{-1/2}} & \gamma = 3 \\ s^{-3/2} & \gamma > 3 \end{cases} . \tag{8.61}$$

In summary, the four cascading failure models discussed in this chapter predict a critical regime where the avalanche size distribution follows a power law. A summary of the avalanche exponents obtained for these models is provided in Table 8.3.



MODELS	α_{ER}	α_{SF}
Failure Propagation Model	1.5	
Overload Model	1.5	-
BTW Sandpile Model	1.5	γ/(γ-1)
Branching Process	1.5	γ/(γ-1)

Table 8.3

The avalanche exponents in the explored models

The avalanche exponents for the four models supporting cascading failures. Here α_{ER} corresponds to the avalanche exponent if the underlying network is random; α_{SF} describes avalanches propagating on a scale-free network.

ADVANCED TOPICS 8.G
ATTACK AND ERROR TOLERANCE OF REAL NETWORKS

In this section we explore the attack and error curves for the ten reference networks discussed in Tables 4.1 and Eq. 8.2. The corresponding curves are shown in Fig. 8.33. Their inspection reveals several patterns, confirming the results discussed in this chapter:

- For all networks the error and attack curves separate, confirming the Achilles' Heel property: real networks are robust to random failures but are fragile to attacks.

- The degree of separation between the error and attach curves depends on the underlying degree heterogeneity and the average degree of each network. For example, for the citation and the actor networks we observe a very large f_c for attacks, at of 0.5 and 0.75, respectively. This is mainly because these networks have an unusually large $\langle k \rangle$, with $\langle k \rangle = 20.8$ for citations and $\langle k \rangle = 83.7$ for the actor network. Their high robustness to attacks is attributed to their high link density.

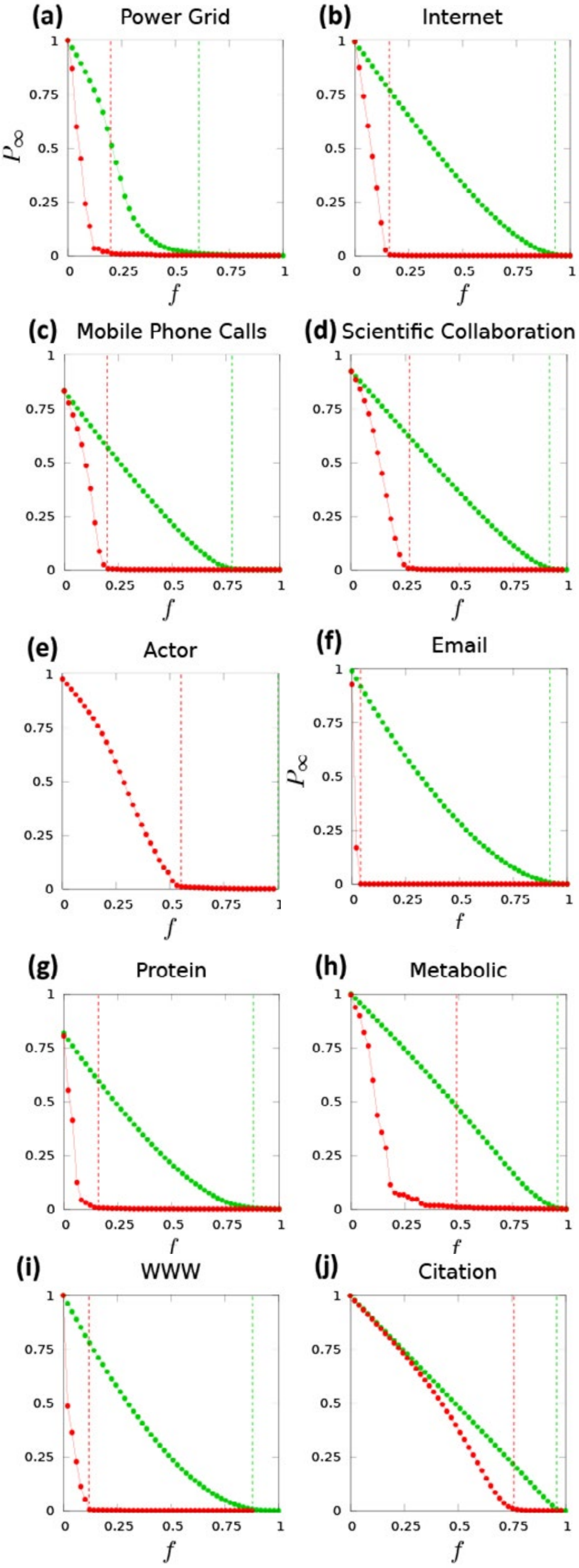


Figure 8.32

Error and attack curves for reference networks

The error (green) and attack (red) curves for the ten reference networks listed in **Table 4.1**. The green vertical line corresponds to the estimated f_c for errors, while the red line corresponds to f_c for attacks. We estimated f_c as the point where the giant component first drops below 1% of its original size. This procedure in most systems accurately captures the point where P_∞ drops to zero. The only exception is for the metabolic network, for which $f < 0.25$, but a small cluster persists, pushing the reported f_c to $f_c \simeq 0.5$. Another way to detect the critical point fc is to plot the size of the second largest component in function of the fraction of deleted nodes *f*. For infinite networks S_2 diverges at f_c; for finite networks we do not observe a true phase transition, but S_2 has a maximum at f_c. Hence we can estimate the critical point f_c by searching for this maximum.

ADVANCED TOPICS 8.H
THE OPTIMAL DEGREE DISTRIBUTION

The purpose of this section is to derive the bimodal distribution that simultaneously optimizes a network's topology against attacks and failures, as discussed in SECTION 8.6 [36]. Let us assume, as we did in Eq. 8.17, that the degree distribution is bimodal, consisting of two delta functions:

$$p_k = (1-r)\delta(k-k_{\min}) + r\delta(k-k_{\max})\,. \tag{8.62}$$

First, we calculate the total threshold, f^{tot}, as a function of r and k_{max} for a fixed $\langle k\rangle$. To obtain analytical expressions for f_c^{rand} and f_c^{targ}, we start by calculating the moments of the bimodal distribution Eq. 8.60,

$$\begin{aligned}\langle k\rangle &= (1-r)k_{\min} + rk_{\max}\\ \langle k^2\rangle &= (1-r)k_{\min}^2 + rk_{\max}^2 = \frac{(\langle k\rangle - rk_{\max})^2}{1-r} + rk_{\max}^2\,.\end{aligned} \tag{8.63}$$

Inserting these into Eq. 8.7 we obtain

$$f_c^{rand} = \frac{\langle k\rangle^2 - 2r\langle k\rangle k_{\max} - 2(1-r)\langle k\rangle + rk_{\max}}{\langle k\rangle^2 - 2r\langle k\rangle k_{\max} - (1-r)\langle k\rangle + rk_{\max}} \tag{8.64}$$

To determine the threshold for targeted attack, we must consider the fact that we have only two types of nodes, an r fraction of nodes having the larger degree k_{max} and the remaining (1 – r) fraction has the smaller degree k_{min}. Hence hub removal can either remove all hubs (case (i)), or only some fraction of them (case (ii)):

(i) $f_c^{\text{targ}} > r$. In this case all hubs have been removed, hence all nodes left after the targeted attack have degree k_{min}. We therefore obtain

$$f_c^{t\arg} = r + \frac{1-r}{\langle k\rangle - rk_{\max}}\left\{\langle k\rangle\frac{\langle k\rangle - rk_{\max} - 2(1-r)}{\langle k\rangle - rk_{\max} - (1-r)} - rk_{\max}\right\}\,. \tag{8.65}$$

(ii) $f_c^{\text{targ}} < r$. In this case the removed nodes are all from the higher degree group, leaving behind some k_{max} nodes. Hence we obtain

$$f_c^{t\arg} = \frac{\langle k\rangle^2 - 2r\langle k\rangle k_{\max} + rk_{\max}^2 - 2(1-r)\langle k\rangle}{k_{\max}(k_{\max}-1)(1-r)}. \quad (8.66)$$

With the thresholds Eq. 8.64 - Eq. 8.66, we can now evaluate the total threshold f_c^{tot} given by Eq. 8.16. We can obtain an expression for the optimal value of k_{max} as a function of r by determining the value of k for which f_c^{tot} is maximal. Using Eq. 8.64 and Eq. 8.66, we find that for small r the optimal value of k_{max} can be approximated by

$$k_{\max} \sim \left\{ \frac{2\langle k\rangle^2 (\langle k\rangle - 1)^2}{2\langle k\rangle - 1} \right\}^{1/3} r^{-2/3} \equiv Ar^{-2/3}. \quad (8.67)$$

Using this result and Eq. 8.14, for small r

$$f_c^{tot} = f_{\underset{-}{c}} - \frac{3\langle k\rangle}{A^2} r^{1/3} + O(r^{2/3}). \quad (8.68)$$

Thus f tot c approaches the theoretical maximum when r approaches zero. For a network of N nodes, the maximum value of f_c^{tot} is obtained when $r = 1/N$, being the smallest value consistent with having at least one node of degree k_{max}. Given this r the equation determining the optimal k_{max}, representing the size of the central hubs, is [36]

$$k_{\max} = AN^{2/3}, \quad (8.69)$$

where A is defined in Eq. 8.67.

HOMEWORK

1. We have seen in SECTION 8.2 that the value of p_c decreases with th lattice dimension: for a simple cubic lattice, representing the three dimensional version of a square lattice, we have $p_c = 0.2488$, less than half of $p_c = 1/2$ for two dimensional square lattice.

   Can you offer an intuitive explanation why does p_c decrease with the lattice dimension?

BIBLIOGRAPHY

[1] R. Albert, H. Jeong, and A.-L. Barabási. Attack and error tolerance of complex networks. Nature 406: 378, 2000.

[2] D. Stauffer and A. Aharony, Introduction to Percolation Theory. Taylor and Francis. London, 1994.

[3] A. Bunde and S. Havlin. Fractals and Disordered Systems. Springer, 1996.

[4] B. Bollobás, O. Riordan. Percolation. Cambridge University Press. Cambridge, 2006.

[5] S. Broadbent and J. Hammersley. Percolation processes I. Crystals and mazes. Proceedings of the Cambridge Philosophical Society 53: 629, 1957.

[6] M. Molloy and B. Reed. Random Structures and Algorithms 6:161, 1995.

[7] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. Resilience of the Internet to random breakdowns. Phys. Rev. Lett. 85: 4626, 2000.

[8] D. S. Callaway, M. E. J. Newman, S. H. Strogatz. and D. J. Watts. Network robustness and fragility: Percolation on random graphs. Phys. Rev. Lett. 85: 5468–5471, 2000.

[9] R. Cohen, K. Erez, D. ben-Avraham and S. Havlin. Breakdown of the Internet under intentional attack. Phys. Rev. Lett 86: 3682, 2001.

[10] B. Bollobás and O. Riordan. Robustness and Vulnerability of Scale-Free Random Graphs. Internet Mathematics 1:2003.

[11] D.N. Kosterev, C.W. Taylor and W.A. Mittlestadt. Model Validation of the August 10, 1996 WSCC System Outage. IEEE Transactions on Power

Systems 14:1999.

[12] C. Labovitz, A. Ahuja and F. Jahasian. Experimental Study of Internet Stability and Wide-Area Backbone Failures. Proceedings of IEEE FTCS, Madison, WI, 1999.

[13] A. G. Haldane and R. M. May. Systemic risk in banking ecosystems. Nature 469: 351-355, 2011.

[14] T. Roukny, H. Bersini, H. Pirotte, G. Caldarelli and S. Battiston. Default Cascades in Complex Networks: Topology and Systemic Risk. Scientific Reports 3: 2759, 2013.

[15] G. Tedeschi, A. Mazloumian, M. Gallegati, and D. Helbing. Bankruptcy cascades in interbank markets. PLoS One 7: e52749, 2012.

[16] D. Helbing. Globally networked risks and how to respond. Nature 497:2013.

[17] I. Dobson, B. A. Carreras, V. E. Lynch and D. E. Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. CHAOS 17: 026103, 2007.

[18] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In Proceedings of thefourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 65-74, 2011.

[19] Y. Y. Kagan. Accuracy of modern global earthquake catalogs. Phys. Earth Planet. Inter. 135:173, 2003.

[20] M. Nagarajan, H. Purohit, and A. P. Sheth. A Qualitative Examination of Topical Tweet and Retweet Practices. In ICWSM, 2010.

[21] P. Fleurquin, J.J. Ramasco and V.M. Eguiluz. Systemic delay propagation in the US airport network. Scientific Reports 3: 1159, 2013.

[22] B. K. Ellis, J. A. Stanford, D. Goodman, C. P. Stafford, D.L. Gustafson, D. A. Beauchamp, D. W. Chess, J. A. Craft, M. A. Deleray, and B. S. Hansen. Long-term effects of a trophic cascade in a large lake ecosystem. PNAS. 108: 1070, 2011

[23] V. R. Sole, M. M. Jose. Complexity and fragility in ecological networks. Proc. R. Soc. Lond. B 268:2039, 2001.

[24] F. Jordán and I. Scheuring. Can keystones help in background extinction? preprint, 2000.

[25] S.L. Pimm and P. Raven. Biodiversity: Extinction by numbers. Nature 403: 843, 2000.

[26] World Economic Forum, Building Resilience in Supply Chains. World Economic Forum, 2013.

[27] Joint Economic Committee of US Congress. Your flight has been delayed again: Flight delays cost passengers, airlines and the U. S. economy billions. Available at http://www.jec.senate.gov, May 22. 2008.

[28] I. Dobson, B. A. Carreras, and D. E. Newman. A loadingdependent model of probabilistic cascading failure. Probability in the Engineering and Informational Sciences 19: 15, 2005.

[29] D. Watts. A simple model of global cascades on random networks. PNAS 99: 5766, 2002.

[30] K.-I. Goh, D.-S. Lee, B. Kahng, and D. Kim. Sandpile on scale-free networks. Phys. Rev. Lett. 91:148701, 2003.

[31] D.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim. Sandpile avalanche dynamics on scale-free networks. Physica A, 338: 84, 2004.

[32] M. Ding and W. Yang. Distribution of the first return time in fractional Brownian motion and its application to the study of onoff intermittency. Phys. Rev. E. 52: 207-213, 1995.

[33] A. E. Motter and Y.-C. Lai. Cascade-based attacks on complex networks. Physical Review E 66: 065102, 2002.

[34] Z. Kong and Edmund M. Yeh. Resilience to Degree-Dependent and Cascading Node Failures in Random Geometric Networks. IEEE Transactions on Information Theory 56: 5533, 2010.

[35] G. Paul, S. Sreenivas, an and H. E. Stanley. Resilience of complex networks to random breakdown. Phys. Rev. E 72, 056130, 2005.

[36] G. Paul, T. Tanizawa, S. Havlin, and H. E. Stanley. Optimization of robustness of complex networks. European Physical Journal B 38: 187–191, 2004.

[37] A.X. C. N. Valente, A. Sarkar, and H. A. Stone. Two-peak and three-peak optimal complex networks. Phys. Rev. Lett. 92: 118702, 2004.

[38] T. Tanizawa, G. Paul, R. Cohen, S. Havlin, and H. E. Stanley. Optimization of network robustness to waves of targeted and random attacks. Phys. Rev. E 71: 047101, 2005.

[39] A.E. Motter. Cascade control and defense in complex networks. Phys. Rev. Lett. 93: 098701, 2004.

[40] A. Motter, N. Gulbahce, E. Almaas, A.-L. Barabási. Predicting synthetic rescues in metabolic networks. Molecular Systems Biology 4: 1-10,

2008.

[41] R.V. Sole, M. Rosas-Casals, B. Corominas-Murtra, and S. Valverde. Robustness of the European power grids under intentional attack. Phys. Rev. E 77: 026102, 2008.

[42] R. Albert, I. Albert, and G.L. Nakarado. Structural Vulnerability of the North American Power Grid. Phys. Rev. E 69: 025103 R, 2004.

[43] C.M. Schneider, N. Yazdani, N.A.M. Araújo, S. Havlin and H.J. Herrmann. Towards designing robust coupled networks. Scientific Reports 3: 1969, 2013.

[44] A.-L. Barabási. Linked: The New Science of Networks. Plume, New York, 2002.

[45] C.M. Song, S. Havlin, H.A and Makse. Self-similarity of complex networks. Nature 433:392, 2005.

[46] S.V. Buldyrev, R. Parshani, G. Paul, H.E. Stanley and S. Havlin. Catastrophic cascade of failures in interdependent networks. Nature 464: 08932, 2010.

[47] R. Cohen, D. ben-Avraham and S. Havlin. Percolation critical exponents in scale-free networks. Phys. Rev. E 66: 036113, 2002.

[48] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Anomalous percolation properties of growing networks. Phys. Rev. E 64: 066110, 2001.

[49] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 64: 026118, 2001.

[50] R. Cohen and S. Havlin. Complex Networks: Structure, Robustness and Function. Cambridge University Press. Cambridge, UK, 2010. [51] P. Bak. How Nature Works: The Science of Self-Organized Criticality. New York, Copernicus, 1996.

[52] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: an explanation of noise. Phys. Rev. Lett. 59: 381, 1987.