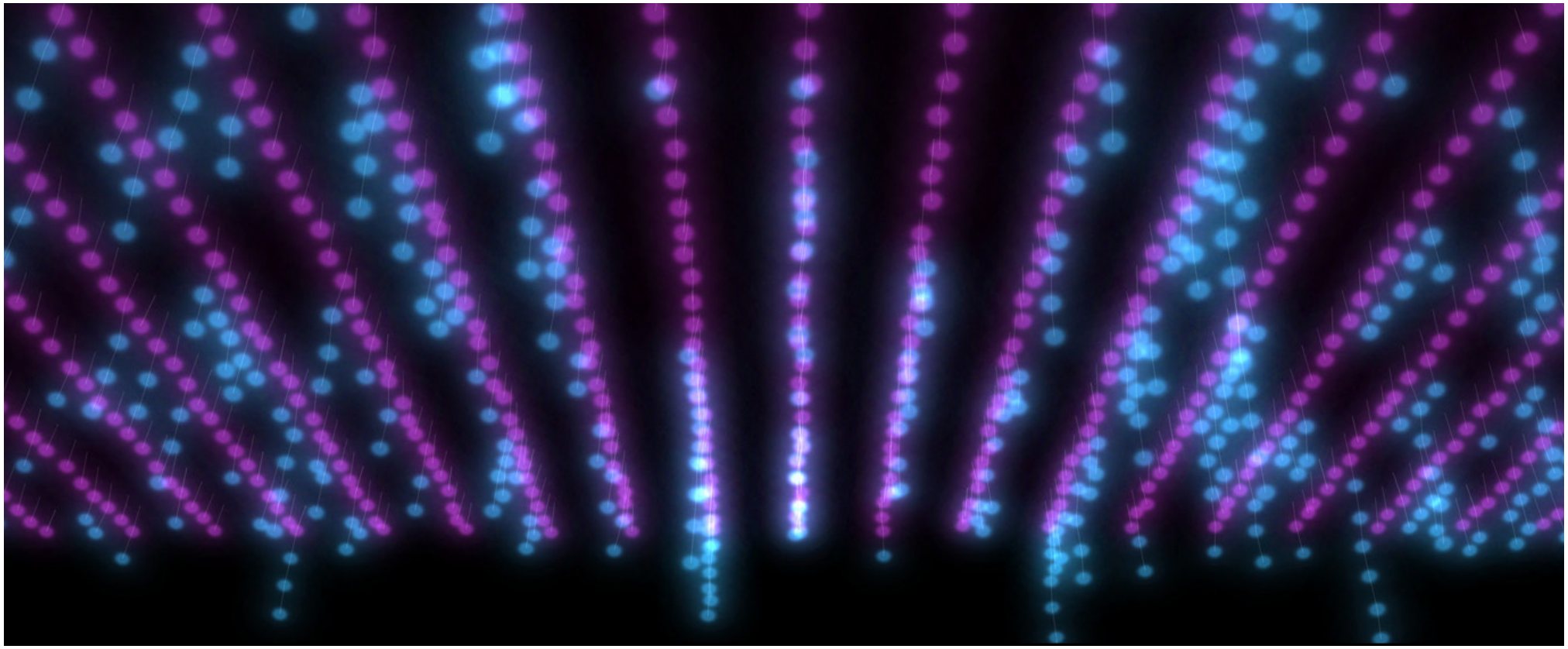
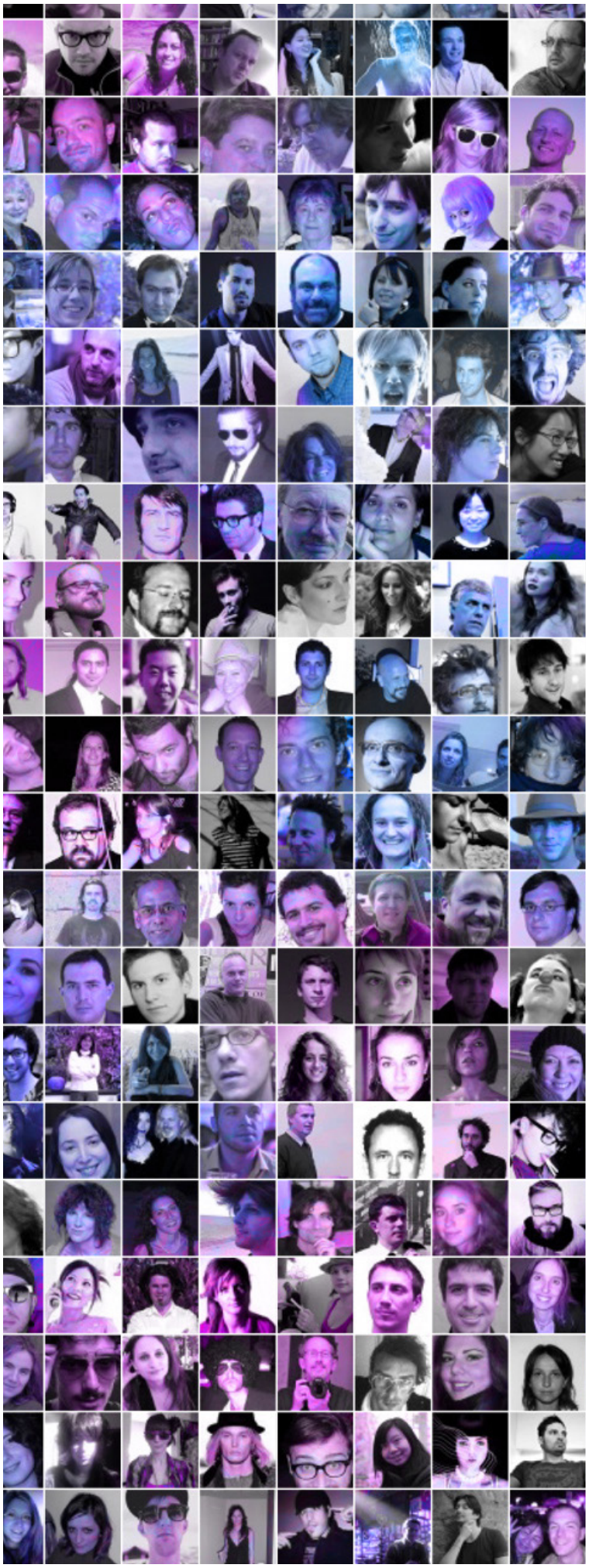


Network Science

Albert-László Barabási

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

PDF Version:
November 2012



CHAPTER 1
INTRODUCTION

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.

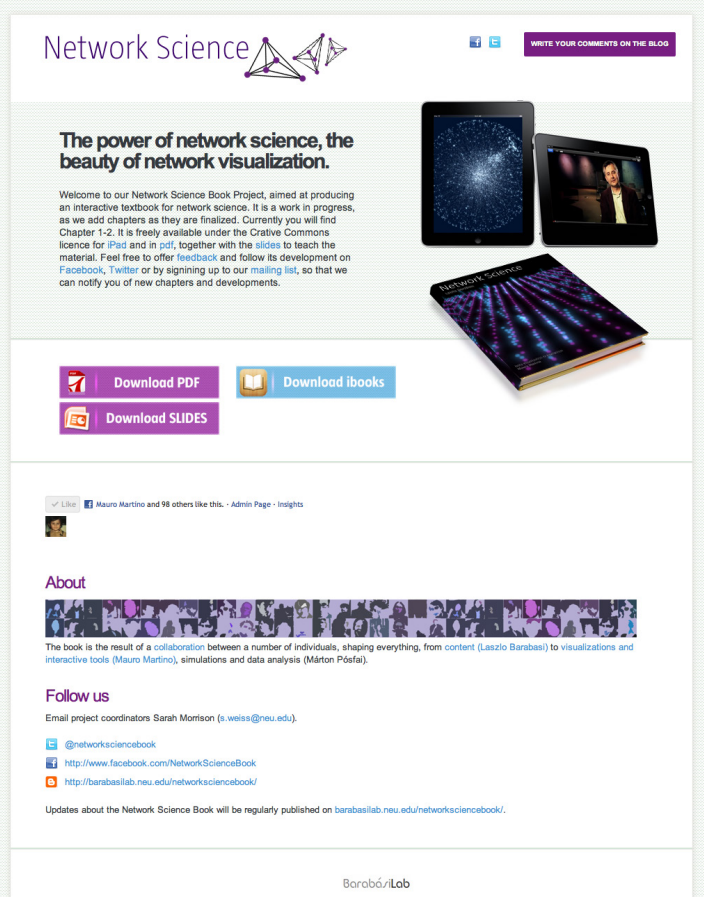


Image 1.1 http://barabasilab.com/networksciencebook

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.

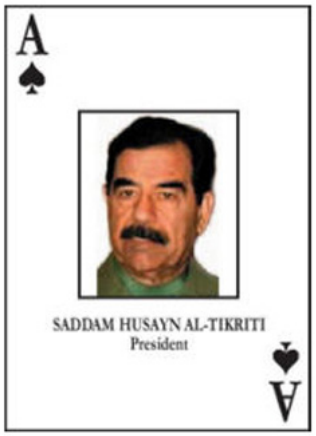


Image 1.2a
The network of Saddam Hussein.

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].

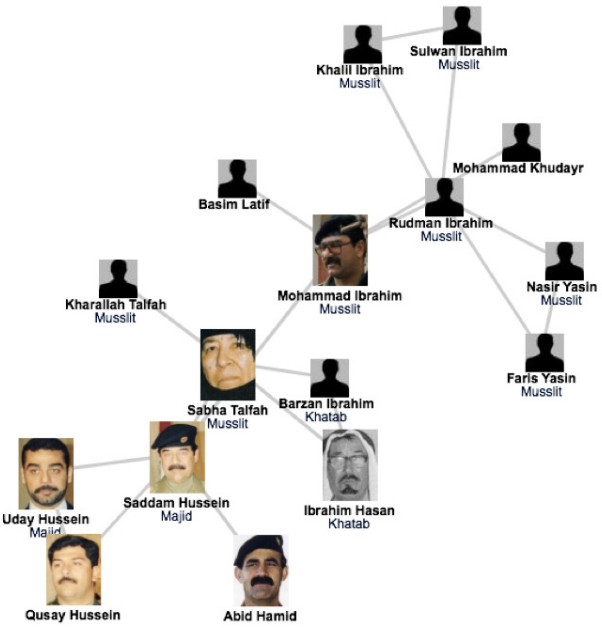


Image 1.2b
The network of Saddam Hussein.

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

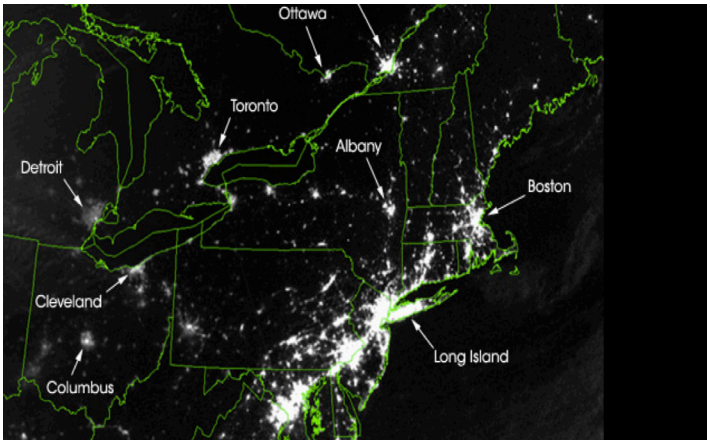
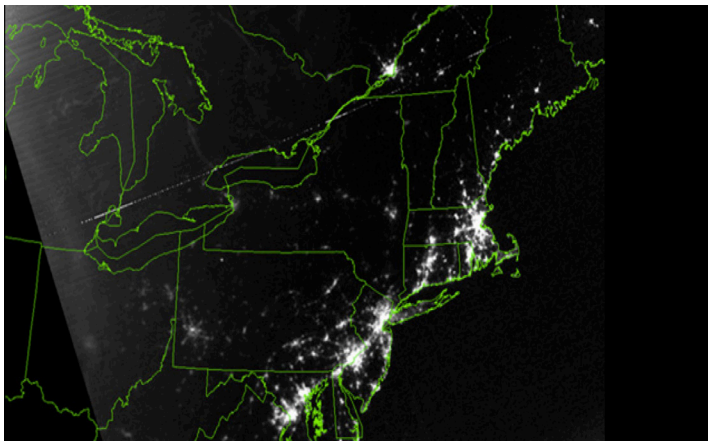


Image 1.3a, 1.3b
2003 North American blackout.

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.

NETWORKS AT THE HEART OF COMPLEX SYSTEMS

"I think the next century will be the century of complexity."

Stephen Hawking

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 mod-

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.

ern 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.

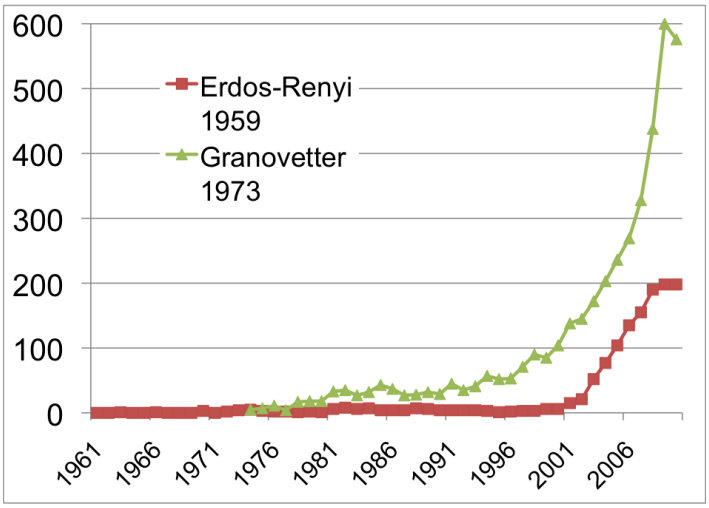


Image 1.5
The emergence of network science.

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 evo-

lution; 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

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

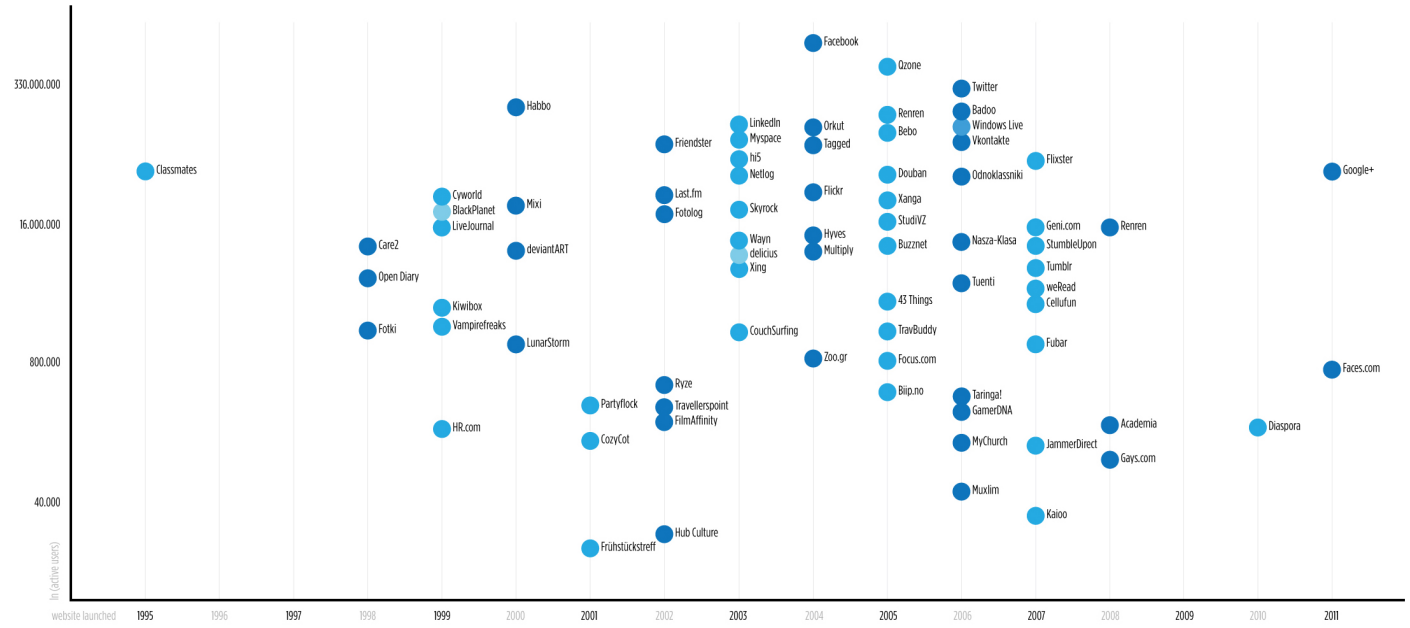


Image 1.6
The rise of social networking.

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

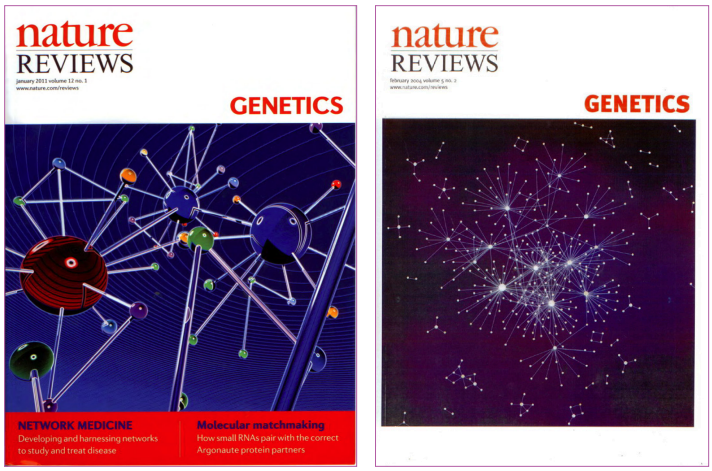


Image 1.7a, 1.7b

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.

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-

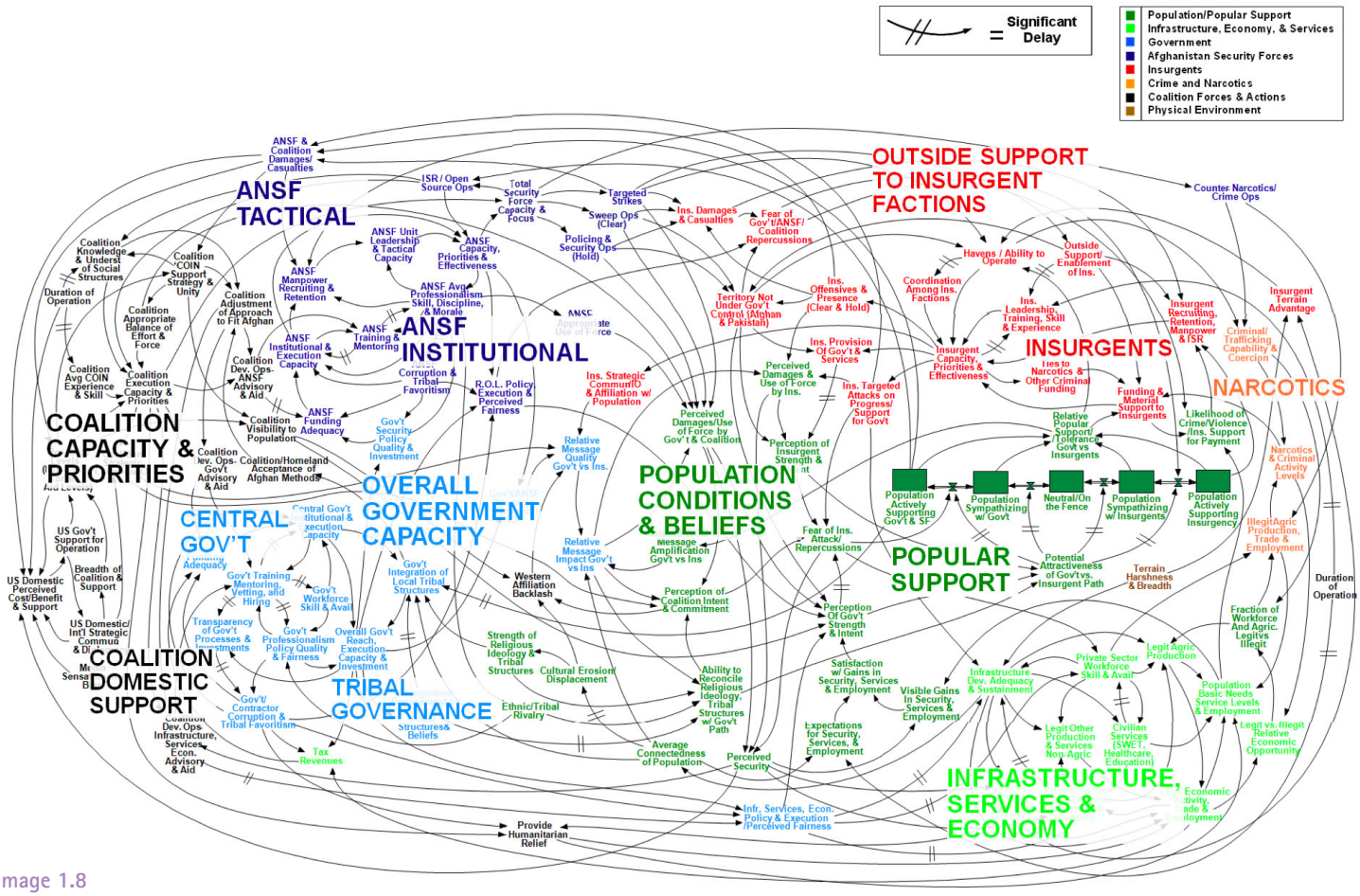


Image 1.8

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.

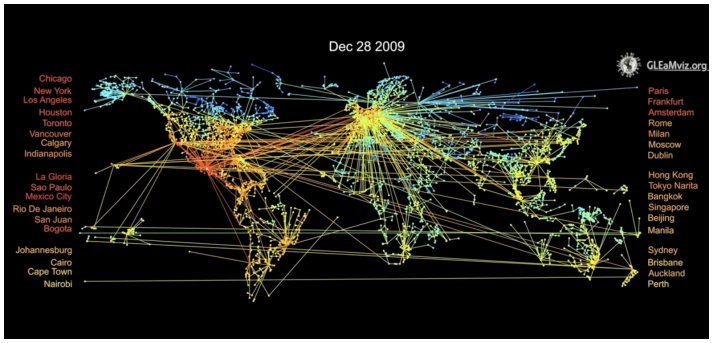


Image 1.9
Predicting the H1N1 epidemic.

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.

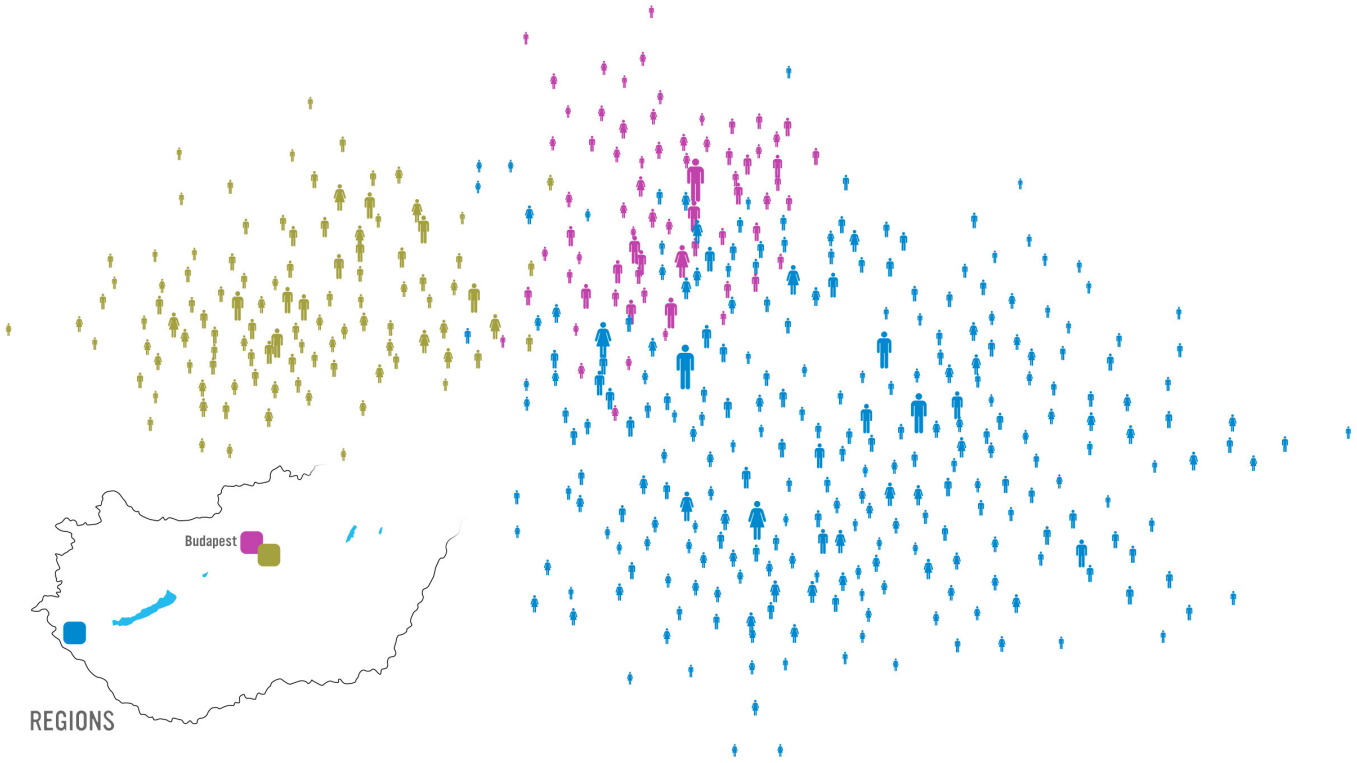


Image 1.10a

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.

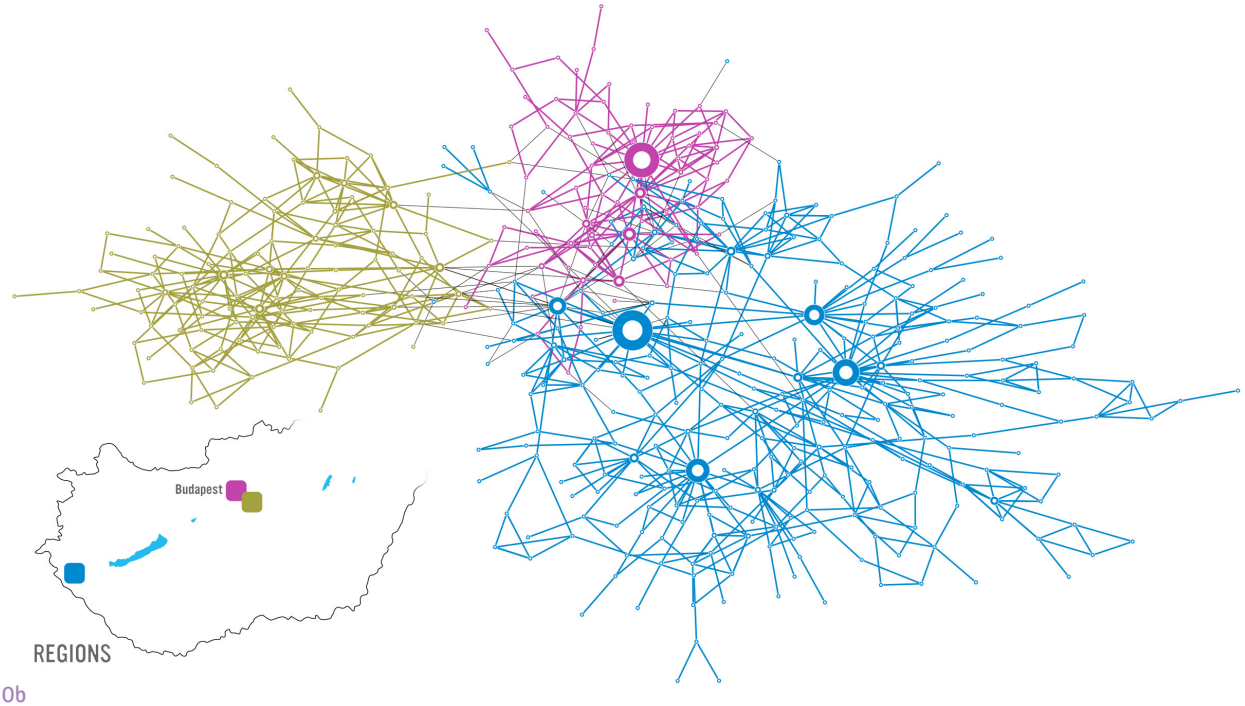


Image 1.10b

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

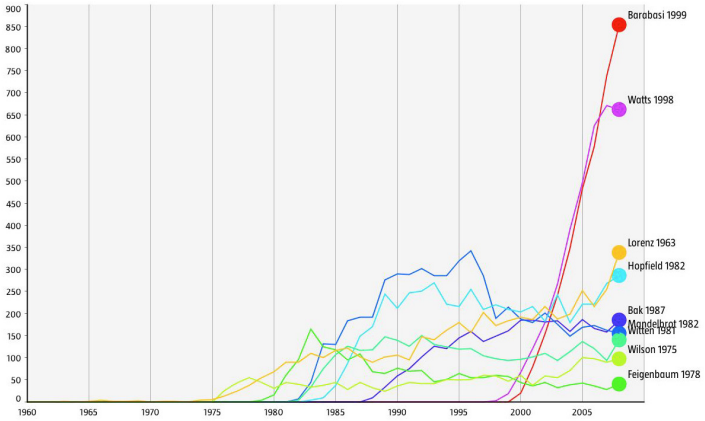


Image 1.11

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.

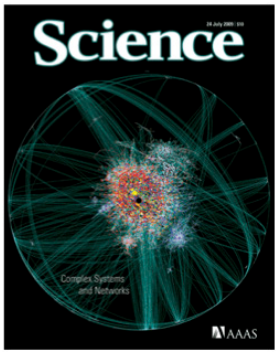


Image 1.12

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

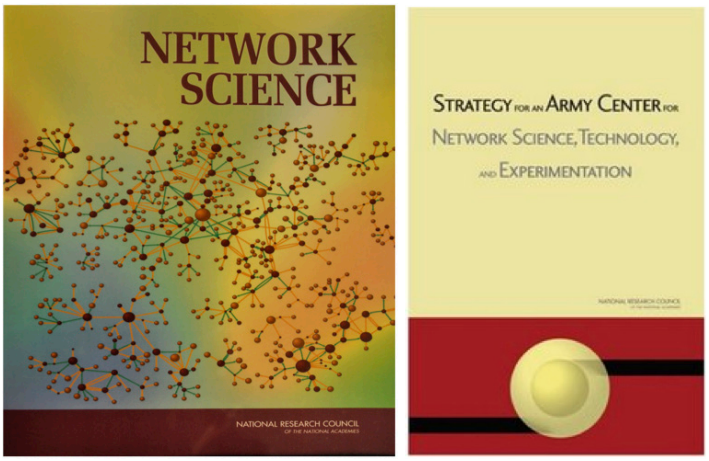


Image 1.13
National Research Council Reports.

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

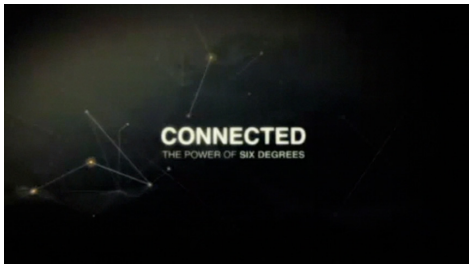


Image 1.14
Connected.

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

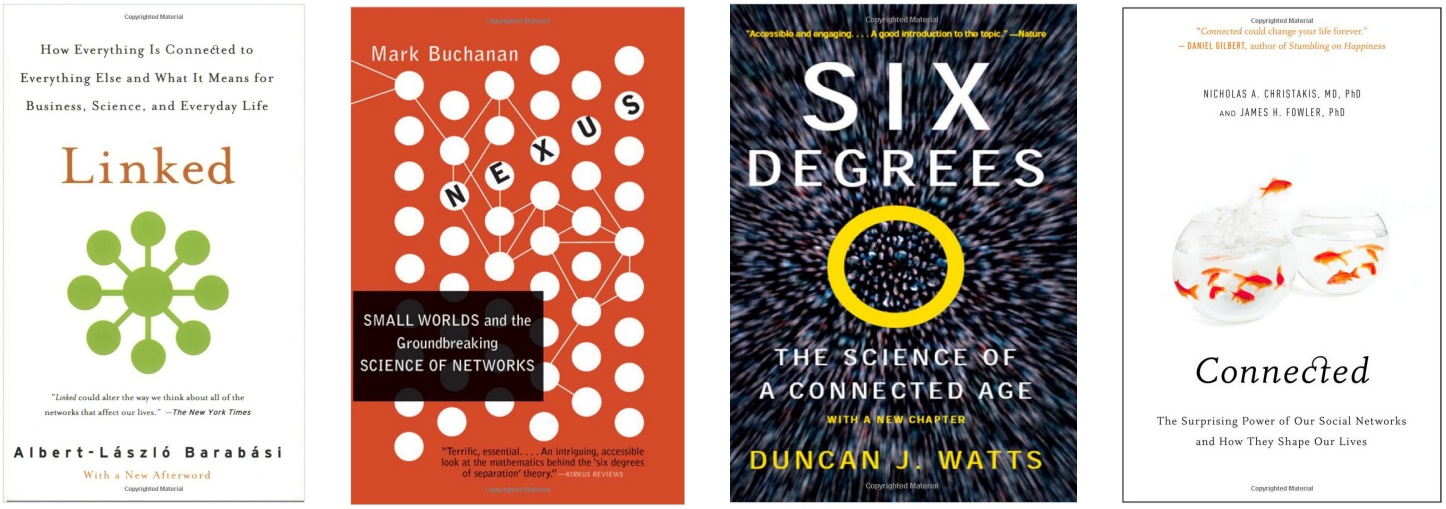


Image 1.15
Wide impact.

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.

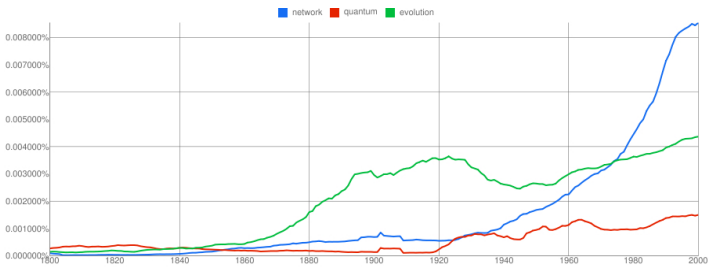


Image 1.16

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

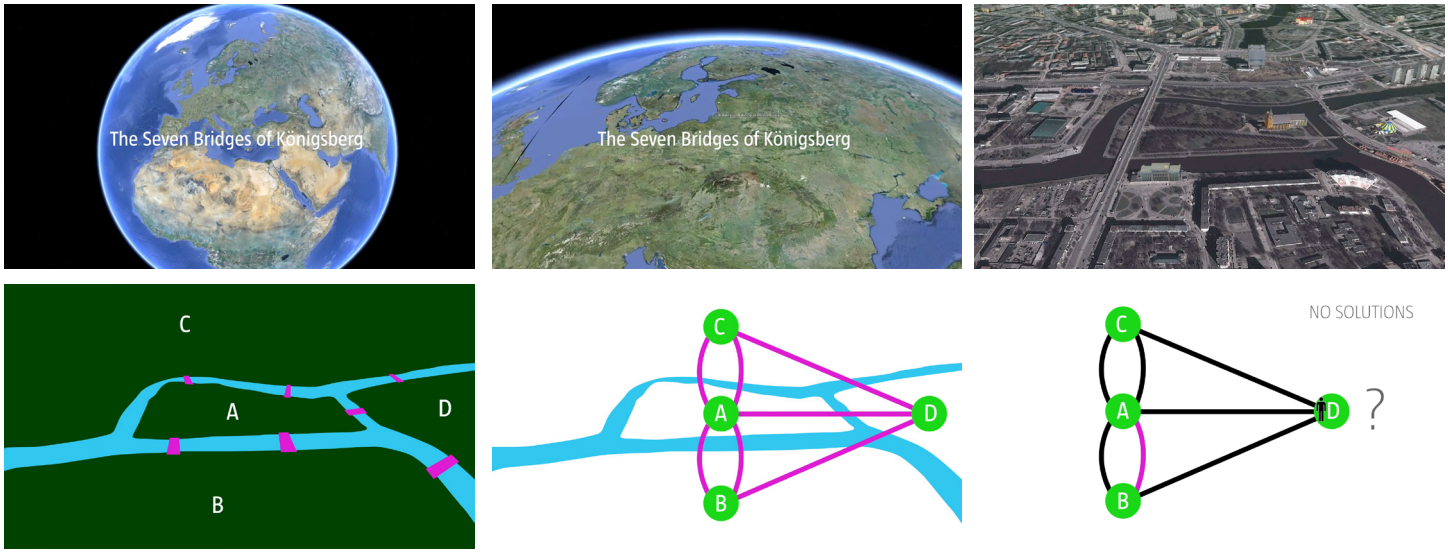


Image 2.1
The bridges of Königsberg.

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 mathe-

matical 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.

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.

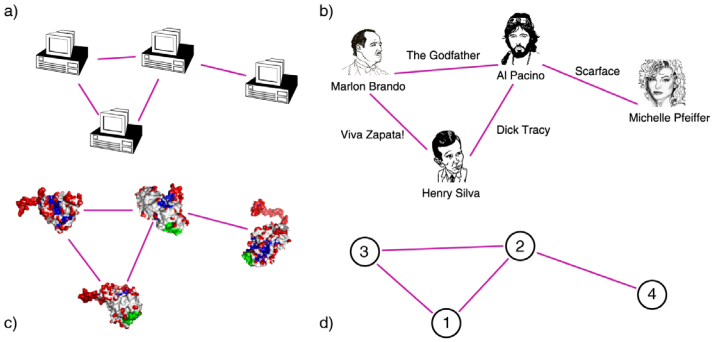


Image 2.3
Real systems of quite different nature can have the same network representation.

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).

Box 2.1

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

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 $i = 1, 2, \ldots, N$. 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.

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 $\langle k \rangle = \langle k_{in} \rangle = \langle k_{out} \rangle$.

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.

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

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

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 $N_k = N p_k$.

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.

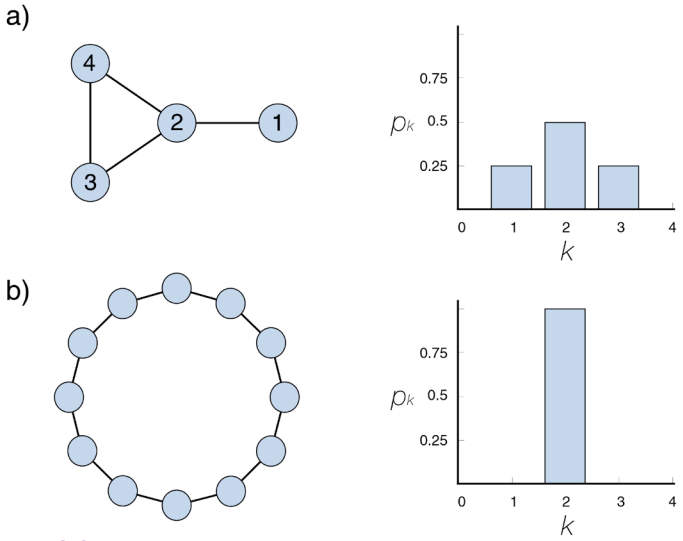


Image 2.4a

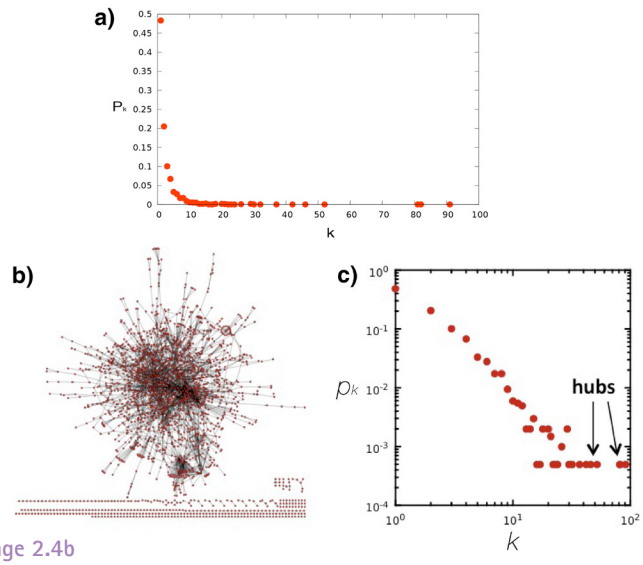


Image 2.4b

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 $p_k = \delta (k - 2)$.

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 ($N \simeq 7 \times 10^9$) 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),

$$L_{\max} = \left(\begin{array}{c} N \\ 2 \end{array} \right) = \frac{N(N-1)}{2} \qquad (11)$$

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.

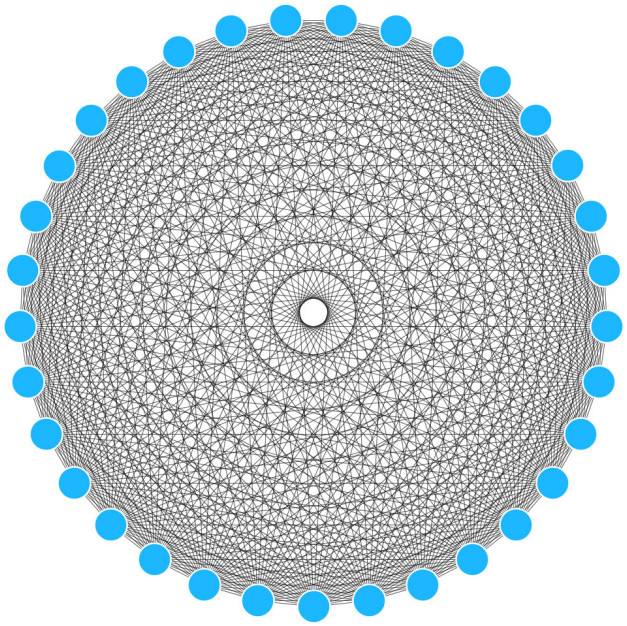


Image 2.5
Complete graph.

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 $\langle k \rangle = N - 1$.

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

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

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

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

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.

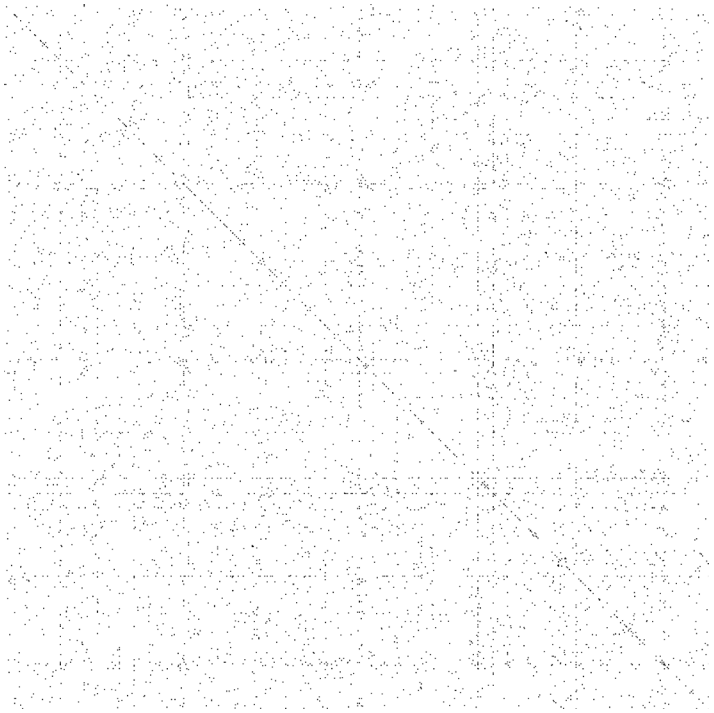


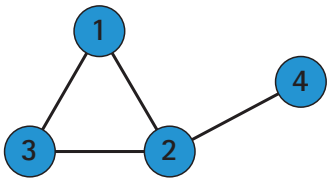
Image 2.6
The adjacency matrix is typically sparse.

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



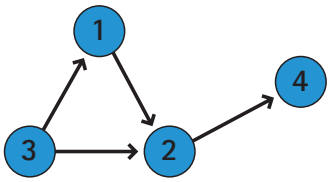
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

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

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, ($\langle k \rangle$), expressed in terms of the elements of the adjacency matrix.

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

$$A_{ij} = w_{ij} \quad . \qquad (15)$$

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.

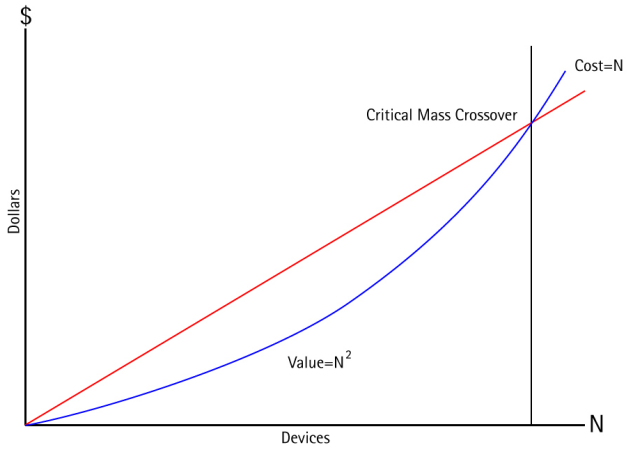


Image 2.8

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

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).

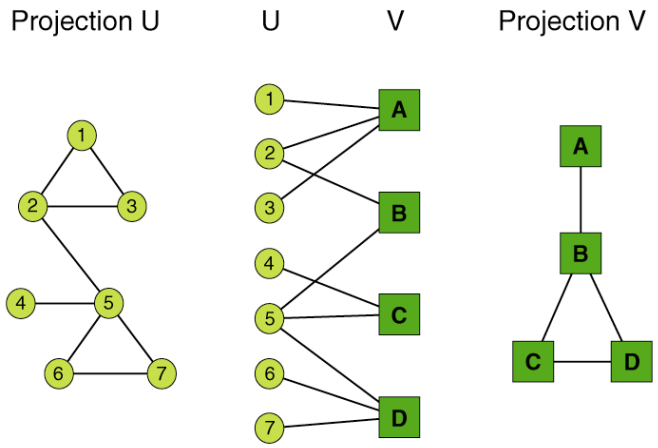


Image 2.9a
Bipartite network.

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.

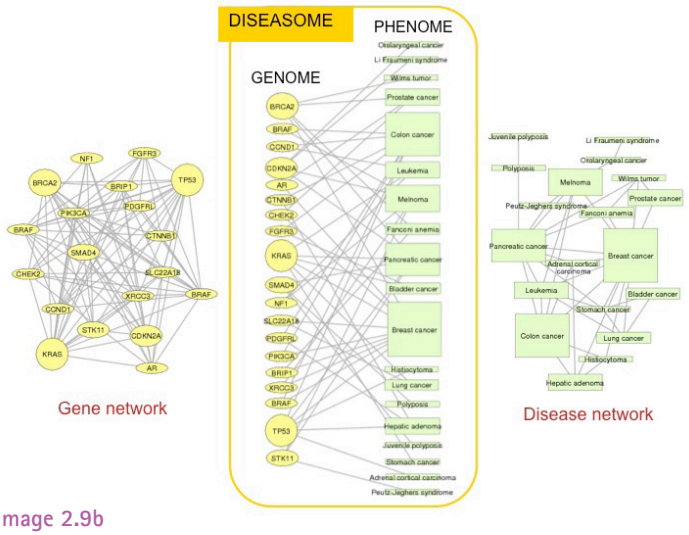


Image 2.9b
Bipartite network.

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])

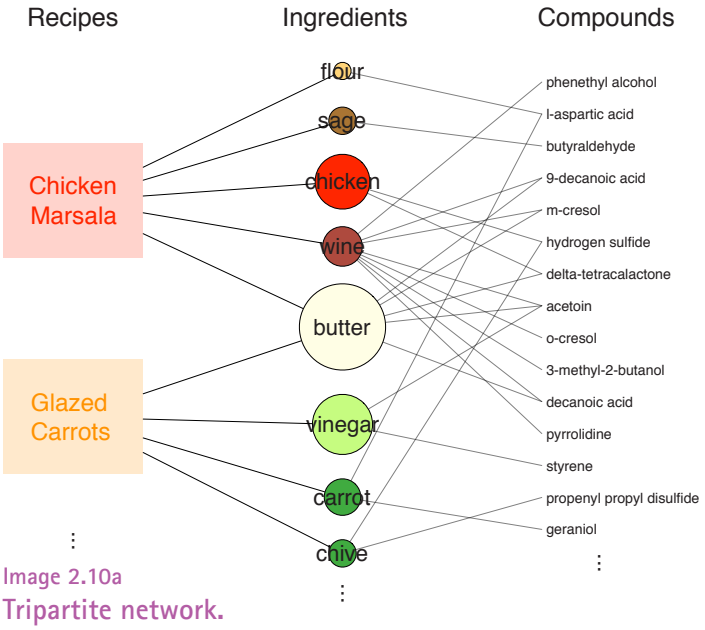


Image 2.10a

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.

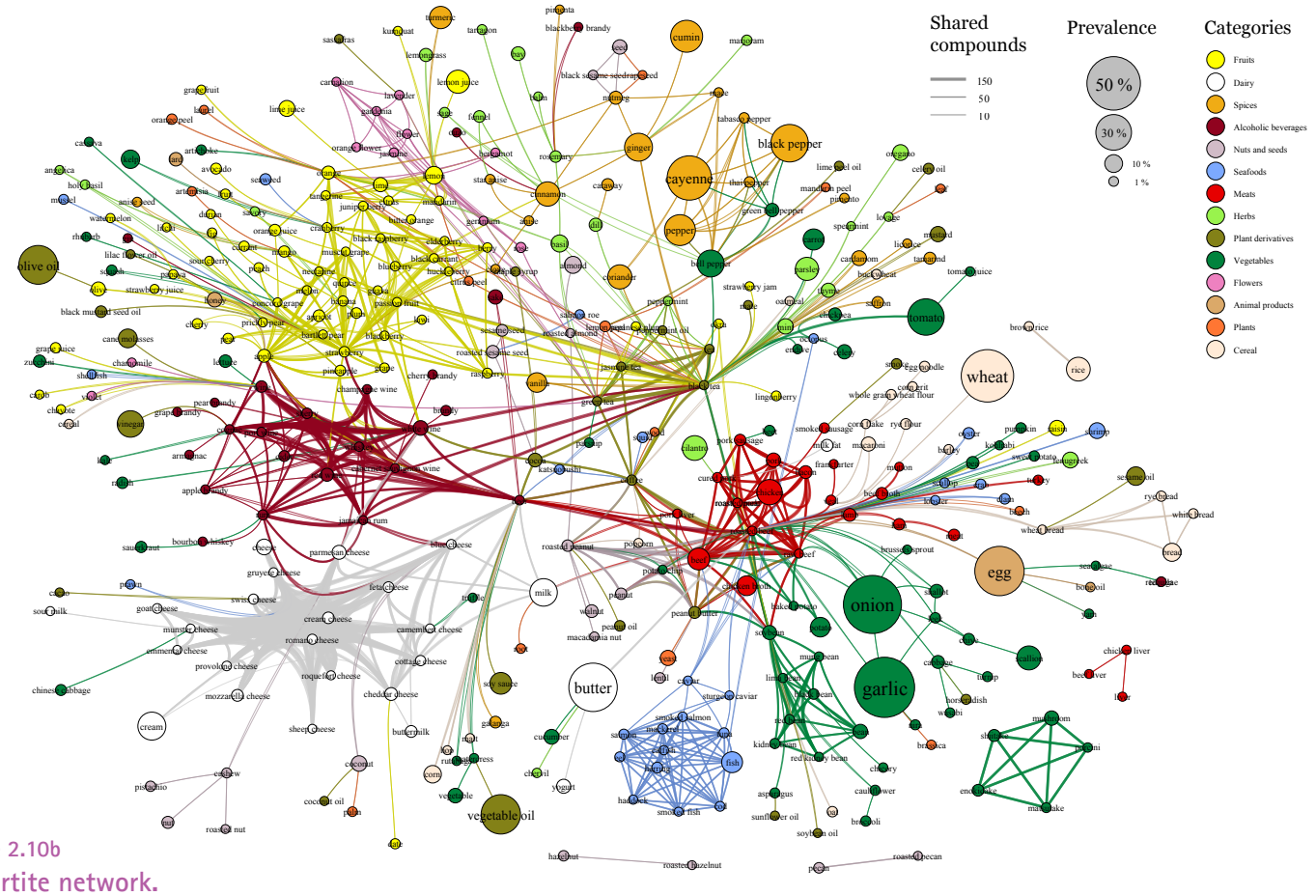


Image 2.10b

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]).

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.

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-

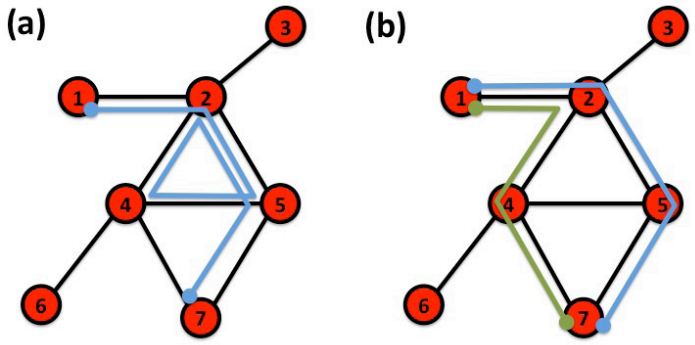


Image 2.11
The adjacency matrix is typically sparse.

(a) A path between nodes i_0 and i_n is an ordered list of n links $P_d = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_n\text{-}1, i_n),\}$.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.

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

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

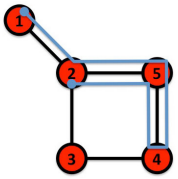
where [...]ij denotes the $(ij)^{th}$ element of a matrix.

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

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

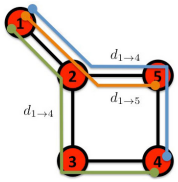
Equation (17) holds for both directed and undirected networks and can be generalized to multigraphs as well. The distance between nodes i and j is the path with the smallest d for which $N_{ij}^{(d)} > 0$. Despite the mathematical elegancy of Eq. (17), faced with a large network, it is more efficient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

Image 2.12
Pathology.



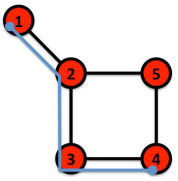
Path

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.



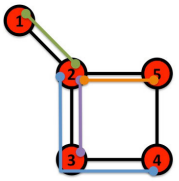
Shortest Path

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



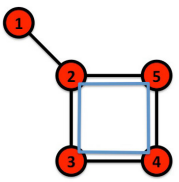
Diameter

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



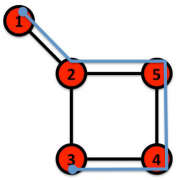
Average Path Length

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



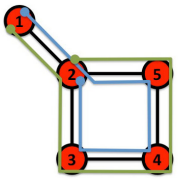
Cycle

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



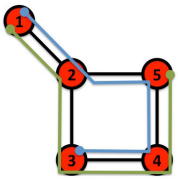
Self-avoiding Path

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

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



Hamiltonian Path

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

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.

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

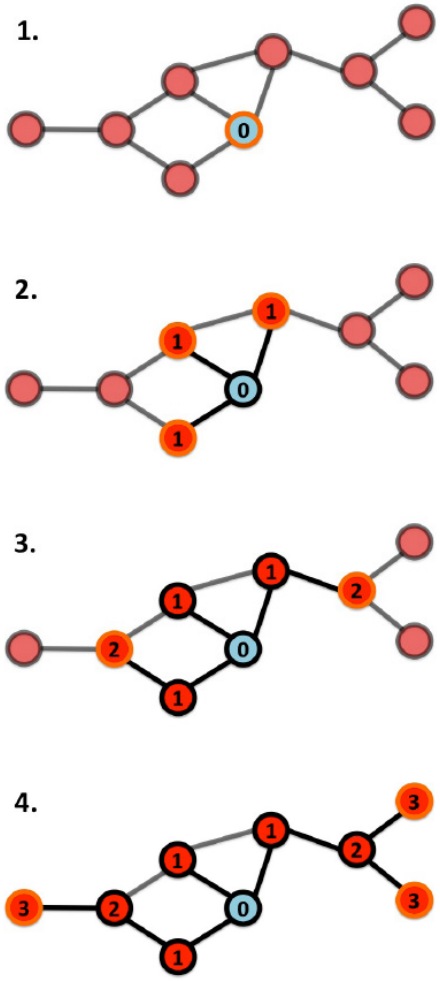


Image 2.13

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 $d_{03} = 3$.

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).

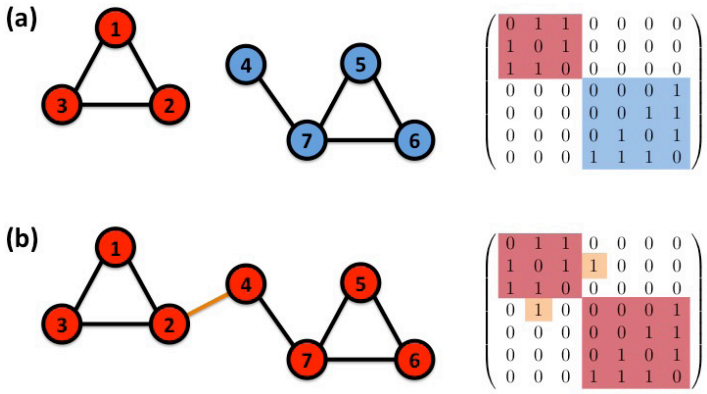


Image 2.14
Connected and disconnected networks.

(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 $n \rightarrow n + 1$. 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.

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],

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

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.

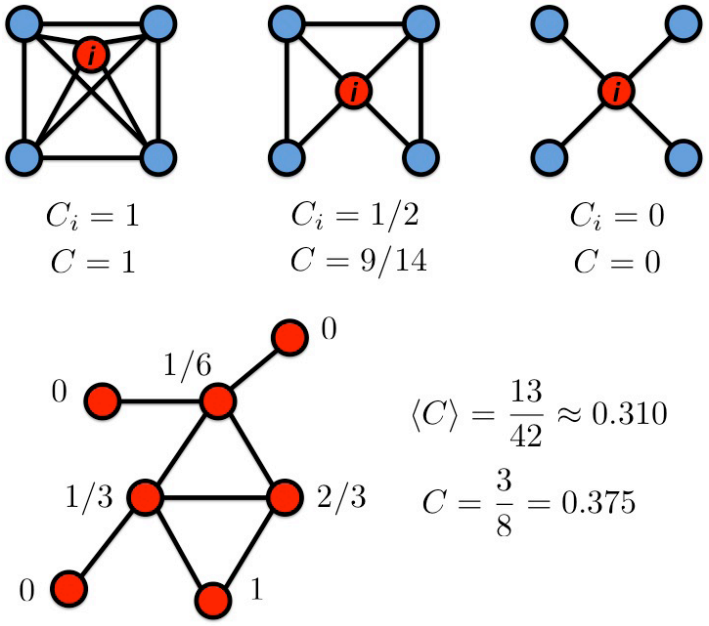


Image 2.15
Clustering Coefficient.

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.

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.

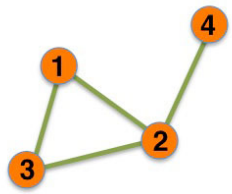
The breath first search algorithm will also convince us that the protein interaction network is not connected, but consists of 185 components, shown as isolated clusters in Image 2.17a. The largest, called the giant component, contains 1,647 of the 2,018 nodes; all other components are tiny compared 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 $\langle C \rangle$ =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



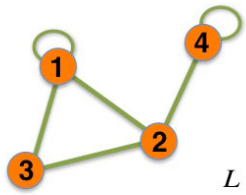
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

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

Self-interactions



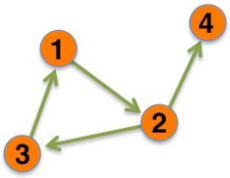
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

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

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



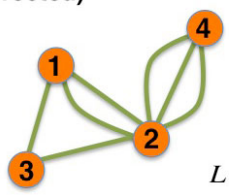
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

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

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

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

Multigraph (undirected)



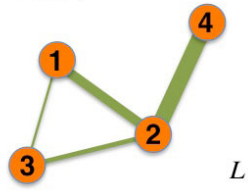
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

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

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

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

Weighted (undirected)



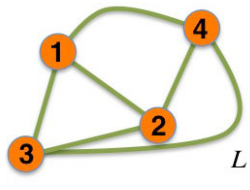
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

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

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

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.

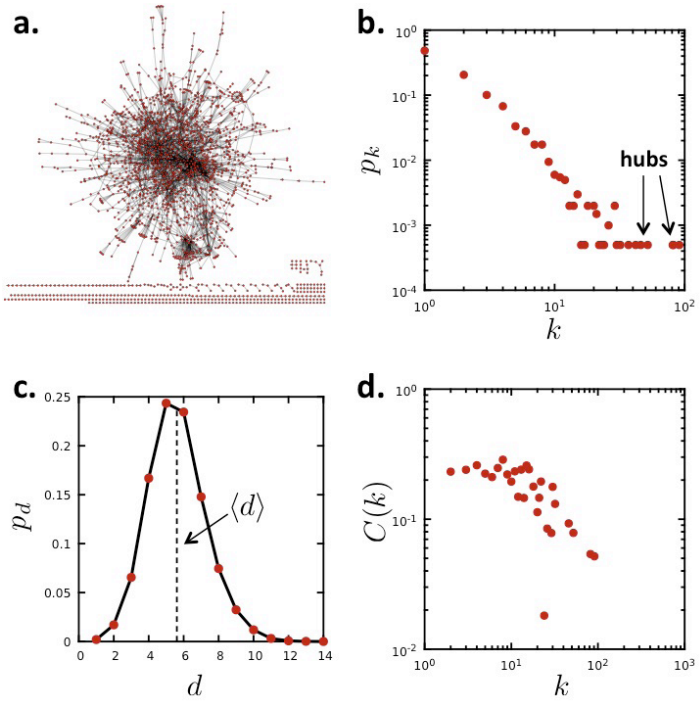


Image 2.16

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 $\langle d \rangle = 5.61$.

(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.

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 $C = 0.375$.

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.

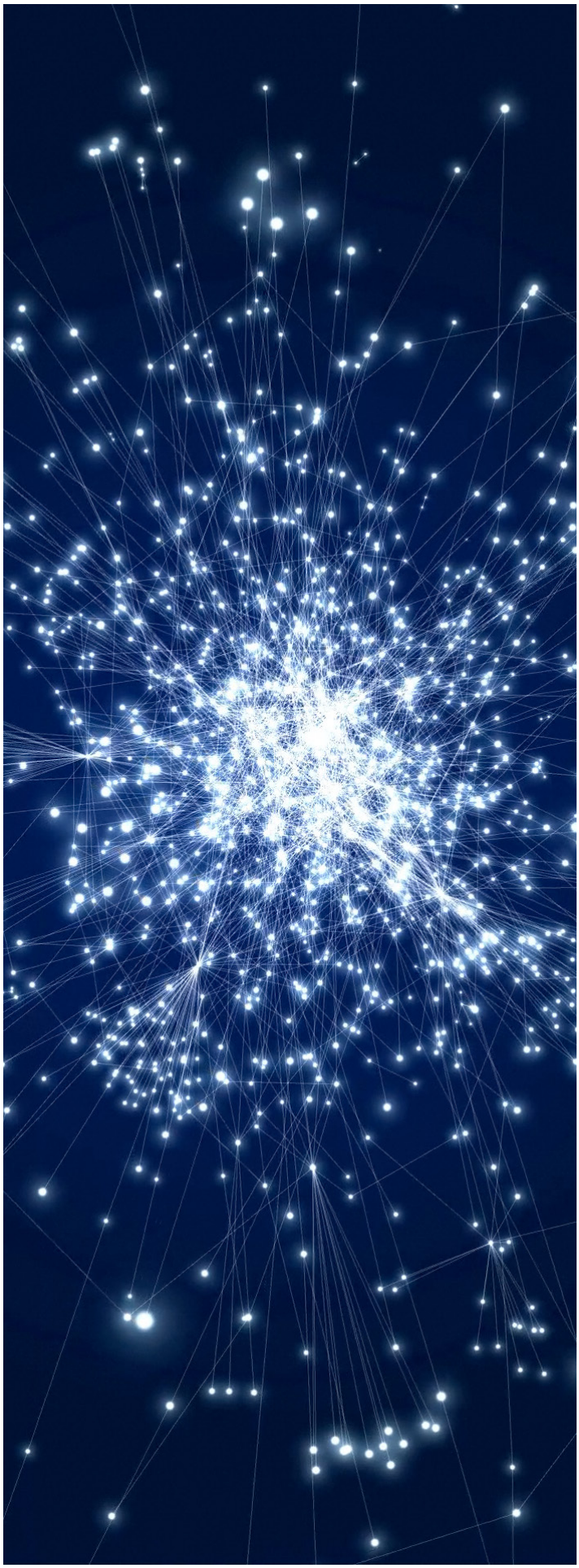
[10] R. D. Luce and A. D. Perry. *A method of matrix analysis of group structure*. Psychometrika, 14:95–116, 1949.

[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

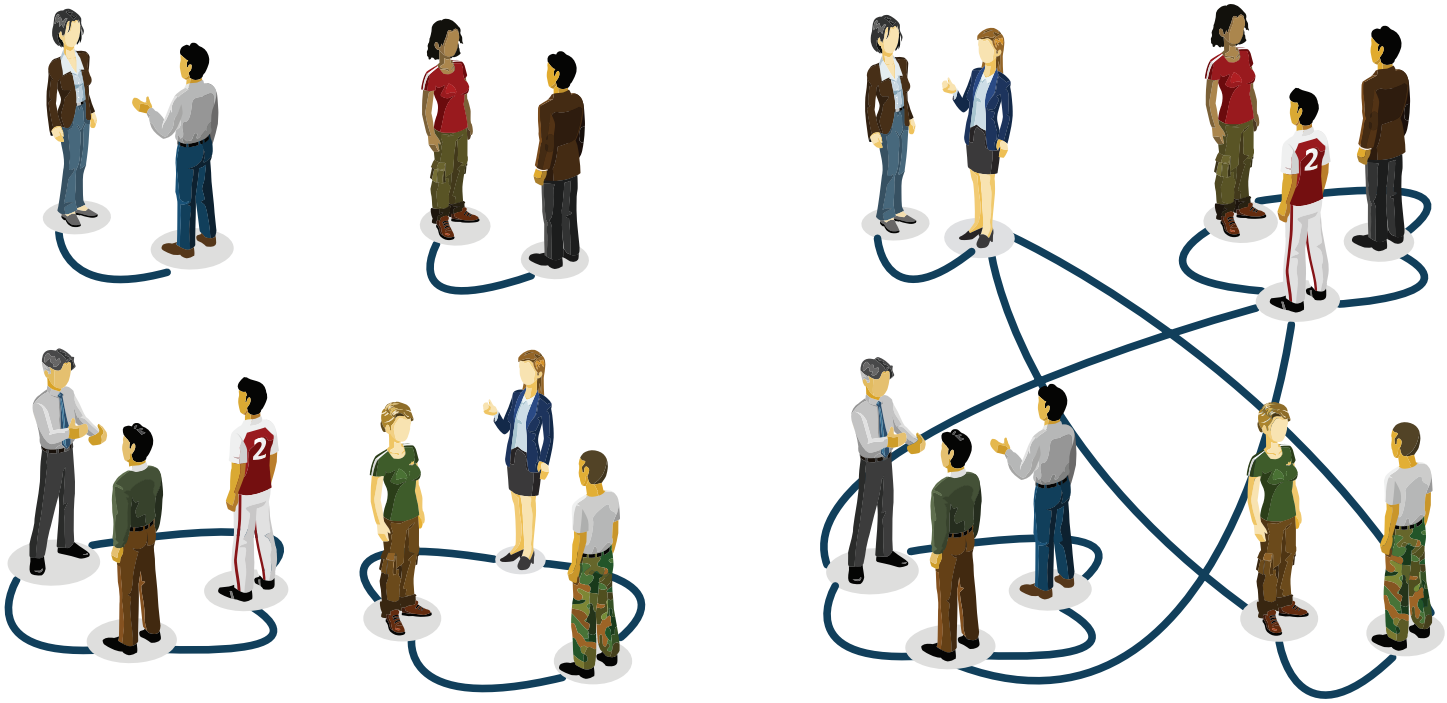


Image 3.1
From a cocktail party to random networks.

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.

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.

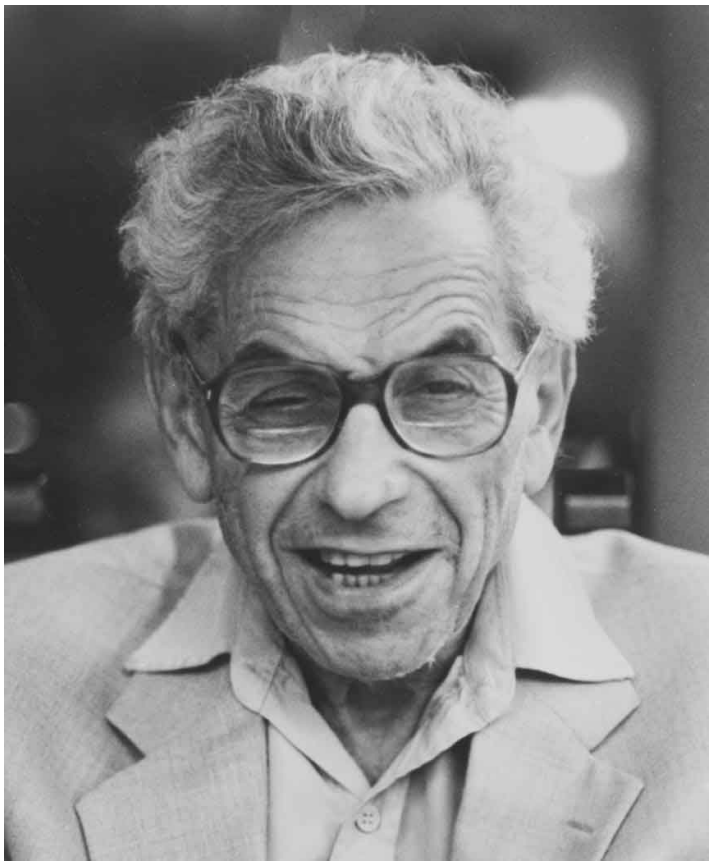


Image 3.2a
Pál Erdős (1913–1996)

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.

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, $\binom{\binom{N}{2}}{L}$ 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

$$p_L = \binom{\binom{N}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}. \quad (1)$$

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

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

Hence $\langle L \rangle$ 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

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

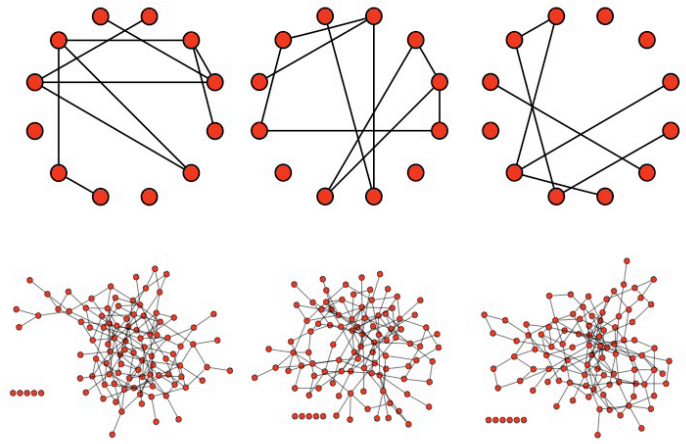


Image 3.3
Random networks are truly random.

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 $\langle k \rangle$ 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 $\langle L \rangle = 0$ to L_{max} and the average degree of a node increases from $\langle k \rangle = 0$ to $\langle k \rangle = N-1$.

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

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

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

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.

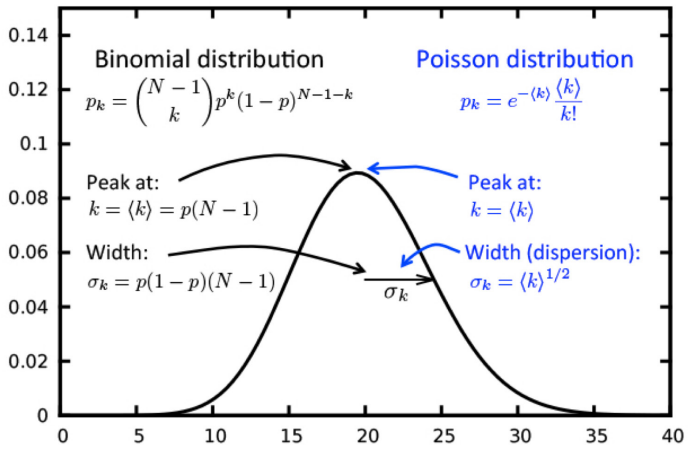


Image 3.4a

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 $N \gg \langle k \rangle$, 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, $\langle k \rangle$.

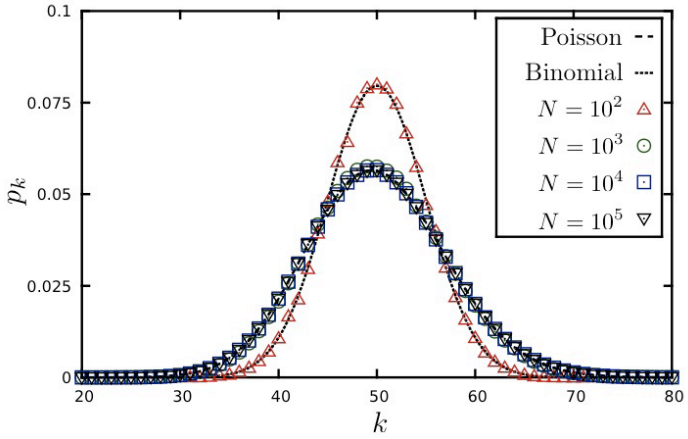


Image 3.4b

Degree distribution is independent of the network size.

The degree distribution of a random network with average degree $\langle k \rangle = 50$ 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, $N \gg \langle k \rangle$, 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 $\langle k \rangle \approx 1{,}000$. 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).

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.

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.*

- 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.

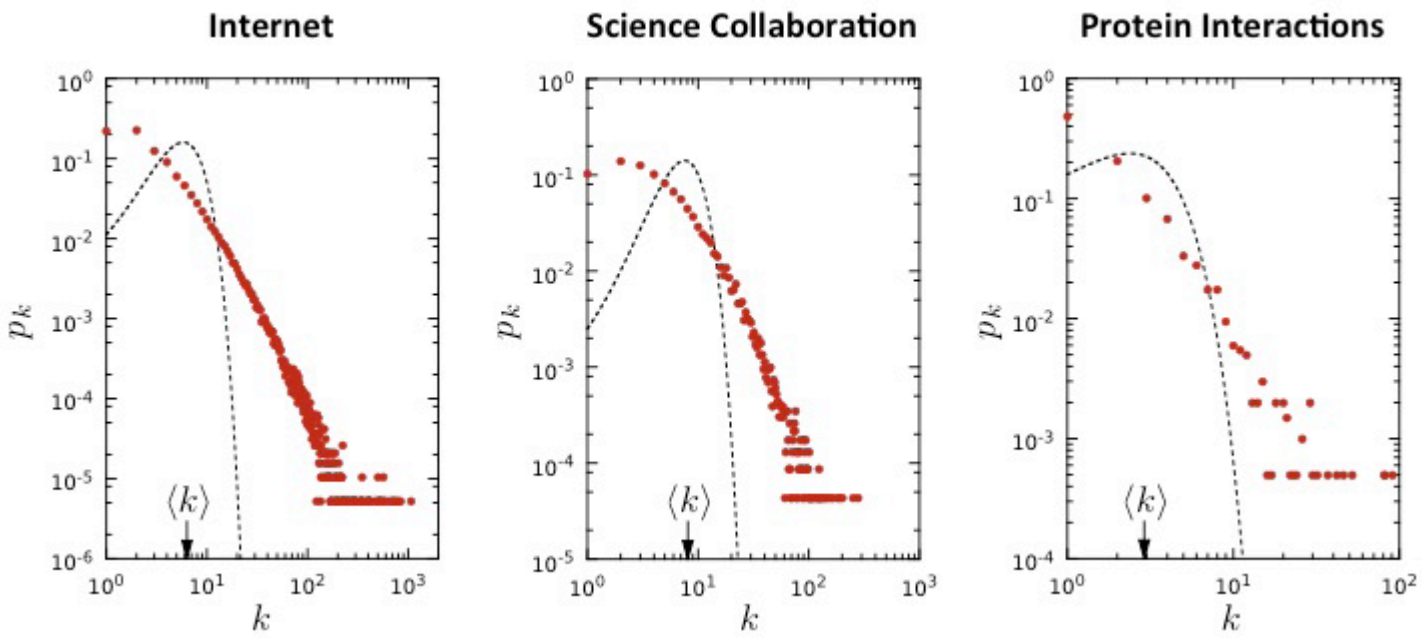
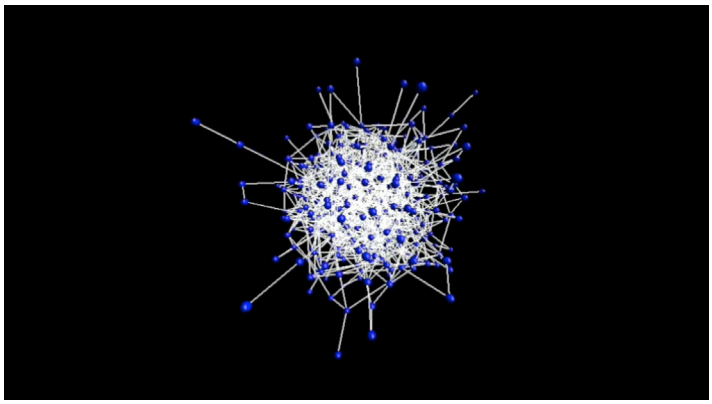


Image 3.5

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.

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 $\langle k \rangle$. 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 $\langle k \rangle$ from 0 to $N-1$. Yet, as Image 3.6a indicates, this is not the case: N_G / N remains zero for small $\langle k \rangle$, indicating the lack of a giant component for a range of $\langle k \rangle$ values. Once $\langle k \rangle$ 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 $\langle k \rangle$ allow us to distinguish four topologically distinct regimes (Image 3.6), each with its unique characteristics:

(a) Subcritical regime: $0 < \langle k \rangle < 1$, ($p < \frac{1}{N}$).

For $\langle k \rangle = 0$ the network consists of N isolated nodes. Increasing $\langle k \rangle$ is equivalent with adding $N\langle k \rangle = pN(N-1)/2$ 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 $\langle k \rangle$ 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 $\langle k \rangle < 1$ the largest cluster is a tree

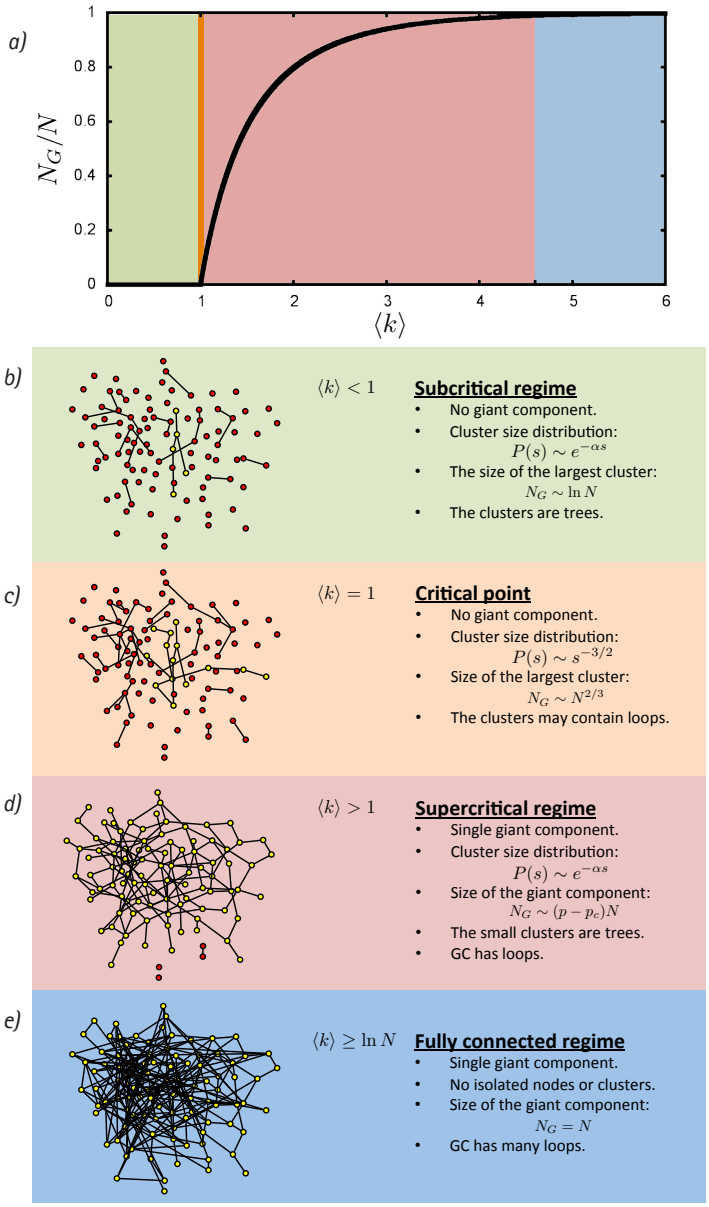


Image 3.6
Evolution of a random network.

(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 $N_G \approx N$. 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 $-\infty$ 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 $N \to \infty$. 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.

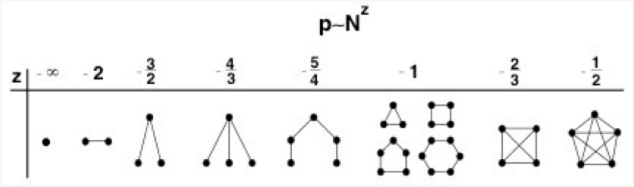


Image 3.7

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

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 $\langle k \rangle > \ln N$ 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 $\langle k \rangle > \ln N$, 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	$\langle k \rangle$	$\ln N$
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 $\ln N$. A giant component is expected for $\langle k \rangle > 1$ and all nodes should join the giant component for $\langle k \rangle \geq \ln N$. While for all networks $\langle k \rangle > 1$, for most $\langle k \rangle$ is under the $\ln N$ 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 $\langle k \rangle > \ln N$), or we expect the network to be fragmented into multiple components (if $\langle k \rangle < \ln N$). For social networks this would mean that $\langle k \rangle \geq \ln(7 \times 10^9) \approx 22.7$. 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 $\langle k \rangle \approx 1{,}000$ 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.

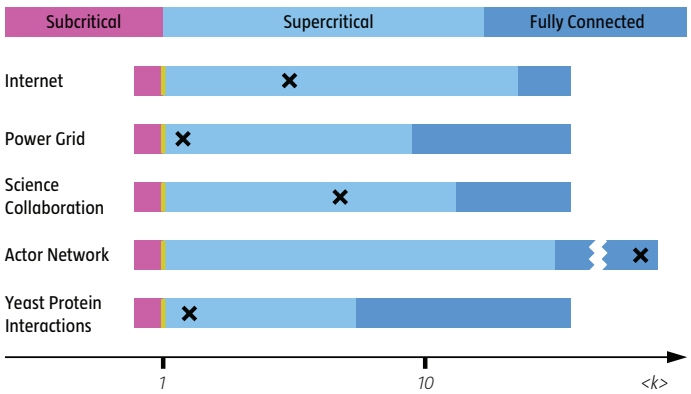


Image 3.8

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 $\ln N$, 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.

SMALL WORLD PROPERTY

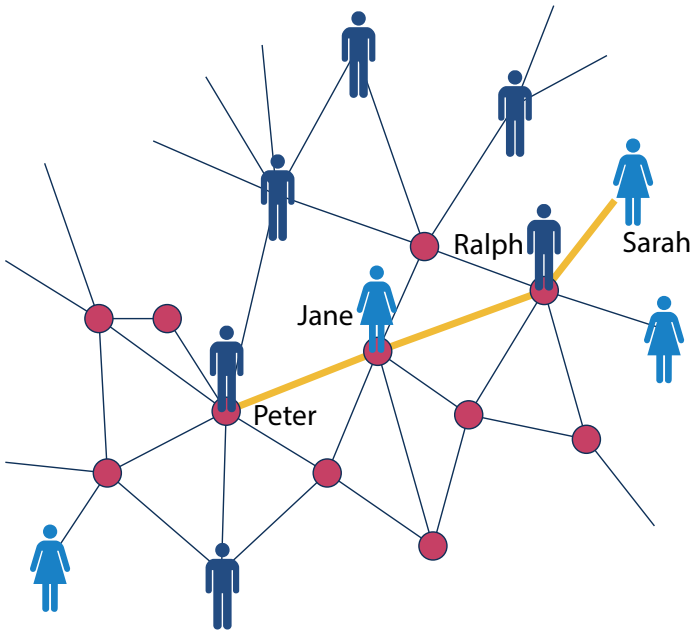


Image 3.9

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

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

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.

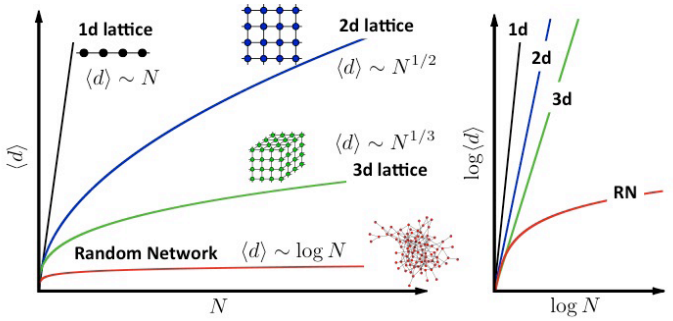


Image 3.10

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: $d_{max} \sim \langle d \rangle \sim N$.
- For a square lattice $d_{max} \sim \langle d \rangle \sim N^{1/2}$.
- For a cubic lattice $d_{max} \sim \langle d \rangle \sim N^{1/3}$.
- In general, for a d-dimensional lattice we have $d_{max} \sim \langle d \rangle \sim N^{1/d}$.

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 $\langle d \rangle$ 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 $(7 \times 10^9)^{1/2} = 83{,}666$. 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).

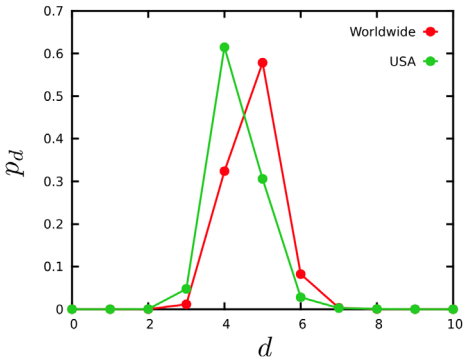


Image 3.11

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 $N \simeq 7 \times 10^9$ and $\langle k \rangle \simeq 10^3$, we obtain

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

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 $\langle d \rangle$ for several real networks, finding that despite the diversity of these systems and the significant differences between them in terms of N and $\langle k \rangle$, Eq. (19) offers a reasonable approximation to the empirically observed $\langle d \rangle$.

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

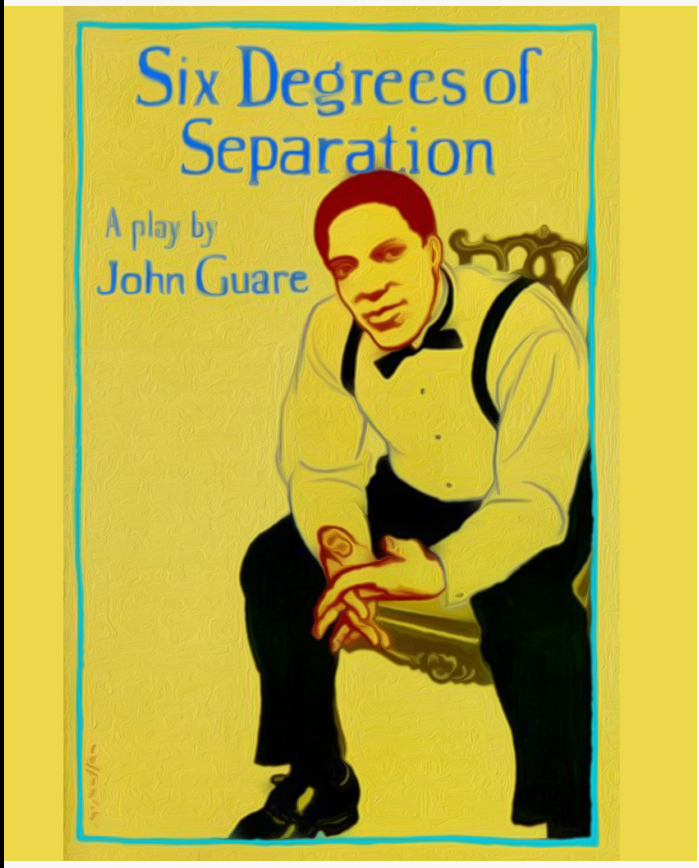


Image 3.14
Six Degrees of Separation.

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

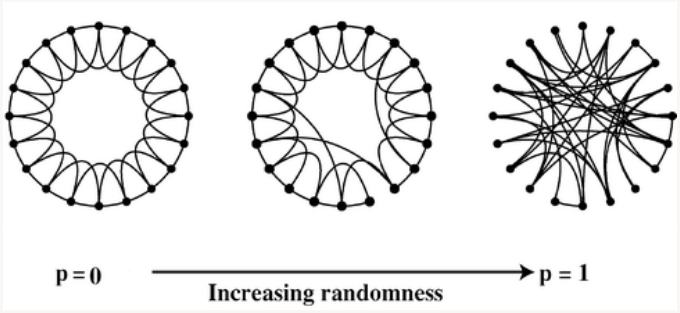


Image 3.15
Watts-Strogatz model.

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]).

Box 3.6c

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 3.6d

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.

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.

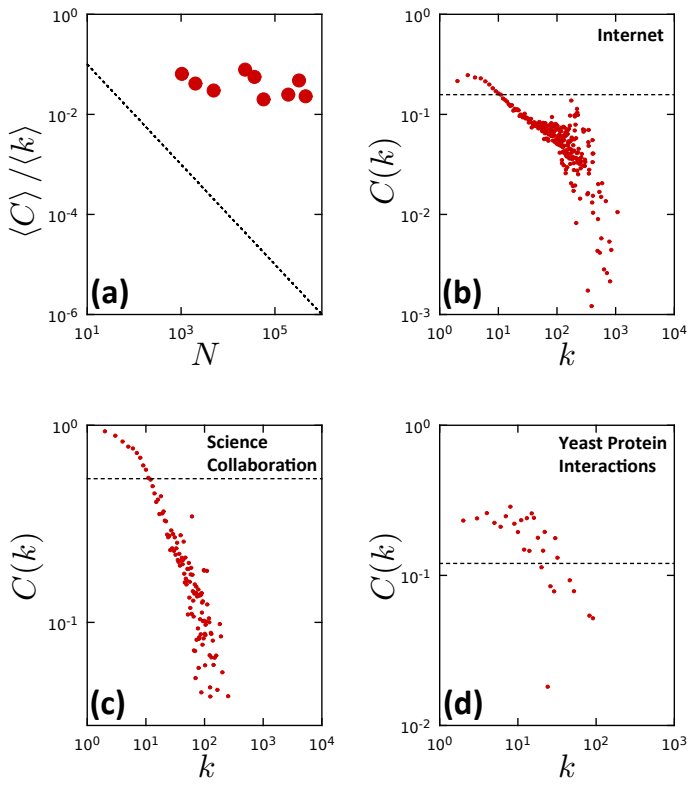


Image 3.16
Clustering in real networks.

(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.

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 $\langle d \rangle \sim logN / log\langle k \rangle$, 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).

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.

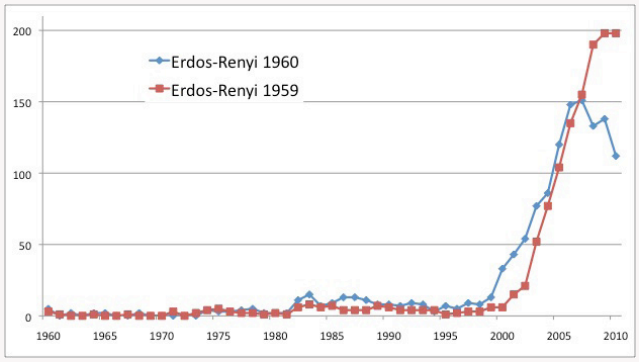


Image 3.17

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.

Box 3.7

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 ($k \ll N$), P_k has the Poisson form

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

- *Giant component* (N_G) :

$\langle k \rangle < 1$: no giant component ($N_G \sim \ln N$)

$1 < \langle k \rangle < \ln N$: one giant component and disconnected clusters

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

$\langle k \rangle > \ln N$: 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

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

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 $N \gg k$, 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 $\simeq$ 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 $k_{min} = 816$.

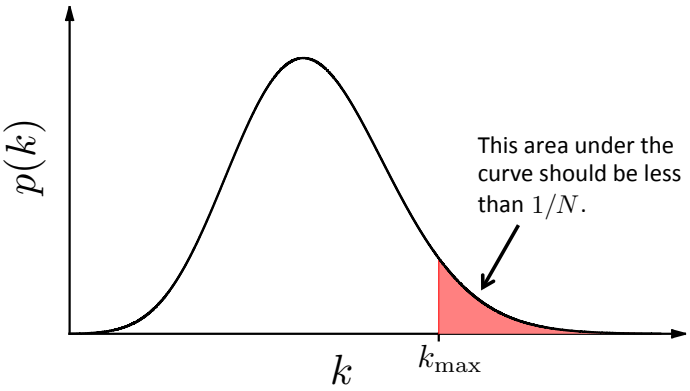


Image 3.18

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 $\langle k\rangle = 1$ (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 $\langle k\rangle$ (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 $\langle k\rangle$. To have a nonzero solution, the obtained curve must intersect with the dotted diagonal, representing the left hand side of Eq. (32). For small $\langle k\rangle$ the two curves intersect each other only for $S = 0$, indicating that for small $\langle k\rangle$, the size of the giant component is zero. Only when $\langle k\rangle$ exceeds a threshold value, does a non-zero solution emerge.

To determine the value of $\langle k\rangle$ 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 $\langle k\rangle = 1$.

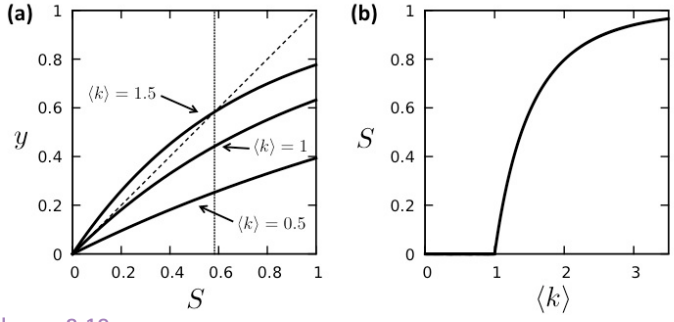


Image 3.19

Graphical solution for the size of the giant component.

(a) The three curves in the left panel show $y = 1-exp[-\langle k\rangle S]$ for various $\langle k\rangle$. The diagonal dashed line corresponds $y = S$, and the intersection of the dotted and continuous lines provides the solution to Eq. (32), $S = 1-exp[-\langle k\rangle S]$. 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 $\langle k\rangle$ 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 ‹k›, 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 ‹k› values.

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

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

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

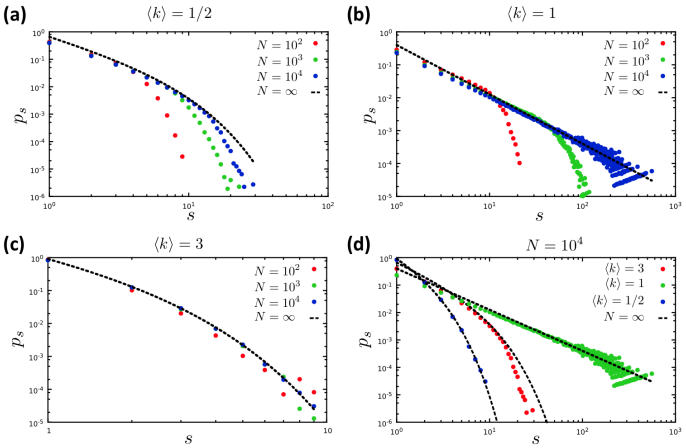


Image 3.20
Component size distribution.

Component size distribution in a random network, p_s, excluding the giant component. *(a)-(c)* shows p_s for different ‹k› 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 ‹k›. 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.

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).

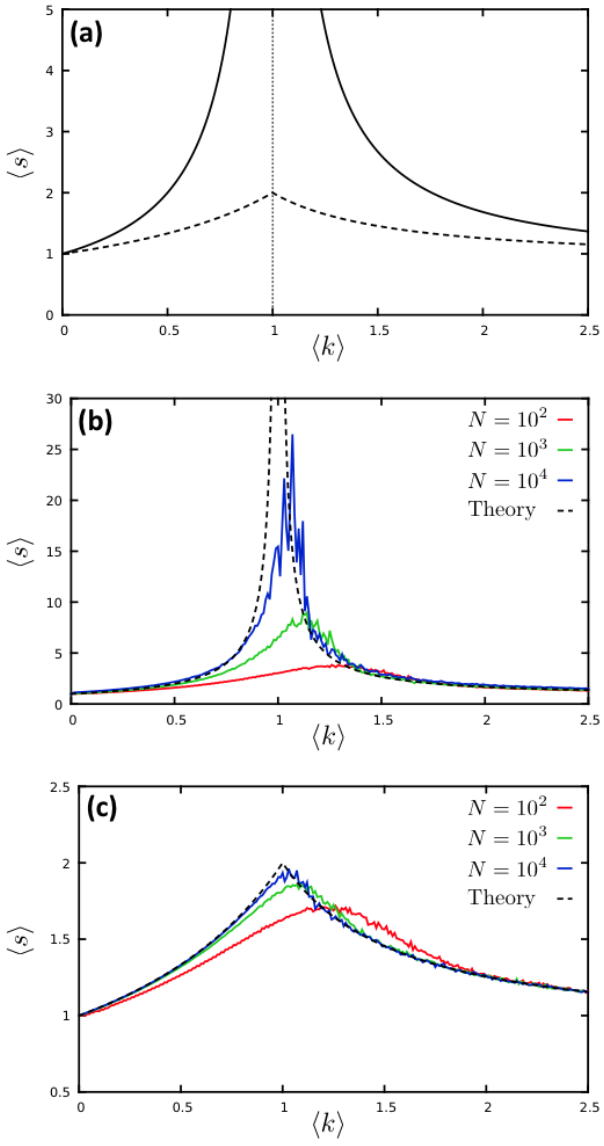


Image 3.21

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).

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

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

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

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

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

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 $\langle s \rangle$ diverges as we approach $\langle k \rangle = 1$ (Advanced Topics 3.D).

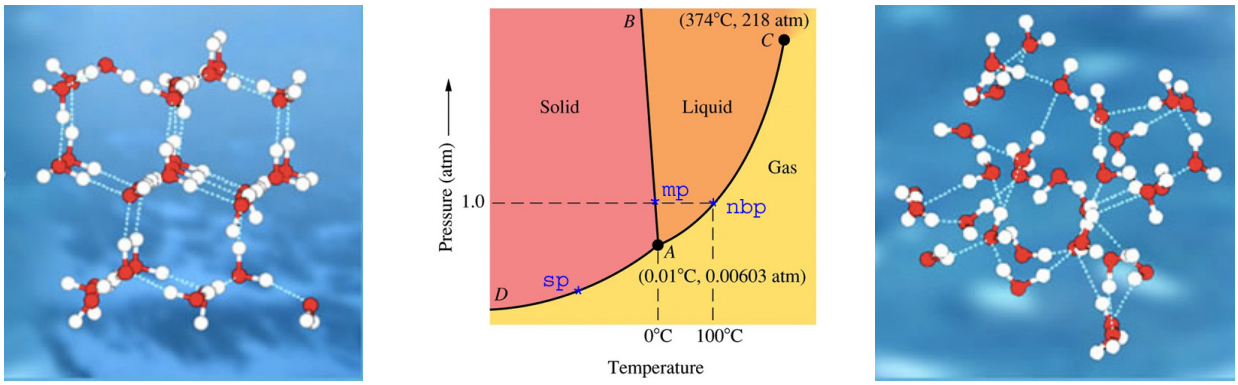


Image 3.22a

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/

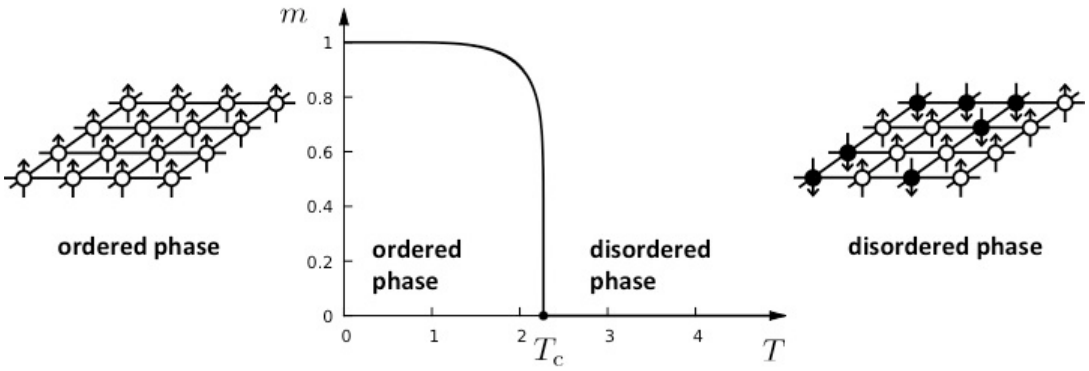


Image 3.22b

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 $\langle k\rangle \longrightarrow o$ 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 $\langle k\rangle \longrightarrow \infty$ 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.