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] Hence, the two most fundamental questions to answer when using networks are:

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

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

For any particular system, there may be multiple ways of answering these questions. For instance, in a social network, where vertices are people, we might define multiple different *types* of edges, which might represent different kinds of social interactions. An edge might represent friendship, or appearing together in a photograph, or being willing to borrow money from the other person, etc. Similarly, in a biological networks in which nodes are genes, an edge might represent a regulatory interaction, a binding affinity between the corresponding proteins, a similarity in terms of evolutionary history, etc. For this reason, how we answer the two fundamental questions can greatly shape the *kind* of descriptive, predictive, or causal questions we can approach about the underlying system.

The table below lists several examples of real-world networks, along with corresponding answers to these two fundamental questions. It also includes a few examples of different answers to these questions for a single system, and the different kinds of networks that gives rise to.

[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 $k > 2$, and then we call the object a *hypergraph* to denote this "higher-order" interaction. Examples of hypergraphs include collaboration networks like actors who appear in a film together or scientists who coauthor a paper together. For most of these notes, however, an edge will be defined as a pairwise interaction.

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	person	friendship (various)
	sexual network	person	intercourse
economic	employment network	person or job	employment-in
	international trade	country	trade flow
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 these questions, it's important to remember that a network is simply a **representation** or a **description** of some underlying system. In some cases, a network representation is a better approximation than in others, e.g., a road network is a fairly good description of a system of roads and a power grid is a fairly good representation of 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 domains. 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. This particular domain taxonomy 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. Networks of economic interactions, e.g., economic 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. 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. And finally, transportation networks capture systems of physical transportation, such as roads, railroads, airplanes, ships, etc., but they can also represent animal transportation systems, e.g., ant trails.

1.1 Graph properties of networks

Because networks are a way of representing an underlying system, there are many variations on the basic idea of a set of V nodes and E pairwise interactions. A given network will thus have a particular set of **graph properties**, which help define the specific aspects of the underlying system that the network captures.

For concreteness, we define a graph or network as $G = (V, E)$, where V is the set of vertices, and E is the set of edges. Each edge is a pair $i, j \in V$ such that $(i, j) \in E$.

1.1.1 Simple graphs

The most basic kind of network is called a *simple* graph, which has the following properties:

1. edges are **undirected**: a connection $(i, j) \in E$ implies a connection $(j, i) \in E$
2. edges are **unweighted**: edges are "binary," and have no weight value assigned to them
3. there are no self-loops: no edge connects a vertex to itself $(i, i) \notin E$
4. there are no annotations on the nodes: nodes are just uniquely indexed.

The following figure shows an example of a simple graph, on the left, and a more exotic (non-simple) graph on the right, which provides some examples of how additional information can be stored in a graph representation.

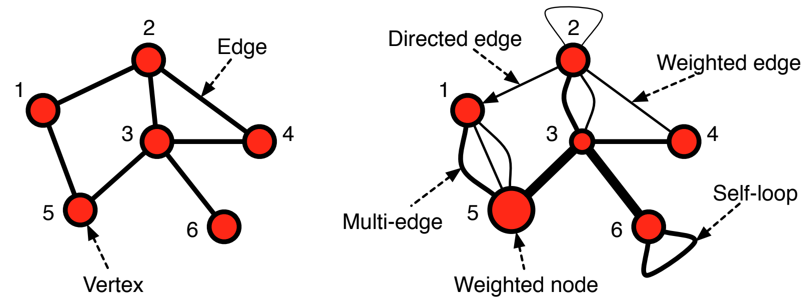


Figure 1: A simple graph (unweighted, undirected, no self-loops), and a more exotic network.

1.1.2 Non-simple graphs

When we relax one of the graph properties of a simple graph, we get a network representation that can capture additional kinds of information about the underlying system. A simple graph is called "simple" mainly because it is the closest to the basic definition of a set of discrete entities V and their pairwise interactions.

This table lists a number of common graph properties, arranged by whether the property is a function of an *edge*, a *node*, or the whole *network*. Properties with a symbol next to them ($\circ$, $\bullet$, $\star$, $\diamond$) represent grouped properties, such that a network can only have one property from that group.

edge	*node*	*network*
unweighted $\circ$	metadata	sparse $\star$
weighted $\circ$	attributes	dense $\star$
signed $\circ$	locations	bipartite $\diamond$
undirected $\bullet$	state variables	projection $\diamond$
directed $\bullet$		acyclic
multigraph		temporal
timestamps		multiplex
		hypergraph

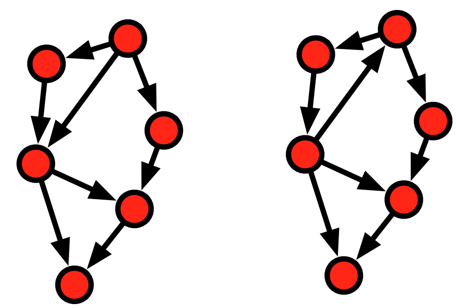
For instance, a **multigraph** relaxes the prohibition against repeated connections (and generally also the constraint on self-loops), meaning that for at least one pair $i, j \in V$, there exists a multiplicity of edges $(i, j) \in E$. If vertices represent cities, and edges represent driving paths between a pair of cities, then a multigraph will be a reasonable representation because there can be several distinct such paths between a pair of cities. Similarly, in a network of neuron cells, two neurons can have multiple synapses and we might wish to represent each such connection as a distinct edge.

In the rest of this section, we'll define and give some examples of these non-simple graph properties.

Networks with edge and node attributes
Many modern networks have auxiliary data associated with either the edges or the nodes. Edges are often **weighted**, meaning they are annotated with a scalar value or edge weight, which might represent the frequency of interaction or the interaction's strength. Edges might also be **signed** $\{-1, +1\}$ representing some amount of inhibition or activation in a biological system. Edges annotations can be arbitrarily complicated, extending to a whole vector of attributes, a list of "tags," or just a "color" or other kind of categorical variable. When the edge attributes denote a discrete point in time for that edge to exist, we say the network is a **temporal** network. In contrast, if an edge attribute denotes a continuous duration in time, we say the network has **timestamps**.

Nodes can also have **attributes** or **metadata** attached to them, and these can be categorical variables (sometimes called "labels"), single scalars or vectors representing **state variables**, or even spatial coordinates or **locations** in some metric space.[3] For example, if nodes are cities, node attributes might include the city's population and GPS coordinates. In a social network, node metadata may include age, sex, and location. In a protein-interaction network, node attributes might include the molecular weight or Gene Ontology functional labels.

Directed networks
If edges can be asymmetric, we call them **directed**, meaning that the edge (i, j) can occur independently of (j, i). Such directed edges are sometimes called *arcs*, but not always. The World Wide Web is a familiar directed network: webpages are nodes, and hyperlinks are the directed edges. Many biological networks are directed, including gene regulation and neural activation.



directed acyclic graph   directed graph

A network is **acyclic** if it is a directed network that contains no cycles, i.e., for all choices of i, j, if there exists a path $i \to \cdots \to j$ then there does not exist a path in the reverse direction

[3] It has become trendy to refer to such information as "ground truth" when one is trying to predict missing values on the nodes. However, this is wrong—node annotations are just more data, and thus should not be treated as absolute truth under any circumstances. We'll revisit this idea when we talk about community detection.

$j \to \cdots \to i$. All trees are acyclic undirected networks. When we allow directionality in the edges, non-trees can be acyclic. For instance, a citation network, in which published papers are vertices and paper i connects to paper j if i cites j in its bibliography, is a kind of acyclic directed network (at least in theory; in practice, some cycles exist). Food webs representing predation, however, are not always acyclic: some species predate themselves (cannibalism), and some pairs of species predate each other (a 2-cycle).

Bipartite networks and one-mode projections

To be a k-partite graph, where k is an integer, the following must be true. The set of vertices V is composed of k distinct classes of nodes (e.g., producers and consumers) *and* only nodes of different classes interact (consumers only interact with producers, and vice versa).[4] The simplest and most common form of such graph is the **bipartite** network, where $k = 2$. A popular type of bipartite graph is the actor-film network, in which actors and films represent the two classes, and actors connect to the films in which they play a part.[5]

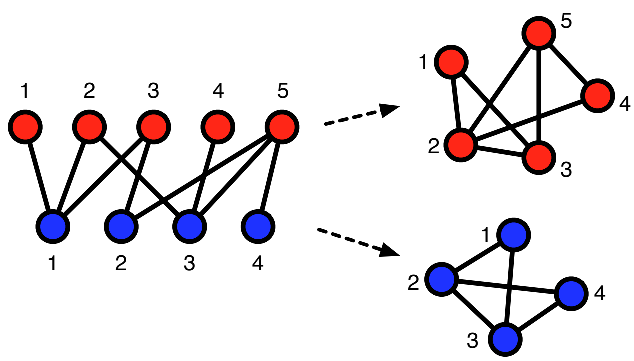


Figure 2: Example of a bipartite graph and its two one-mode projections.

Sometimes, we prefer not to work with a k-partite graph and would instead like to work with a network in which all the nodes are of the same class. This conversion is called a **one-mode projection**. In every k-partite graph, there are k one-mode projections. And, in a one-mode projection, two vertices are connected if and only if they share a neighbor in k-partite graph. For instance, to derive the actor-collaboration network from the actor-film network, we add an edge between a pair of actors i, j if they ever appeared in a film together. This procedure is equivalent to saying i, j are connected in the projection if there exists a path of length 2 in the actor-film. Projections

[4] Mathematically, a bipartite network can be defined in this way: $V = A \cup B$ where $A \cap B = \emptyset$, and $\forall_{(i,j)\in E}\,((i \in A) \wedge (j \in B)) \vee ((i \in B) \wedge (j \in A))$.

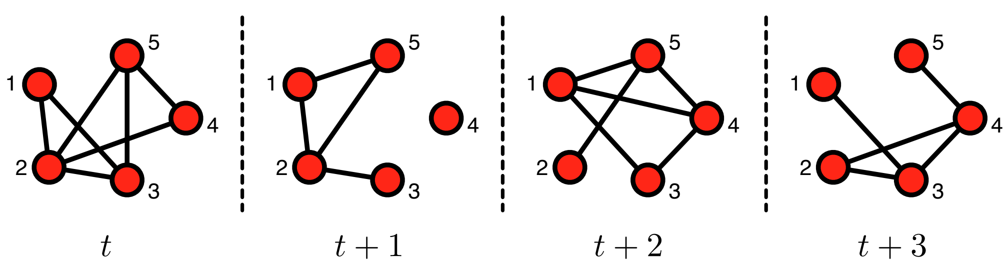
[5] If there are multiple classes of vertices, but edges can exist within each class, then we do not call it a k-partite graph. Instead, it is simply an annotated network with mixed node types.

can also be weighted, so that (i, j) in the projection is given a weight w_{ij} that corresponds to its multiplicity as a result of the projection procedure, e.g., w_{ij} would count the number of movies in which the actors i, j appeared together.

Warning: an important consequence of the one-mode projection procedure is the construction of cliques, i.e., a subgraph of size ℓ in which every pair of nodes is connected. Every node i that is being "projected through," meaning i will not be present in the projection, is represented in the projected graph as a clique of size ℓ, because all pairs of its neighbors are exactly distance two away from each other. For instance, all actors in a particular film will be joined in a clique in the one-mode actor projection.

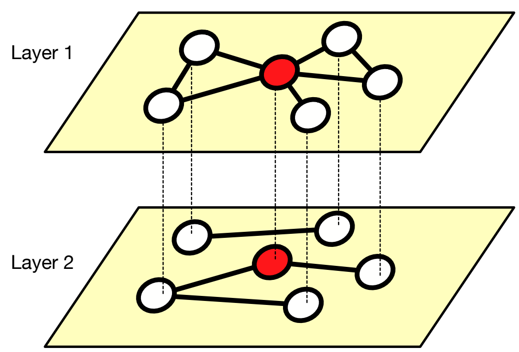
Temporal, dynamic and evolving networks.
The networks described so far are *static*, meaning that the vertices and edges do not change over time. Graphs that do change over time are an important class of networks. For example, in citation networks, new vertices join the network continuously and each time a new vertex joins, it creates new edges representing citations to older papers. If a network varies over time, we call it a *temporal* or *dynamic* or *evolving* network.

A **temporal** network typically refers to a kind of edge-annotated network in which the annotations represent points in discrete time. We often talk about temporal networks as being a sequence of network "snapshots" $A^{(t_1)}, A^{(t_2)}, \ldots$, where the superscript t_i indexes the passage of discrete time. For instance, all the interactions observed among friends on Monday, and then Tuesday, etc. It may not always be obvious in the data, but a crucial question is whether A^t represents only the interactions present at time t or the aggregation of all interactions between t and $t + 1$. Often, it is the latter.



A **time-stamped** network is one in which edges are annotated with the duration of continuous time in which they existed, e.g., (i, j, t_1, t_2) might represent a phone call or face-to-face interaction that began at time t_1 and ended at time t_2.

Multiplex networks, spatial networks, and hypergraphs.
A network in which edges are marked by which "layer" they exist in is called a **multiplex** or **multilayer** network. These networks are used to represent a system in which there are multiple types of interactions, and we store the connectivity of each type in a different "layer" of the multiplex network. A temporal network is a special kind of multiplex network, where these layers form a temporal (ordered) sequence. Crucially, there can dynamics on each vertex that govern which layer some kind of interaction occurs on, so multiplex networks are not merely a special kind of graph in which edges are annotated by different colors or layer numbers.



Spatial networks are a special kind of node-annotated network, in which the annotations represent the node's location in some d-dimensional space. This graph property is most common in transportation networks, e.g., as road and city networks, airport transportation networks, oil and gas distribution networks, shipping networks, etc., but can also appear in social networks. *Planar* graphs are a special case of spatial networks, in which the nodes are embedded on a 2-dimensional surface and edges do not cross.

Hypergraphs are another type of network, in which edges denote the interaction of more than two vertices, e.g, $E \subseteq V \times V \times V$. Scientific collaboration graphs can be represented as a hypergraph, in which each "edge" is the set of coauthors on a scientific article. However, collaboration networks are more commonly represented as bipartite graphs, in which scientists and papers form two sets of vertices, and scientist-nodes are connected to all the paper-nodes on which they are authors.

2 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. In this effort, we may 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.

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*.

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.

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.

Finally, *causal* modeling aims to identify specific cause-and-effect relations that involve networks, such as asking whether being more or less centrally located in a network *causes* better access to information. Often, 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. Biologists and social scientists may use network experiments to tease out causality, e.g., by knocking out or inducing some change in an edge. Finally, mathematicians and physicists tend to favor mathematical models, where causal behavior is embedded in the mechanics of the math, e.g., differential equations or stochastic processes, and then the predictions of the mathematical model are compared with empirical data.

3 At home

1. Peruse or skim Chapters 1–5 in *Networks* (background material)
2. Read Chapter 6.1–6.10 in *Networks*