ELTE | FACULTY OF INFORMATICS

Network Science Hálózattudomány

IPM-18fatNSEG

IPM-22fpiNSEG

IPM-24ATNSEG

Dr. Tamás Orosz

associate professor, Ph.D., habil.

Department of Data Science and Engineering

Lecture 1, September 12, 2024

Agenda

- Welcome & Introduction
- What is Network Science?
- History of Network Science
- Key Concepts
- Network Types
- Real-world Applications
- Conclusion & Q&A

Introduction of Network Science by Albert-László Barabási

https://www.youtube.com/watch?v=RfgjHoVCZwU
https://networksciencebook.com

Introduction to Network Science

- Network Science is an interdisciplinary field that has gained immense importance in the realm of informatics,
  - offering a powerful framework to understand, analyze, and solve complex problems arising in various domains.
- Rooted in the groundbreaking work of researchers like Albert-László Barabási,
  - Network Science focuses on the study of networks – systems composed of interconnected nodes and edges.

Understanding Network Science

- At its core, Network Science seeks to unravel the intricate web of connections that underlie diverse phenomena, from social interactions to biological processes and technological infrastructures.
- This approach transcends traditional disciplinary boundaries, making it an invaluable tool for MSc Informatics students as it bridges the gap between computer science, mathematics, sociology, and many other fields.

Key Significance of Network Science in Informatics

1. Complex Systems Analysis

Complex Systems Analysis:

In the age of big data, informatics professionals grapple with increasingly complex systems. Network Science equips students with the analytical tools to make sense of intricate structures, revealing emergent properties and patterns within these systems.

Key Significance of Network Science in Informatics

2. Data Mining and Analysis

Data Mining and Analysis:

- Networks are everywhere, and they encapsulate essential information about relationships, dependencies, and interactions.
- Informatics professionals can leverage Network Science to extract valuable insights from large datasets, aiding decision-making processes in various industries.

Key Significance of Network Science in Informatics

3. Social Network Analysis

Social Network Analysis:

- With the rise of social media and online platforms, the study of social networks has become indispensable.
- Network Science offers techniques to analyze information flow, identify influencers, and understand the dynamics of online communities.

Key Significance of Network Science in Informatics

4. Epidemiology and Disease Modeling

Epidemiology and Disease Modeling: I

In an era marked by global health challenges, Network Science plays a critical role in modeling the spread of diseases and devising effective intervention strategies. Informatics professionals can use these models to inform public health policies.

Key Significance of Network Science in Informatics

6. Cybersecurity

Cybersecurity:

- Network Science is instrumental in detecting and mitigating cybersecurity threats.
- Informatics experts can employ network analysis to identify vulnerabilities, track malware propagation, and enhance network security.

Key Significance of Network Science in Informatics

7. Machine Learning and AI

Machine Learning and AI:

- Network Science intersects with machine learning and artificial intelligence.
- Students can apply network-based algorithms to tasks such as recommendation systems, image analysis, and natural language processing.

Key Significance of Network Science in Informatics

8. Innovation and Collaboration

Innovation and Collaboration:

Network Science sheds light on innovation processes and collaboration dynamics. Informatics professionals can harness this knowledge to foster innovation within organizations and optimize teamwork.

Key Significance of Network Science in Informatics

Multidisciplanirity

- Network Science is a multidisciplinary field that holds immense promise for MSc Informatics students.
- Its capacity to uncover hidden patterns and connections in complex systems makes it a valuable tool in various informatics applications. By delving into the principles of Network Science, students can gain a deeper understanding of the world's interconnectedness and harness this knowledge to address real-world challenges effectively.
- Whether you are interested in data analysis, social networks, or cybersecurity, Network Science offers a robust framework to unlock new insights and drive innovation in the field of informatics.

Origin of Network Science

- Network Science, as a field of study, has evolved over centuries, and its contemporary form owes much to the pioneering work of several key figures.
- Understanding the historical context and origin of Network Science provides valuable insights, as it sheds light on the foundational concepts and principles that underpin the field.
- Among these key figures, the contributions of Paul Erdős, Leonhard Euler, and Duncan Watts and Steven Strogatz have played crucial roles in shaping Network Science into what it is today.

Seven Bridges of Königsberg

The roots of Network Science can be traced back to the 18th century when the Swiss mathematician Leonhard Euler made a groundbreaking contribution to the field. At the time, the city of Königsberg in Prussia posed a famous problem known as the "Seven Bridges of Königsberg." Euler's solution to this problem marked the birth of graph theory, a fundamental precursor to Network Science.

The Origin of Graph Theory

- Key Contributions: Paul Erdős, a Hungarian mathematician known for his prolific work in number theory and combinatorics, furthered the field.
- He introduced the concept of "random graphs" in the 1950s, laying the groundwork for the study of complex networks.
- Erdős' work on random graphs initiated the exploration of network properties and their implications, providing an early glimpse into the world of Network Science.

Leonhard Euler: The Seven Bridges Problem and Graph Theory

In 1736, Euler tackled the Seven Bridges of Königsberg problem, which involved finding a route that would cross each of the city's seven bridges once and only once. This problem, though seemingly unrelated to networks, led Euler to develop the fundamental concepts of graph theory.

Key Contributions: Euler introduced the idea of abstracting the city's landmasses and bridges into nodes and edges. He then demonstrated that the problem could be solved by determining whether a graph was traversable without retracing any edges, a concept now known as an Eulerian path. Euler's work laid the groundwork for the systematic study of networks, establishing the essential concept of graph theory.

Duncan Watts and Steven Strogatz: The Emergence of Small-World Networks

- In the late 1990s, Network Science experienced a resurgence of interest, thanks in part to the work of Duncan Watts and Steven Strogatz.
- Key Contributions: Watts and Strogatz introduced the concept of "small-world networks." They discovered that many real-world networks exhibited a striking property where nodes were not randomly connected but displayed both local clustering and long-distance connections.
- This revelation challenged traditional random graph models and led to a deeper understanding of the structure and dynamics of networks. Their "Watts-Strogatz model" demonstrated how networks could transition from regular to small-world structures through the rewiring of connections.

Timeline of Important Milestones in Network Science

1736: Leonhard Euler solves the Seven Bridges of Königsberg problem, laying the foundation for graph theory.

1950s: Paul Erdős introduces the concept of random graphs.

1998: Duncan Watts and Steven Strogatz publish their seminal paper on small-world networks.

2000s: The study of complex networks gains momentum, with researchers exploring scale-free networks, network robustness, and dynamics.

2006: Albert-László Barabási publishes "Linked: The New Science of Networks," popularizing Network Science in the mainstream.

Summary

- The historical context and origin of Network Science provide a rich backdrop as they delve into this interdisciplinary field.
- From the initial musings of Euler on the bridges of Königsberg to the groundbreaking work of Erdős, and later, Watts and Strogatz, Network Science has grown into a vibrant area of research with far-reaching applications.
- By tracing the development of the field and understanding the contributions of key figures, students can appreciate the profound impact of Network Science on modern informatics and its ongoing relevance in analyzing and understanding complex systems.

Nodes and Edges in Network Science

- Network Science is a multidisciplinary field that seeks to understand the structure, dynamics, and behavior of complex systems through the lens of networks.
- At the heart of network analysis are two fundamental components:

  nodes (vertices) and edges (links).
- We will delve into the definitions and principles of nodes and edges, and how they are interconnected, drawing from the Network Science theory of Albert-László Barabási.

Nodes and Edges

Nodes (Vertices): Nodes, also known as vertices, are the fundamental building blocks of a network. Each node represents an individual entity or element within the system you are studying. These entities can be diverse, ranging from people in a social network to websites on the internet, proteins in a biological network, or cities in a transportation network.

Edges (Links): Edges, often referred to as links or connections, represent the relationships or interactions between nodes. They signify how nodes are connected to each other within the network. Edges can take various forms, such as friendships between individuals in a social network, hyperlinks between web pages, biochemical interactions between proteins, or physical connections between cities in a transportation network.

Significance of Nodes and Edges

<u>Nodes as Entities</u>:

Nodes serve as the entities or objects of interest within a network. For instance, in a social network, nodes can represent individuals, and in a computer network, they can represent devices like computers or routers. Understanding the characteristics and attributes of nodes is crucial for network analysis, as it allows us to answer questions about the roles and properties of individual components.

Edges as Relationships

- Edges represent the connections or relationships between nodes. These connections are often defined by specific criteria relevant to the network under study.
- For instance, in a transportation network, edges may represent direct roads between cities, while in a citation network, they may signify references between scientific papers. The relationships captured by edges can vary widely, and they are essential for understanding how information, goods, or influence flows within a system.

How Nodes Are Connected by Edges

The Structure of Networks

- In Network Science, networks can have varying degrees of complexity. Some networks are simple, while others are highly intricate.
- The way nodes are connected by edges defines the structure of a network, and this structure can provide valuable insights into the system being studied.

Types of Connections

Nodes can be connected by edges in different ways, leading to various types of networks:

- Undirected Networks: In undirected networks, edges do not have a direction, meaning that the relationship between two nodes is symmetric. For example, in a friendship network, the connection between two friends is mutual.
- Directed Networks: In directed networks, edges have a specific direction, indicating that the relationship flofrom one node to another. For instance, in a web page hyperlink network, the links point from one page to another.

Types of Connections

- Weighted Networks: Some networks assign weights to edges, indicating the strength or importance of the connection. In a transportation network, the weight of an edge may represent the travel time between two cities.
- Multi-edge Networks: Networks can also have multiple edges connecting the same pair of nodes, each representing a distinct type of relationship or interaction.

Visualizing Networks

- To comprehend how nodes are connected by edges, network scientists often use visual representations.
- Graphs or network diagrams provide a visual snapshot of the network's structure, making it easier to identify patterns, hubs (highly connected nodes), and clusters (groups of densely connected nodes).

Summary

- Nodes and edges are the foundational elements of Network Science.
- Nodes represent entities or objects, while edges signify relationships or interactions within a network.
- Understanding how nodes are connected by edges is essential for analyzing complex systems, uncovering hidden patterns, and gaining insights into various domains, from social networks to biological systems and beyond.
- These concepts will enable you to harness the power of Network Science in your studies and future endeavors in the field of informatics.

Degrees in Network Science

- In the realm of Network Science, the concept of "degree" is fundamental to understanding the structure and behavior of networks.
- We will explore what degree means in the context of networks, differentiate between in-degree and out-degree, and provide real-world examples to illustrate these concepts, drawing from the Network Science theory of Albert-László Barabási.

Introduction to Degree in Networks

Definition of Degree

In the context of network analysis, "degree" refers to the number of edges connected to a node. It quantifies the extent to which a node is connected within the network. Degree is a critical measure because it helps us identify the importance, influence, and role of individual nodes within a network.

In-Degree and Out-Degree

In-Degree

In-degree is a measure that focuses on the incoming edges of a node. It tells us how many other nodes are directly connected to a specific node. In simple terms, in-degree reveals how many connections are directed towards a particular node.

In-Degree and Out-Degree

<u>Out-Degree</u>

- Out-degree is a measure that considers the outgoing edges from a node.
- It indicates how many other nodes are directly reachable from a specific node.
- In essence, out-degree reveals how many connections originate from a particular node.

Real-World Examples

Social Networks

- In-Degree: In a social network like Facebook, in-degree would represent the number of friend requests or connections directed towards a user. The higher the in-degree, the more popular or influential the user may be.
- Out-Degree: Out-degree, in this context, would represent the number of friends or connections a user has. A high out-degree indicates an individual who actively reaches out to others and forms connections.

Real-World Examples

Web Links

- In-Degree: In the World Wide Web, in-degree would signify the number of incoming hyperlinks pointing to a specific web page. A webpage with a high in-degree is considered influential or authoritative.
- Out-Degree: Out-degree would represent the number of hyperlinks on a webpage that lead to other pages. High out-degree implies that a webpage serves as a hub, directing users to various resources.

Real-World Examples

Transportation Networks

- In-Degree: In a transportation network, in-degree could indicate the number of incoming roads or transportation links leading to a city. Cities with high in-degree are major transportation hubs.
- Out-Degree: Out-degree, in this context, would represent the number of roads or transportation links originating from a city. Cities with high out-degree may have extensive transportation networks.

Summary

Transportation Networks

The concept of degree is a fundamental metric in Network Science that quantifies the connectivity and importance of nodes within a network. By differentiating between in-degree (incoming connections) and out-degree (outgoing connections), we gain a nuanced understanding of how nodes interact with their neighbors.

Real-world examples in social networks, the World Wide Web, and transportation networks demonstrate the practical applications of these concepts.

The concept of degree will empower you to analyze and leverage network structures effectively in various domains, from social media analysis to optimizing transportation systems.

Introduction to Clustering Coefficient

Definition of Clustering Coefficient

In network analysis, the clustering coefficient measures the extent to which nodes in a network tend to cluster together.

It quantifies the degree of local interconnectedness or "cliquishness" within the network. In other words, the clustering coefficient helps us understand how likely it is for a node's neighbors to be connected to each other.

Importance in Network Analysis
Identifying Local Structure

- The clustering coefficient provides insights into the local structure of a network.
- It helps us identify nodes or regions where connections are concentrated, potentially indicating the presence of communities or functional modules within the network.
- This information is invaluable for understanding the network's organization and dynamics.

Importance in Network Analysis

- A low clustering coefficient for a node can indicate a departure from the expected local structure.
- Such nodes might play unique or pivotal roles in the network.
- etecting nodes with unusually low clustering coefficients can be crucial in identifying outliers, potential vulnerabilities, or points of interest in various applications, including social networks and transportation systems.

Examples of Clustering Coefficients
High Clustering Coefficient

- Consider a social network within a close-knit community or a group of friends. In such a network, individuals are likely to be friends with not only each other but also with their mutual friends.
- This leads to a high clustering coefficient, as many connections form triangles within the network.
- High clustering coefficients often characterize tightly connected, cohesive groups.

Examples of Clustering Coefficients

Low Clustering Coefficient

- In contrast, imagine a transportation network with airports as nodes.
- If we calculate the clustering coefficient for a specific airport node, we might find that its connections do not create many triangles.
- This low clustering coefficient suggests that the airport's destinations are not well-connected to each other. In this context, low clustering may indicate less local interaction or direct flights between destinations.

Summary

- Theclustering coefficient is a valuable metric in Network Science that quantifies the local interconnectedness of nodes within a network.
- Its significance lies in revealing the network's local structure, identifying anomalies, assessing robustness, and aiding in various real-world applications.
- Understanding the clustering coefficient equips you with a powerful tool for analyzing and optimizing complex networks in fields ranging from social networks to transportation systems and beyond.

Introduction to the Small World Phenomenon

What is the Small World Phenomenon?

- The small world phenomenon, also known as the "small world effect" or "six degrees of separation," refers to the idea that in large networks, even when most nodes are not directly connected, it only takes a few intermediate connections to link any two nodes.
- In essence, it suggests that the world is smaller and more interconnected than we might initially perceive.

Six Degrees of Separation

- The "six degrees of separation" concept posits that any two people on Earth can be connected through a chain of acquaintances, with an average of only six intermediary connections.
- In other words, you are, on average, six introductions away from anyone else in the world. This concept extends beyond social networks to various types of networks.

Real-World Examples of the Small World Phenomenon
Social Networks

Real-World Example:

- The "Kevin Bacon Game" is a popular illustration of the small world phenomenon. It challenges players to connect any actor to Kevin Bacon within six or fewer films.
- This game demonstrates how interconnected the film industry is, with actors being linked through a surprisingly small number of connections.

Real-World Examples of the Small World Phenomenon
Internet Connectivity

Real-World Example:

- The small world phenomenon is also evident in the structure of the World Wide Web. With billions of web pages, it's remarkable how a few clicks can connect any two web pages.
- The famous experiment by Jon Kleinberg in the late 1990s, known as the "Hubs and Authorities" algorithm, revealed that even in this vast digital landscape, web pages are typically within a few clicks of each other.

Real-World Examples of the Small World Phenomenon

Disease Spread

Real-World Example:

- The small world phenomenon has profound implications in epidemiology.
- Diseases can spread rapidly through a population due to the interconnectedness of social networks.
- A person can become infected and, within a few steps, transmit the disease to individuals who may be geographically distant but linked through a web of contacts.

Key takeaways

1. Network Science Definition: Network Science is an interdisciplinary field focused on studying complex systems represented as networks, encompassing various disciplines.

2. Historical Origins: The field of Network Science has historical roots, with Euler's graph theory, Erdős' random graphs, and Watts-Strogatz's small-world model as key milestones.

3. Interdisciplinary Nature: Network Science integrates concepts from mathematics, computer science, sociology, biology, and more, making it a powerful tool for understanding real-world systems.

4. Nodes and Edges: Networks consist of nodes (vertices) representing entities and edges (links) representing connections or relationships between them.

Key takeaways

5. Degree: Degree measures node connectivity, with in-degree focusing on incoming connections and out-degree on outgoing connections. It's a key metric for node importance.

6. Real-World Examples of Degree: In social networks, high in-degree indicates popularity, while high out-degree suggests an active connector. In web networks, in-degree indicates popularity, and out-degree indicates influence.

7. Clustering Coefficient: This metric reveals the local connectivity patterns in networks, helping identify communities and assess robustness.

8. Importance of Clustering Coefficient: It aids in community detection, resilience analysis, and understanding the network's structure at a local level.

Key takeaways

9. High Clustering Coefficient Examples: Tight-knit friend groups in social networks, closely interacting proteins in biological networks.

10. Low Clustering Coefficient Examples: Random networks, where connections are distributed randomly.

11. Small World Phenomenon: This concept, exemplified by "six degrees of separation," highlights the idea that any two people in the world are connected by a surprisingly small number of intermediaries.

12. Real-World Examples of the Small World Phenomenon: The Kevin Bacon Game in film networks, web pages connected through few clicks on the internet, disease spread in social networks.

Key takeaways

13. Significance in Informatics: Network Science is valuable for MSc Informatics students as it aids in analyzing complex systems, data mining, social network analysis, and understanding the interconnected world.

14. Network Visualization: Graphs and diagrams help visualize network structures, aiding in analysis and interpretation.

15. Community Detection: Understanding clustering and communities in networks is crucial for identifying functional units or groups of entities.

16. Robustness Analysis: Assessing a network's resilience to failures or attacks is vital for system design and security.

Key takeaways

17. Interconnectedness: The small world phenomenon highlights the surprising interconnectedness of nodes in diverse networks, a concept applicable in various domains.

18. Real-World Applications: Network Science finds applications in fields such as social media analysis, disease modeling, transportation optimization, and more.

19. Collaboration and Innovation: Understanding network dynamics can foster collaboration and innovation within organizations and research communities.

Key takeaways

20. Power of Networks: Network Science equips informatics professionals with powerful tools to navigate the complexities of interconnected systems in the digital age.

21. Continuous Evolution: The field of Network Science continues to evolve, with ongoing research and new discoveries shaping our understanding of complex networks.

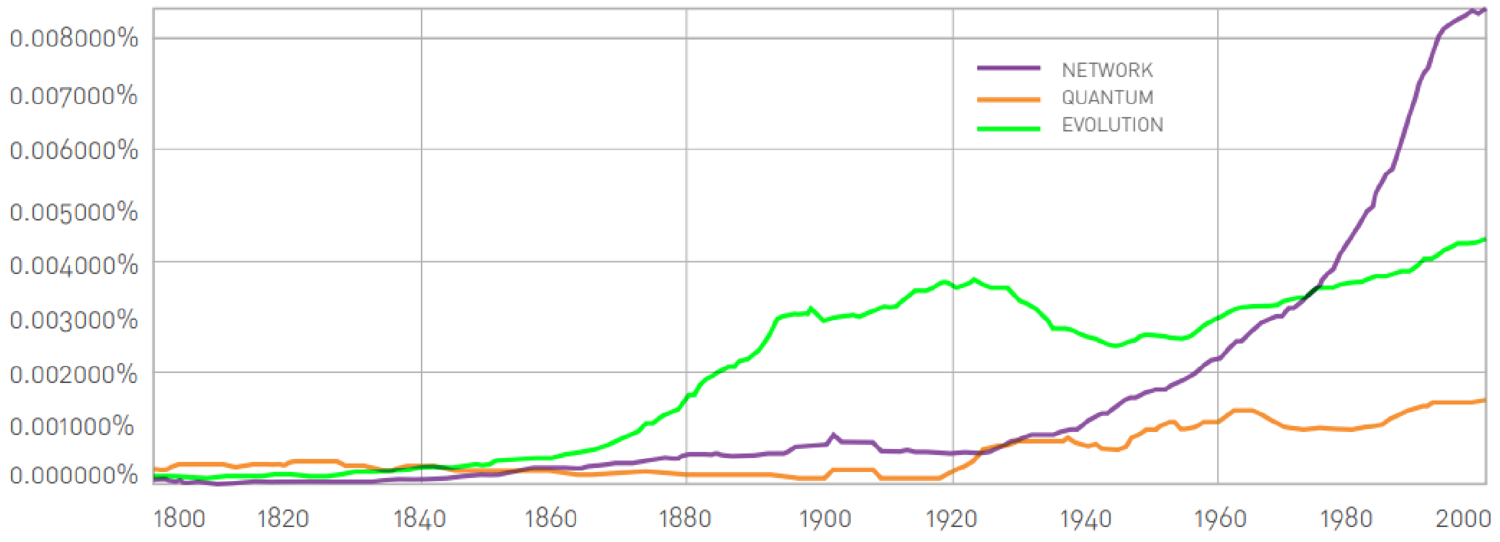
22. Cross-Disciplinary Insights: Embracing the interdisciplinary nature of Network Science can lead to novel insights and solutions in informatics and beyond.

Thank you for your attention!

Dr. Tamás Orosz
associate professor, Ph.D., habil.

NETWORK SCIENCE
Lecture 2

We learnt…



The frequency of use of the words evolution, quantum, and networks in books since 1880.

The plot indicates the exploding societal awareness of networks in the last decades of the 20th century, laying the ground for the emergence of network science.

The plots were generated by Google’s ngram platform, calculating the fraction of books published in a year

that mention evolution, quantum or networks.

Discussion

1.1. Networks Everywhere

- List three different real networks and state the nodes and links for each of them.

1.2. Your Interest

Tell us of the network you are personally most interested in. Address the following questions:

(a) What are its nodes and links?

(b) How large is it?

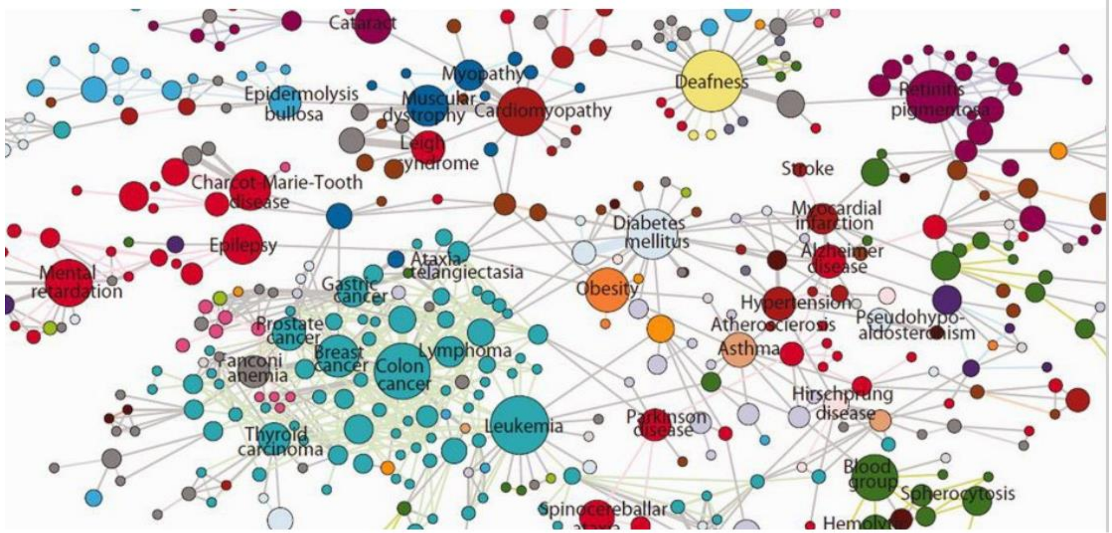
(c) Can be mapped out?

(d) Why do you care about it?

1.3 Impact

In your view what would be the area where network science could have the biggest impact in the next decade? Explain your answer

Human Disease Network



The Human Disease Network, whose nodes are diseases connected if they have common genetic origin. Published as a supplement of the Proceedings of the National Academy of Sciences, the map was created to illustrate the genetic interconnectedness of apparently distinct diseases.

With time it crossed disciplinary boundaries, taking up a life of its own. The New York Times created an interactive version of the map and the London-based Serpentine Gallery, one of the top contemporary art galleries in the world, have exhibited it part of their focus on networks and maps. It is also featured in numerous books on design and maps.

The Bridges of Königsberg

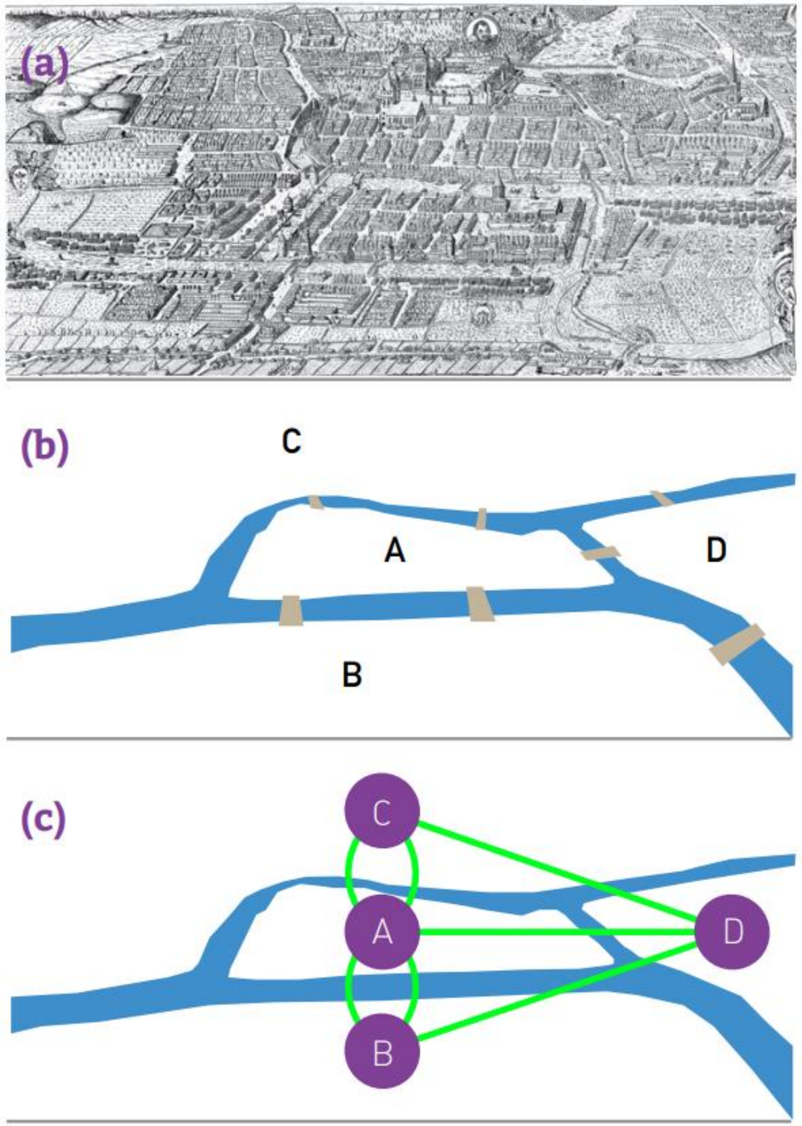


Figure The Bridges of Königsberg

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 1735 in Königsberg, the capital of Eastern Prussia, 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 to the mainland the elegant island Kneiphof, caught between the two branches of the Pregel. The remaining two crossed the two branches of the river (Figure). 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.

The Bridges of Königsberg

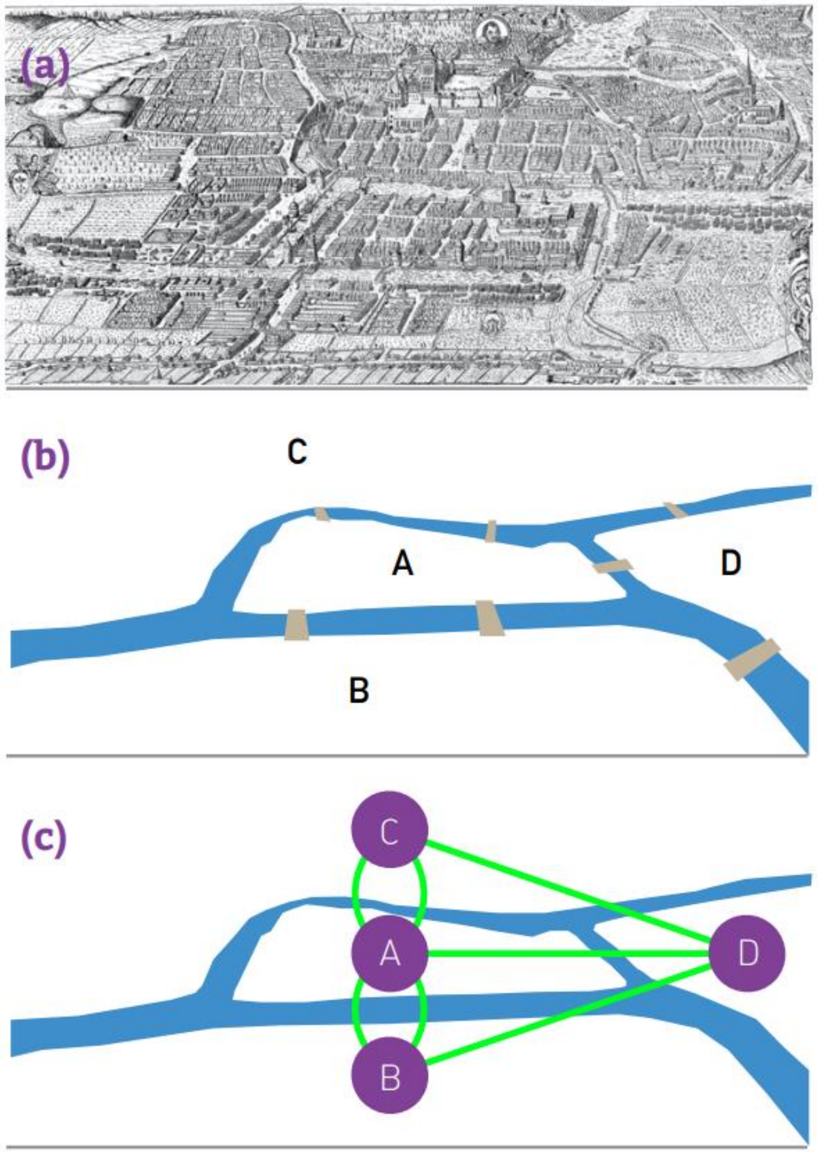


Figure The Bridges of Königsberg

Euler represented each of the four land areas separated by the river with letters A, B, C, and D (Figure). 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 find yourself having no unused link for you to leave it. A walking 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 four nodes with an odd number of links, A, B, C, and D, so no path could satisfy the problem.

The Bridges of Königsberg

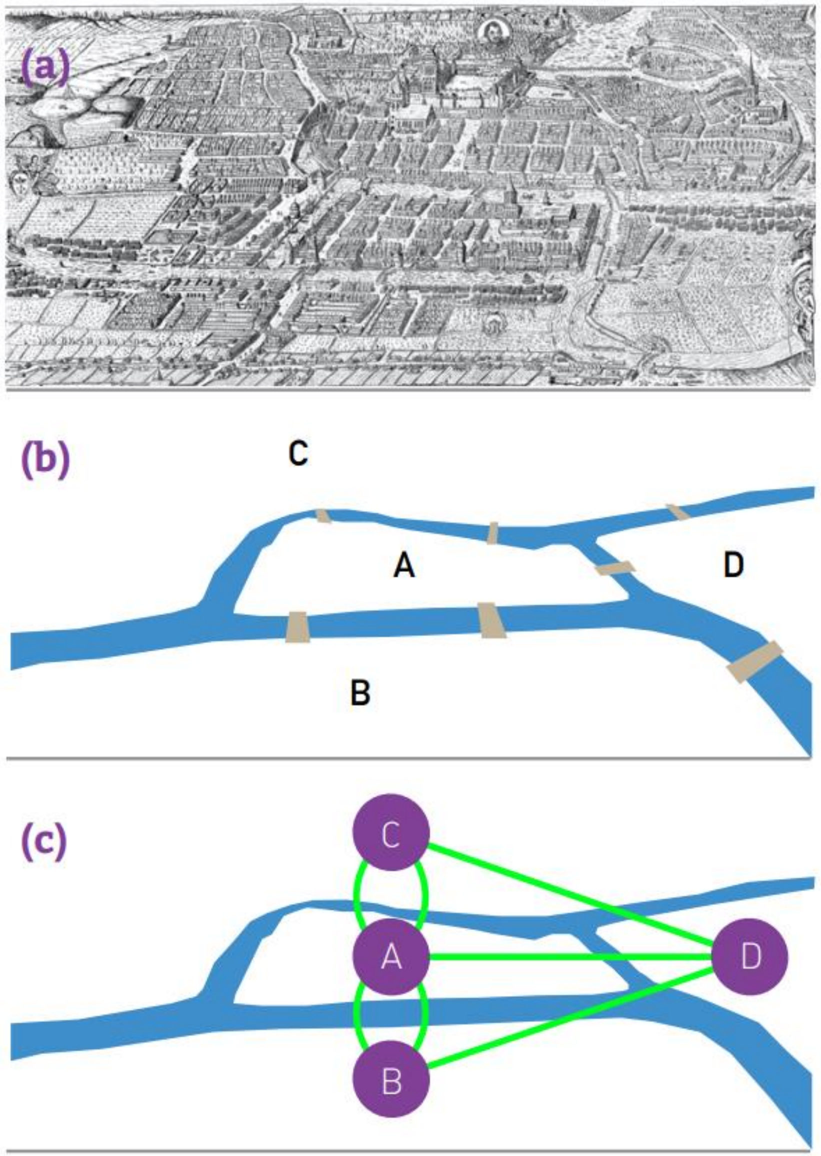


Figure The Bridges of Königsberg

a) A contemporary map of Königsberg (now Kaliningrad, Russia) during Euler's time.

b) A schematic illustration of Königsberg's four land pieces and the seven bridges across them.

c) (c) Euler constructed a graph that has four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. He then showed that there is no continuous path that would cross the seven bridges while never crossing the same bridge twice. The people of Königsberg gave up their fruitless search and in 1875 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. Consequently we should be able to find the desired path. **Homework**: Can you find one yourself?

The Bridges of Königsberg

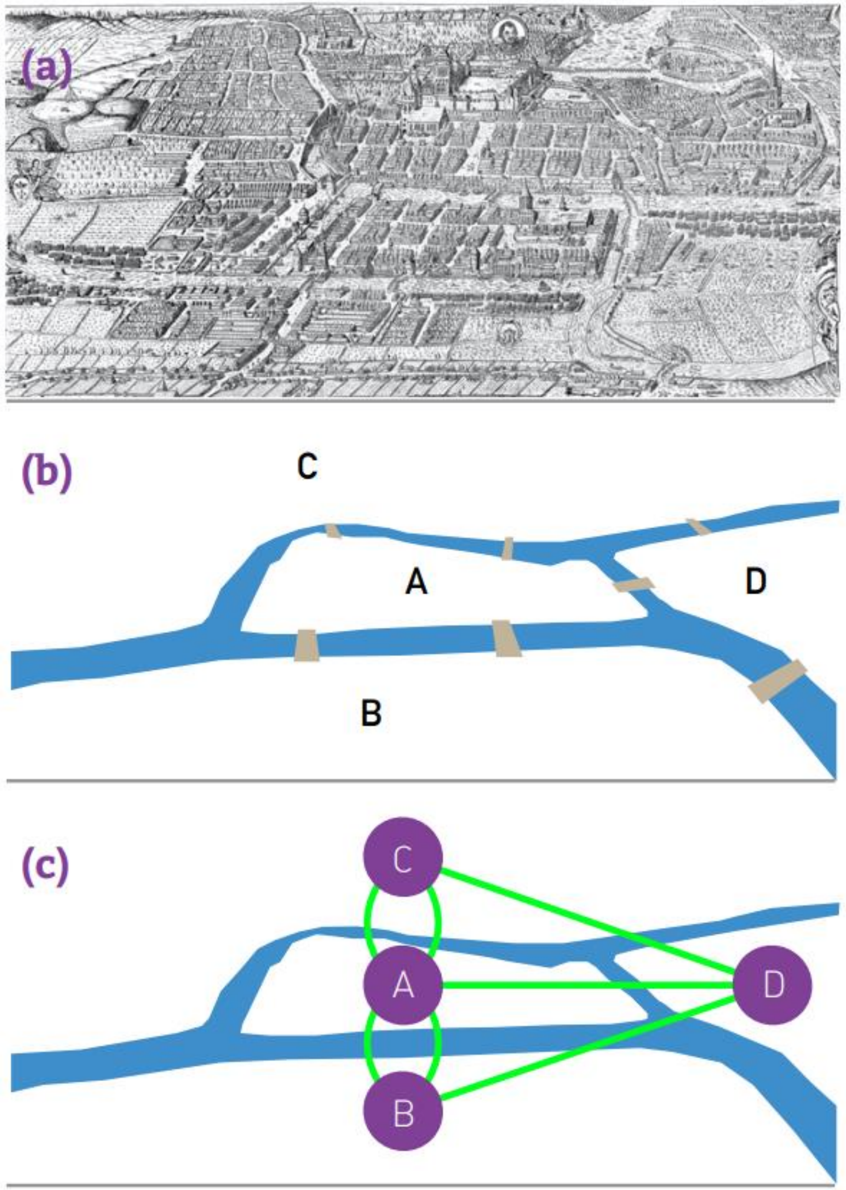


Figure The Bridges of Königsberg

Euler's proof was the first time someone solved a mathematical problem using a graph. For us the proof has two important messages: The first is that some problems become simpler and more tractable 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 encoded in their structure that limit or enhance their behavior.

To understand the many ways networks can 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. We learn how to represent a network as a graph and introduce the elementary characteristics of networks, from degrees to degree distributions, from paths to distances and learn to distinguish weighted, directed and bipartite networks. We will introduce a graph-theoretic formalism and language.

NETWORKS AND GRAPHS

If we want to understand a complex system, we first need to know how its components interact with each other. In other words we 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. This network representation offers a common language to study systems that may differ greatly in nature, appearance, or scope. Indeed, as shown in Figure 2.2, three rather different systems have exactly the same network representation.

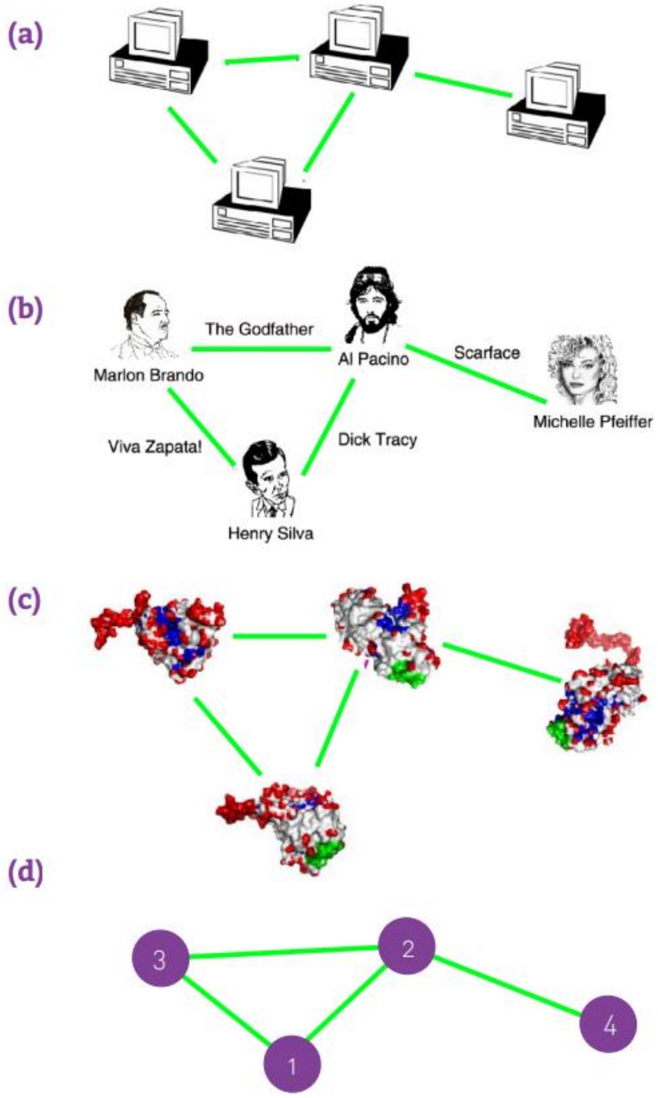


Figure 2.2
Different Networks, Same Graph

Basic network parameters

- Number of nodes, or N, represents the number of components in the system. We will often call N the size of the network. To distinguish the nodes, we label them with i = 1, 2, ..., N.

- Number of links, which we denote with L, represents the total number of interactions between the nodes. 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 networks shown in Figure 2.2 have N = 4 and L = 4.

Directed or undirected links of a network

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 have undirected links, like romantic ties: if John dates Janet, Janet also dates John, or like 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; it is called 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).

System representation by Network

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

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

(b) 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.

System representation by Network (2)

(c) 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.

(d) 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 have different uses and purposes.

- We can also build networks that may be valid from a graph theoretic perspective, but may have little practical utility. For example, if we link all individuals with the same first name, Johns with Johns and Marys with Marys, we do obtain a well-defined graph, whose properties can be analyzed with the tools of network science. Its utility is questionable, however.
- 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.
- Throughout our investigations we will use ten networks to illustrate the tools of network science. These reference networks span 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 in the E. coli metabolism, to almost half million nodes in the citation network.

They cover several areas where networks are actively applied, representing ‘canonical’ datasets frequently used by researchers to illustrate key network properties. As we indicate, some of them are directed, others are undirected.

We are discussing 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.

DEGREE, AVERAGE DEGREE, AND DEGREE DISTRIBUTION

- If we want to understand a complex system, we first need to know how its components interact with each other. In other words we 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.
- This network representation offers a common language to study systems that may differ greatly in nature, appearance, or scope. Indeed, as shown in Figure 2.2, three rather different systems have exactly the same network representation.

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.

DEGREE

We denote with ki the degree of the i_{th} node in the network. For example, for the undirected networks shown in Figure 2.2 we have $k_1=2$, $k_2=3$, $k_3=2$, $k_4=1$.

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

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i \cdot$$

.

Here the 1/2 factor corrects for the fact that in the sum each link is counted twice. For example, the link connecting the nodes 2 and 4 in Figure 2.2 will be counted once in the degree of node 1 and once in the degree of node 4.

AVERAGE DEGREE

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

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N}$$

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

$$\boldsymbol{k_i = k_i^{in} + k_i^{out}}$$

- For example, on the WWW the number of pages a given document points to represents its outgoing degree, k_{out}, and the number of documents that point to it represents its incoming degree, k_{in}. The total number of links in a directed network is

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$

- The 1/2 factor seen earlier is now absent, as for directed networks the two sums separately count the outgoing and the incoming degrees. The average degree of a directed network is

$$\langle k^{\text{in}} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{\text{in}} = \langle k^{\text{out}} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{\text{out}} = \frac{L}{N}$$

DEGREE DISTRIBUTION

- 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 network with N nodes the degree distribution is the normalized histogram is given by

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

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

$$N_k = Np_k.$$

Degree distribution (2)

- The degree distribution has assumed a central role in network theory following the discovery of scale-free networks. One reason 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

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

- The other reason is that the precise functional form of pk determines many network phenomena, from network robustness to the spread of viruses.

ADJACENCY MATRIX

A complete 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 Figure 2.2 is uniquely described by listing

- its four 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

ADJACENCY MATRIX

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, $A_{ij} = A_{ji}$. The degree ki 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.

$$k_i = \sum_{j=1}^{N} A_{ji} = \sum_{i=1}^{N} A_{ji} \ .$$

ADJACENCY MATRIX

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

$$k_i^{\text{in}} = \sum_{j=1}^{N} A_{ij}, \qquad k_i^{\text{out}} = \sum_{j=1}^{N} A_{ji}.$$

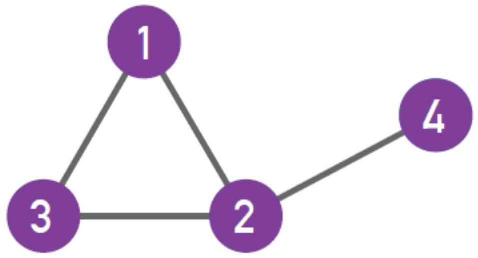
- Given that in an undirected network the number of outgoing links equals the number of incoming links, we have

$$2L = \sum_{i=1}^{N} k_i^{\text{in}} = \sum_{i=1}^{N} k_i^{\text{out}} = \sum_{ij}^{N} A_{ij}.$$

- 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 $A_{ji} = 1$, a link pointing from i to j.

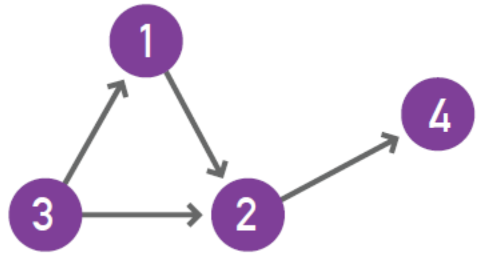
ADJACENCY MATRIX

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



Undirected network

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



Directed network

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

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 nervous system of a living organism, has $N = 302$ neurons (nodes).
- In contrast the human brain is estimated to have about a hundred billion $N \approx 1011$) neurons. The genetic network of a human cell has about 20,000 genes as nodes; the social network consists of eight billion individuals ($N \approx 8\times10^9$) and the WWW is estimated to have over many trillions web documents ($N > 1012$).

Number of Links

- These wide differences in size are noticeable, which are reflected by 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), while others are only samples, representing a subset of the full network (like the WWW or the mobile call graph).
- The number of links also varies widely. In a network of N nodes the number of links can change between L = 0 and Lmax, where

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

is the total number of links present in a complete graph of size N In a complete graph each node is connected to every other node.

L in real networks

In real networks L is much smaller than L_{max}, reflecting the fact that most real networks are sparse.

We call a network *sparse* if $L \ll L_{max}$.

- For example, the WWW graph in has about 1.5 million links.
- Yet, if the WWW were to be a complete graph, it should have Lmax ≈ 5x1010 links
- Consequently the web graph has only a 3×10^{-5} fraction of the links it could have.

WEIGHTED NETWORKS

So far we discussed only networks for which all links have the same weight, i.e. $A_{ij} = 1$. 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 individuals 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 as

$$A_{ij} = w_{ij} .$$

- Most networks of scientific interest are weighted, but we can not always measure the appropriate weights. Consequently we often approximate these networks with an unweighted graph. We predominantly focus on unweighted networks, but whenever appropriate, we discuss how the weights alter the corresponding network property.

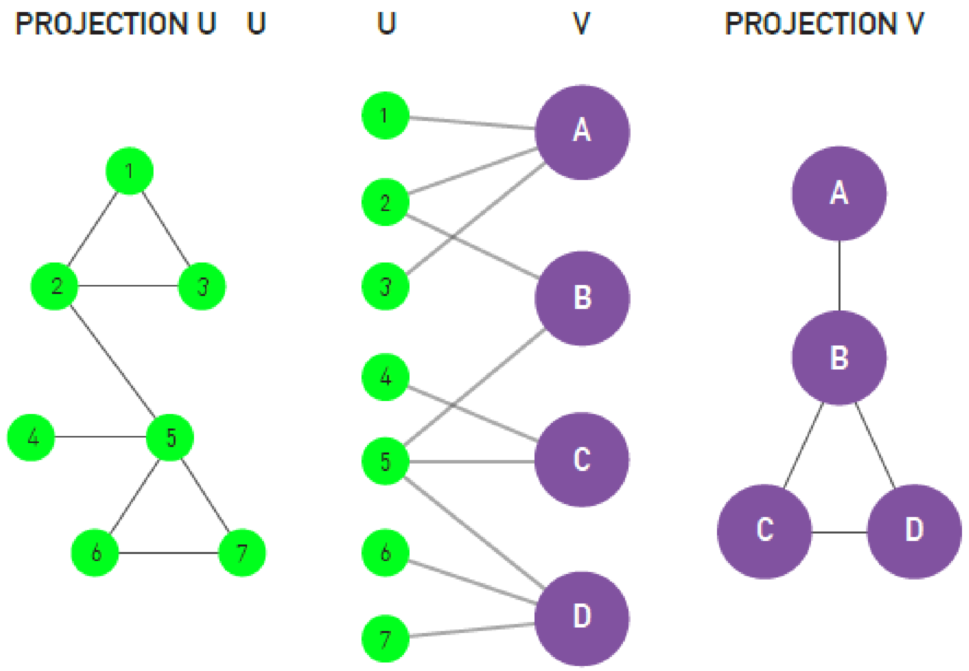
BIPARTITE NETWORKS

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 green and the V-nodes purple, then each link must connect nodes of different colors

We can generate two projections for each bipartite network.

- The first projection connects two U-nodes by a link if they are linked to the same V-node in the bipartite representation.
- The second projection connects the V-nodes by a link if they connect to the same U-node

BIPARTITE NETWORKS



- 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 is connected to an actor if the actor plays in that movie. One projection of this bipartite network is the actor network, in which two nodes are connected to each other if they played in the same movie.
- The other connected if they share at least one projection is the movie network, in which two movies are actor in their cast.

PATHS AND DISTANCES

- Physical distance plays a key role in determining the interactions between the components of physical systems. For example, the distance between two atoms in a crystal or between two galaxies in the universe determine the forces that act between them.
- In networks distance is a challenging concept. Indeed, what is the distance between two webpages, or between two individuals who do not know each other?
- The physical distance is not relevant here: Two webpages could be sitting on computers on the opposite sides of the globe, yet, have a link to each other.
- At the same time two individuals that live in the same building may not know each other.

PATHS AND DISTANCES

- In networks physical distance is replaced by *path length*.

A *path* is a route that runs along the links of the network.

A path's *length* represents

- the number of links the path contains.
- Note that some texts require that each node a path visits is distinct.
- In network science paths play a central role.

SHORTEST PATH

- The shortest path between nodes i and j is the path with the fewest number of links The shortest path is often called the distance between nodes i and j, and is denoted by d_{ij}, or simply d. We can have multiple shortest paths of the same length d between a pair of nodes 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 $d_{ij} \neq d_{ji}$. Furthermore, in a directed network the existence of a path from node i to node
- In real networks we often need to determine the distance between two j does not guarantee the existence of a path from j to i. nodes.
- For a small network, this is an easy task. For a network with 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. In practice we use the breadth first search (BFS) algorithm for this purpose.

NETWORK DIAMETER

- The *diameter* of a network, denoted by d_{max}, is the maximum shortest path in the network.
- In other words, it is the largest distance recorded between *any* pair of nodes.

AVERAGE PATH LENGTH

- The 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

$$d = \frac{1}{N(N-1)} \sum_{\substack{i,j=1,N \\ i \neq j}} d_{i,j} \, .$$

CONNECTEDNESS

- A phone would be of limited use as a communication device if we could not call any valid phone number; email would 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 network behind the phone or the Internet must be capable of establishing a path between any two nodes. This is in fact the key utility of most networks: they ensure connectedness.

CONNECTEDNESS (2)

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

CONNECTEDNESS

- A network is connected if all pairs of nodes in the network are connected.
- A network is disconnected if there is at least one pair with $d_{ij} = \infty$.
- 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.

CLUSTERING COEFFICIENT

- The clustering coefficient captures the degree to which the neighbors of a given node link to each other. For a node *i* with degree *ki* the *local clustering coefficient* is defined as

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

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.
- C_i is the probability that two neighbors of a node link to each other.

Consequently 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 link density: The more densely interconnected the neighborhood of node i, the higher is its local clustering coefficient.

CLUSTERING COEFFICIENT

- The degree of clustering of a whole network is captured by the *average clustering coefficient*, $\langle C \rangle$, representing the average of C_i over all nodes $i = 1, ..., N$,

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

- In line with the probabilistic interpretation $\langle C \rangle$ is the probability that two neighbors of a randomly selected node link to each other.

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \, .$$

Bibliography

1. ALBERT-LÁSZLÓ BARABÁSI: NETWORK SCIENCE

http://www.networksciencebook.com/

2. Filippo Menzer, Santo Furtando, and Clayton A. Davis: A First Course in Network Science

Network Science

Which are the most important details so far…?

<u>Basic Definitions</u>

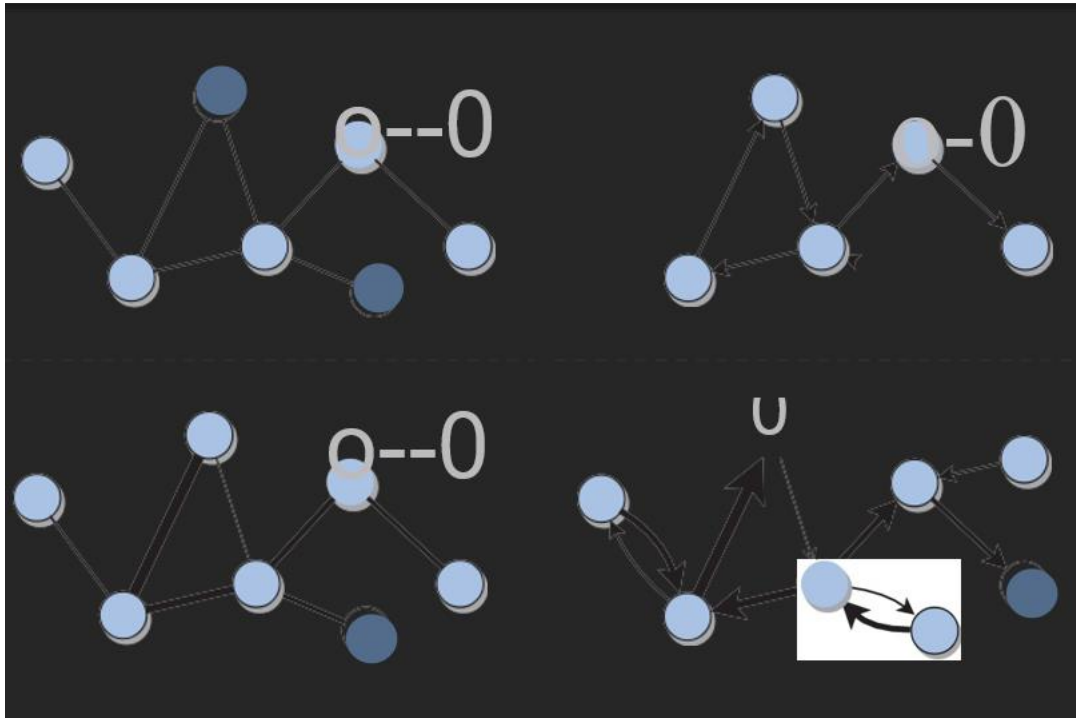
- **Network** (graph): set of elements called **nodes**
  - ❑ Set of connections between pairs of nodes called **links**
  - ❑ Two nodes are **connected** (or adjacent)if there is a link between them
- **Definition of the Network**:
  - ❑ A network **G** has two parts, a set of **N** elements, called **nodes** or **vertices**, and a set of **L** pairs of nodes, called **links** or **edges**.
  - ❑ The **link (i,j)** joins the nodes **i** and **j**.

Undirected or directed networks

- A **directed network** is also called a ***digraph***. In directed networks, links are called ***directed links*** and the order of the nodes in a link reflects the direction:
  - ❑ The link (i,j) goes from the source node i to the target node j
- In **undirected networks**, all links are bi-directional and
  - ❑ The order of the two nodes in a link does not matter.
- A network can be **unweighted** or **weighted**.
  - ❑ In a weighted network, links have associated weights: the weighted link (i,j,w) between nodes i and j has weight w.
  - ❑ A network can be both directed and weighted, in which case it has directed weighted links.

Undirected or directed networks

- Graphical representations of undirected, directed, and weighted networks. The circles represent the nodes.
- Pairs of adjacent nodes are connected by a line (link) or arrow (directed link).
- Arrows indicate the direction of the links.
- The thickness of a link represents its weight in weighted networks.



Density and Sparsity

- The **maximum number of links** in a network is bounded by the possible number of distinct connections among the nodes of the system. The maximum number of links is therefore given by the number of pairs of nodes.
- A network with the maximum number of links, in which all possible pairs of nodes are connected by links, is called a **complete network**.
- The maximum number of links in an undirected network with N nodes is the number of distinct pairs of nodes: $\boldsymbol{L_{Max} = \binom{N}{2} = N * (N-1)/2}$

Density and Sparsity (2)

- The density of a network with *N* nodes and *L* links is $\boldsymbol{d = L/L_{Max}}$.
  - ❑ In an undirected network:

$$\boldsymbol{d = L/L_{Max} = \frac{2L}{N(N-1)}}$$

  - ❑ In a directed network:

$$\boldsymbol{d = L/L_{Max} = \frac{L}{N(N-1)}}$$

Density and Sparsity (3)

- In a complete network, **$d = 1$** by definition, since **$L = L_{max}$**.
- In a sparse network **$L << L_{max}$**, $d<<1$.
- When a network grows very large, we can observe how the number of links increases as a function of the number of nodes.
- The network is **sparse** if the number of links grows proportionally to the number of nodes **$L \sim N$**, or even slower. If instead the number of links grows faster, e.g. quadratically with network size (**$L \sim N^2$**), then we say that the network is **dense**.
- **Path:** a sequence of arcs in a network that can be traced continuously without retracing any arc.

Subnetworks

- A subset of a network, which is itself a network is called a **subnetwork** (or subgraph).
- A subnetwork is obtained by **selecting** a **subset** of the **nodes** and **all** of the **links** among these nodes.
- A special type of subnetwork is the **ego network** of a node, which is the subnetwork consisting of the chosen node - called the ego - and its neighbors. Ego networks are often studied in social network analysis.

Degree

- The **degree** of a node is its **number** of **links**, or neighbors. We use **k_i** to denote the degree of node i.
- A node with no neighbors, has degree zero (k = 0) and is called a **singleton**.
- The **average degree** of a network is denoted by **<k>**. It is an important property and is related (directly proportional) to its density.
- The average degree of a network is defined as

$$< \boldsymbol{k} > = \frac{\sum_i \boldsymbol{k_i}}{\boldsymbol{N}}$$

Network representations

- To store/retrieve a network in/from a computer file or memory, we need a way to formally represent its nodes and links.
- There are several possible network representations.
- The simplest is the **adjacency matrix**, an N x N matrix in which each element represents the link between the nodes indexed by the corresponding row and column.

Handling Networks in Code

- **Network analysis** and **visualization tools**, and **libraries** are available
  - to handle networks in many programming languages, such as **Gephi**
  - **NetworkX** (networkx.github.io; https://networkx.org/), a Python package for the creation, manipulation, and study of the structure, dynamics, and function of networks.

    NetworkX provides data structures, algorithms, measures, and generators for networks, as well as rudimentary visualization facilities.

Connectedness and Components

- To relate network structure and function, it is useful to consider the ***connectedness*** of a network. The connectedness defines many properties of a network's physical structure.

- The number of links in a network is bound by the number of nodes. This is an **upper bound**; there is **no lower bound**, as a network might have no links at all.

- The **higher the density**, the greater the chances that the network is connected.

- The **fewer the links** and the **lower the density**, the higher the chances that the network is disconnected.

Six Degrees of Separation

- The name of the game Six Degrees of Kevin Bacon is inspired by the concept of ***six degrees of separation***.
- The idea is the same as that of a small world: any two people in the world are connected by a short chain of acquaintances. In other words, social networks have a short diameter and an even shorter average path length. The number "six" in the expression originated from Hungarian author Frigyes Karinthy in the 1920s, and some credit goes to Italian inventor Guglielmo Marconi as well for coming up with the same idea 20 years before, in the early 1900s. However, what made the "six degrees" expression famous was an experiment conducted by psychologist Stanley Milgram in the 1960s, which provided the first empirical evidence of small worlds.

Summary

1. A network is made up of two sets of elements: the nodes and links connecting pairs of nodes.

2. A subnetwork is a subset of the network including some of its nodes and all of the links among them.

3. In directed networks, links have a direction. There may be a link from node 1 to node 2, and not necessarily one from node 2 to node 1. In undirected networks, links are reciprocal.

Summary (2)

4. In weighted networks, links have associated weights that represent connection attributes like importance, similarity, distance, traffic, etc. In unweighted networks, all links are the same.
5. Multilayer networks have different types of nodes and links, divided into interconnected layers. If the nodes are the same in each layer, the multilayer network is called a multiplex.
6. The density of a network is the fraction of node pairs that are connected. A network is complete if all pairs of nodes are connected, so that the density is one. Most real networks are sparse, meaning that they have very small density.

Summary (3)

7. The degree of a node is the number of neighbors. In directed networks, nodes have in-degree and out-degree measuring the number of incoming and outgoing links, respectively. If the network is weighted, the strength of a node is the sum of the weights of its links. The nodes of weighted directed networks have in-strength and out-strength.
8. Adjacency lists and edge lists are efficient representations to store sparse networks.
9. NetworkX is a popular and convenient programming library to code networks in the Python language.

Homework

1. Consider a network with N nodes. Given a single link, what is the maximum number of nodes that link can connect? Given a single node, what is the maximum number of links that can connect to that node?
2. Consider a directed network of N nodes. Now consider the total in-degree (i.e. the sum of the in-degree over all nodes in the network). Compare this to the analogous total out-degree. Which of the following must hold true for any such network?
   a. Total in-degree must be less than total out-degree
   b. Total in-degree must be greater than total out-degree
   c. Total in-degree must be equal to total out-degree
   d. None of these hold true in all instances

3. Consider an undirected network with N nodes. What is the maximum number of links this network can have?
4. Consider a Twitter retweet network, where users are nodes and we want to show how many times a given user has retweeted another user. What link type best captures this relation?
    a. Undirected, unweighted
    b. Undirected, weighted
    c. Directed, unweighted
    d. Directed, weighted

Trees

- A special class of undirected, connected networks such that the deletion of any one link will disconnect the network into two components. Such graphs are called trees.
- The number of links in a tree is $\boldsymbol{L = N - 1}$.

Proof:

Start with a network with $\boldsymbol{N = 2}$ nodes, which needs $\boldsymbol{L = 1}$ link to be connected. Then, as we add one node at a time, we must add a link to connect the new node to some existing node. So the number of links is always equal to the number of nodes minus one. Removing any link will disconnect at least one node.

Trees (2)

- **Trees have no cycles**.

Proof:

Trees cannot have cycles by contradiction: if a tree had a cycle, we could remove at least one link of the cycle without disconnecting it.

Therefore the network would not be a tree - a contradiction. Because there are no cycles, given any pair of nodes, there is only a single path connecting them.

Trees (3)

- **Trees are hierarchical**.

You can pick any node in a tree and call it a root. Each node in a tree is connected to a parent node (toward the root) and to one or more children nodes (away from the root).

The exceptions are the root, which has no parent, and the so-called leaves of the tree, which have no children.

Hubs

hub: a center around which other things revolve or from which they radiate; a focus of activity, authority, commerce, transportation, etc.

- If you have traveled on a plane, you have traversed an important network - the air transportation network. Nodes are airports and links are direct flights between them. While most air ports are rather small, a few major ones (e.g. Atlanta, Chicago, Denver) have daily flights to hundreds or even thousands of destinations.
- Similarly, in social communities there are individuals who are much more visible and influential than others; and on the Web there are some very popular sites, such as google.com, while most sites are unknown to most.

Heterogeneity

Heterogeneous networks present a wide variability in the properties and roles of their elements - nodes and/or links.

This reflects the diversity present in the complex systems described by networks. In air transportation networks, social networks, the Web, and many other net works, a clear source of heterogeneity is the degree of the nodes:

a few nodes have many connections (Atlanta, Google, Obama), while most nodes have few.

Centrality Measures

1. **Degree: t**he degree of a node is the number of neighbors of that node.

- In the example of the US airport network, the degree of a node (airport) is the number of other airports reachable from it via direct flights.
- In a social network, the degree of a node (individual) is the number of social links connecting the node to others. For instance, in a coauthorship network, the degree is the number of collaborators.High-degree nodes in social net works are people with many connections - whether because they are sociable, sought after, or simply eager to collaborate, these nodes seem to be important in some sense.

<u>Therefore, the degree is a very natural measure of centrality in social networks.</u>

The **average degree** of a network indicates how connected the nodes are on average. The average degree may not be representative of the actual distribution of degree values. This is the case when the nodes have heterogeneous degrees, as in many real-world networks.

Closeness

Another way to measure the centrality of a node is by determining how "close" it is to the other nodes.

This can be done by summing the distances from the node to all others.

If the distances are short on average, their sum is a small number and we say that the node has **high centrality**.

This leads to the definition of **closeness centrality**, which is simply the inverse of the sum of distances of a node from all others.

Betweenness

- Many phenomena taking place in networks are based on diffusion processes.
- Examples include the transmission of information across a social network, the traffic of goods through a port, and the spread of epidemics in the network of physical contacts between the individuals of a population.
- This has suggested a third notion of centrality, called
- **betweenness**: a node is the more central, the more often it is involved in these processes.

Robustness

- A system is robust if the failure of some of its components does not affect its function. For instance, an airplane keeps flying if one of its engines stops working. In general, robustness depends on which components fail and on the extent of the damage.
- How to define the robustness of a network?
- Nodes can describe a broad variety of entities, such as people, routers, proteins, neurons, websites, and airports. In such a high-level representation, it is not straightforward to define the failure of a node, which depends on the specific type of network. But if we assume that a node stops working somehow, we can ask how the structure and consequently the function of the network changes without that node and all of its links.

Standard robustness test for networks

- The standard robustness test for networks consists of checking how the connectedness is affected as more and more nodes are removed, along with all of their adjacent links.
- To estimate the amount of disruption following node removal, scholars compute the relative size of the giant component (i.e. the ratio of the number of nodes in the giant component to the number of nodes initially present in the network).

Standard robustness test for networks

- Let us suppose that the initial network is connected.
- In this case the giant component coincides with the whole network, so its relative size is one. If the removal of a subset of nodes does not break it into disconnected pieces, the proportion of nodes in the giant component just decreases by the fraction of removed nodes.
- If, however, the node removal breaks the network into two or more connected components, the size of the giant component may drop substantially.
- As the fraction of removed nodes approaches one, the few remaining nodes are likely distributed among tiny components, so the proportion of nodes in the giant component is close to zero.

Resources

- Menczer, Filippo, Santo Fortunato, and Clayton A. Davis. A first course in network science. Cambridge University Press, 2020.

NETWORK SCIENCE

RANDOM NETWORKS

Erdős Number

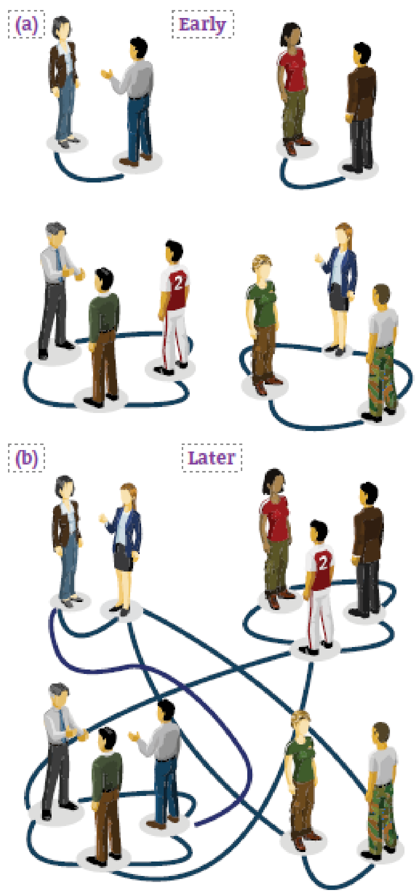
The Hungarian mathematician Pál Erdős authored hundreds of research papers, many of them in collaboration with other mathematicians. His relentless collaborative approach to mathematics inspired the Erdős Number, which works like this:

Erdős' Erdős number is 0. Erdős' coauthors have Erdős number 1. Those who have written a paper with someone with Erdős number 1 have Erdős number 2, and so on.

If there is no chain of co-authorships connecting someone to Erdős, then that person's Erdős number is infinite. Many famous scientists have low Erdős numbers: Albert Einstein has Erdős Number 2 and Richard Feynman has 3. The collaborators of Pál Erdős, as drawn in 1970 by Ronald Graham, one of Erdős' close collaborators. As Erdős' fame rose, this image has achieved an iconic status.

Imagine organizing a party for a hundred guests who initially do not know each other. Offer them wine and cheese and you will soon see them chatting in 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, your expensive wine appears to be safe, as she only had time to meet a few others so far.

The guests will continue to mingle, however, 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 there is an invisible path from John to Mary through Mike. As time goes on, the guests will be increasingly interwoven by such elusive links. With that the secret of the unlabeled bottle will pass from Mary to Mike and from Mike to John, 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 your fine wine would be left for you to enjoy once the guests are gone.

Yet, you would be wrong. We 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 emerge that will allow the information to reach all of them. Hence in no time everyone will be enjoying the better wine.

THE RANDOM NETWORK MODEL

- Network science aims to build models that reproduce the properties of real networks. Most networks we encounter do not have the comforting regularity of a crystal lattice or the predictable radial architecture of a spider web. Rather, at first inspection they look as if they were spun randomly. Random network theory embraces this apparent randomness by constructing and characterizing 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 decide where to place the links between the nodes so that we reproduce the complexity of a real system. In this respect the philosophy behind a random network is simple: We assume that this goal is best achieved by placing the links randomly between the nodes. That takes us to the definition of a random network:
- A random network consists of N nodes where each node pair is connected with probability p.

THE RANDOM NETWORK MODE

To construct a random network we follow these steps:

1) Start with N isolated nodes.

2) Select a node pair and generate a random number between 0 and 1. If the 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 after 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 these networks. In their honor a random network is called the Erdős-Rényi network

Anatol Rapoport (1911-2007), a Russian immigrant to the United States, was the first to study random networks. Rapoport's interests turned to mathematics after realizing that a successful career as a concert pianist would require a wealthy patron. He focused on mathematical biology at a time when mathematicians and biologists hard-ly spoke to each other. In a paper written with Ray Solomonoff in 1951, Rapoport demonstrated that if we increase the average degree of a network, we observe an abrupt transition from disconnected nodes to a graph with a giant component.

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, they merged probability theory and combinatorics with graph theory, establishing random graph theory, a new branch of mathematics.

The random network model was independently introduced by Edgar Nelson Gilbert (1923-2013) 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 founders of random graph theory.

NUMBER OF LINKS

Each random network generated with the same parameters N, p looks slightly different. Not only the detailed wiring diagram changes between realizations, but so does 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,

$$\binom{\frac{N(N-1)}{2}}{L},$$

counting the number of different ways we can place L links among N(N-1)/2 node pairs.

NUMBER OF LINKS

We can therefore write the probability that a particular realization of a random network has exactly L links as

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

It is a binomial distribution, the expected number of links in a random graph is

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

NUMBER OF LINKS (2)

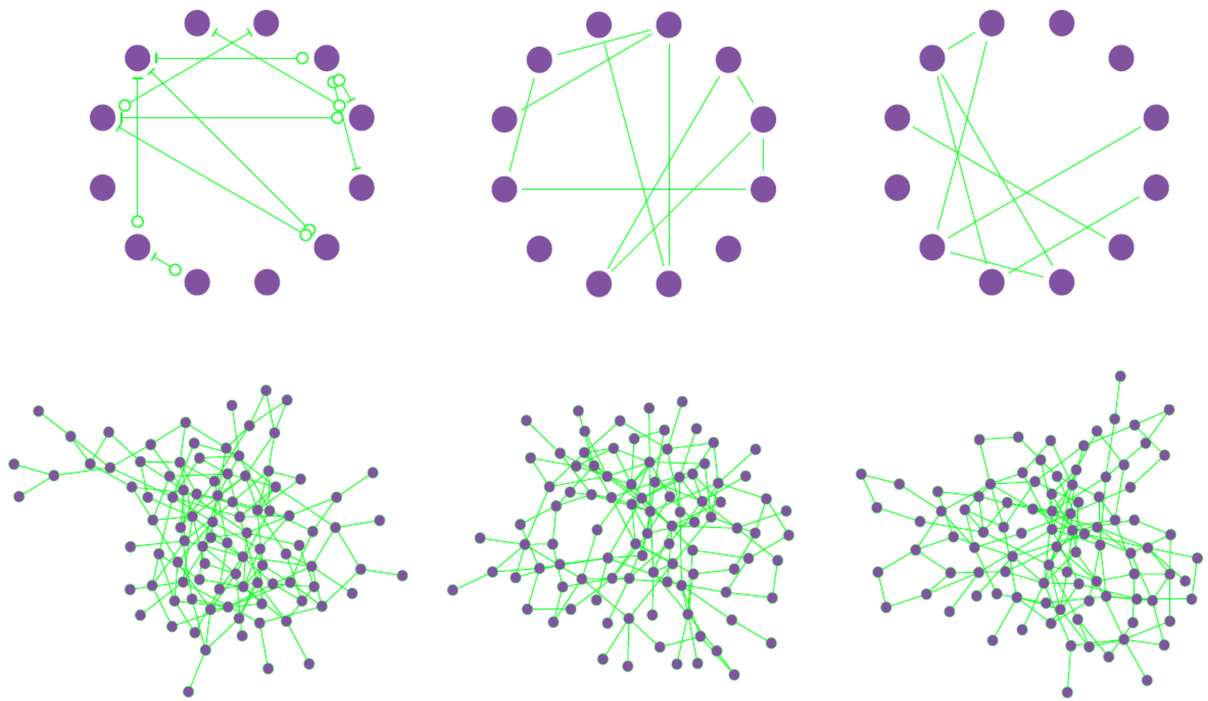
Hence $<L>$ is the product of the probability p that two nodes are connected and the number of pairs we attempt to connect, which is $L_{max} = N(N - 1)/2$

We obtain the average degree of a random network

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

Hence $<k>$ is the product of the probability p that two nodes are connected and (N-1), which is 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 varies between realizations. Its expected value is determined by N and p. If we increase p a random network becomes denser: The average number of links increase linearly from $<L> = 0$ to $L_{\max}$ and the average degree of a node increases from $<k> = 0$ to $<k> = N-1$.



Random Networks are Truly Random

Top Row

Three realizations of a random network generated with the same parameters p=1/6 and N=12. Despite the identical parameters, the networks not only look different, but they have a different number of links as well (L=10, 10, 8).

Bottom Row

Three realizations of a random network with p=0.03 and N=100. Several nodes have degree k=0, shown as isolated nodes at the bottom.

DEGREE DISTRIBUTION

In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links. These differences are captured by the degree distribution, p_k, which is the probability that a randomly chosen node has degree k. We derive p_k for a random network and discuss its properties.

BINOMIAL DISTRIBUTION

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

- 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

$$\binom{N-1}{k}.$$

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

- **BINOMIAL DISTRIBUTION**

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

- The shape of this distribution depends on the system size N and the probability p. The binomial distribution allows us to calculate the network's average degree <k>, recovering, as well as its second moment $\langle k^2 \rangle$ and variance σ_k.

- **POISSON DISTRIBUTION**

Most real networks are sparse, meaning that for them $\langle k \rangle \ll N$. In this limit the degree distribution is well approximated by the Poisson distribution

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

which is often called, the degree distribution of a random network.

The binomial and the Poisson distribution describe the same quantity, hence they have similar properties.

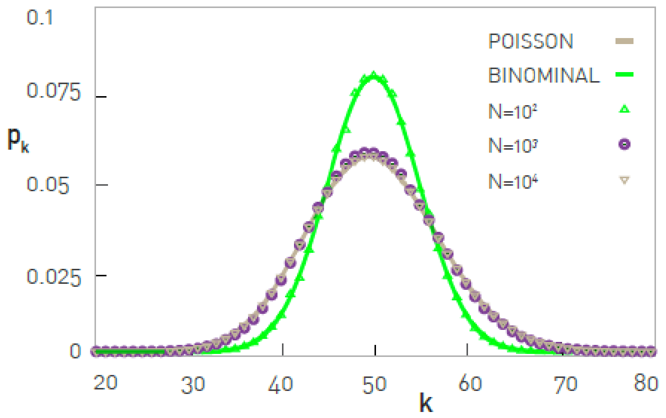
- Both distributions have a peak around <k>. If we increase p the network becomes denser, increasing <k> and moving the peak to the right.
- The width of the distribution (dispersion) is also controlled by p or <k>. The denser the network, the wider is the distribution, hence the larger are the differences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

When we use the Poisson form, we need to keep in mind that:

• The exact result for the degree distribution is the binomial form, thus the Poisson expression $p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$, represents only an approximation valid in the <k> ≪ N limit. As most networks of practical importance are sparse, this condition is typically satisfied.

• The advantage of the Poisson form is that key network characteristics, like <k>, <k^2> and σ_k , have a much simpler form), depending on a single parameter, <k>.

• The Poisson distribution in does not explicitly depend on the number of nodes N. Therefore, predicts that the degree distribution of networks of different sizes but the same average degree <k> are indistinguishable from each other:



In summary, while 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.

REAL NETWORKS ARE NOT POISSON

- As the degree of a node in a random network can vary between 0 and N-1, we must ask, how big are the differences between the node degrees in a particular realization of a random network?
- That is, can high degree nodes coexist with small degree nodes? We address 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 random society may not be as far fetched 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, prompting us to assume that $\langle k \rangle \approx 1{,}000$. Using the results obtained so far about random networks, we arrive to a number of intriguing conclusions about a random society of $N \simeq 7 \times 10^9$ of individuals

REAL NETWORKS ARE NOT POISSON

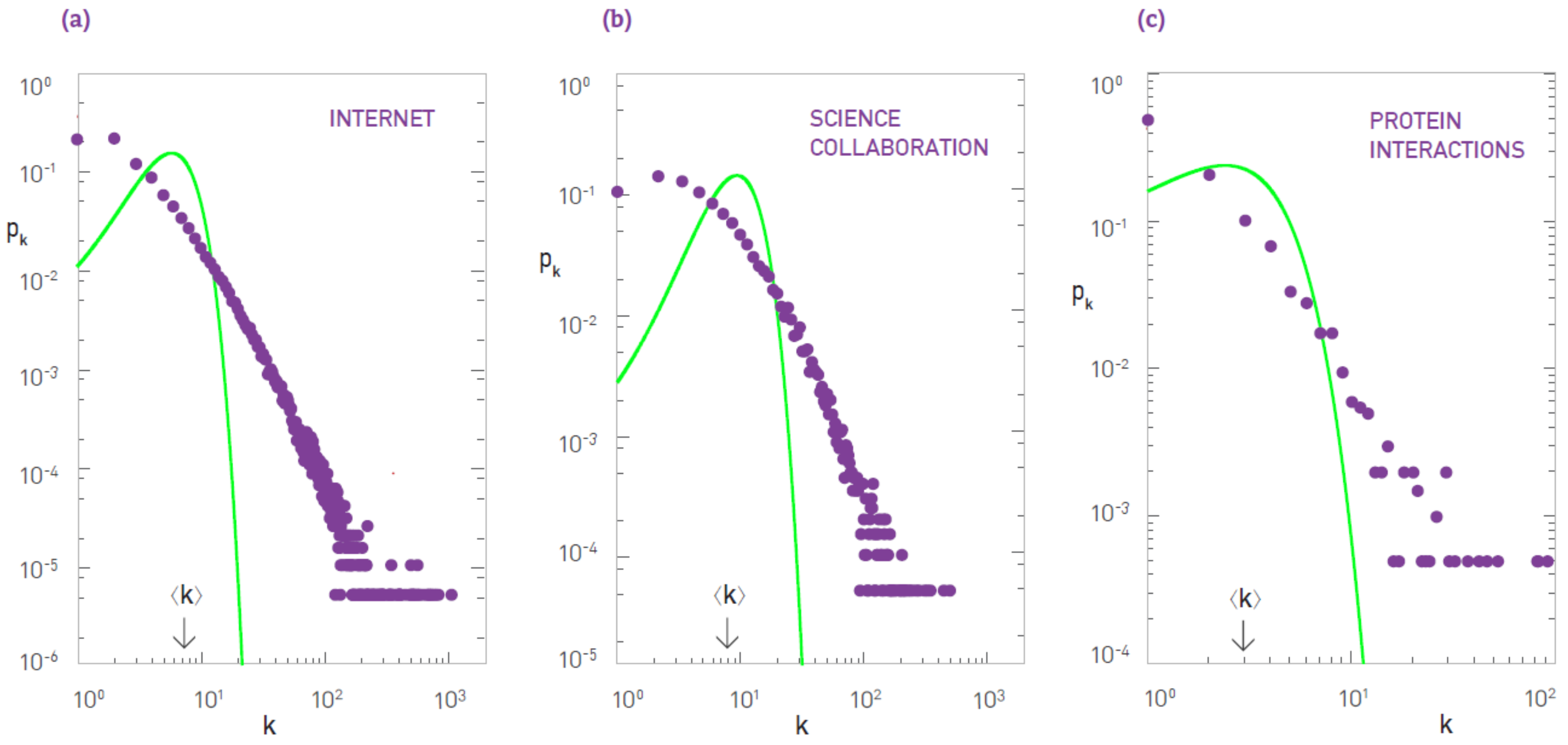
- The most connected individual (the largest degree node) in a random society is expected to have k_{max} = 1,185 acquaintances.
- The degree of the least connected individual is k_{min} = 816, not that different from k_{max} or $<k>$.
- The dispersion of a random network is $\sigma_k = <k>^{1/2}$, which for $<k>$ = 1,000 is σ_k = 31.62. This means that the number of friends a typical individual has is in the $<k> \pm \sigma_k$ range, or between 968 and 1,032, a rather narrow window.

Taken together, in a random society all individuals are expected to have a comparable number of friends. Hence if people are randomly connected to each other, we lack outliers: There are no highly popular individuals, and no one is left behind, having only a few friends. This suprising conclusion is a consequence of an important property of random networks: *in a large random network the degree of most nodes is in the narrow vicinity of* $<k>$*.*

REAL NETWORKS ARE NOT POISSON

- This prediction blatantly conflicts with reality. Indeed, there is extensive evidence of individuals who have considerably more than 1,185 acquaintances. For example, US president Franklin Delano Roosevelt's appointment book has about 22,000 names, individuals he met personally. 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. To understand the origin of these discrepancies we must compare the degree distribution of real and random networks. We show the degree distribution of three real networks, together with the corresponding Poisson fit.

REAL NETWORKS ARE NOT POISSON (2)



The degree distribution of the (a) Internet, (b) science collaboration network, and (c) protein interaction network. The green line corresponds to the Poisson prediction, obtained by measuring <*k*> for the real network and then plotting.

The significant deviation between the data and the Poisson fit indicates that the random network model

underestimates the size and the frequency of the high degree nodes, as well as the number of low degree nodes. Instead the random network model predicts a larger number of nodes in the vicinity of <*k*> than seen in real networks.

THE EVOLUTION OF A RANDOM NETWORK

- The cocktail party we encountered at the beginning of this chapter captures a dynamical process: Starting with N isolated nodes, the links are added gradually through random encounters between the guests. This corresponds to a gradual increase of p, with striking consequences on the network topology. To quantify this process, we first inspect how the size of the largest connected cluster within the network, N_G, varies with $<k>$. Two extreme cases are easy to understand:

- For $p = 0$ we have $<k> = 0$, hence all nodes are isolated. Therefore the largest component has size $N_G = 1$ and $N_G/N \rightarrow 0$ for large N.

- For $p = 1$ we have $<k> = N\text{-}1$, hence the network is a complete graph and all nodes belong to a single component. Therefore $N_G = N$ and $N_G/N = 1$.

THE EVOLUTION OF A RANDOM NETWORK (2)

- One would expect that the largest component grows gradually from $N_G = 1$ to $N_G = N$ if $<k>$ increases from 0 to N-1. Yet, this is not the case: N_G/N remains zero for small $<k>$, indicating the lack of a large cluster. Once $<k>$ exceeds a critical value, N_G/N increases, signaling the rapid emergence of a large cluster that we call the *giant component*.
- Erdős and Renyi in their classical 1959 paper predicted that the condition for the emergence of the giant component is

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

- In other words, we have a giant component if and only if each node has on average more than one link.
- The fact that we need at least one link per node to observe 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.

THE EVOLUTION OF A RANDOM NETWORK (3)

- We can express <k> in terms of *p* using, obtaining

$$p_c = \frac{1}{N-1} \approx \frac{1}{N},$$

- Therefore the larger a network, the smaller *p* is sufficient for the giant component.
- The emergence of the giant component is only one of the transitions characterizing a random network as we change <*k*>. We can distinguish four topologically distinct regimes, each with its unique characteristics:

THE EVOLUTION OF A RANDOM NETWORK (4)

- **Subcritical Regime**: **$0 < \langle k \rangle < 1$** (p <1/N)
- The critical point separates the regime where there is not yet a giant component ($\langle k \rangle < 1$) from the regime where there is one ($\langle k \rangle > 1$). At this point the relative size of the largest component is still zero. Indeed, the size of the largest component is $N_G \sim N^{2/3}$. Consequently N_G grows much slower than the network's size, so its relative size decreases as $N_G/N \sim N^{-1/3}$ in the $N \to \infty$ limit.
- Note, however, that in absolute terms there is a significant jump in the size of the largest component at $\langle k \rangle = 1$. For example, for a random network with $N = 7 \times 10^9$ nodes, comparable to the globe's social network, for $\langle k \rangle < 1$ the largest cluster is of the order of $N_G \simeq \ln N = \ln (7 \times 10^9) \simeq 22.7$. In contrast at $\langle k \rangle = 1$ 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 and at the critical point the largest component contains only a vanishing fraction of the total number of nodes in the network.
- In summary, at the critical point most nodes are located in numerous small components. The power law form indicates that components of rather different sizes coexist. These numerous small components are mainly trees, while the giant component

$$p_c = \frac{1}{N-1} \approx \frac{1}{N},$$

THE EVOLUTION OF A RANDOM NETWORK (5)

- **Supercritical Regime**: $<k> > 1$ ($p > 1/N$)

- This regime has the most relevance to real systems, as for the first time we have a giant component that looks like a network. The giant component contains a finite fraction of the nodes. The further we move from the critical point, a larger fraction of nodes will belong to it.

- In summary in the supercritical regime numerous isolated components coexist with the giant component.

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

THE EVOLUTION OF A RANDOM NETWORK (6)

- **Connected Regime: ‹k› > lnN** (p > lnN / N)
- For sufficiently large p the giant component absorbs all nodes and components, hence $N_G \simeq N$. In the absence of isolated nodes the network becomes connected. The average degree at which this happens depends on N as $\langle k \rangle = \ln N$.
- Note that when we enter the connected regime the network is still relatively sparse, as $\ln N / N \to 0$ for large N. The network turns into a complete graph only at $\langle k \rangle = N - 1$.
- In summary, the random network model predicts that the emergence of a network is not a smooth, gradual process: The isolated nodes and tiny components observed for small <k> collapse into a giant componentthrough a phase transition.
- The discussion follows an empirical perspective, fruitful if we wish to compare a random network to real systems. A different perspective, with its own rich behavior, is offered by the mathematical literature.

REAL NETWORKS ARE SUPERCRITICAL

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

- 1) Once the average degree exceeds <k> = 1, a giant component should emerge that contains a finite fraction of all nodes. Hence only for <k> > 1 the nodes organize themselves into a recognizable network.
- 2) For <k> > lnN all components are absorbed by the giant component, resulting in a single connected network.

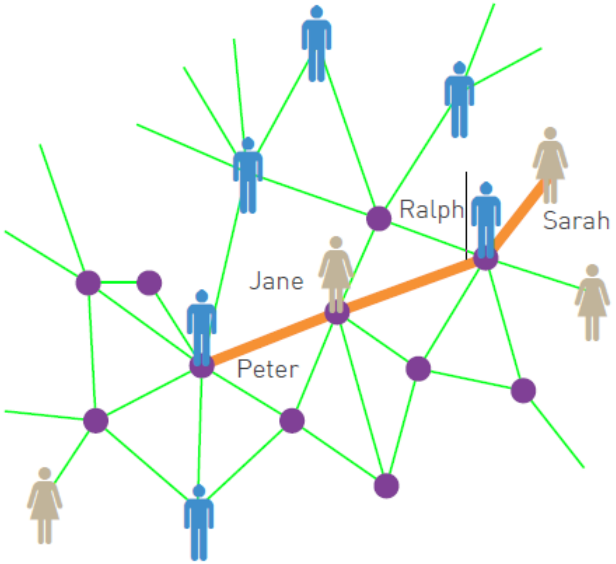
Do real networks satisfy the criteria for the existence of a giant component, i.e. <k> > 1? And will this giant component contain all nodes for <k> > lnN, or will we continue to see some disconnected nodes and components?

REAL NETWORKS ARE SUPERCRITICAL

We find that most real networks are in the supercritical regime.

Therefore these networks are expected to have a giant component, which is in agreement with the observations. Yet, this giant component should coexist with many disconnected components, a prediction that fails for several real networks. Note that these predictions should be valid only if real networks are accurately described by the Erdős-Rényi model, i.e. if real networks are random. In the followings, as we learn more about the structure of real networks, we will understand why real networks can stay connected despite failing the $k > \ln N$ criteria.

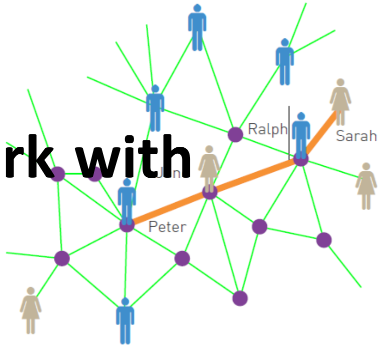
SMALL WORLDS



- The *small world phenomenon*, 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 (Figure on the right hand side) 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 states, however, that even individuals who are on the opposite side of the globe can be connected to us via a few acquaintances.

In the language of network science the small world phenomenon implies that the distance between two randomly chosen nodes in a network is 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?

SMALL WORLDS (2)



Both questions are answered by a simple calculation. Consider a random **network with average degree <*k*>. A node in this network has on average:**

<*k*> nodes at distance one (*d*=1).

<*k*>2 nodes at distance two (*d*=2).

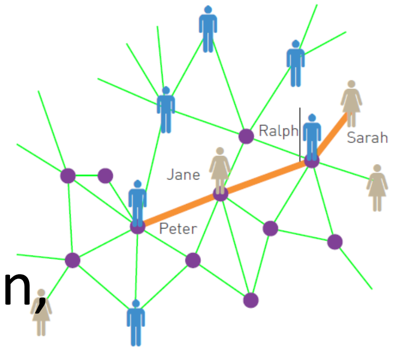
<*k*>3 nodes at distance three (*d* =3).

...

<*k*>d nodes at distance *d*.

For example, if <*k*> ≈ 1,000, which is the estimated number of acquaintences an individual has, we expect 106 individuals at distance two and about a billion, i.e. almost the whole earth’s population, at distance three from us.

SMALL WORLDS (3)



In summary the small world property has not only ignited the public's imagination, but plays an important role in network science as well. 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.

In the coming chapters we will see that in real networks we encounter systematic deviations, forcing us to replace it with more accurate predictions. Yet the intuition offered by the random network model on the origin of the small world phenomenon remains valid.

CLUSTERING COEFFICIENT

- The degree of a node contains no information about the relationship between a node's neighbors. Do they all know each other, or are they perhaps isolated from each other? The answer is provided by the local clustering coefficient C_i, that measures the density of links in node *i*'s immediate neighborhood: $C_i = 0$ means that there are no links between *i*'s neighbors $C_i = 1$ implies that each of the *i*'s neighbors link to each other.

- 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

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

CLUSTERING COEFFICIENT (2)

- Thus the local clustering coefficient of a random network is

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

This Equation makes two predictions:

(1) For fixed $<k>$, the larger the network, the smaller is a node’s clustering coefficient. Consequently a node's local clustering coefficient C_i is expected to decrease as $1/N$. Note that the network's average clustering coefficient, $<C>$ also follows this formula.

(2) The local clustering coefficient of a node is independent of the node’s degree.

In summary, we find that the random network model does not capture the clustering of real networks. Instead real networks have a much higher clustering coefficient than expected for a random network of similar N and L. An extension of the random network model proposed by Watts and Strogatz addresses the coexistence of high $<C>$ and the small world property. It fails to explain, however, why high-degree nodes have a smaller clustering coefficient than low-degree nodes

REAL NETWORKS ARE NOT RANDOM

Since its introduction in 1959 the random network model has dominated mathematical approaches to complex networks. The model suggests that the random-looking networks observed in complex systems should be described as purely 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.

As the interactions between our proteins are governed by the strict laws of biochemistry, for the cell to function its chemical architecture cannot be random. Similarly, in a random society an American student would be as 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.

REAL NETWORKS ARE NOT RANDOM (2)

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. We can do this, taking advantage of the fact that random network theory makes a number of quantitative predictions:

- **Distribution**

A random network has a binomial distribution, well approximated by a Poisson distribution in the $k \ll N$ limit. Yet, as shown, the Poisson distribution fails to capture the degree distribution of real networks. 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, however, do not satisfy the $\langle k \rangle > \ln N$ condition, implying that they should be broken into isolated clusters. Some networks are indeed fragmented, most are not.

REAL NETWORKS ARE NOT RANDOM (3)

- **Average Path Length**

Random network theory predicts that the average path length follows, a prediction that offers a reasonable approximation for the observed path lengths. Hence the random network model can account for the emergence of small world phenomena.

- **Clustering Coefficient**

In a random network the local clustering coefficient is independent of the node's degree and <*C*> depends on the system size as 1/*N*. In contrast, measurements indicate that for real networks *C(k)* decreases with the node degrees and is largely independent of the system size.

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 networks.

THE SCALE-FREE PROPERTY

- **World Wide Web:**

  a network whose nodes are documents and the links are the uniform resource locators (URLs) that allow us to "surf" with a click from one web document to the other.

- **Importance of the World Wide Web:**

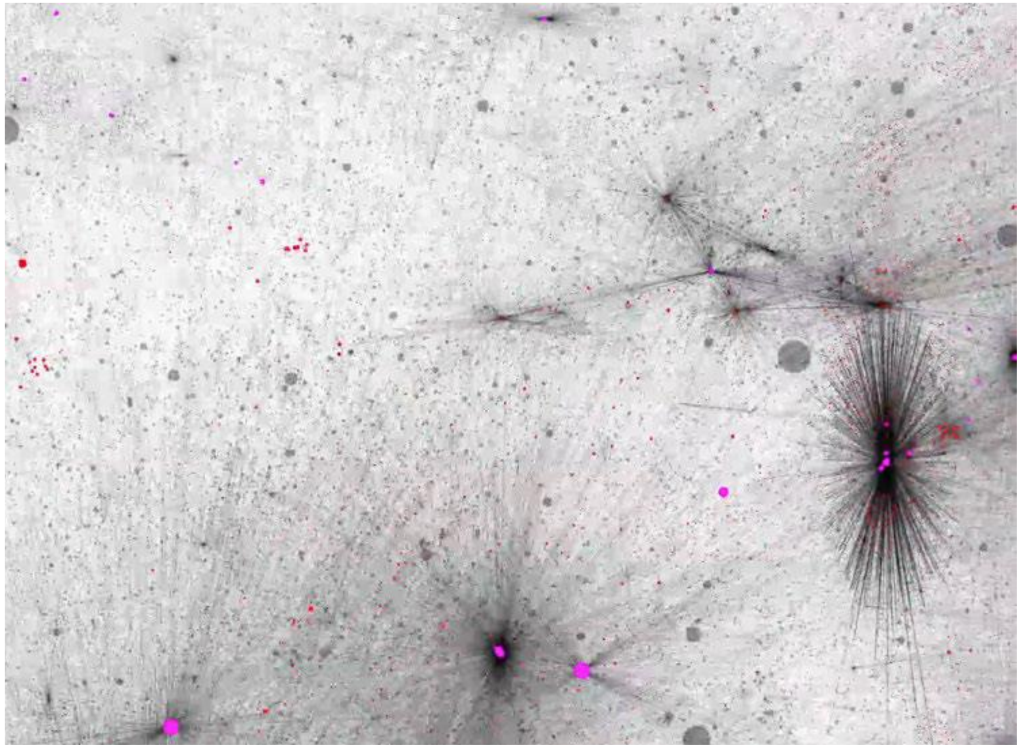
  WWW played in the development of network theory: it facilitated the discovery of a number of fundamental network characteristics and became a standard testbed for most network measures.

CRAWLER

- A software called a crawler can be used to map out the Web's wiring diagram.
- A crawler can start from any web document, identifying the links (URLs) on it. Next it downloads the documents these links point to and identifies the links on these documents, and so on. This process iteratively returns a local map of the Web. Search engines like Google or Bing operate crawlers to find and index new documents and to maintain a detailed map of the WWW.
- The purpose of the map was to compare the properties of the Web graph to the random network model. Indeed, **in 1998** there were reasons to **believe** that **the WWW** could be well **approximated by** a **random network**.

Differences between real and random network

- **In a random network** highly connected nodes, or **hubs, are** effectively **forbidden**.
- In contrast, **in the Figure** numerous **small-degree nodes coexist with** a few **hubs**, nodes with an exceptionally large number of links.



DEGREE DISTRIBUTION OF REAL NETWORKS

- H**ubs** are not unique to the Web, but we encounter them **in most real networks**. They represent a signature of a deeper organizing principle that we call the scale-free property.
- No matter what network property we are interested in, from communities to spreading processes, it must be inspected in the light of the network's degree distribution.

POWER LAWS AND SCALE-FREE NETWORKS

- If the WWW were to be a random network, the degrees of the Web documents should follow a Poisson distribution.

- However, the Poisson form offers a poor fit for the WWW's degree distribution. Instead on a log-log scale the data points form an approximate straight line, suggesting that the degree distribution of the WWW is well approximated with

$$p_k \sim k^{-\gamma}$$

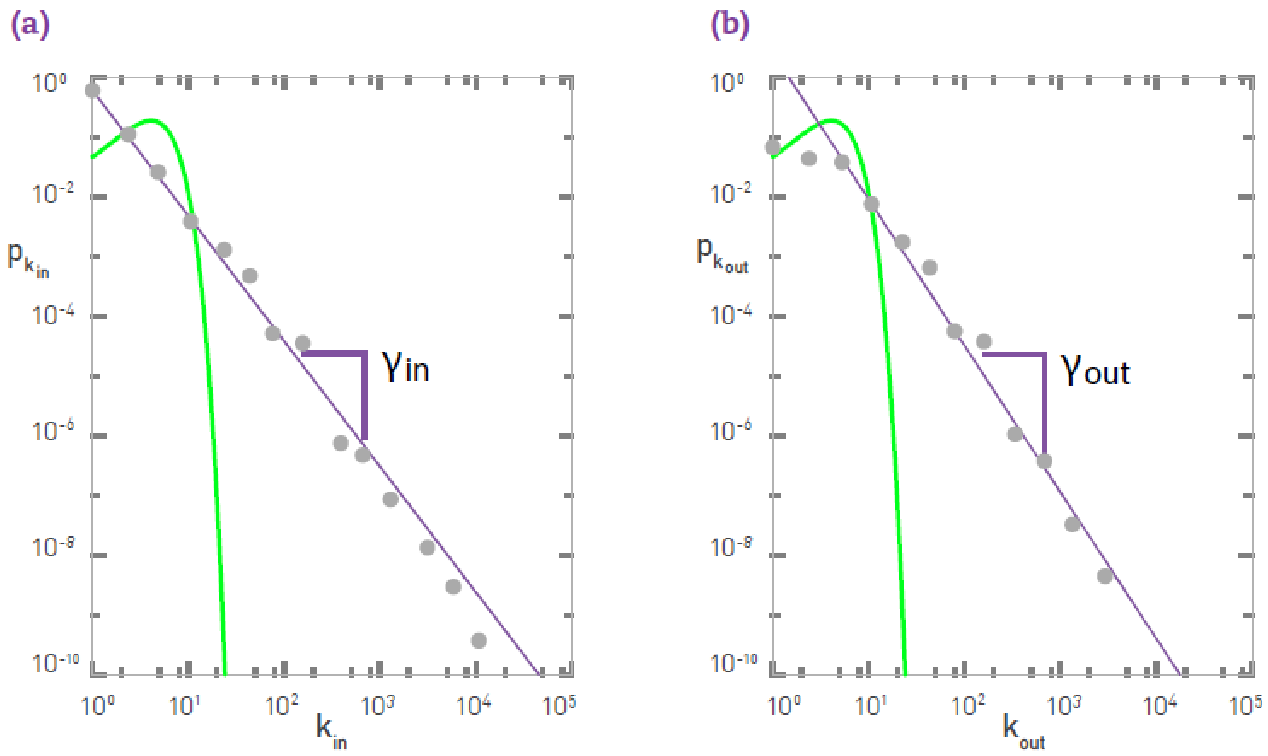
- The **equation is called** a **power law distribution** and the exponent **γ** is its **degree exponent**.

Logarithm plot

- If we take a logarithm of the equation, we obtain

$$\log\, p_k \sim -\gamma \cdot \log\, k$$

- **log** p_k is expected to depend linearly on **log** k, the slope of this line being the degree exponent **γ**.



in-degree and out-degree

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

$$\boldsymbol{p_{k_{in}} \sim k^{-\gamma_{in}}} \text{ and } \boldsymbol{p_{k_{out}} \sim k^{-\gamma_{out}}}$$

- where $\boldsymbol{\gamma_{in}}$ and $\boldsymbol{\gamma_{out}}$ are the degree exponents for the in- and out-degrees, respectively.
- In general $\boldsymbol{\gamma_{in}}$ can differ from $\boldsymbol{\gamma_{out}}$. For example, we got $\boldsymbol{\gamma_{in} \approx 2.1}$ and $\boldsymbol{\gamma_{out} \approx 2.45}$.

Definition of the scale-freee network

- The **empirical results document the existence of** a **network** whose degree distribution is quite **different from** the Poisson distribution characterizing **random networks**.
- We will call such networks *scale-free*, defined as:

A scale-free network is a network whose degree distribution follows a power law.

Discrete and the continuum formalisms of power-law distribution

Discrete Formalism (1)

As node degrees are positive integers, $\boldsymbol{k}$ **= 0, 1, 2, ...,** the discrete formalism provides the probability p_k that a node has exactly k links

$$\boldsymbol{p_k = C \cdot k^{-\gamma}}$$

The constant C is determined by the normalization condition

$$\boldsymbol{\sum_{k=1}^{\infty} p_k = 1}$$

$$\boldsymbol{C \cdot \sum_{k=1}^{\infty} k^{-\gamma} = 1}$$

Discrete Formalism (2)

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\xi(\gamma)}$$

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

$$p_k = \frac{k^{-\gamma}}{\xi(\gamma)}$$

Continuum formalism

In analytical calculations it is often convenient to assume that the degrees can have any positive real value. In this case we write the power-law distribution as

$$\boldsymbol{p(k) = C \cdot k^{-\gamma}}$$

Using the normalization condition

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

$$\boldsymbol{C = \frac{1}{\int_{k_{min}}^{\infty} k^{-\gamma} dk} = (\gamma - 1){k_{min}}^{\gamma-1}}$$

Continuum formalism (2)

Therefore in the continuum formalism the degree distribution has the form

$\boldsymbol{p(k) = (\gamma - 1) \cdot {k_{min}}^{\gamma-1} \cdot k^{-\gamma}}$, where $\boldsymbol{k_{min}}$ is the smallest degree for which the power law holds:

$$\boldsymbol{p_k = \frac{k^{-\gamma}}{\xi(\gamma)}}$$

The physical interpretation of the continuum formalism:

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

is the probavility that a randomly chosen node has degree between $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$.

In summary, networks whose degree distribution follows a power law are called scale-free networks. If a network is directed, the scale-free property applies separately to the in- and out-degrees. To mathematically study the properties of scale-free networks, we can use either the discrete or the continuum formalism. The scale-free property is independent of the formalism we use.

ELTE | FACULTY OF INFORMATICS

Network Science

The scale-free property

Dr. Tamás Orosz
Ph.D., habil.

Lecture 5, 10/10/2024

THE SCALE-FREE PROPERTY

- **World Wide Web:**

  a network whose nodes are documents and the links are the uniform resource locators (URLs) that allow us to “surf” with a click from one web document to the other.

- **Importance of the World Wide Web:**

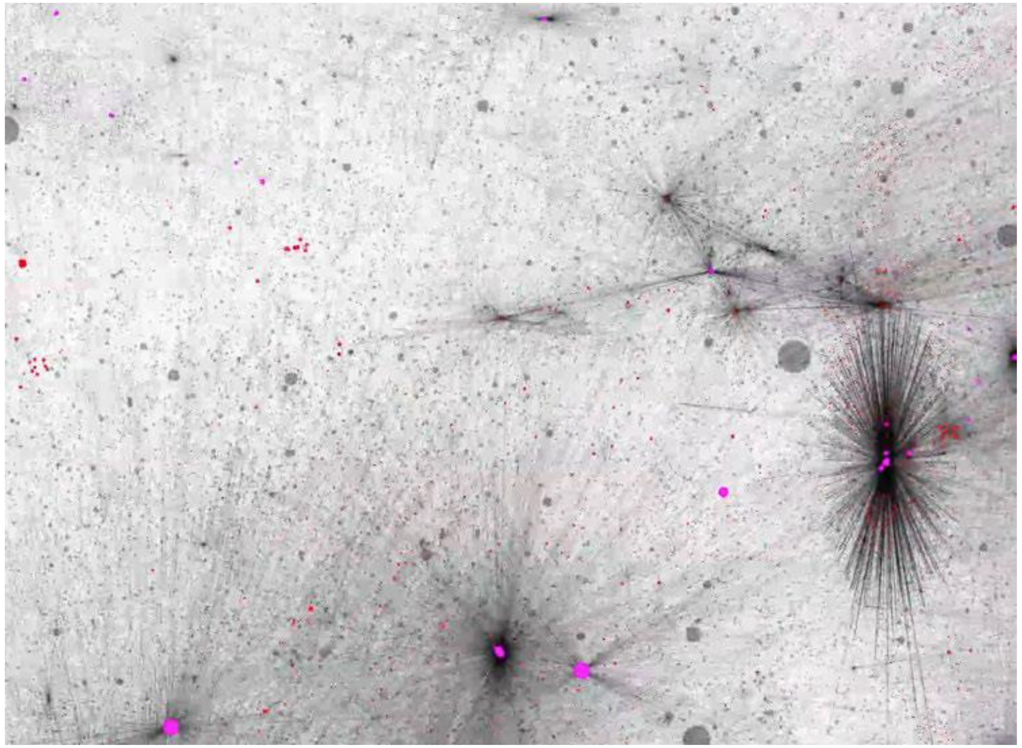
  WWW played in the development of network theory: it facilitated the discovery of a number of fundamental network characteristics and became a standard testbed for most network measures.

CRAWLER

- A software called a crawler can be used to map out the Web's wiring diagram.
- A crawler can start from any web document, identifying the links (URLs) on it. Next it downloads the documents these links point to and identifies the links on these documents, and so on. This process iteratively returns a local map of the Web. Search engines like Google or Bing operate crawlers to find and index new documents and to maintain a detailed map of the WWW.
- The purpose of the map was to compare the properties of the Web graph to the random network model. Indeed, **in 1998** there were reasons to **believe** that **the WWW** could be well **approximated by** a **random network**.

Differences between real and random network

- **In a random network** highly connected nodes, or **hubs, are** effectively **forbidden**.
- In contrast, **in the Figure** numerous **small-degree nodes coexist with** a few **hubs**, nodes with an exceptionally large number of links.



DEGREE DISTRIBUTION OF REAL NETWORKS

- H**ubs** are not unique to the Web, but we encounter them **in most real networks**. They represent a signature of a deeper organizing principle that we call the scale-free property.
- No matter what network property we are interested in, from communities to spreading processes, it must be inspected in the light of the network's degree distribution.

POWER LAWS AND SCALE-FREE NETWORKS

- If the WWW were to be a random network, the degrees of the Web documents should follow a Poisson distribution.
- However, the Poisson form offers a poor fit for the WWW's degree distribution. Instead on a log-log scale the data points form an approximate straight line, suggesting that the degree distribution of the WWW is well approximated with

$$p_k \sim k^{-\gamma}$$

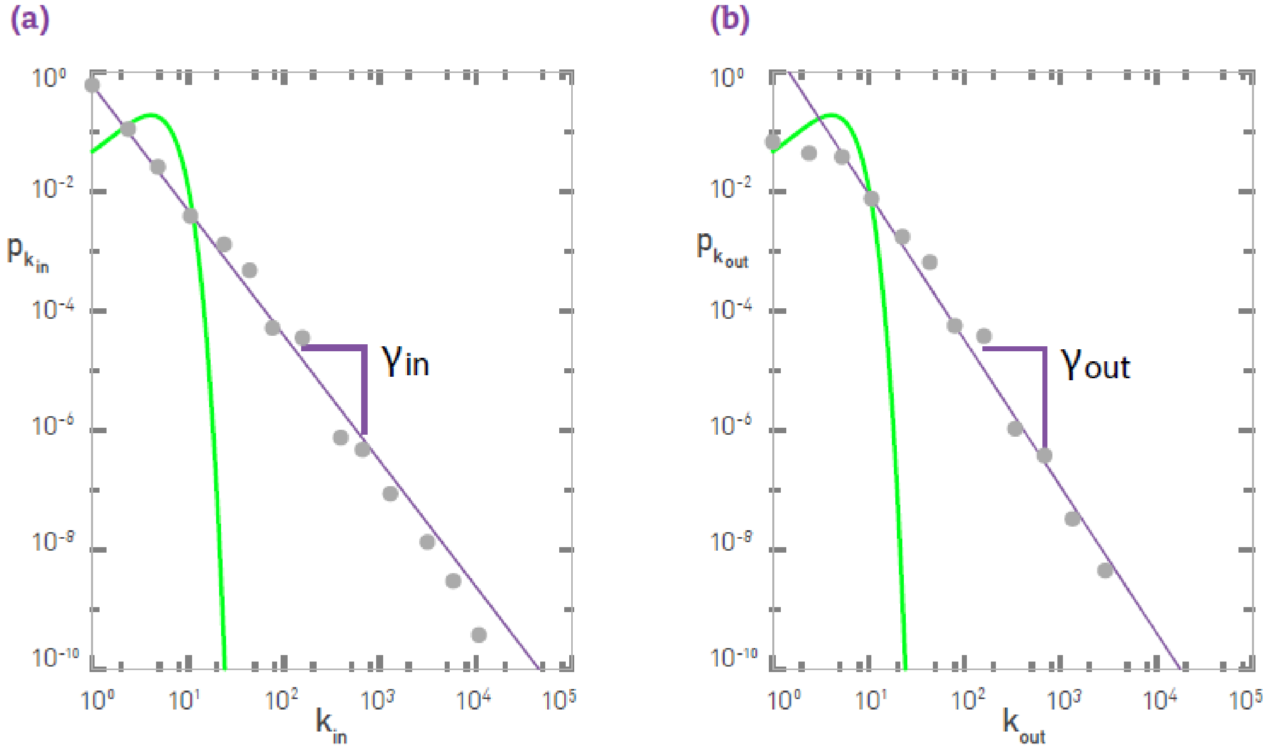
- The **equation is called** a **power law distribution** and the exponent **γ** is its **degree exponent**.

Logarithm plot

- If we take a logarithm of the equation, we obtain

$$\log p_k \sim -\gamma \cdot \log k$$

- **log** $\boldsymbol{p_k}$ is expected to depend linearly on **log** $\boldsymbol{k}$, the slope of this line being the degree exponent **γ**.



in-degree and out-degree

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

$$\boldsymbol{p_{k_{in}} \sim k^{-\gamma_{in}}} \textbf{ and } \boldsymbol{p_{k_{out}} \sim k^{-\gamma_{out}}}$$

- where $\boldsymbol{\gamma_{in}}$ and $\boldsymbol{\gamma_{out}}$ are the degree exponents for the in- and out-degrees, respectively.
- In general $\boldsymbol{\gamma_{in}}$ can differ from $\boldsymbol{\gamma_{out}}$. For example, we got $\boldsymbol{\gamma_{in} \approx 2.1}$ and $\boldsymbol{\gamma_{out} \approx 2.45}$.

Definition of the scale-freee network

- The **empirical results document the existence of** a **network** whose degree distribution is quite **different from** the Poisson distribution characterizing **random networks**.

- We will call such networks *scale-free*, defined as:

A scale-free network is a network whose degree distribution follows a power law.

Discrete and the continuum formalisms of power-law distribution

Discrete Formalism (1)

As node degrees are positive integers, $\boldsymbol{k}$ **= 0, 1, 2, ...,** the discrete formalism provides the probability p_k that a node has exactly k links

$$\boldsymbol{p_k = C \cdot k^{-\gamma}}$$

The constant C is determined by the normalization condition

$$\sum_{k=1}^{\infty} p_k = 1$$

$$C \cdot \sum_{k=1}^{\infty} k^{-\gamma} = 1$$

Discrete Formalism (2)

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\xi(\gamma)}$$

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

$$p_k = \frac{k^{-\gamma}}{\xi(\gamma)}$$

Continuum formalism

In analytical calculations it is often convenient to assume that the degrees can have any positive real value. In this case we write the power-law distribution as

$$\boldsymbol{p(k) = C \cdot k^{-\gamma}}$$

Using the normalization condition

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

$$\boldsymbol{C = \frac{1}{\int_{k_{min}}^{\infty} k^{-\gamma} dk} = (\gamma - 1){k_{min}}^{\gamma-1}}$$

Continuum formalism (2)

Therefore in the continuum formalism the degree distribution has the form

$\boldsymbol{p(k) = (\gamma - 1) \cdot k_{min}{}^{\gamma-1} \cdot k^{-\gamma}}$, where $\boldsymbol{k_{min}}$ is the smallest degree for which the power law holds:

$$\boldsymbol{p_k = \frac{k^{-\gamma}}{\xi(\gamma)}}$$

The physical interpretation of the continuum formalism:

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

is the probavility that a randomly chosen node has degree between $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$.

In summary, networks whose degree distribution follows a power law are called scale-free networks. If a network is directed, the scale-free property applies separately to the in- and out-degrees. To mathematically study the properties of scale-free networks, we can use either the discrete or the continuum formalism. The scale-free property is independent of the formalism we use.

ELTE | FACULTY OF INFORMATICS

Degree Colleration

Dr. Tamás Orosz
Ph.D., habil.

Lecture 7, 11/06/2024

Introduction to Degree Correlation in Network Science

- Definition and importance of degree correlation in network science.
- Relationship between nodes based on degree connectivity.
- Degree correlation as a measure of network structure.
- Importance of understanding assortative and disassortative mixing in networks.
- Brief outline of topics: assortativity, measuring correlations, structural cutoffs, and impact on real networks

Degree Correlation Types

- Assortative Networks: Hubs link to other hubs; small-degree nodes link to small-degree nodes.
- Disassortative Networks: Hubs avoid linking to each other, preferring small-degree nodes.
- Neutral Networks: No specific tendency for nodes to connect by degree.
- Examples of each type in real-world networks (e.g., social networks, biological networks).

Measuring Degree Correlations

- Explanation of the Degree Correlation Matrix (e_{ij}).
- The role of matrix analysis in identifying correlation types.
- How the degree correlation function (k_{inn}) measures connectivity tendencies.
- Use of correlation coefficients to quantify assortative or disassortative trends.
- Importance of visual and mathematical tools for correlation measurement.

Assortativity Coefficient

- Definition and mathematical formula of the assortativity coefficient (r).
- Range of r values: Positive for assortative, negative for disassortative.
- Assortativity coefficient as a summary measure for network type.
- Examples of r values for common networks: social (positive), biological (negative).
- Implications of high or low r values in network analysis.

Assortative Networks in Depth

- Characteristics of assortative networks.
- Examples: Social networks where hubs link to other hubs.
- Assortative mixing leads to increased robustness in network structure.
- Impact on information or disease spread within assortative networks.
- Contrast with disassortative structures and their vulnerability.

Disassortative Networks in Depth

- Characteristics of disassortative networks.
- Examples: Biological and technological networks where hubs connect to low-degree nodes.
- Hub and spoke model in disassortative networks.
- Consequences on network vulnerability and resilience.
- How disassortativity affects information and resource flow.

Neutral Networks

- Definition of neutral networks without correlation preference.
- Role of randomness in node connections.
- Examples: Certain power grids and infrastructure networks.
- Importance of neutral networks in theoretical models.
- Analysis methods for neutral networks and implications for resilience.

Degree Correlation Matrix (e_{ij})

- **Purpose of the degree correlation matrix in identifying patterns.**
- **Use of e_{ij} to predict connectivity likelihood between degrees.**
- **Visualization techniques for matrix interpretation.**
- **Application of matrix analysis in large network data.**
- **Advantages and limitations of the degree correlation matrix.**

Degree Correlation Function (k_{inn})

- Description of the degree correlation function and its calculation.
- How k_{inn} varies with node degree in assortative vs. disassortative networks.
- Implications of k_{inn} for predicting network behavior.
- Visualization of k_{inn} trends in assortative and disassortative networks.
- Practical use cases of k_{inn} in network science.

Real-World Examples of Degree Correlations

- Real-world networks displaying assortative mixing (e.g., scientific collaborations).
- Networks with disassortative mixing (e.g., biological networks).
- Neutral networks in practical applications (e.g., power grids).
- Degree correlation as a predictor of network behavior in various domains.
- Significance of these patterns for network stability and resilience

Assortative Mixing Patterns

- **Assortative mixing in networks: Hubs prefer connecting to hubs.**
- **Impact on network cohesion and robustness.**
- **Observed in social networks and collaboration networks.**
- **Allows clusters or communities to form naturally.**
- **Enhances network resilience to random failures.**

Disassortative Mixing Patterns

- Disassortative mixing: Hubs avoid connecting to each other.
- Leads to a "hub and spoke" network structure.
- Common in biological and technological networks.
- Helps distribute load across the network efficiently.
- Increased vulnerability to targeted hub removal.

Neutral Mixing Patterns

- **No preference in connection patterns by degree.**
- **Network structure remains random and unbiased.**
- **Observed in infrastructure networks, e.g., power grids.**
- **Neutral networks offer a baseline for theoretical analysis.**
- **Insights into the impact of non-assortative configurations.**

Structural Cutoffs in Networks

- Concept of structural cutoff in network theory.
- When hub connectivity is limited by network design.
- Structural cutoffs are more common in simple networks.
- Effect on assortative and disassortative mixing.
- Practical examples: actor networks, certain biological networks.

Degree Correlation and Structural Constraints

- Constraints in network topology impact correlation.
- Structural cutoffs limit possible hub connections.
- Degree correlation may be influenced by physical or design limits.
- Consequences for network resilience and vulnerability.
- Strategies to analyze networks with structural constraints.

Correlation in Social Networks

- High degree of assortativity in social networks.
- Hubs (high-degree nodes) form closely knit communities.
- Examples: friendship networks, professional collaborations.
- Implications for information spread and community resilience.
- Social networks as a model for robust communication systems.

Correlation in Biological Networks

- Biological networks often display disassortative patterns.
- Hubs connect to low-degree nodes, avoiding other hubs.
- This structure helps manage biological stress and load.
- Examples: protein interaction networks, metabolic networks.
- Disassortative structure aids in stability under targeted disruptions.

Correlation in Technological Networks

- Technological networks often disassortative by design.
- Hubs maintain connectivity with peripheral nodes, not other hubs.
- Example: the Internet, power grids.
- Load balancing and fault tolerance in disassortative tech networks.
- Role in protecting critical hubs in network infrastructure.

Measuring Assortativity with Pearson's Coefficient

- Pearson's coefficient (r) quantifies assortative or disassortative tendencies.
- Positive values indicate assortative networks, negative for disassortative.
- r = 0 implies a neutral network without degree correlation.
- Helps classify networks and predict structural behavior.
- Important for comparing different network types.

Degree Correlation Function

- Degree correlation function measures average neighbor degree.
- knn(k)k_{nn}(k)knn(k) increases with k in assortative, decreases in disassortative networks.
- Indicator of clustering tendencies and mixing patterns.
- Analysis of network's assortative/disassortative nature.
- Practical applications: epidemiology, traffic, and network robustness.

Assortative Networks and Community Formation

- Assortative networks facilitate natural clustering.
- Clusters form around hubs, strengthening local connections.
- Enhances robustness through tightly-knit community structures.
- Examples: scientific collaboration networks, school social networks.
- Role in promoting resilience and knowledge-sharing.

Disassortative Networks and Hierarchical Structure

- Disassortative networks often have a hierarchical structure.
- Central hubs link to numerous low-degree nodes.
- Creates a hub-and-spoke topology ideal for control and distribution.
- Typical in transportation and power networks.
- Vulnerability in case of targeted attacks on hubs.

Implications of Neutral Mixing on Network Stability

- Neutral mixing yields random connectivity, lacking degree bias.
- Stabilizes the network without preferential connection patterns.
- Less prone to clustering or hierarchical tendencies.
- Neutral networks are typical in randomized models and theoretical studies.
- Useful for understanding fundamental network behaviors.

Visualizing Degree Correlations

- Graphs and heat maps used to display degree correlations.
- Visualization aids in identifying assortative or disassortative tendencies.
- Degree correlation matrix helps map connection likelihoods.
- Clusters on heat maps reveal assortative connections.
- Visualization as a critical tool in network analysis.

Degree Correlation Matrix (e_{ij}) Analysis

- Degree correlation matrix represents link probabilities.
- e_{ij} values show probability of connection between nodes of degrees i and j.
- High values along the diagonal indicate assortative mixing.
- Off-diagonal high values suggest disassortative mixing.
- Matrix insights aid in understanding complex network structures

Correlation Matrix in Social Networks

- In assortative social networks, hubs tend to link to each other.
- Degree correlation matrix shows strong diagonal patterns.
- Facilitates community formation and resilient social structures.
- Practical use in studying collaborative and friendship networks.
- Enables targeted network intervention strategies.

Correlation Matrix in Biological Networks

- Disassortative patterns in biological networks show off-diagonal clustering.
- Hubs avoid each other, connecting to low-degree nodes.
- Matrix reveals distribution and interaction patterns of proteins or cells.
- Highlights resilience against random node failures.
- Useful for predicting biological network stability.

Impact of Assortativity on Information Spread

- Assortative networks enhance rapid information spread.
- Clusters and hubs support efficient communication within groups.
- Resilient to information isolation or disruption.
- Examples in social media and professional collaboration.
- Important for strategies in public health and marketing.

Impact of Disassortativity on Containment and Control

- Disassortative networks help control spread by isolating hubs.
- Hubs with few inter-hub connections slow down transmission.
- Useful in managing epidemics or controlling information flow.
- Example: Limiting high-risk interaction in biological networks.
- Important for containment strategies and infrastructure stability.

Degree Correlation Function Visualization

- Visualization of knn(k)k_{nn}(k)knn(k) reveals trends in node connectivity.
- Upward trend for assortative networks, downward for disassortative.
- Highlights patterns of connectivity across degrees.
- Valuable in social and biological network studies.
- Practical for analyzing clustering and spread patterns.

Friendship Paradox in Assortative Networks

- Friendship Paradox: On average, your friends have more connections than you do.
- Common in social networks where hubs have many connections.
- Indicates clustering around high-degree nodes.
- Highlights the social advantage of being connected to hubs.
- Important for understanding social influence and network dynamics.

Network Robustness and Degree Correlation

- Assortative networks show robustness to random failures.
- Disassortative networks are vulnerable if hubs are targeted.
- Degree correlation impacts network resilience.
- Structural planning benefits from understanding correlation effects.
- Applications in network design and infrastructure protection.

Impact of Degree Correlation on Epidemic Spread

- Assortative networks can speed up epidemic spread.
- Clustering aids rapid transmission among connected nodes.
- Disassortative networks can slow spread, isolating high-degree nodes.
- Practical for epidemiology and disease modeling.
- Network structure guides containment strategies.

Correlation Coefficient Applications

- Degree correlation coefficient helps classify network types.
- Positive values (assortative) vs. negative (disassortative).
- Key metric in social, biological, and technological networks.
- Used for predictive analysis and structural insights.
- Supports targeted interventions and resilience planning.

Degree Correlation in Scientific Collaboration Networks

- Collaboration networks often assortative: high-degree nodes link.
- Hubs represent prolific collaborators or institutions.
- Clustering promotes efficient knowledge sharing.
- Increases resilience to member turnover.
- Assortativity encourages formation of research communities.

Degree Correlation in Infrastructure Networks

- Infrastructure networks often disassortative for stability.
- Hubs connect to low-degree nodes, minimizing inter-hub links.
- Limits damage in case of targeted attacks.
- Typical in power grids, transportation, and telecoms.
- Ensures stable operation with high resilience to overload.

Role of Degree Correlation in Network Evolution

- Network growth affects degree correlation over time.
- Preferential attachment increases assortative tendencies.
- Natural evolution can lead to distinct assortative/disassortative patterns.
- Studying evolution helps understand structural stability.
- Applications in predicting future network behaviors.

Generating Correlated Networks for Analysis

- Algorithms simulate assortative and disassortative structures.
- Generated networks help test correlation theories.
- Tools for modeling network resilience and failure responses.
- Enables custom network analysis for targeted research.
- Important for theoretical and applied network science.

Modeling Real-World Networks Using Correlation Data

- Real-world network models use observed correlation patterns.
- Degree correlation data aids in accurate simulations.
- Model-based insights into network behaviors under stress.
- Enables predictive analysis for resilience and adaptation.
- Key for infrastructural and social network planning.

Challenges in Measuring Degree Correlations

- Accurate measurement requires comprehensive data.
- Limitations in visual and mathematical correlation tools.
- Structural complexity affects precision of correlation metrics.
- Ongoing research aims to refine measurement techniques.
- Importance of advanced tools in large-scale network studies.

Advanced Techniques in Correlation Analysis

- Statistical models for more precise correlation measurement.
- Use of machine learning for pattern recognition in networks.
- Network simulation tools to model hypothetical changes.
- Importance of advanced techniques in complex networks.
- Key for analyzing large, evolving networks like social media.

Influence of Correlation on Network Robustness

- Assortative networks more resilient to random failures.
- Disassortative networks vulnerable to targeted attacks on hubs.
- Degree correlation informs network stability strategies.
- Vital for designing resilient communication and utility networks.
- Robustness influenced by degree correlation structure.

Predicting Network Behavior with Correlation Data

- Degree correlation helps forecast network response to disruptions.
- Example: Predicting epidemic spread in social networks.
- Models based on correlation data inform public health policies.
- Practical uses in risk management and infrastructure.
- Importance in proactive network maintenance.

Degree Correlation in Disease Modeling

- Network structure affects pathogen transmission rates.
- Assortative networks facilitate rapid within-group spread.
- Disassortative structures can slow disease spread.
- Applications in epidemic forecasting and intervention planning.
- Importance for targeted containment efforts.

Social Influence and Correlation in Online Networks

- Degree correlation affects influence dynamics on social media.
- Hubs amplify trends by linking to other hubs.
- Assortative structure creates echo chambers around key topics.
- Analysis useful for understanding misinformation spread.
- Implications for content moderation and policy.

Network Science

Lecture 6

17 October, 2024

The Barabási-Albert Model

A Detailed Analysis of Scale-Free Networks

Understanding the structure of complex networks (1)

The Barabási-Albert model represents a key breakthrough in understanding the structure of complex networks:

- It explains why many real-world networks, like the internet or social networks, are scale-free.
- Scale-free networks have a few highly connected nodes, or hubs, and many nodes with fewer connections.
- This model starts with a small number of connected nodes and grows by adding new nodes over time.
- When new nodes are added, they are more likely to connect to highly connected nodes, a process called preferential attachment.
- This process explains the "rich-get-richer" phenomenon seen in many systems, where a few entities become dominant.
- The model reveals how hubs emerge naturally over time, without any central control.

Understanding the structure of complex networks (2)

- Real-world networks, like airline routes or the World Wide Web, show similar hub structures.
- These hubs play a critical role in the network’s resilience, ensuring the network stays connected even if most nodes fail.
- The model also explains why these networks are vulnerable to targeted attacks on hubs.
- The average distance between any two nodes in a scale-free network is relatively small, a property known as the "small-world effect."
- This property allows information or diseases to spread quickly across the network.
- In technology, this model helps explain the structure of the internet, where a few websites have most of the links.
- In biology, it shows why certain proteins are more essential than others in cellular networks.

Understanding the structure of complex networks (3)

- Social networks exhibit similar characteristics, with influential individuals (hubs) playing central roles in information flow.
- Mathematically, the degree distribution of a Barabási-Albert network follows a power law, which is a key feature of scale-free networks.
- The model has been extended to explain networks with directed links, weighted links, and evolving topologies.
- Understanding the dynamics of this model helps improve network design and predict network failures.
- Its applications range from epidemiology, where it models the spread of diseases, to economics, where it helps understand market dynamics.
- The Barabási-Albert model is one of the cornerstones of modern network science and has influenced various disciplines.
- Future work involves exploring how networks evolve under different conditions and improving the robustness of these systems against failures.

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

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

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

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

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

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

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

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

To understand why so *different* systems converge to a *similar* architecture we need to first understand the mechanism responsible for the emergence of the scale-free property.

Given the diversity of the systems that display the scale-free property, the explanation must be simple and fundamental.

The answers will change the way we model networks, forcing us to move from describing a network’s topology to modeling the evolution of a complex system.

Networks Expand Through the Addition of New Nodes

The random network model assumes that we have a *fixed* number of nodes, N.

Yet, *in real networks the number of nodes continually grows thanks to the addition of new nodes*.

Few examples

1. In 1991 the WWW had a single node, the first webpage build by Tim Berners-Lee, the creator of the Web.

Today the Web has over a trillion documents, an extraordinary number that was reached through the continuous addition of new documents by millions of individuals and institutions

2. The collaboration and the citation network continually expands through the publication of new research papers.

3. The actor network continues to expand through the release of new movies.

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

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

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

(A) Growth

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

(B) Preferential Attachment

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

These two differences between real and random networks, *growth* and *preferential attachment*, play a particularly important role in shaping a network's degree distribution.

THE BARABÁSI-ALBERT MODEL

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

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

(A) Growth

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

B) Preferential attachment

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

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

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

While most nodes in the network have only a few links, a few gradually turn into hubs. These hubs are the result of a *rich-gets-richer phenomenon*: Due to preferential attachment new nodes are more likely to connect to the more connected nodes than to the smaller nodes. Hence, the larger nodes will acquire links at the expense of the smaller nodes, eventually becoming hubs.

In summary, the Barabási-Albert model indicates that two simple mechanisms, *growth* and *preferential attachment*, are responsible for the emergence of scale-free networks.

The origin of the power law and the associated hubs is a *rich-gets-richer phenomenon* induced by the coexistence of these two ingredients. To understand the model's behavior and to quantify the emergence of the scale-free property, we need to become familiar with the model's mathematical properties.

DEGREE DYNAMICS

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

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

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

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

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

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

Therefore

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

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

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

We obtain

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

We call β the *dynamical exponent* and has the value $\beta = \frac{1}{2}$.

The previous equation offers a number of predictions:

1. The degree of each node increases following a power-law with the same dynamical exponent β =1/2 Hence all nodes follow the same dynamical law.
2. The growth in the degrees is sublinear (i.e. $\beta < 1$). This is a consequence of the growing nature of the Barabási-Albert model: Each new node has more nodes to link to than the previous node. Hence, with time the existing nodes compete for links with an increasing pool of other nodes.

DEGREE DISTRIBUTION

The distinguishing feature of the networks generated by the Barabási-Albert model is their power-law degree distribution. We can calculate the functional form of p, helping us understand its origin.

A number of analytical tools are available to calculate the degree distribution of the Barabási-Albert network. The simplest is the *continuum theory*. It predicts the degree distribution,

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

with

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

THE ABSENCE OF GROWTH OR PREFERENTIAL ATTACHMENT

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

MODEL A

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

(A)Growth

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

(B)Preferential Attachment

The probability that a new node links to a node with degree k_i is $\Pi(k_i) = \frac{1}{(m_0 + t - 1)}$.

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

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

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

a much slower growth than the power law increase. Consequently the degree distribution follows an exponential

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

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

MODEL B

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

(B) Preferential Attachment

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

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

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

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

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

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

In summary, the absence of preferential attachment leads to a growing network with a stationary but exponential degree distribution.

In contrast the absence of growth leads to the loss of stationarity, forcing the network to converge to a complete graph.

This failure of Models A and B to reproduce the empirically observed scale-free distribution indicates that growth and preferential attachment are simultaneously needed for the emergence of the scale-free property.

MEASURING PREFERENTIAL ATTACHMENT

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

Preferential attachment relies on **two** distinct hypotheses:

Hypothesis 1

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

Hypothesis 2

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

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

In summary, we can detect the presence (or absence) of preferential attachment in real networks. The measurements show that the attachment probability depends on the node degree. We also find that while in some systems preferential attachment is linear, in others it can be sublinear.

NON-LINEAR PREFERENTIAL ATTACHMENT

The observation of sublinear preferential attachment raises an important question: What is the impact of this nonlinearity on the network topology? To answer this we replace the linear preferential attachment and calculate the degree distribution of the obtained *nonlinear Barabási-Albert model*.

The behavior for $\alpha=0$ is clear: In the absence of preferential attachment we are back to Model A. Consequently the degree distribution follows the exponential.

For $\alpha = 1$ we recover the Barabási-Albert model, obtaining a scale-free network with degree distribution.

THE ORIGINS OF PREFERENTIAL ATTACHMENT

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

Why does $\Pi(k)$ depend on k?

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

In the past decade we witnessed the emergence of two philosophically different answers to these questions. The first views preferential attachment as the interplay between random events and some structural property of a network.

These mechanisms do not require global knowledge of the network but rely on random events, hence we will call them *local* or *random* mechanisms. The second assumes that each new node or link balances conflicting needs, hence they are preceeded by a cost-benefit analysis. These models assume familiarity with the whole network and rely on optimization principles, prompting us to call them *global* or *optimized* mechanisms.We discuss both approaches.

LOCAL MECHANISMS

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

Link Selection Model

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

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

Link Selection: We select a link at random and connect the new node to one of the two nodes at the two ends of the selected link.

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

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

$$q_k = Ckp_k \cdot$$

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

Copying Model

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

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

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

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

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

- *Social Networks*: The more acquaintances an individual has, the higher is the chance that she will be introduced to new individuals by her existing acquaintances. In other words, we "copy" the friends of our friends. Consequently without friends, it is difficult to make new friends.
- *Citation Networks*: No scientist can be familiar with all papers published on a certain topic. Authors decide what to read and cite by "copying" references from the papers they have read. Consequently papers with more citations are more likely to be studied and cited again.
- *Protein Interactions*: Gene duplication, responsible for the emergence of new genes in a cell, can be mapped into the copying model, explaining the scale-free nature of protein interaction networks.

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

OPTIMIZATION

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

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

DIAMETER AND CLUSTERING COEFFICIENT

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

Diameter

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

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

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

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

Clustering coefficient

The clustering coefficient of the Barabási-Albert model

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

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

SUMMARY

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

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

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

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

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

- The model predicts γ=3 while the degree exponent of real networks varies between 2 and 5
- Many networks, like the WWW or citation networks, are directed, while the model generates undirected networks.
- Many processes observed in networks, from linking to already existing nodes to the disappearance of links and nodes, are absent from the model.
- The model does not allow us to distinguish between nodes based on some intrinsic characteristics, like the novelty of a research paper or the utility of a webpage.

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

We can show, these limitations can be systematically resolved.

Network Science

Lecture 8

Introduction to Network Robustness

- Definition of network robustness and its importance.
- Key areas where robustness matters: biology, engineering, and social systems.
- Networks as a resilience mechanism in complex systems.
- Exploring robustness in technological, social, and ecological networks.
- Objective: Understand how network structure influences resilience.

Role of Networks in Ensuring Robustness

- ❑Networks provide stability in diverse systems.
- ❑Robustness in cellular, ecological, and social networks.
- ❑How network topology impacts system stability.
- ❑Examples of resilient network structures.
- ❑Importance of studying networks for robustness insights.

Understanding Failures in Networks

- ❑Networks withstand random failures to some extent.
- ❑Importance of node and link integrity in networks.
- ❑How isolated failures can impact network performance.
- ❑Example: Failures in social, biological, and communication networks.
- ❑Investigating how different networks respond to disruptions

Types of Network Failures

- Random vs. targeted node failures.
- Effects of random failures on network fragmentation.
- Targeted attacks disrupt network hubs, causing breakdowns.
- Different failure impacts on assortative and disassortative networks.
- Examples of robustness testing in various network types.

Introduction to Percolation Theory

- Percolation theory explains network fragmentation.
- Examines node and link removals in networks.
- Identifies critical thresholds for network breakdown.
- Applications in understanding network robustness.
- Helps predict fragmentation points under failures.

Key Concepts in Percolation Theory

- Average cluster size in fragmented networks.
- Order parameter as an indicator of network integrity.
- Critical point (pc) where a giant component emerges.
- Correlation length between nodes in connected clusters.
- Mathematical tools for analyzing network stability.

Network Fragmentation and Critical Thresholds

- Fragmentation occurs when critical thresholds are reached.
- High node removal rates cause network breakdown.
- Impact of threshold variation by network type.
- Examples: Robustness of Internet vs. social networks.
- Network type determines resilience under random failures.

Robustness of Scale-Free Networks

- ❑ Scale-free networks show unique resilience properties.
- ❑ High robustness under random failures.
- ❑ Gradual degradation instead of abrupt fragmentation.
- ❑ Scale-free structure helps maintain connectivity.
- ❑ Importance of hubs in scale-free network stability.

Attack Tolerance in Networks

- Scale-free networks resilient to random failures but vulnerable to attacks.
- Targeted hub removals disrupt network stability.
- Real-world applications: Internet, biological networks.
- Balancing resilience with vulnerability in network design.
- Study of attack tolerance as a key network property.

Cascading Failures in Networks

- ❑ Cascading failures caused by interconnected dependencies.
- ❑ Examples: Power grid, financial systems, social networks.
- ❑ Trigger events can lead to large-scale disruptions.
- ❑ Understanding triggers and spread patterns in failures.
- ❑ Cascade modeling for resilience planning.

Examples of Cascading Failures

- Blackouts in power grids as cascading failures.
- Financial crises spread through interconnected institutions.
- Denial-of-service attacks in networked communication systems.
- Ecosystem collapses from species extinction.
- Importance of mitigating cascading risks in critical networks.

Modeling Cascading Failures

- Cascades modeled through percolation and branching theories.
- Factors: Network topology and node dependency.
- Models simulate failure spread through various network types.
- Insights into preventing widespread disruptions.
- Applications in power, financial, and infrastructure networks.

Percolation and Network Breakdown

- ❑ Network stability assessed by percolation thresholds.
- ❑ Removing nodes affects cluster size and connectivity.
- ❑ Thresholds indicate when a network fragments.
- ❑ Key in understanding robust and fragile points.
- ❑ Examples of percolation effects on different network types

Inverse Percolation in Robustness Analysis

- Inverse percolation studies network fragmentation.
- Assesses network vulnerability by removing nodes.
- Identifies critical node density for stability.
- Relevant in maintaining large networks like the Internet.
- Helps quantify resilience against node failures.

Role of Critical Thresholds in Resilience

- Critical threshold (fc) marks stability loss point.
- Below fc, the network retains a giant component.
- Above fc, network disintegrates into isolated clusters.
- fc influenced by network topology and degree distribution.
- Applications in network design for robustness.

Network Types and Critical Thresholds

- Random failures impact peripheral nodes, preserving hubs.
- Targeted attacks on hubs cause faster fragmentation.
- Scale-free networks more resilient to random failure.
- Example: Internet withstands random failures but not targeted attacks.
- Analysis aids in identifying key vulnerabilities.

Scale-Free Networks' Resistance to Random Failures

- Scale-free topology limits impact of random node removals.
- High-degree nodes remain intact in random failure scenarios.
- Small nodes removed without major connectivity loss.
- Example: Social networks maintain resilience with low-degree node loss.
- Relevance for designing failure-resistant networks.

Understanding Attack Tolerance in Networks

- Attack tolerance measures resilience against targeted attacks.
- Hubs in scale-free networks make them vulnerable to attacks.
- Removing a few hubs can collapse the entire network.
- Example: Biological networks vulnerable to hub protein removal.
- Network resilience requires balancing attack and failure tolerance.

Percolation in Scale-Free Networks

- Scale-free networks display unique percolation properties.
- Gradual fragmentation under random node removal.
- Lack of critical point in infinite scale-free networks.
- Random removal affects primarily low-degree nodes.
- Example: Scale-free network models show extended robustness.

Attack Tolerance: Scale-Free vs. Random Networks

- ❑ Random networks show similar resilience to random failures and attacks.
- ❑ Scale-free networks are more attack-sensitive.
- ❑ Targeted hub removal disrupts scale-free networks more effectively.
- ❑ Network design implications for attack tolerance.
- ❑ Importance of diversifying node connections for stability.

Cascade Effects in Power Grids

- Power grid failures often spread due to network dependencies.
- Initial faults lead to load redistribution and further failures.
- Cascading blackouts exemplify network vulnerability.
- Structural improvements mitigate cascade risks.
- Key lessons for resilient infrastructure design.

Denial-of-Service Attacks as Cascading Failures

- Router failures can overload neighboring routers.
- Denial-of-service (DoS) attacks exploit these vulnerabilities.
- Traffic rerouting in network increases pressure on functioning nodes.
- DoS attack prevention requires understanding cascade mechanisms.
- Mitigating overload risks in communication networks.

Financial Networks and Cascading Crises

- Interconnected banks spread risk during economic downturns.
- 2008 financial crisis as an example of network cascade.
- Default risks cascade across linked institutions.
- Network design informs crisis prevention in financial systems.
- Importance of risk diversification in reducing contagion.

Empirical Patterns in Cascading Failures

- Observed patterns in blackouts, DoS attacks, and financial crises.
- Small events often coexist with major cascading failures.
- Power-law distribution in failure magnitudes.
- Identifying high-risk connections in complex systems.
- Empirical data informs prevention and recovery strategies.

Modeling Network Cascades with Power Laws

- Power law helps predict cascading event sizes.
- Applies to various domains: power, finance, social media.
- Large, rare events alongside frequent minor ones.
- Power law distribution aids in risk assessment.
- Tool for analyzing potential cascade scenarios.

Power Grid Cascading Failures: North America Example

- Example of North American blackout cascades.
- Small local failures triggered large-scale disruptions.
- Importance of grid design for mitigating failure impact.
- Lessons for resilience in critical infrastructure.
- Data on blackout events informs future improvements

Information Cascades in Social Media

- Information spreads quickly in highly connected social networks.
- Cascade effect seen in viral tweets or shared content.
- Twitter as a case study for digital information cascades.
- Large cascades are rare but impactful.
- Implications for managing misinformation and digital influence.

Role of Tectonic Network Cascades in Earthquakes

- Earthquakes cause cascades along fault lines.
- Triggered stress releases impact neighboring areas.
- Studied as a network cascade in geology.
- Implications for predicting and mitigating earthquake impact.
- Network theory aids in seismic event prediction.

Preventing Cascading Failures in Critical Networks

- Redundancy design reduces failure propagation.
- Isolating high-risk nodes to prevent cascade spread.
- Enhancing node resilience in critical hubs.
- Planning for rapid containment during initial failures.
- Key for infrastructure stability and resilience.

Twitter Cascade Dynamics

- Twitter's follower network facilitates rapid content spread.
- URL sharing tracks digital cascade paths.
- Large retweet chains create significant information cascades.
- Data insights: Small events common, large cascades rare.
- Key for understanding social influence on digital platforms.

Earthquake Cascades and Network Patterns

- Fault lines interact in cascade-like patterns.
- Earthquake network: high interdependency among fault segments.
- Sudden shifts can propagate across geological networks.
- Example: Haiti earthquake as a major cascading failure.
- Seismic networks studied for resilience and risk mapping.

Universal Patterns in Cascading Failures

- Cascading events share common distribution patterns.
- Large disruptions follow power-law distributions.
- Observed in systems from power grids to social media.
- Helps predict rare, high-impact events.
- Insights applied across fields for risk management.

Impact of Node Degree on Cascade Potential

- High-degree nodes often trigger larger cascades.
- Networks with densely connected hubs more vulnerable.
- Degree distribution influences cascade size.
- Scale-free networks exhibit cascading sensitivity.
- Important for evaluating network robustness.

Modeling Cascades: Key Approaches

- Percolation and branching models simulate cascades.
- Network structure and node behavior central to models.
- Predict failure spread based on node connections.
- Cascading model applications in infrastructure and finance.
- Helps anticipate and control potential large-scale failures.

Failure Propagation Model

- Each node has a failure threshold.
- Neighboring node failures increase risk of spreading.
- Models applied to social influence and opinion spread.
- Can predict cascade reach and impact.
- Useful in social networks, economic systems, and technology.

Subcritical, Supercritical, and Critical Regimes

- Subcritical: Limited, contained cascades.
- Supercritical: Failures spread widely, risking total collapse.
- Critical: Unpredictable, varied cascade sizes.
- Critical regime particularly challenging to manage.
- Highlights need for network planning in sensitive thresholds.

Branching Model for Cascade Analysis

- Nodes fail based on cascading “branches.”
- Simple model to capture basic cascade dynamics.
- Predicts how failures spread in hierarchical patterns.
- Used in electrical grids, communication systems.
- Foundation for understanding larger cascading events.

Branching Model in Random Networks

- Network structure affects branching cascade growth.
- High average degree leads to large-scale cascades.
- Key for assessing risks in random networks.
- Cascade size distribution follows power law in critical regimes.
- Practical applications in supply chains and transport networks.

Critical Thresholds in Network Cascades

- Cascades intensify near critical percolation threshold.
- Below threshold: Limited failure spread.
- Above threshold: Widespread breakdowns.
- Threshold analysis essential in infrastructure design.
- Key for planning resilient network structures.

Scaling Behavior in Cascading Failures

- Cascade size scales predictably with network characteristics.
- Larger networks require different containment strategies.
- Scaling laws assist in resilience planning.
- Application in managing complex system disruptions.
- Insights aid global infrastructure and critical system planning.

Cascading Blackouts: Power Grid Case Study

- Blackouts show cascade effects in electrical networks.
- Example: 2003 Northeast blackout in North America.
- Load redistribution leads to sequential failures.
- Network design improvements reduce blackout risk.
- Case study in critical infrastructure management.

Preventing Information Overload in Communication Networks

- High traffic rerouting can lead to node overload.
- Network load-balancing prevents denial-of-service cascades.
- Distributed design reduces overload risks.
- Examples: Internet, cloud networks.
- Critical for maintaining stable digital services.

Economic Networks and Systemic Risk

- Banks linked through financial assets and obligations.
- Failures cascade, spreading default risks.
- Example: Global financial crisis.
- Diversification lowers systemic risk in financial networks.
- Essential for economic stability and regulation.

Network Design for Cascade Prevention

- Redundant pathways reduce single points of failure.
- Distributed structure limits cascade reach.
- Strategic hub protection enhances resilience.
- Sector-specific design: telecommunications, energy, finance.
- Planning essential in high-stakes, interconnected networks.

Complexity and Criticality in Network Cascades

- Increased complexity raises cascade risk near critical points.
- Criticality in networks correlates with interconnectedness.
- Complex systems studied for managing tipping points.
- Critical for preventing systemic failures.
- Cross-disciplinary insights for infrastructure and technology.

Percolation in Social Networks

- Social networks maintain resilience to random failures.
- Highly connected nodes central to information spread.
- Removal of hubs disrupts connectivity significantly.
- Cascade dynamics inform social network management.
- Useful for digital media, public health, and marketing.

Social Networks and Misinformation Spread

- Cascading nature of information increases misinformation spread.
- Targeted misinformation disrupts public discourse.
- Social network resilience requires addressing cascade risks.
- Key role of network structure in containment strategies.
- Critical for information integrity in digital age.

Digital Networks and Security

- Cyber-attacks exploit network vulnerabilities.
- Distributed denial-of-service (DDoS) attacks create cascades.
- Cascade management essential in digital security.
- Effective load balancing and redundancy mitigate risks.
- Security practices in highly connected networks.

Resilience through Network Redundancy

- Redundant links enhance network stability.
- Protects against node and link failures.
- Examples: Power grid redundancies, internet pathways.
- Cost-effective method for enhancing robustness.
- Practical in infrastructure and service continuity.

Robustness in Biological Networks

- Biological systems rely on robust protein and metabolic networks.
- Redundant pathways increase resilience to mutations.
- Hubs central to network connectivity.
- Example: Protein-protein interaction networks.
- Insights aid drug discovery and disease management.

Hub Vulnerability in Biological Systems

- Hubs essential for cellular function but vulnerable to attacks.
- Protein hubs critical in cell signaling and metabolism.
- Drug targeting of hubs impacts disease-causing organisms.
- Challenge: Balancing therapeutic efficacy with resilience.
- Example: Antibiotic targeting strategies.

Applications of Network Robustness in Medicine

- Robustness concepts applied to cellular networks.
- Drug discovery uses network analysis for target identification.
- Understanding cascade effects in cell death pathways.
- Resilience helps in managing disease mutations.
- Network-based approach to medical treatment.

Cascading Failures in Ecological Networks

- Species interdependence creates cascade vulnerability.
- Loss of keystone species disrupts entire ecosystems.
- Example: Impact of predator loss in food webs.
- Cascade insights inform conservation strategies.
- Key for biodiversity and environmental stability.

Supply Chain Networks and Cascade Risks

- Supply chain disruptions lead to production delays.
- Highly interconnected suppliers increase cascade potential.
- Example: Automotive industry during natural disasters.
- Network design optimizes resilience in supply chains.
- Critical for efficient and reliable global production.

Internet Resilience and Redundant Pathways

- Redundant paths maintain connectivity during disruptions.
- Example: Internet’s robustness to router failures.
- Redundancy mitigates cascading impacts of localized outages.
- Essential for reliable global communication.
- Backbone of modern digital infrastructure.

Scale-Free Networks in Technology

- Scale-free nature of tech networks enhances robustness.
- Example: Internet backbone, telecom networks.
- High-degree hubs support network connectivity.
- Vulnerability to targeted attacks on hubs.
- Security implications for digital infrastructure.

Network Fragility in Financial Crises

- Financial markets show cascading failures during crises.
- Interconnected assets create systemic vulnerabilities.
- Risk management addresses fragile connections.
- Crisis containment requires quick intervention.
- Implications for policy and economic planning.

Telecommunication Network Resilience

- Redundant systems ensure service continuity.
- Importance of load balancing in peak traffic.
- Structural stability crucial for digital communication.
- Example: Cellular network redundancy.
- Network design supports reliable user experience.

Addressing Cascade Risks in Health Networks

- Health systems rely on robust communication and resource networks.
- Failure in critical nodes impacts service delivery.
- Pandemic management as a cascade problem.
- Redundancy in health supply chains reduces risk.
- Vital for effective response to health crises.

The Role of Degree Distribution in Network Stability

- Degree distribution affects network robustness.
- Scale-free networks more resilient to random failures.
- Importance of hubs in maintaining connectivity.
- Random networks lack hub centrality, impacting resilience.
- Key for designing stable network structures.

Supply Chain Redundancy for Resilience

- Redundant suppliers reduce production cascade risks.
- Geographic diversification enhances supply chain stability.
- Example: Tech manufacturing during natural disasters.
- Network modeling helps predict supply disruptions.
- Essential for global production reliability.

Mitigating Risk in Power Distribution Networks

- Grid design with redundant lines limits blackouts.
- Distributed energy sources improve resilience.
- Example: Localized grids as backup systems.
- Essential for stable electricity access.
- Network planning critical for energy security.

Network Science in Epidemic Control

- Epidemic spread influenced by social network structure.
- Targeting hubs reduces transmission risk.
- Contact tracing and isolation based on network theory.
- Models support pandemic management strategies.
- Effective for public health interventions.

Role of Hubs in Network Traffic Management

- High-degree nodes central in internet traffic routing.
- Hub failure risks cascade in traffic disruptions.
- Network load balancing reduces stress on hubs.
- Key for stable internet and communication services.
- Design strategies enhance digital network resilience.

Environmental Networks and Ecosystem Stability

- Ecosystems resilient through diverse, interconnected species.
- Loss of keystone species triggers cascading extinctions.
- Network analysis aids conservation efforts.
- Example: Food webs and biodiversity protection.
- Network science informs environmental resilience planning.

Application of Network Theory in Infrastructure Resilience

- Resilience through redundant connections in critical networks.
- Infrastructure networks require strategic design.
- Examples: Transport, water, and power systems.
- Network theory guides infrastructure planning.
- Essential for national and regional stability.

Enhancing Cybersecurity with Network Design

- Disassortative networks reduce risk concentration.
- Network segmentation prevents cyberattack spread.
- High-risk nodes receive additional security layers.
- Insights from network science inform security protocols.
- Key for protecting critical information systems.

Community Resilience through Network Robustness

- Social networks support resilience during crises.
- Strong local connections aid in emergency response.
- Examples: Community support in natural disasters.
- Resilient community networks support recovery.
- Importance for social network-based emergency planning.

Understanding Network Vulnerability through Node Importance

- Critical nodes impact network stability disproportionately.
- Importance of high-degree and central nodes.
- Removing critical nodes disrupts network connectivity.
- Vulnerability analysis key for network planning.
- Applied in cybersecurity, power grids, and transport.

Balancing Efficiency and Robustness in Networks

- Trade-offs between fast communication and resilience.
- Dense networks enable quick spread but increase risk.
- Robust design adds redundancy, limiting efficiency.
- Applications in corporate, digital, and transport networks.
- Key for optimizing networks under resource constraints.

Network Interdependencies in Cascading Failures

- Interconnected networks compound cascade risks.
- Example: Power and communication network dependencies.
- Failures in one network impact others.
- Risk management addresses inter-network links.
- Essential for critical infrastructure planning.

Using Network Simulation for Risk Assessment

- Simulation models predict cascade scenarios.
- Virtual testing identifies vulnerabilities.
- Applications in finance, infrastructure, and social media.
- Real-time analysis supports crisis management.
- Tools improve network robustness preemptively.

Complexity and Resilience in Multilayer Networks

- Multilayer networks interact across domains.
- Resilience requires securing each layer.
- Example: Transport and logistics network interdependencies.
- Complex interactions add resilience challenges.
- Multilayer analysis essential for holistic planning.

Distributed Networks for Decentralized Resilience

- Decentralized networks spread risk across nodes.
- Examples: Blockchain, mesh networks.
- Eliminates single points of failure.
- Distributed design supports security and resilience.
- Key for future decentralized technology development.

Network Robustness in Urban Infrastructure

- Urban networks require robust transportation, utilities.
- Redundancy critical for disaster preparedness.
- Urban planning integrates network resilience principles.
- Examples: Public transit, water supply systems.
- Network science applied in resilient city design.

Emergency Preparedness with Network Models

- Network theory aids emergency response planning.
- Modeling evacuation routes, resource allocation.
- Simulation of crisis scenarios for preparedness.
- Social networks support communication during crises.
- Network resilience crucial for disaster management.

Network Effects in Viral Marketing

- Social networks amplify marketing campaigns.
- Targeting hubs maximizes reach.
- Cascade modeling predicts viral spread.
- Applications in digital and social media marketing.
- Network science key to influence strategy.

Network-based Drug Discovery in Medicine

- Identifying protein interactions for disease treatment.
- Network approach maps cellular processes.
- Drug targets in critical nodes for therapeutic effects.
- Applied in cancer, infectious disease research.
- Network medicine guides targeted therapy.

Impact of Redundancy in Wireless Networks

- Redundancy prevents single-point wireless failures.
- Distributed nodes maintain connectivity in network loss.
- Mesh network design for fault tolerance.
- Applications in remote and emergency wireless setups.
- Network science enhances wireless resilience.

Humanitarian Networks for Crisis Relief

- Robust networks for aid distribution.
- Resilient communication networks in disaster zones.
- Network-based approach for resource logistics.
- Examples: Food, water, medical supply chains.
- Essential for efficient humanitarian response.

Improving Healthcare Access with Network Analysis

- Analyzing healthcare provider distribution.
- Identifying underserved areas via network gaps.
- Network improvements enhance healthcare access.
- Applied in rural and urban health planning.
- Key for efficient healthcare network development.

Digital Payment Networks and Security

- Security risks in interconnected payment systems.
- Robustness through multi-layer encryption and redundancy.
- Mitigating cascading failure in transaction networks.
- Blockchain applied in secure, decentralized finance.
- Key for resilient global payment infrastructure.

Role of Network Theory in Climate Adaptation

- Modeling ecosystem networks for climate resilience.
- Analyzing species interdependence in climate impact.
- Network planning aids conservation efforts.
- Critical for biodiversity and ecosystem management.
- Helps inform climate adaptation strategies.

Utility of Network Theory in Traffic Management

- Analyzing road networks for congestion resilience.
- Distributed traffic flow through alternative routes.
- Traffic load balancing reduces peak-time overload.
- Network science applied in smart city transport.
- Key for efficient urban traffic systems.

Building Resilient Food Supply Networks

- Networked supply chains manage food production.
- Redundancy ensures supply during disruptions.
- Geographic distribution reduces local impact risks.
- Examples: Food security in natural disasters.
- Network planning for sustainable food access.

Exploring Ecosystem Services with Network Models

- Mapping interactions in ecosystem service networks.
- Protecting keystone species for service resilience.
- Examples: Pollination, nutrient cycling, water purification.
- Network analysis aids in ecosystem conservation.
- Essential for sustainable ecosystem service management.

Financial Stability through Network Theory

- Networked connections impact systemic financial risks.
- Diversification strategies to reduce contagion.
- Network analysis aids in policy and regulation.
- Example: Banking network structure during crises.
- Key for managing financial system resilience.

Public Transport Networks and Cascade Management

- Redundant routes prevent transport network breakdowns.
- Load balancing maintains service during peak demand.
- Network modeling optimizes public transport systems.
- Examples: Subway, bus network resilience.
- Essential for urban transport planning.

Network Analysis for Smart Grid Stability

- Smart grids integrate renewable energy with network robustness.
- Distributed energy resources improve grid resilience.
- Load management reduces cascade blackout risk.
- Network science supports stable energy distribution.
- Key for future renewable energy integration.

Educational Networks for Knowledge Sharing

- Academic collaboration forms robust knowledge networks.
- Assortative connections enhance knowledge flow.
- Network analysis supports educational resource planning.
- Virtual networks in online learning environments.
- Importance of robust academic collaboration networks.

Network Robustness in Climate Data Networks

- Distributed sensors maintain data collection in failures.
- Resilient climate data networks track environmental change.
- Network redundancy ensures consistent data flow.
- Applications in climate modeling and forecasting.
- Key for accurate climate adaptation data.

Economic Networks and Trade Stability

- Trade networks maintain stability through diverse routes.
- Geographic diversification mitigates disruption impact.
- Example: Global supply chain resilience in trade networks.
- Network science informs economic trade policy.
- Essential for global economic stability.

Power Grids and Decentralized Energy Networks

- Decentralized grids enhance local resilience.
- Renewable integration improves energy network stability.
- Distributed resources reduce central dependency.
- Example: Microgrids in disaster-prone areas.
- Important for future energy security.

Smart Cities and Integrated Network Planning

- Interconnected urban systems enhance resilience.
- Transportation, utilities, and emergency networks.
- Network science informs smart city design.
- Examples: Traffic, public services, and infrastructure.
- Key for sustainable urban development.

Cascading Effects in Logistic Networks

- Supply chain failures cascade through logistic networks.
- Examples: Transportation delays impact production.
- Network redundancy minimizes cascade effects.
- Applied in global logistics and distribution planning.
- Critical for maintaining supply chain stability.

Network Science in Biodiversity Conservation

- Mapping species networks for ecosystem stability.
- Network design helps protect biodiversity.
- Keystone species resilience supports entire ecosystems.
- Conservation efforts guided by network analysis.
- Key for sustaining biodiversity under environmental threats.

Cyber-Physical Systems and Network Resilience

- Integration of physical and digital networks.
- Smart infrastructure and IoT connectivity.
- Resilience planning critical in cyber-physical systems.
- Network science enhances digital-physical security.
- Key for future smart infrastructure.

Real-World Application of Network Robustness

- Robust networks in communication, finance, and health.
- Examples: Redundancy, load balancing, and decentralized design.
- Network resilience in daily life and global systems.
- Continuous adaptation to maintain stability.
- Importance of network science in complex systems.

Conclusion and Key Insights on Network Robustness

- Robustness as a core property for all networks.
- Cascading failure prevention across disciplines.
- Importance of degree distribution and redundancy.
- Network science enables resilient systems in society.
- Future directions in research and practical applications.

Network Science: Community Detection and Analysis

Communities, Networks, and Advanced Detection Algorithms

Why Are Communities Important?

- Communities reveal the hidden structure of networks.
- Key applications: Social, biological, and technological networks.
- Improve understanding of human behavior and interactions.
- Example: Belgium’s linguistic communities derived from call data.
- Communities drive information flow and stability in networks.
- Foundational concept in network science research.

Key Concepts of Communities

- Definition: Groups of nodes with high internal connectivity.
- Communities form naturally in networks with structured interactions.
- Connectedness Hypothesis: Nodes in a community form connected subgraphs.
- Density Hypothesis: More internal than external links for community nodes.
- Examples: Friend groups, work clusters, biological pathways.
- Real-world impact: Understanding and predicting network behavior.

Real-World Examples of Communities

- Social Networks: Friend circles, professional groups.
- Biological Systems: Protein interaction networks, disease modules.
- Technological Networks: Internet resilience and structural optimization.
- Case Study: Zachary's Karate Club split due to internal conflict.
- Communities play roles in collaboration, competition, and co-evolution.

The Belgium Linguistic Case

- Researchers analyzed Belgium’s mobile call data.
- Discovered two clusters: Flemish (Dutch-speaking) and Walloon (French-speaking).
- Communities correlated strongly with language groups.
- Central finding: Communities form around cultural and social lines.
- Visual representation: Node clusters based on call patterns.
- Implication: Social behavior influences network structure.

Applications in Social Networks

- Communities help map human relationships and dynamics.
- Use Cases: Identifying influencers, predicting social movements.
- Examples: Social media clusters, professional networks.
- Help understand group behaviors and preferences.
- Algorithm Testing: Zachary's Karate Club as a benchmark.
- Real-life impact: Better-targeted policies and interventions.

Biological Applications of Community Detection

- Biological networks reveal communities as functional modules.
- Disease Module Hypothesis: Proteins linked to the same disease interact more.
- Metabolic networks show clusters of reactions (e.g., E. coli pathways).
- Communities help identify drug targets in molecular systems.
- Understanding disease propagation through connected modules.
- Example: Pyrimidine metabolism as a network module.

Technological Network Applications

- Internet: Identify critical hubs and improve resilience.
- Belgium’s mobile call network study: Language-based clusters.
- Enhancing network efficiency through community-based designs.
- Prevent cascading failures by understanding interdependencies.
- Applications in traffic optimization and smart city planning.
- IoT and sensor networks: Decentralized community frameworks.

Defining Communities

- Clique Definition: All nodes in a community are interconnected.
- Strong vs. Weak Communities:
  - Strong: Each node has more internal links than external.
  - Weak: Aggregate internal links exceed external links.
- Communities are locally dense and connected subgraphs.
- Visualization: Dendrograms, adjacency matrices.
- Examples: Social groups, overlapping affiliations.
- Challenges: Diverse definitions for different applications.

Connectedness and Density Hypothesis

- Connectedness Hypothesis: Nodes in a community form a connected graph.
- Density Hypothesis: Internal links dominate over external links.
- Both hypotheses underpin community detection algorithms.
- Examples: Work clusters, biological reaction networks.
- Limitations: Not all dense regions represent meaningful communities.
- Frameworks evolve based on network size and purpose.

Social Network Case Studies

- Zachary's Karate Club: Conflict split community into two.
- Girvan-Newman algorithm accurately predicted the split.
- Workplace communities: Interactions driven by professional links.
- Online networks: Social media clusters show clear groupings.
- Predicting social influence through community detection.
- Applications: Marketing, public relations, organizational design.

Biological Insights into Communities

- Metabolic networks: Clusters of molecules enable cellular functions.
- Disease studies: Genes often belong to overlapping communities.
- Lee Hartwell's concept of functional modules in cells.
- Identifying densely connected clusters in biological systems.
- Visualizing metabolic pathways as hierarchical structures.
- Community detection aids in personalized medicine.

The Hypotheses of Community Structures

- H1: Community structure encoded in network wiring diagrams.
- H2: Communities exhibit connectedness and density.
- H3: Random networks lack inherent communities.
- Testing hypotheses through modularity analysis.
- Deviations from random behavior highlight true communities.
- Algorithms refine these hypotheses for real-world networks.

Graph Partitioning vs. Community Detection

- **Graph Partitioning:** Predefined subgraph numbers and sizes.
- **Community Detection:** Natural group identification without prior assumptions.
- Example: Integrated circuit design to minimize wire crossings.
- Key difference: Exploratory nature of community detection.
- Graph partitioning focuses on computational efficiency.
- Applications: Chip design, parallel computing, load balancing.

Algorithms for Community Detection

- Hierarchical Clustering: Builds communities from local similarities.
- Girvan-Newman Algorithm: Divisive method using edge betweenness.
- Louvain Method: Efficient modularity optimization for large networks.
- Clique Percolation: Finds overlapping communities.
- Random walk-based algorithms explore probabilistic paths.
- Each algorithm adapts to specific network types and sizes.

Hierarchical Clustering in Community Detection

- Two approaches: **Agglomerative** and **Divisive.**
- **Agglomerative Methods:** Start with individual nodes, merge similar ones.
- **Divisive Methods:** Start with the full network, split by removing weak links.
- Output: Dendrogram representing nested structures.
- Examples: Hierarchical modules in biological and social networks.
- Advantage: Captures hierarchical relationships naturally.

Agglomerative Hierarchical Clustering

- Example: Ravasz Algorithm for metabolic networks.
- Steps:

1. Calculate node similarity (e.g., shared neighbors).
2. Merge the most similar nodes or groups.
3. Recompute similarities for merged groups.

- Repeat until a single community forms.
- Dendrogram helps visualize community hierarchy.
- Applications: Biological modules, nested organizations.

Divisive Hierarchical Clustering

- Example: Girvan-Newman Algorithm.

- Steps:

1. Compute edge betweenness (number of shortest paths).
2. Remove highest betweenness edge.
3. Recompute betweenness after each removal.

- Continue until isolated communities emerge.

- Case Study: Zachary's Karate Club split.
- Advantage: Identifies key inter-community links.

Modularity in Community Detection

- **Modularity (M):** Measures quality of a network partition.
- M = Observed links - Expected links in random network.
- Higher modularity indicates stronger community structure.
- Random hypothesis: Random networks lack communities.
- Applications: Social media, citation networks, infrastructure.
- Challenges: Resolution limit and ambiguous partitions.

Example of Modularity Optimization

- Case Study: Collaboration network in physics research.
- Louvain Algorithm: Detected ~600 communities.
- Subgroups aligned with scientific fields (e.g., condensed matter).
- Key metric: Modularity M = 0.713 for the network.
- Applications: Identifying influential research groups.
- Visualization: Community heatmaps and modular dendrograms.

Louvain Algorithm Overview

- Designed for scalability: Works efficiently on large networks.
- **Steps:**
  1. Assign each node to its own community.
  2. Optimize modularity locally by merging communities.
  3. Reassign merged communities into a single node.
  4. Repeat until modularity stabilizes.
- Example: Applied to Facebook graph with millions of nodes.
- Advantages: Speed and accuracy for real-world applications.

Girvan-Newman Algorithm

- Divisive method focusing on edge betweenness.
- Central edges connecting communities are removed.
- Steps:
  1. Compute betweenness for all edges.
  2. Remove edge with highest betweenness.
  3. Repeat until all edges are removed.
- Example: Zachary's Karate Club case study.
- Visualization: Hierarchical dendrograms of splits.

Overlapping Communities

- Definition: Nodes belong to multiple groups simultaneously.
- Real-world examples:
  - Scientists in professional and social communities.
  - Genes involved in multiple diseases.
- Clique Percolation Method (CPM): Identifies overlapping k-cliques.
- Visualization: Shared subgraphs highlight overlaps.
- Applications: Social dynamics, biological functions.
- Challenge: Representing multi-community memberships.

Clique Percolation Method (CPM)

- **Steps:**
    1. Identify all k-cliques in a network.
    2. Build adjacency matrix for k-cliques sharing (k-1) nodes.
    3. Find connected components of overlapping k-cliques.
    - Output: Overlapping communities as subgraphs.
- Example: Social media tags and overlapping interests.
- Challenges: Computationally intensive for dense networks.

Applications of Overlapping Communities

- Biological Systems: Genes linked to multiple diseases.
- Social Networks: Individuals in multi-group affiliations.
- Technological Systems: Overlapping infrastructure and dependencies.
- Real-world example: Multi-affiliation networks (e.g., LinkedIn and Facebook).
- Advantages: Captures realistic complexity.
- Limitations: Ambiguity in defining overlap thresholds.

Challenges in Community Detection

- **Scalability:** Handling large-scale networks efficiently.
- **Resolution Limit:** Detecting small communities.
- **Overlap Representation:** Defining multi-community memberships.
- **Validation:** Limited ground truth in real-world networks.
- **Noise:** Real-world data is often incomplete or inaccurate.
- Future direction: Hybrid algorithms for dynamic networks.

Future Directions in Network Science

- Adaptive algorithms for evolving networks.
- Integration of dynamic data streams.
- Real-time community detection.
- Improved modularity metrics for scalability.
- Advanced visualization techniques for complex networks.
- Applications: Smart cities, personalized medicine, IoT.

Hierarchical Modularity in Networks

- Nested structures form naturally in real-world networks.
- Examples: Organizational charts, metabolic pathways.
- Quantitative signature: Decreasing clustering coefficient with node degree.
- Insight: Hubs connect distinct communities.
- Applications: Multi-layered biological systems, infrastructure networks.
- Visualization: Heatmaps and hierarchical trees.

Practical Tools for Community Detection

- Software: Gephi, NetworkX, Pajek.
- Algorithms: Louvain, Girvan-Newman, Infomap.
- Visualization Tools: Dendrograms, adjacency matrices.
- Hands-on applications in academic and industrial settings.
- Strengths and weaknesses of each tool.
- Emerging frameworks for interdisciplinary use.

Real-World Case Studies

- Social Networks: Community dynamics on Facebook and Twitter.
- Biological Networks: Functional modules in gene interaction maps.
- Technological Systems: Optimization of telecom infrastructure.
- Citation Networks: Identifying leading researchers and topics.
- Key takeaway: Community detection enhances understanding across domains.

Social Media and Community Detection

- Social media clusters reflect real-life connections.
- Influencers: Central nodes within communities.
- Applications: Targeted advertising, content personalization.
- Examples: Facebook’s friend groups, Twitter hashtags.
- Algorithms: Detecting topics, trends, and echo chambers.
- Challenges: Rapidly changing dynamics in networks.

Technological Networks Applications

- Internet infrastructure resilience through modular design.
- Traffic flow optimization in smart cities using community detection.
- Overlapping communities in IoT sensor networks.
- Example: Mapping telecom networks for failure prevention.
- Applications in supply chain and logistics networks.
- Future directions: Autonomous systems and real-time analysis.

Citation Networks and Research Communities

- Citation patterns reveal clusters of research areas.
- Case Study: Physics collaboration networks.
- Modularity detects distinct scientific subfields.
- Influential nodes: Researchers with high citation counts.
- Applications: Identifying trends and emerging fields.
- Visualization: Co-authorship graphs and topic clusters.

Dynamic Community Detection

- Real-world networks evolve over time.
- Examples: Social media trends, disease outbreak patterns.
- Dynamic algorithms: Tracking community changes.
- Challenges: Scalability and real-time processing.
- Future applications: Predicting network behavior.
- Tools: Dynamic modularity, temporal clustering.

Advanced Algorithms for Large-Scale Networks

- Louvain Algorithm: Efficient modularity optimization.
- Infomap: Entropy-based clustering for large datasets.
- Random walk-based approaches: Probabilistic node exploration.
- Challenges: Balancing accuracy and computational efficiency.
- Emerging methods: Hybrid and multi-level algorithms.
- Applications: Multi-layered network systems.

Girvan-Newman Algorithm in Depth

- Focuses on edge betweenness to split communities.
- Removes edges that act as bridges between clusters.
- Case Study: Zachary’s Karate Club.
- Visualization: Dividing the network step-by-step.
- Applications: Identifying critical connections.
- Limitations: Computationally intensive for large networks.

Modularity Optimization Challenges

- **Resolution Limit:** Difficulty in detecting small communities.
- **Ambiguity:** Multiple partitions with similar modularity scores.
- Randomness in optimization algorithms.
- Adjustments for weighted or directed networks.
- Strategies: Multi-resolution approaches and fine-tuning.
- Future developments: Combining modularity with other metrics.

Overlapping Community Challenges

- Defining clear overlaps without redundancy.
- Example: Genes in multiple disease modules.
- Social networks: Multi-group affiliations.
- Threshold selection for overlaps in cliques.
- Visualization: Venn diagrams, heatmaps.
- Applications: Complex systems and interdisciplinary networks.

Metrics Beyond Modularity

- **Conductance:** Ratio of external to internal edges in a community.
- **Density:** Measure of internal link tightness.
- **Silhouette Coefficient:** Quality of clustering compared to neighbors.
- **Normalized Cut:** Used for spectral clustering approaches.
- Combining metrics for more comprehensive insights.
- Applications: Robust evaluations in diverse networks.

Hybrid Models in Community Detection

- Combining modularity and random walk methods.
- Multi-layer networks: Integrating data from multiple sources.
- Examples: Social + professional networks.
- Challenges: Balancing complexity and interpretability.
- Applications: Systems biology, supply chain networks.
- Tools: Multi-level algorithms and scalable frameworks.

Temporal Networks and Community Dynamics

- Networks change over time due to new links or nodes.
- Dynamic algorithms track temporal evolution.
- Example: Tracking social media trends over months.
- Applications in epidemiology: Disease spread models.
- Challenges: Real-time computation on dynamic graphs.
- Tools: Temporal modularity and streaming algorithms.

Network Visualization Techniques

- Tools: Gephi, Cytoscape, Pajek.
- Graphical elements: Node-link diagrams, adjacency matrices.
- Highlighting communities with colors and shapes.
- Dynamic visualizations for temporal networks.
- Case Study: Collaboration networks in academic research.
- Importance: Making complex data comprehensible.

Practical Applications of Network Science

- Social media analysis for targeted marketing.
- Epidemiology: Mapping disease outbreak clusters.
- Infrastructure: Resilient design in technological systems.
- Academia: Analyzing co-authorship and citation networks.
- Business: Market segmentation and consumer profiling.
- Future directions: Cross-domain applications.

Case Study - Zachary's Karate Club

- 34 members split due to organizational conflict.
- Analysis revealed two main clusters.
- Algorithms successfully predicted real-world split.
- Visualization: Color-coded communities pre- and post-split.
- Implications: Understanding organizational dynamics.
- Lessons: Real-world validation of community detection methods.

Case Study - E. coli Metabolic Network

- Hierarchical clustering revealed functional modules.
- Identified pyrimidine metabolism as a key module.
- Biological insights: Nested community structures.
- Visualization: Dendrogram reflecting biochemical roles.
- Applications: Pathway analysis and drug discovery.
- Demonstrates scalability of network algorithms.

Conclusion - The Power of Communities

- Communities are key to understanding complex systems.
- Applications span social, biological, and technological fields.
- Algorithms are evolving to handle dynamic and overlapping networks.
- Visualization tools enhance interpretability and insight.
- Network science connects theory with real-world impact.
- Future focus: Scalability, dynamic networks, and hybrid models.

Introduction to Advanced Algorithms

- Advanced techniques for large-scale networks.
- Balancing computational efficiency with accuracy.
- Addressing challenges: Resolution limit, scalability, and overlaps.
- Emerging methods: Louvain, Infomap, random walk algorithms.
- Applications in multi-layer and dynamic networks.
- Importance: Handling real-world complex systems.

Louvain Method: Step-by-Step

- Assign each node to its own community.
- Merge communities to maximize modularity locally.
- Reassign merged communities as single nodes.
- Repeat until modularity stabilizes.
- Advantages: Fast, scalable, suitable for massive networks.
- Application example: Analyzing social media graphs.

Infomap Algorithm

- Based on information theory and entropy.
- Goal: Minimize the description length of a random walker's path.
- Output: Optimal community partitions.
- Advantages: High accuracy for overlapping networks.
- Challenges: Computationally demanding for dense networks.
- Application: Mapping information flow in web graphs.

Random Walk-Based Algorithms

- Explore network by probabilistically traversing links.
- High probability of staying within dense community regions.
- Example: Personalized PageRank for local clustering.
- Applications: Search engines, recommendation systems.
- Advantages: Captures natural flow dynamics in networks.
- Limitations: Requires fine-tuning for optimal performance.

Overlapping Community Detection

- Real-world networks exhibit overlaps (e.g., social groups).
- Methods: Clique Percolation, link clustering, and probabilistic approaches.
- Example: Overlapping cliques in collaboration networks.
- Visualization: Shared subgraphs and Venn diagrams.
- Applications: Social networks, biological systems, and IoT.
- Challenges: Defining thresholds for overlaps.

Combining Algorithms for Hybrid Approaches

- Integrating multiple detection techniques.
- Example: Modularity optimization + random walks.
- Hybrid models for multi-layered and temporal networks.
- Benefits: Improved scalability and adaptability.
- Applications: Supply chains, energy grids, and traffic systems.
- Tools: Frameworks supporting algorithm customization.

Temporal Networks and Their Dynamics

- Networks change over time due to evolving connections.
- Algorithms for tracking dynamic communities.
- Example: Disease spread models in real-time.
- Visualization: Animated graphs and temporal clustering.
- Applications: Epidemics, social media trends, and finance.
- Challenges: High computational demand for live updates.

Case Study - Social Media Networks

- Clustering users based on interactions and content.
- Examples: Twitter hashtag communities, Facebook groups.
- Algorithms used: Louvain, Infomap, and sentiment clustering.
- Insights: Identifying influencers and echo chambers.
- Applications: Targeted marketing and political campaigns.
- Challenges: Rapidly changing network structure.

Case Study - Biological Pathways

- Communities in protein-protein interaction networks.
- Applications: Drug target discovery and disease linkage.
- Example: Hierarchical modules in E. coli metabolism.
- Tools: Dendrograms and adjacency matrices for visualization.
- Insights: Nested structures reveal functional relationships.
- Future directions: Personalized medicine through network analysis

Community Detection in Citation Networks

- Mapping academic influence and collaboration.
- Communities reveal topical clusters in research.
- Example: Physics collaboration network with modularity M = 0.713.
- Visualizing co-authorship and citation patterns.
- Applications: Identifying leading researchers and trends.
- Tools: Bibliometric analysis and modularity optimization.

Dynamic Visualization Techniques

- Node-link diagrams with temporal animations.
- Tools: Gephi, Cytoscape, and custom graph APIs.
- Highlighting community evolution over time.
- Example: Tracking hashtags across social platforms.
- Benefits: Clearer insights into changing networks.
- Challenges: Handling large datasets visually.

Combining Insights Across Domains

- **Social Networks:** Understanding behaviors and trends.
- **Biological Networks:** Disease pathways and drug discovery.
- **Technological Systems:** Infrastructure optimization.
- Bridging disciplines through shared methodologies.
- Tools: Multi-layer networks and hybrid algorithms.
- Future opportunities: Cross-domain collaboration.

The Future of Community Detection

- Incorporating machine learning for prediction.
- Adaptive algorithms for dynamic, multi-layer networks.
- Real-time processing for IoT and sensor systems.
- Visualization advancements for large-scale graphs.
- Emphasis on interdisciplinary research and applications.
- Focus: Scalability, accuracy, and real-world impact.