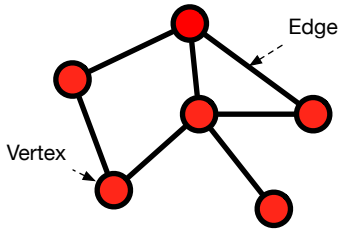
1 What are networks?

A **network** or a **graph** is a collection of discrete entities and the set of interactions among them. We call the entities **vertices** or **nodes** (or sometimes sites or actors), and we call the interactions **edges** or **links** (or sometimes bonds or ties). Any system that we can describe as being composed of identifiable nodes and definable links can be modeled and analyzed as a network.[1]



When using networks, the two most fundamental questions to answer are:

1. *What is a vertex?*
   The answer defines the set V of discrete entities or objects, among which edges exist.

2. *What is an edge?*
   The answer defines the set E of *pairwise* interactions[2] among the vertices, i.e., $E \subseteq V \times V$.

But, for any particular system—say, social interactions among people—there are often multiple ways of answering these two questions. In a social network where vertices are people, we can define multiple *types* of edges, each denoting a different kind of social interaction, such as trust, or friendship, or intimacy, or even just appearing together in a photograph. In a biological network where nodes are genes, an edge might denote a regulatory interaction, or a binding affinity between the corresponding proteins, or even a similarity in terms of evolutionary history. How we answer the two fundamental questions shapes the *kind* of descriptive, predictive, or causal questions we can ask about the underlying system.

[1] Historically, the study of graphs stretches back at least as far as Euler and his 1736 solution to the famous Königsberg Bridge puzzle. Prior to the 20th century, graphs were mainly the domain of mathematicians, and thus the term "graph" has a somewhat mathematical connotation to it. *Graph theory*, for instance, is a branch of mathematics concerned with the mathematical properties of different mathematical families of graphs. During most of the 20th century, sociologists were the main developers of *social network analysis*, which has a more empirical connotation. In the very late 20th century, in part because the computer revolution made it easier to measure, store, and analyze large network data sets, *network science* emerged as an interdisciplinary field, drawing on sociology, computer science, statistics, machine learning, and statistical physics for methods, and with applications in nearly every imaginable field, from science to the humanities.

[2] An "edge" can also be defined as a k-wise interaction, for $k > 2$. Such a network is called a *hypergraph*, denoting these "higher-order" interactions. Examples of hypergraphs include networks of actors and the films they appear in, and scientists and the papers they coauthor. In these notes, edges are typically only pairwise.

Networks are everywhere

Networks appear in nearly every domain of natural, social, and artificial phenomena. The table below illustrates a small portion of that diversity, and the variety of answers to the two fundamental questions. Note that in several cases, for the same underlying system, we can answer the questions differently, producing different kinds of network representations of a single system.

domain	*network*	*vertex*	*edge*
biological	metabolic network	metabolite	metabolic reaction
	protein-interaction network	protein	bonding
	gene regulatory network	gene	regulatory effect
	drug interactions	drug	*in vivo* health interaction
	connectome	neuron	synapse
	physiology	muscles and bones	physical attachment
	pollination network	plants and pollinators	pollination
	food web	species	predation or resource transfer
social	friendship network (offline)	person	friendship, trust, etc.
	friendship network (online)	account	"friendship," follow, etc.
	proximity network	person	physical proximity
	sexual network	person	intercourse
	coauthorships	authors	collaboration
	fictional	character	co-appearance
	animal behavior	animals	interaction
economic	hiring network	workers and jobs	hired into
	international trade	country	trade flow
	purchasing	users and items	purchased
	board of directors	directors and boards	sits on
	inventions	inventors and patents	authored
information	software	function	function call
	World Wide Web	web page	hyperlink
	documents	article, patent, legal case	citation
	artifacts	item, document, concept	relatedness or similarity
	language	word	adjacency in text
technological	Internet (1)	computer	IP network adjacency
	Internet (2)	autonomous system (AS)	GBP connection
	digital circuits	logic gates	wire
	power grid	generating or relay station	transmission line
transportation	rail system	rail station	railroad tracks
	road network (1)	intersection	pavement
	road network (2)	named road	intersection
	airport network	airport	non-stop flight

Networks are models

When answering the two fundamental questions, it's important to remember that a network is a **representation** or a **description** of an underlying system. Sometimes, a network representation is a better approximation than in others, e.g., a network can be a fairly good description of both a system of roads and a system of power transmission lines. But, a network is probably a poor

representation of the stars in a galaxy, and captures only some aspects of friendships among people. Similarly, in molecular signaling networks, some signals are mediated by conglomerations of several proteins, each of which can have its own independent signaling role. A network representation might be a poor model of the underlying signaling system because proteins can interact with other proteins either individually or in groups, and that behavior is difficult to represent as simple pairwise interactions. Throughout the use and study of networks, it is important to keep this fundamental point in mind: networks are models.

Network have structure
In the table above, each example network is also tagged by one of six scientific **domains**: biological, social, economic, technological, information, or transportation. These are not mathematical categories, but are rather a rough taxonomy of the kind of underlying system the network models. The domain labels answer the question of what kind of phenomenon are the nodes and edges modeling? The six-domain taxonomy used here originates from the *Index of Complex Networks* (`icon.colorado.edu`), which is a large index of network datasets, organized by domain and **subdomain**, e.g., online vs. offline for social networks.

Biological networks, for instance, include networks of molecules, genes, cells, tissues, and entire species, and are studied across nearly all life-science fields, e.g., molecular biology, microbiology, developmental biology, physiology, neuroscience, ecology, and evolutionary biology.

Social networks include all different kinds of social interactions among people or organizations, except for those that are explicitly economic in nature.

Economic networks represent economic interactions, e.g., financial transactions, preferences, and relationships, get their own economic networks category.

Information networks is a broad category, including both web graphs, software graphs, and document networks, all of which are defined by citation-like interactions, as well as semantic networks, where edges denote abstract or ontological relationships. This category also includes networks based on pairwise similarity or relatedness scores that do not obviously fall into some other category.

Technological networks capture systems fundamentally grounded in technology, and especially computer technology, such as the Internet or various other kinds of electronic communication networks.

Finally, *transportation networks* capture systems of physical movement, such as roads, railroads, airplanes, ships, etc., but they can also represent animal transportation systems, e.g., ant trails.

2 Networks are special. . . and dangerous

To understand what makes a network data different from other types of data, e.g., vector or point-cloud data, or even time-series data, let us consider what makes networks potentially *dangerous*. In this section, we will use an exploration of three dimensions of the *ethics* of working with network data to understand the unique risks and benefits of networks.

The big picture. Network data inherits all the ethical considerations of any other type of empirical data. But crucially, networks pose additional ethical complexities because network data—the edges—are explicitly non-independent with respect to the nodes: if we add or remove an edge, we may change all manner of other characteristics of the network. This non-independence implies that information about or control over one node provides information about or control over other nodes. That is, ethical concerns are not easily isolated down to single nodes alone.

2.1 Networks are maps, and maps *reveal*

The process of constructing a network data set is one of measurement and then assembly. What you have at the end of that measurement process a *map* of a system, one that gives you, the map-holder, the power to more clearly see the full environment—the whole system—in ways that no individual within the network can.

[[Figure showing network as map]]

Measuring a network, e.g., recording social interactions among people or writing down all the metabolic pathways in an organism, is an act of power. Particularly when nodes are people, measuring the network may reveal information that some individuals may prefer to remain hidden (to protect themselves, or protect an advantage). It may also reveal information that could benefit all, or just some. By measuring and recording the interactions in a network form, localized information is made global, persistent, shareable, and analyzable. In much the same way that maps gave the map holder power during any age of exploration, networks are maps, and constructing one tends to reveal patterns of connection and proximity that were formerly hidden.

Networks that relate to people can raise more clear ethical concerns, even in the very act of measuring them. For instance, sexual networks, where nodes are people and edges represent sexual interactions, can be useful for studying the spread of sexually transmitted infections. But assembling such a network typically requires asking individuals to share a list of their sexual partners with you, a researcher, and potentially whomever later has access to the data, which poses privacy risks to the individuals and to their partners. Similarly, contact tracing is a network reconstruction approach intended to combat the spread of respiratory diseases like COVID. But, it requires contacts to name others with whom they have interacted, and some interactions may be more sensitive

to others. Examples of other kinds of potentially sensitive social network interactions abound: criminal, secret or just clandestine interactions, medical interactions, genetic relatedness, economic transactions,[3] mobile phone calls, and many more. In fact, even deciding which types of interactions to record (and which to ignore) can be sensitive. (Consider what aspects of a person's medical history should or should not be recorded, and the risk that goes along with recorded information being reused, misused, or disclosed.) A useful guide to reasoning about ethics and measurement is to ask, What makes a social interaction sensitive? And, sensitive for whom?

Other networks may seem ethically free because of their subject matter. Species in a food web have no expectation of privacy, and don't care about being represented in a network. And similarly for a metabolic network, in which nodes are metabolites and edges are enzyme-catalyzed metabolic reactions. These networks exist as if they are natural objects to be studied. But, does this make them free of ethical questions? Could knowing the detailed structure of a network be used to intentionally disrupt its natural function, e.g., a local food web, or the local waterways? What if the measurement of the edges requires the sacrifice of living organisms? Etc.

Discussion questions:

- Under what circumstances can it be ethical to use network data that was obtained illegally, e.g., data sets that are leaked, or shared without consent?
- Are there types of interactions that are too sensitive to even record as a network?
- What kinds of risks to individuals are created by recording a network that includes them?
- What obligations do researchers have to the individuals in these networks?
- What are the ethical tradeoffs of using digital trace information to passively observe network interactions vs. measuring interactions by obtaining informed consent from individuals?
- Under what circumstances should a researcher share vs. protect network data?

2.2 Networks leak information

In many networks, nodes are annotated with *attributes*, i.e., metadata that lists a node's unique characteristics. In social networks, these can be physical characteristics like eye color or cancer risk, or social characteristics like breakfast preferences or voting behavior. In molecular networks, attributes may denote molecular weight, charge structure, conformation, etc.

Not all attributes of nodes correlate across edges in a network (a pattern called "homophily"), but a great many do. And when they do, we can indirectly learn something about one particular node's attributes merely by examining the revealed attributes of its neighbors. That is, homophily

[3] As of January 2025, Venmo transactions are public, by default.

implies a kind of *guilt by association*, without needing to directly observe the individual's variable. Information about a node leaks across its edges because edges are more likely to occur if two nodes' attributes are correlated.

[[Figure showing guilt by association attack]]

In most social settings, we embrace a principle of mutual autonomy across individuals—we don't get to tell our associates what information they can or cannot reveal about themselves, and they don't get to tell us the same. When this autonomy combines with the global view that a network provides of a social context, the mere presence of an edge allows our friends, family, and colleagues to probabilistically reveal information about us, and vice versa. Our association makes us a threat to our associates' private variables; in other words, *privacy is a network effect*. Information about a node that we gain by examining its neighbors is thus distinct from the information we might gain by directly inspecting a given node, since the node likely has more control over what it personally reveals, but not over what its associates reveal.

Examples of private social information that can be "leaked" in this way include sexual orientation, disease exposure, genetic predispositions, medical treatments, political views, religious beliefs, location information, economic activity, criminal activity, lies, etc. (Can you think of more?) Of course, not all variables that exhibit homophily on a network pose specific risks if revealed; although a person may not disclose their favorite breakfast cereal, it's unlikely that inferring or revealing it would harm that individual. And, for a given variable, not all risks are the same across individuals; the risks of being outed are greater for individuals living in cultures hostile to non-heterosexual behavior. And, not all variables exhibit homophily.

The strength of homophily on any particular variable at the global level, and its covariance with other variables, sets the baseline rate at which information about it leaks across edges in a network. The ubiquity of homophily implies that all information leaks on networks. In settings where networks evolve in "real time," these information leaks can enable efficient surveillance of a whole population from a relatively small number of "sentry" nodes (as in disease surveillance) or compromised nodes (as in surveillance for control).

Discussion questions:

- What distinguishes "public" vs. "private" data?
- Are some types of information okay to leak? If so, what types?
- How might we assess the degree to which some information leaks or not?
- What obligations do researchers have if they recover sensitive hidden information via network analysis? (Can you think of an example?)

- What ethical obligations do network researchers bear when developing methods that can be used to recover hidden (private) information?
- What are characteristics of network data that increase the risk of network information leaks?
- What limitations should social networking companies self-impose on using their network data to infer private information, i.e., facts about their users that their users have not disclosed?

2.3 Networks are re-identifiable

Like all data, network data can be persistent, transferrable, and recombinable. They can be copied and distributed widely, stored for long periods of time in multiple places, stripped of associated contextual information, lost, or combined with new information. How a network data set might be used long after it is collected is almost unforeseeable, and that makes it difficult to assess the potential harms or benefits that go along with difference choices for recording network information. Because networks are relational, every node in a network data set is potentially re-identifiable within an anonymized network, using information leaks via the edges, or via the unique arrangement of edges around it.

In science, we often take it as a principle that data should be open and shared, so that past results can be replicated, and new results can be obtained by carrying out new analyses on old data. This idea reflects the interests of the scientific community. But if there are risks to the individuals *in the data*, this principle can (but not always) collide with the interests of individuals, to remain anonymous. Networks pose special risks for re-identification of individuals because of information leakage, because the particular set of neighbors and their attributes greatly increases the uniqueness, and uniqueness makes it easier to re-identify each node in a network. Molecular networks are often reconstructed by combining information from multiple individuals, and re-identifying those individuals may be easier if fewer contributed their data or if some contributors are more unusual than others.

Methods for de-identifying network data is an area of active research, and how well a network can be anonymized depends deeply on which details must be preserved and which can be obscured. Even erasing all node attribute information may not be sufficient to anonymize a network! If an attacker has access to the network before it is anonymized, they could create a unique pattern of connections that they could look for after release in order to re-identify the perturbed nodes. Too much obfuscation may alter the results that come from running the same network analysis on the de-identified data (do you see how?), while too little may allow some or all nodes to be re-identified, even after anonymization. A persistent challenge is our natural failure to imagine risks and harms, especially from adversarial situations or unforeseen downstream uses or unanticipated "side" information.

Discussion questions:

- How does reidentification risk vary with the size of the network?
- Should de-identification aim to protect all members of the network, or just most of them?
- Which nodes are likely the most easily re-identified?
- Can an individual give informed consent for risks associated with re-identifying their neighbors?
- Under what circumstances should a researcher share vs. protect network data?
- What technical means can be used to reduce re-identification risk in networks?
- Under what conditions is simple node-level anonymity acceptable?
- What are characteristics of a network data set that increase the risk of reidentification?

2.4 Ethics and network analysis

In fact, network analysis itself poses ethical conundrums, depending on what network insights it produces and the actions or interventions those insights facilitate that would not have been possible before. For example, the large and growing literature on methods to predict missing network information, such as missing links or missing node attributes, is often motivated by addressing accuracy problems with network measurement. But, these methods are based on the tendency for information to leak across edges, and can enable the recovery of information that was intentionally hidden or omitted. Improving methods for predicting missing information can facilitate or amplify ethical problems in their use.

In fact, we might argue that *all* network analyses induce ethical questions because they provide a privileged view into the global structure of the system and enable network interventions. Here are some examples. Studies of different strategies for removing nodes or edges from a network in order to disrupt its connectivity may enable new kinds of attacks on social, biological, economic, transportation, or technological networks. Network-based vaccination strategies, which prioritize nodes with certain structural features or network positions over others, raise questions of vaccine equity and moral hazards. Methods for "aligning" two different networks in order to match nodes that exist in both could be used to create sophisticated structural re-identification attacks. And, any work on the "controllability" of a network's dynamics poses questions about how such control might be used. What other examples can you think of?

3 Four flavors of network analysis and modeling

There are four general approaches in network analysis and modeling: exploratory, explanatory, predictive, and causal.

Exploratory analysis is typically descriptive in nature, and its central goal is to produce a clear view of the kinds of statistical patterns that exist in a network. This approach is largely unsupervised, in the sense that we may not know exactly what we are looking for, or what is interesting about a network's structure. We often use statistical summaries of the network's structure, such as the degree distribution, its community structure, node-level measures like centrality scores or measures of degree assortativity, and more. With a new data set, we often start with exploratory analysis, even if our ultimate interest is in understand what degrees of freedom underlie and explain whatever patterns we may find. That is, the first step is often to identify what patterns are worth explaining in the first place.

The outcome of good exploratory analyses is typically one or more hypotheses about potential causal effects or underlying mechanisms that relate to a network's structure. Exploratory analysis cannot itself test those hypotheses, but it can use *null models* as a way of deciding if some pattern is interesting, e.g., by asking whether a pattern observed in a real-world network is distinguishable from "noise," which is typically operationalized through some kind of *random graph model*. We will explore this topic more in Lectures 2, 3, and 5.

Good exploratory analysis requires *creativity* (to imagine what shape the network might have, and why), *mathematical intuition* (to know what kinds of shapes are possible, and even plausible), *algorithmic tools* (to know how to see that shape and to extract it from the data), and *statistical rigor* (to show that the shape is real and not a clever illusion). Good exploratory analysis finds new and interesting patterns within empirical data, and generates questions to be addressed through a more hypothesis-driven approach.

Many exploratory network analysis tasks can be reduced to the following kind of model. We imagine that an edge (i, j) exists with probability

$$p_{ij} = \Pr(i \to j \,|\, x_i, x_j, \gamma_i, \gamma_j, \theta_{ij}) \ , \tag{1}$$

where each x represent a set of vertex-level observed attributes, each γ represents a set of vertex-level latent (unobserved) attributes, and θ is some latent attributes of the pairing of i and j. For instance, consider a pair of individuals i and j on Facebook. Each person's x contains the attributes they disclose about themselves on Facebook (age, sex, location, etc.). Their γ represents all attributes not disclosed on Facebook (including attributes that Facebook does not ask about), and θ represents latent attributes of the pair (family relationship, work relationship, etc.).

Facebook has many reasons for wanting to know p_{ij}, but they may not care about why it takes that value or how that value changes over time. But, if they knew a functional representation of Eq. (**??**) for their network, they could do many powerful things, including inferring missing attributes and

predicting missing links. The goal of exploratory analysis is, to a large extent, estimating a low-dimensional form for Eq. (**??**), i.e., a form that depends on many fewer variables than the number of vertices or edges in the network. The more compact the form, the simpler the shape of the network.

Explanatory analysis typically seeks to "explain" some observed pattern as being driven by some other, hopefully more fundamental variable. In social network analysis in sociology, this often takes the form of explaining how some attribute of a node correlates with the network structural patterns that surround that node. For instance, explaining a node's wealth as a function of its central position in the network, or explaining a node's large influence or special behavior via its network connections.

The simplest version of explanatory analysis is to convert the network itself into additional node-level features that encode a node's network characteristics, and then carry out traditional explanatory modeling, i.e., regression, between the "independent" variables and whatever dependent variable we are trying to explain. There are, however, many technical details that make this task non-trivial, most of which stem from the fact that each node's network characteristics are not independent—they are all derived from the same underlying network. In the social sciences, tools like exponential random graph models or stochastic actor-oriented models are network-based methods for explanatory modeling (but, beware [4]).

Predictive modeling aims to construct a predictive model of either node attributes (including future state variables) or structural features, using other network information as the input. Predictive modeling often uses machine learning tools to do its work, and these tools can be classification, regression, or probabilistic (often Bayesian) models.

For instance, recommendation algorithms, like on Netflix or Amazon, are really a kind of link prediction algorithm: given a set of nodes attribute representing user preferences and product characteristics, and a set of past connections between users and products, predict which connections are missing; these missing connections are the new product recommendations. If we're making recommendations only among the users (i.e., predicting missing links on the network of users), then it's like the "People You May Know" feature on Facebook, the "Suggestions for you" feature on Instagram, and the "Who to follow" feature on Twitter.

Causal modeling aims to identify cause-and-effect relations that involve networks, such as asking whether being more centrally located in a network *causes* better access to information, or whether a particular social behavior, e.g., buying something or clicking on an ad, is caused by influence

[4] Under common specifications, these "explanatory" modeling approaches can have significant pathologies that can make their outputs scientifically useless, e.g., see Shalizi and Rinaldo, "Consistency under sampling of exponential random graph models." *Annals Statistics* **41**, 508-535 (2013).

from friends' behavior or not.

Generally, causal modeling comes in several flavors. Statisticians and machine learners favor causal inference models that can be applied to an observed network and its dynamics. These techniques can be very complicated in part because they need to isolate and model the many different paths along which influence can travel from one node to another. Nevertheless, these techniques are essential when the goal is making causal claims, but the data are a network, and it is either impractical or unethical to conduct controlled experiments.

In contrast, biologists and some social scientists often favor *network experiments* to tease out causality, e.g., by knocking out an edge or otherwise inducing some change in it and then observing the subsequent effects. Network experiments can be very expensive, because the unit of replication across experiments is the entire network.

Finally, mathematicians and physicists tend to favor mathematical models and *network simulations*, where causal behavior is expressed via a mathematical mechanism, e.g., differential equations or stochastic processes, and then the predictions of the mathematical model are compared with empirical data. While these models can establish sufficiency, i.e., they assume, if the world works like this model, then its consequences are such and such, but they cannot establish necessity, i.e., they say little about alternative explanations.

Good causal modeling (and good explanatory modeling) is often hypothesis driven, meaning that we already have in mind some notion of why and how an effect of interest comes about. But genuinely establishing causality often requires *creativity* (to imagine how a network's shape could lead to the behavior of interest), *mathematical intuition and rigor* (to know what kinds of mechanisms are possible and to show their consequences), *numerical tools* (to simulate the mechanism and analyze its results), and *statistical rigor* (to show that the hypothesis is supported or not). Good hypothesis-driven analysis identifies and demonstrates believable causes for real effects, and shows that these explanations are better than simple alternatives.