Chapter 2

Network data across fields

We seek to understand network data in part because *networks are everywhere*. Whenever a system has interacting elements, a network may be a useful representation of it. Even when there is no obvious network structure, often network structure emerges from the interactions in the system. As a result, most complex systems can be described as networks. The network framework makes itself useful.

For instance, consider again the problem from Ch. 1 of predicting the 3D structure of a protein. Structure prediction has been a famous outstanding challenge for biologists until recently, when DeepMind's AlphaFold made an incredible breakthrough [226]. AlphaFold has been touted[1] as the most impactful contribution of deep learning because finding the 3D structure of a protein has been a bottleneck in the investigation of protein functions and interactions. At first glance, there is no obvious network structure we can see in this problem—it's about understanding the complex energy landscape of various possible 3D molecular structures, right? Indeed, at the basic level it is about finding a 3D configuration of a long chain of amino acids. But at the same time, the problem is about effectively predicting the *long-range* connections between amino acids. A pair of amino acids can be separated by hundreds of other acids in the chain, yet be neighbors in the folded protein. So, how should we represent this "neighboring" relationship? Yes, you can represent it as a weighted network between amino acid residues! One of the key components of AlphaFold's machinery is exactly predicting this network structure.

Beyond proteins, networks are pervasive throughout biology and in fact all fields of research. In this chapter we'll discuss examples of networks and the data used to measure them through biology, neuroscience, sociology, economics, and more. We'll also identify several ***focal points***—network data we will take with us throughout this book as working examples and use cases. These networks were chosen because they collectively capture most of the variations that we want to cover and because they are easy to grasp and understand.

[1] AlphaFold's success surprised many. Scientists have struggled with the structure prediction problem for nearly 50 years. "I never thought I'd see this in my lifetime," said John Moult, cofounder of the protein structure prediction competition that AlphaFold won [258]. In 2022, DeepMind announced that AlphaFold had solved the structures of over 200 million proteins, nearly every protein known to exist [90]. And as successful as it has been and promises to be, AlphaFold is not the end of the story, but the beginning [316].

https://doi.org/10.1017/9781009212601.004 Published online by Cambridge University Press

2.1 Biology

Networks come into play at all levels of biological research. Organismal processes are fundamentally governed by network processes, and a major research effort of biologists has been gathering, at scale, the data necessary to describe such networks. Here are some examples.

Networks in the cell

We can find many networks across a wide range of scales in biological systems, from interaction networks between microscopic molecules to ecological networks between living organisms. Let's talk about some of these networks.

First, consider the essential parts of biology's *central dogma*. All of the most critical molecules in living organisms—DNAs, RNAs, and proteins—interact with each other. As mentioned earlier, each protein molecule can be thought of as a weighted network between amino acids, connected by the long-range connections created when folding the protein. These proteins are *translated* from RNA molecules, which are the result of the *transcription* process from DNA molecules. All of these processes are executed by tiny protein machines called protein complexes.

Now we already have many choices on how to operationalize the network between these molecules. We can focus only on the physical interactions between individual proteins (e.g., will these two proteins stick together or not?), or we can focus on the underlying *genes* by considering interactions between genes *mediated* by the proteins. The former is called the *protein–protein interaction network* (PPI) and one example of the latter is the *genetic regulatory network* (GRN). A gene can either promote or inhibit the expression of another gene, depending on what the protein that it generates can do.

Proteins perform many different tasks in living organisms. Some proteins regulate the production of other proteins. Some act as catalysts for the synthesis or breakdown of molecules, governing whether a certain metabolic reaction is possible or not. Some proteins are for the structure of the cells. Some proteins carry messages.

Protein networks capture interactions in a very real sense: molecules fitting together to make complexes. But networks can be more phenomenological as well. Genetic interaction networks are one example. In these networks, two genes are connected based on how the deletion of one or the other (or both) affects the organism. If knocking out both genes produces a surprisingly devastating impact on the organism, then we can say that the two genes are *interacting*, maybe through a process of compensation.

Other molecules synthesized and controlled by the activities of the organism's genes and proteins are worth studying as networks. The *metabolic network* captures how metabolites are transformed along synthesis pathways. Here nodes represent metabolites and links represent the transformation of one metabolite into another as part of a synthesis pathway. Often times diseases manifest as dysfunction of these pathways, and comparing the networks with and without disease can guide us to understand the disease's effects and even inform possible drug targets or other medical treatments.

In all these networks, we are reducing the complex interactions that involve DNA, RNA, and proteins and other molecules by focusing only on one type of element at a time because it simplifies the picture a lot. It is also possible to model different types

of elements together, which may be closest to reality, but this also introduces a lot of complexity to our data. And of course, with so many possible interactions, the data we need is vast and, as rich as current data are, we are still just beginning to see the larger picture of the cell.

Network neuroscience

Going up the scale, we can examine interactions between cells. One of the most obvious examples would be the neuronal network of the brain. A brain is a dynamic network of neurons that communicate with each other using chemical and electrical signals.

The brain can be modeled in many different ways. At the base level, there is an actual network of neurons that are connected by synapses. This is of course very difficult to measure because the synapses are microscopic. Yet surprisingly, scientists have brain network data at the level of individual neurons, using methods to collect data about individual connections between neurons. For instance, the connectome project takes pictures of extremely thin slices of a brain, which are computationally aligned to map out connection structure from across the slices. Other methods use multi-electrode arrays and calcium imaging to show the spiking of neurons over time. And scientists have developed methods to infer the network structure based on the timing of spikes across different neurons.

Perhaps the most famous example of a complete neuron-level brain network is that of *C. elegans*, a small (about 1 mm) nematode. It does not have many cells and neurons, and nematodes with identical genes develop the same neuronal network. In heroic effort, scientists have mapped every cell in the nematode, including the neurons and the connections between the neurons. They froze the worm and created very thin slices across the body and painstakingly followed every neuron across slices. This work, in combination with other mapping of *C. elegans* genetics, development, and more, led to the 2002 Nobel Prize for Physiology or Medicine.

Other methods can map the brain network but only at a larger scale. Instead of working at individual neurons, we can think of brain regions that roughly correspond to certain brain functions as nodes. Now each node can be millions or billions of neurons. But what are the edges? Again the edges can be defined in many ways, based on the data available. One way to define and measure edges is from physical connections between regions in the brain. Methods like *diffusion tensor imaging* (DTI) allow us to follow large bundles of axons that wire different parts of the brain together. By inferring these axon bundles from measurements, we can "connect" brain regions. This network is often called the "structural connectome."

But there is also a completely different way to measure connections, called the "functional connectome." Here, instead of examining physical connections, we measure the activities of brain regions with *functional magnetic resonance imaging* (fMRI). If two regions fire together often, it probably means that they are functionally connected. This is the idea behind the functional brain network.

Network neuroscience continues to expand its data. Gene expression data is now often incorporated into network studies, and brain networks are often compared to sociodemographic information using large-population brain imaging studies such as the massive ABCD (Adolescent Brain Cognitive Development) study [97].

Networked ecology

Networks enter into ecology in a variety of ways. One example is *trophic networks*, more commonly known as food webs. In an ecosystem's trophic network, nodes represent species and links represent the "exchange of carbon"—a pleasant euphemism that often means members of one species *eat* members of the other. Food webs have an intriguing hierarchical structure starting from apex predators at the top and flowing down to organisms that survive on detritus—leftover, usually decaying, organic matter—or on photosynthesis. Capturing data on these trophic interactions is challenging: ecologists must often conduct long-running surveys to determine if one species indeed feeds off another. This labor-intensive research leads to precious data, especially when compounded by the critical need to understand and maintain the health of ecosystems[2] in the face of ongoing climate change.

Another network commonly studied is one of plants and plant pollinators, for example flowers and bees. Here nodes represent organisms (species), either plants or pollinators, and a link exists from a plant to a pollinator if that pollinator is known to pollinate that plant.[3] Such a network is often called a *mutualistic network*. This is fundamentally an experimental network: field observations are made tracking how often different pollinators were caught *in flagrante delicto* with different plants. Combining these observations, perhaps with some data cleaning or other processing, and a plant–pollinator network is made. Studying these networks in different regions and over time tells us about the health of those ecosystems, including their resilience to shocks such as climate change.

2.2 Socioeconomic systems

Our society is also a huge network—our social network. Broadly speaking, nodes in the network are the people in the society and edges are social relationships. But there are countless ways to think about and define specific social networks, especially the social ties. Maybe we are interested in whether two people *know* each other or not; sometimes we want to map the network of *close friends*; or perhaps we wish to understand proximity, the network of individuals who live geographically close to one another. How about family ties? How about shared social groups such as clubs? There are many ways to define a network and map social ties to edges. What definition or definitions we choose should be carefully considered and strongly motivated by the research question at hand.

How are social networks commonly measured? Historically, sociologists relied on surveys, questionnaires, and interviews. Visiting with individuals, they would ask, "Are you friends with X?" or perhaps, "List your 10 closest friends." Already we can see the power of the surveyor: the form of the question can strongly dictate the final network. If they ask, "List the 10 people you spend the most time with" you may get a very different network than asking about close friends. Likewise, why stop at 10 social ties? This will truncate the network in many ways, filtering out casual acquaintances or weak ties. Yet, one of the most famous results in social networks tells us weak ties are

[2] For one, much of our food supply is at stake.

[3] An example of a bipartite network.

https://doi.org/10.1017/9781009212601.004 Published online by Cambridge University Press

important to understand: job seekers often find employment opportunities not from their close, or strong ties, who likely have similar social circles and thus access to the same information, but weak ties [190]. The manual methods for surveying social networks were the backbone of social network research, but the scale of social network makes it difficult to gather sufficient volumes of data.

The rise of both computing and the Internet has changed studies of social networks. Survey data no longer need to be processed by hand but can be analyzed automatically with computers. More importantly, many new sources of data that do not require laborious surveys are available. Mobile phones are very popular devices, and billing records managed by telecommunications companies, tracking who-calls-whom, are a valuable source of communication interactions. And the advent in recent decades of online social networks, platforms where users sign in to share and consume information, give even richer sources of social network data.[4] The social ties are essentially collected automatically from building the friends and "followers" lists of users. Massive amounts of data are now available.

While online social networks are a boon for researchers, they also change the situation in difficult ways. Online social interactions can be quite different from interactions in the real world. A close tie online may not be a close tie in real life, and vice versa. The behaviors people display online may be quite different from real life, and vice versa. The set of users of an online platform may not be representative of all people, often favoring wealthier and more tech-adept people. It may not be possible to tell if one person is using different accounts, or if different accounts actually belong to the same person. Users of the platform may not even be people at all—bots, automated accounts, are quite common. Both measuring people, the nodes, and measuring social interactions, the links, can be fraught. Any inferences about the "real" social network that a researcher draws from the online data may not hold.

Beyond social networks, many other networks play roles in socioeconomic systems. Economic and financial systems are driven by networks at all scales. From the social side, we can study the social relationships between individuals associated with different companies. One way is to build a network where nodes represent company board members and people are connected if they serve on at least one board. Many strategic relationships between companies are associated with shared board members. Most companies are required to publish the memberships of their boards, making this data publicly available.

Another example of a socioeconomic network is a labor flow network. Here nodes consist of job seekers and companies while links exist when a job seeker is employed by a given company. This network evolves in time, tracking the movements between companies as people switch jobs, and can reveal interesting structure among different sectors of the larger labor market [358]. Although not publicly available, data for this network is now tracked by online employment platforms where companies post job openings and job seekers submit applications.

Lastly, the stock market and other financial industries are ripe for network analysis. Using data on worldwide trade, we can build a network between countries based on what the countries are trading and to whom, giving us network insights into global

[4] There can also be some serious privacy concerns with such data (Ch. 3).

trade. Networks can also be extracted from stock market data, where nodes represent traded companies and links exist between companies whose trading prices are, in some manner, heavily associated or correlated over time. (Extracting an underlying network from time series data is a very common task.) This network can reveal connections between companies when their stock prices move together. The banking sector provides even more opportunities for networks. A network between banks, for instance, tracking who holds assets in what, lets us study the stability or robustness of the banking sectors [47]. This became important when recovering from the 2008 financial crisis because it can inform, based on the risk of different asset portfolios, where and how far economic shocks can propagate.

2.3 Other fun networks

Networks arise in countless other contexts. Here are just a few possibilities.

One example is the flavor network (Fig. 1.4). Here nodes represent the ingredients that go into foods, and links exist between ingredients that share chemical compounds. Exploring the structure of this network and how it relates to recipes, sets of foods, may help us discover novel food pairings, new and under-explored recipes [6].

Another network comes from the design of electronic circuit boards and integrated circuits. Here nodes represent electrical components such as resistors and capacitors and links exist between nodes that are electrically connected. Circuit design uses such networks to calculate current flows, voltages, and such, but the network is also spatial in that it must be laid out on a circuit board, and laying out the components so that the electrical connections (the edges) are as short as possible is a challenging design problem. This design is made even more difficult when you realize that edges cannot cross, which may be impossible on a planar circuit board. To lay out such a circuit requires using multiple boards stacked in layers. But using many layers makes the device more expensive to build, and connections between layers are more costly than connections within a layer. This points us to designs that minimize the number of layers and maximize the number of edges within layers.

At the very opposite end of the size range from such microelectronics are infrastructure networks such as the power grid. For the power grid, nodes represent power generators and consumers while links represent electric transmission (load flow) between nodes, often in the form of high-voltage long-distance transmission lines. The modern power grid is exceptionally reliable, but blackouts do occur. The advent of renewable resources such as solar panels and wind farms promises to make controlling the grid more difficult, as these power sources are not controllable like coal or nuclear power plants.

In an entirely different domain, language and linguistics provides fertile ground for networks. Consider networks where the nodes are words. The thesaurus: edges denote words that are synonyms. Word association: directed edges denote words that people respond to when prompted by, "what's the first word you think of when I say the word *X*"? Word co-occurrence: edges denote words that appear next to one another in written documents. This last network has been a major data contributor when constructing *large language models*, machine learning models that can respond to and create convincingly

https://doi.org/10.1017/9781009212601.004 Published online by Cambridge University Press

natural written language.

On the subject of machine learning, *neural networks* have become the dominant method of making predictions. The nodes of a neural network represent areas where data (numbers) are aggregated (summed) and transformed using some type of (nonlinear) *activation function*, in analogy with the "integrate-and-fire" model of biological neurons. Links exist when the output of one node serves as an input to another. These inputs are often modified using a weighted sum, with the weights being parameters that we learn by "training" the network: passing data with known output through the network and examining the network's final output, learning algorithms can adjust the weights to guide the output to match known results. The overall organization of the network is called its *architecture*, and neural networks can be designed to solve many problems by the right choice of architecture. Neural networks can be studied using network science tools and neural networks can be used to study other networks, which we'll explore later in this book.

Of course, the sky is the limit when it comes to networks. Their ubiquity is yet another reason why they are such valuable, important objects of study.

2.4 *Focal Points*: networks used throughout this book

Let's pick some networks for our journey. They should be interesting, representative of certain domains and characteristics, and manageable. These focal networks will be referred to throughout the text, grounding our discussion of real-world issues and practices.

Data for each network is available online at cambridge.org/network-data. Later chapters will work through how to use and study these data.

Zachary's Karate Club

Almost no treatment of networks is complete without some reference to the famous Karate Club [505]. This small network was gathered through surveys by Wayne Zachary during the early 1970s. It captures members of a university martial arts club who interacted heavily outside the club, according to Zachary's data. What's interesting about this network, and what has driven its long-running popularity, is that the members of this club had a disagreement and split into two groups, one focused on the club president and the other focused on the club's karate instructor. These groups are visible in the network's structure prior to the split, making the network a test case for group identification methods.

Plant–pollinator network

Our second focal point is a plant–pollinator network [40]. Here the nodes fall into two groups: pollen-spreading organisms such as bees, and plants who are pollinated by those organisms. Links in the network connect only plant to pollinator, capturing field observations of that pollinator acting to pollinate that plant. This condition, where nodes fall into two groups and links exist only between—not within—the groups, is the

https://doi.org/10.1017/9781009212601.004 Published online by Cambridge University Press

definition of a ***bipartite network***. Bipartite networks such as this one are often studied in ecology. One type of study is to examine differences in the network over time due to climate change, invasive species, and so forth. This particular network was collected from field observations conducted in Spain, and the data includes metadata: the species names associated with each node in the network.

Developer collaboration network

This focal point is a network representing software developers contributing to open source projects hosted by IBM on the GitHub online development platform [27]. Nodes represent developers (identified by their GitHub usernames) and links connect developers who have edited one or more source code files in common, a simple measure of collaboration. We treat the network as weighted by associating with each link a weight counting the number of files commonly edited by the two developers. This makes the network a "projection"[5] of a bipartite network between developers and source code files. This network is also dynamic, evolving over the years 2013–2017.

Flavor network

Mentioned before, this network is derived from a reference text describing what flavor molecules are present in different food ingredients [6]. Food chemists use this reference when devising new flavor additives. But we can use the network to understand better the quality and nature of different recipes (combinations of food ingredients). While cooking is a highly multidimensional process, with preparatory steps, cooking temperature, aroma, and other factors playing important roles in taste, these flavor molecules provide a quantitative starting point to understanding flavor. Indeed, the *pairing hypothesis* states that foods that share many flavor molecules are more likely to taste well together than foods that share few molecules. Testing this hypothesis using a large set of recipes, Ahn et al. [6] found that indeed the hypothesis holds, but more for Western cuisine. East Asian cuisines tend to avoid pairing foods that share compounds. With these data, network analysis can help drive the study of systems gastronomy.

Human Reference Interactome

Our next focal point is HuRI: the Human Reference Interactome [283]. Here nodes in the network represent proteins and links exist between proteins that were observed to interact, according to high-throughout assay experiments. HuRI is the result of a decades-long effort to map out the human *proteome*, the interaction network of human proteins. At the time of this writing, HuRI is the most complete protein–protein interaction (PPI) network to date. Nodes in the network are represented by standardized IDs. A researcher interested in these data can enrich their study with node *metadata*, in

[5] A *projection* of a bipartite network is one where two nodes in the projected network were connected to the same node in the bipartite graph, that common node being absent in the projection. A bipartite network can be projected onto either "side," either set of nodes. For example, a network of film actors where two actors are connected if they costarred in any movies is the projection of an actor–movie network onto the actors.

this case using standard GENCODE gene annotations, a "controlled vocabulary" that biologists use to describe information about the protein.

Malawi Sociometer Network

This network came from a study that asked individuals in a village in Malawi to wear small proximity sensors on their chests as they went about their day-to-day business [353]. These proximity sensors can detect and record the presence of other sensors worn by study participants. Tracking what sensors are near one another and when leads to a contact network between participants that changes over time. Here nodes in the network are study participants (sensor wearers) and a link is noted when the two corresponding sensors have detected one another in close proximity. We treat this network both as a static and a dynamic network (Ch. 15), with the static network made by summing the number of contacts observed between participants over all time. (In other words, an edge in the static network represents the total number of contacts between two individuals.) This focal point also illustrates how gathering and studying network data gives rise to ***ethical concerns*** (Ch. 3): the authors of the original study took care to acquire informed consent from study participants.


We will use a "bolt" symbol (⚡) in the margin when discussing a focal network.

2.5 Summary

All fields of science benefit from gathering and analyzing network data. This chapter has summarized only a small portion of the ways networks are found in research fields thanks to increasing volumes of data and the computing resources needed to work with that data. Epidemiology, dynamical systems, materials science, and many more fields than we can discuss here, use networks and network data. We'll encounter many more examples during the rest of this book.

Bibliographic remarks

Networks pervade biology. For a influential review in the context of cell biology, see Barabási and Oltvai [37]. In the area of neuroscience, readers may be interested in Bassett and Sporns [44] for a review of network neuroscience, or the more expansive *Networks of the brain* [442]. For those interested in ecological studies, consider Pascual and Dunne [361], Proulx et al. [380], and Bascompte [41].

Networks have been a part of sociology from the very beginning, dating all the way back to Jacob Moreno's pioneering work [317]. In many ways, the standard text for social network analysis remains Wasserman and Faust [485]. With the rise of the Internet and new data sources, sociology has kept up, with the new field of computational social science arising [264]. For a exciting general audience overview of social science and these new data, consider Salganik [412].

https://doi.org/10.1017/9781009212601.004 Published online by Cambridge University Press

Readers interested in other areas may wish to consult Ahn et al. [6] for the flavor network study; Baker [31] or the now classic work of Mead and Conway [303] for an overview of circuit board design, known as VLSI (very large scale integration); Chu and Iu [105] for a review of the power grid (and the "smart grid") from a network perspective; or Cong and Liu [114] for a review of human language as a network.

Exercises

2.1 Collecting data on networks is costly, which was especially limiting before computers and computerized data collection. Suppose you are surveying a group of 100 students to learn about the social network of their school. It costs $X = \$10$ to interview each student, during which you ask them to list their 10 closest friends. Later, it costs, $Y = \$2$ to validate each reported social link.

(a) How much will it cost to collect and validate the data? Do interviews or link validations contribute more to the total survey cost?

(b) One student may list another as a friend but the other student may disagree, leading to a social link that is not *reciprocated*. If a link only needs to be checked once, regardless of whether student i listed j as a tie or j listed i, how will the survey's cost change based on how often friendships are reciprocated?

2.2 (**Focal network**) The flavor network captures whether food ingredients share chemical compounds. We could also define a network based on recipes, where nodes represent recipes and links exist between recipes that have common ingredients. While the flavor network itself is not *multilayer* (Sec. 1.4), if a recipe network were brought in, we could think of it as such.

(a) How can we connect the layers together, meaning how can we place links from nodes in one network to nodes in the other?

(b) More generally, would a combined flavor–recipe network be worth studying? Speculate on some ways the second "layer" of the network may relate to the first. What scientific questions can we investigate with this combined network?

2.3 (**Focal network**) The plant–pollinator network is a bipartite network. The developer collaboration network *comes from* a bipartite network. That the two networks share similarities in how they are defined, despite coming from entirely different research areas, is intriguing. Speculate on some similarities and differences between the two networks, think of ways to compare them directly, and describe some hypotheses that may come from drawing a kind of "analogy," broadly speaking, between one network and the other.