Lecture 1: Network and Graph Basics

36-720, Fall 2016
Scribe: Brendan McVeigh

29 August 2016

1 Initial Ideology

A **network** is a configuration of *similar*, *binary* relationships among a group of things, which we will often call "actors" (if we want to sound like sociologists), "nodes" or "vertices" (if we want to sound like mathematicians) or "entities" (if we want to sound spooky). Both the "binary" and the "similar" parts are important.

Binary A *binary* relationship is one between just two people or entities, as "Joey likes Irene", "Irene hates Joey", "Mark collaborates with Duncan", "cats eat voles" or "J. P. Morgan Partners loans money to U.S. Steel". More complex relationships are ones which only make sense with three or more terms, e.g., "Irene is jealous of Joey's feelings for Karl", or "Carnegie guarantees Morgan's loan to U.S. Steel". (Coming up with 4-ary emotional relationships is left as an exercise.) Not all relationships which matter are binary.

Similar Arguably, everything in the world is related *somehow* to everything else. Even if we limit ourselves to binary relationships, if we don't limit ourselves to configurations where the relationships all have a pretty similar nature, we end up including *everything* in a single tangled heap. The result is paranoia, rather than a useful scientific theory. It would be metaphysical and unhelpful to insist on *identical* relationships (think of any two of your friends: do you really have *exactly* the same relationship to both of them?), so there's an element of judgment, as always.

Theories of everything considered unhelpful

In this course, when we talk about "networks" we mean real networks, with things in the real world. When we care about mathematical abstractions over lived reality, we are talking about *graphs*.

A **graph** (as in "graph theory") is a bunch of things (set of **verticies** or **nodes**) V, plus a set of **edges**, **ties** or **links** among them, i.e., a set $E \subseteq V \times V$, a subset of the ordered pairs of V. Figure 1 shows a very basic example.

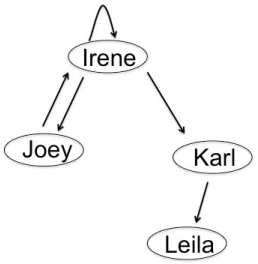

Figure 1: A directed graph with a loop

2 Very Basic Graph Definitions

Graphs *may* have the following characteristics:

- Not all possible links need be present. If all possible links are present, the graph is **complete**.
- Nodes need not be tied to themselves. A binary relation which entities have with themselves is called **reflexive**; in graph-theoretic terms, we speak of **self-loops**. Many graphs are of ir-reflexive relations, and have no self-loops.
- The relationship need not run both ways, i.e., if $(i, j) \in E$, the reverse-directed pair (j, i) may or may not be $\in E$. If $(i, j) \in E$ if and only if (iff) $(j, i \in E$, then the relation is **symmetric**, and we often speak of the graph as being **undirected**. An a-symmetric relation is not *necessarily* anti-symmetric, however.

Graphs can have numeric **weights** or other **attributes** on edges. These might indicate the strength of a relationship, or distinguish different types of relationships among the same nodes. E.g., in a graph of contacts among teachers, we might have two types of edges, one for "socializes with" and another for "seeks advice from", and weights which indicate how often these things happen. We sometimes also allow, formally, for multiple edges between the same pair of nodes (a **multi-graph**), though this can often be represented as a weighted graph.

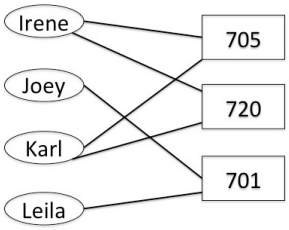

Figure 2: A bipartite graph in which oval nodes indicate students a rectangular nodes indicate classes. Edges indicate that a student is taking the class.

A **simple graph** is one which contains no self-loops and at most one, unweighted, unlabeled edge between any pair of nodes. Simple graphs can be directed or undirected. Much of the mathematical theory for graphs is tailored to simple graphs.

A **bipartite** graph is one where the relationship is between two sharply distinct kinds of things; the nodes can be divided into two kinds, two **parts** (hence "bipartite") or **modes**, and edges are only between nodes in different parts of the graph, never within a part or mode, as in Figure **??**. We might write the node set here as $V = U \cap W$, and insist that $E \subseteq U \times W$. Historically (see note below) the oldest studies of social networks in fact focused on bipartite graphs, with one mode being corporations and the other mode being robber barons (and their lawyers, bankers, etc.) who sat on corporate boards.

3 Some Properties of Graphs

To sum up, the mathematical abstractions we will work with are graphs $G = (V, E)$ where V is set of the vertices or nodes, and $E \subseteq V \times V$ is set of the edges, ties or links. Typically, we insist that E contain no self-loops.

Adjacency matrix

Since computing over sets is usually inconvenient, but computing over numerical arrays is easy, we often prefer to represent graphs as matrices. Say that there are n nodes, i.e., $|V| = n$. Fix an ordering of the nodes. Then the **adjacency matrix** A is the $n \times n$ matrix where $A_{ij} = 1$ if $(i, j) \in E$, otherwise

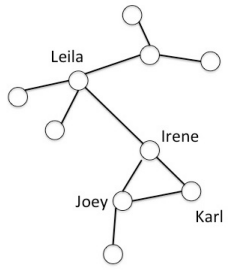

Figure 3: An undirected graph

(i.e., when $(i,j) \notin E$) $A_{ij} = 0$. The graph is undirected when A is symmetric; the absence of self-edges means all the diagonal entries $A_{ii} = 0$.

Once we have the adjacency matrix, we can easily use it to answer questions like "who does i send ties to?" (read the i^{th} row) or "who does i receive ties from?" (read the i^{th} column); as sets, these are i's **in-** and **out- neighborhoods** (its **neighborhood**, in an undirected graph).

Walks and paths

A **walk** is a linked sequence of edges between two nodes, e.g., in Figure 3, Irene-Karl and Irene-Joey-Karl are both walks. In a walk, repeating and back-tracking are both allowed. Higher powers of the adjacency matrix count the number of walks between nodes, i.e., $(A^k)_{ij}$ counts the number of walks of length k between i and j. A stricter notion is that of a **path** between nodes, which is a walk with no repeated vertices.[1]

EXERCISE: Prove that there is a path from i to j if and only if there is a walk from i to j.

Weak and strong connections

In a directed graph, the node i is **weakly connected** to node j when there is a path from i to j. (For completeness, we say that every node is weakly connected to itself.) The nodes i and j are **strongly connected** when there is also a path back to i from j. (Hence, every node is strongly connected to itself.) In an undirected graph, every pair of nodes which is weakly connected is also strongly connected, so we just say that pairs of nodes are **connected**.

As shown in Figure 4 a is weakly connected to every node except f and only strongly connected to b.

Connected components

[1] CRS got this backward during lecture.

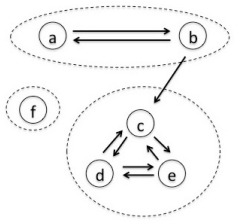

Figure 4: Groups of strong connected nodes are identified by dashed lines. a and b are strongly connected to each other but are only weakly connected to c, d, and e. f is strongly connected only to itself.

Strong connection (or, in an undirected graph, connection) is clearly a symmetric relationship. It is also reflexive, and transitive. (That is, if Irene is connected to Joey, and Joey is connected to Karl, then Irene is connected to Karl.) Since strong connection is reflexive, symmetric and transitive, it is an **equivalence relation**, and the graph partitions into **equivalence classes**, here called **connected components** (Figure 4.). Many graphs for real networks tend to either have one very large connected component, or a multitude of small connected components, for reasons we will explore when we look at our first stochastic models of networks in lecture 4.

Distance

There are multiple ways of defining metrics on graphs, but the overwhelmingly most common one, for undirected graphs, is the **geodesic distance**[2], $d(i,j) \equiv$ length of the shortest path from i to j. Of course $d(i,i) = 0$ for all i (it's a metric!), and we say $d(i,j) = \infty$ if there is no path between i and j.

Subgraphs

One graph $H = (U, F)$ is a **subgraph** of another, $G = (V, E)$, if H's node set is a subset of G's ($U \subseteq V$) *and* H's edge set is also a subset of G ($F \subseteq E$). (Said differently, every *non*-edge in G is also a non-edge in H.) Often we form subgraphs by picking a subset of the nodes of G and then including all the edges

[2] **Geodesy** is the science of measuring the shape of the Earth, hence of determining the shortest distance over the surface of the Earth between two points, hence the "geodesic path" between points in other spaces is the one minimizing some notion of distance intrinsic to the space.

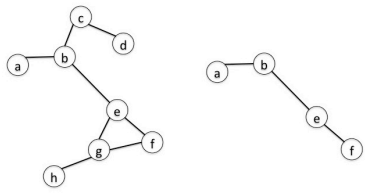

Figure 5: We can create a subgraph of the graph on the left by keeping only nodes a, b, e, and f in addition to the edges between these nodes. The resulting subgraph is shown on the right.

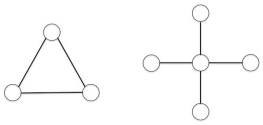

Figure 6: A *triangle* is shown on the left and the right graph is a *4-star*

between them, as done in Figure 5. We need to know abut subgraphs because they give us a very useful way of talking about repeated patterns in the graph, and (sometimes) for dividing the graph into significant parts.

Patterns, isomorphism, subisomorphism, motif

We often want to know whether a particular subgraph of G matches the pattern set by some other graph, or perhaps *how many* subgraphs of G match that pattern. (Figure 6.) "Matching" is intuitively reasonably clear, but we need something more precise for math and/or automation. We say that two graphs are **isomorphic** when there is a one-to-one mapping of their nodes which preserves edges, i.e., $G = (V, E)$ and $M = (U, F)$ are isomorphic when there is a one-to-one mapping $\phi : V \mapsto U$ such that $(i, j) \in E \Leftrightarrow (\phi(i), \phi(j)) \in F$). If a subgraph H of G is isomorphic to M, we say that G is **subisomorphic** to M. The target pattern M being matched is sometimes called a **motif**[3]. If $H \subset G$ is

[3]Note however that some authors only use the word "motif" for patterns which are, in some sense, more common than expected in G.

(sub)isomorphic to M, we can match up the nodes of H one-for-one with nodes of M, so that every edge in M has a corresponding edge in H, and no non-edge in M has a corresponding edge in H.

Examples of motifs: triangles, cliques, stars

As an example, take the **triangle** graph, consisting of three nodes with all possible edges. These are rare among most mathematically possible graphs[4], but very common in actual social networks[5]. "Finding triangles" in G means finding subgraphs in G which are isomorphic to the left side of Figure 6. Triangles generalize to **complete** graphs on more nodes; subgraphs which are (isomorphic to) complete graphs are called **cliques**, and subgraphs which cannot be enlarged without ceasing to be (isomorphic to) complete graphs are called **maximal cliques**[6]. Another graph motif which is often of interest is a **star** pattern, of two (or more) nodes which are only linked through a central node; the right-hand side of Figure 6 shows a **4-star**. Cliques show all-to-all relations, while stars are centralized patterns, so they often have very different functional relationships.

(All of this generalizes in the natural way to directed graphs and directed motifs, of course.)

Degree

The **degree** of a node in an undirected graph is the number of edges it has, $\text{degree}(i) = \sum_j A_{ij}$. The degrees of several nodes in the graph on the left in Figure 5 are

$$\begin{aligned} \text{degree}(a) &= 1 \\ \text{degree}(b) &= 3 \\ \text{degree}(e) &= 3 \\ \text{degree}(g) &= 3 \\ \text{degree}(f) &= 2 \end{aligned}$$

The **degree sequence** of a graph is a vector containing the degrees of each node, typically sorted to be either increasing or decreasing. The **degree distribution** is the probability mass function for all the degrees, i.e., the distribution of degree we would find by picking randomly and uniformly over nodes.

In a directed graph, we need to distinguish between **out-degree**, $\sum_j A_{ij}$, and **in-degree**, $\sum_j A_{ji}$. Note that the sum of all in-degrees must equal the sum of all out-degrees, because each edge comes out of one node and into another.

We will see when we consider network models, that different models can imply very different degree distributions; thus the degree distribution is *a* clue to how the graph formed. I emphasize that it is only *a* clue, because there are, unsurprisingly, some distributions which can be produced by very different

[4] A vague statement; we will be more precise in Lecture 4, on random graphs.

[5] One suggestion for why they are common in social networks is that triangles make *reputation* possible — if Irene and Joey have a common friend Karl, Karl can observe whether Irene treats Joey well or poorly, and just his behavior towards Irene accordingly. There is good reason to think that triangles are especially dense in the social networks of those engaged in risky, long-term endeavors where trust is especially important, e.g., persecuted religious minorities, scientific collaborators or criminals (Tilly, 2005).

[6] Just to be confusing, many writers drop the word "maximal", so that when they say "clique", they mean "set of nodes which are all directly linked to each other, and cannot be enlarged without adding non-edges".

models, and a fair amount of grief, or at least of bad science, has come from people reasoning along the lines "My favorite model produces such-and-such a degree distribution; this graph looks (sort of, in dim light, if you squint) like it has such-and-such a degree distribution; *therefore* the network follows my favorite model".

Network attributes are endogenous

We should close this lightning review of essential ideas from graph theory by noting that degree is not really a property of a *node*, or at the very least a very funny property. In a social network, for instance, where the nodes are people, the nodes have many attributes, like height, weight, age, education, income, etc., which they would have in any social network, or none at all[7]. We can, at least hypothetically, imagine changing these for any given node without changing the network, or even imaging manipulating them for all nodes at once. But "has degree 3" isn't an attribute of an individual node in the same way; for example, it makes no logical sense to imagine increasing the degree of only *one* node. In fact, many properties of nodes-in-networks are like degree in this way (e.g., "is not in any triangle", "is part of a connected component of size 10"). At the very least, this is going to affect our interpretation of how these attributes can work in statistical models. We are really going to want models which deal with the *whole* network, rather than regression-style models with a clear input and output; everything is going to be endogenous and depend on everything else.

4 Historical note

The only decent history of network analysis I know of is Freeman (2004), and it's written by a participant, rather than a proper historian. On an even more amateur basis, I can offer the following points:

- *Graph theory*, in mathematics, began with Euler solving the Konigsberg bridge problem in the 1700s, but this remained basically an isolated toy for a long time.
- Bi-partite networks, of influential people and organizations, were studied by social scientists and reformers (often the same people) from about the 1890s. The oldest *picture* of a social network I know of is a bi-partite graph from 1916, showing prominent Bostonians publicly opposing to the nomination of Louis Brandeis to the Supreme Court, and how they were all linked through a small number of clubs and other organizations (Rauchway, 2008).
- In the 1920s and 1930s, a psychologist named Louis Moreno promulgated the use of mono-partite social network diagrams to depict ties of friendship, rivalry, etc. His ideas were picked up by a small group of sociologists, primarily at Harvard, but with ties to quantitatively-minded social

[7] Though even there, a close inquiry into attributes like "education level" or "income" might show them to be (complicated) social relations.

scientists at Columbia and Chicago. At the latter, there was some intersection with the "mathematical biophysics" group centered on Nicholas Rashevsky[8]. The sociologists kept elaborating the notion of social networks through the 1950s.

- In the 1950s, several mathematical workers — Solomonoff and Rapoport (1951) out of Rashevsky's group; Erdös and Rényi (1960); Gilbert (1959) from Bell Labs — worked out the basic mathematics of random graphs, as we will study in lecture 4. In violation of priority, but in accordance with what sociologists of science call the "Matthew Effect", these came to be called "Erdos-Renyi graphs". This initiated an intense and long-term study of random graphs within pure mathematics.

- Non-random-graphs became objects of mathematical interest in part because of new applications, many of which arose from logistics, (weighted) graphs being natural ways of representing transportation links. Questions like "which links would we have to remove to render the graph disconnected?" were not just points of idle curiosity but of real military (and economic) concern; creating a communications network which would by design be very hard to disconnect was one of the original impulses that led to the Internet[9]. Computer networking was another obvious and important stimulus towards studying networks *in general*.

- Social network analysis emerged as a recognizable sub-field of sociology during the 1970s and 1980s, with its own specialized vocabulary, journals, conferences, internal squabbles, etc.

- Physicists, and certain sorts of physicist-influenced biologists and social scientists, came to an interest in networks during the 1980s and 1990s, largely via studying systems of coupled oscillators, "spin glasses", and neural networks[10]. Watts and Strogatz (1998) marked the point where this interest became detached from studying dynamics *on* networks to studying the structure *of networks*, of all kinds (not just social), and was quickly followed by a flood of work by physicists and ex-physicists.

[8] Rashevsky was one of the most eccentric, and subtly influential, figures in the development of mathematical methods for biological and social phenomena. There is, so far as I know, no proper biography of him, or even a decent article-length study, though he fully deserves one. If you have a chance, ask Prof. Fienberg about him.

[9] One might thus say that the impulse to put everything "on the cloud", i.e., on centralized server farms, undermines the whole *point* of the Internet, but that's another story for another time.

[10] Though the concept of a neural network is much older. Sherrington (1906) already wrote of "networks" of neurons, and the computational properties of such networks were studied by McCulloch and Pitts (1943).

References

Erdös, P. and A. Rényi (1960). "On the Evolution of Random Graphs." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5**: 17–61. Reprinted (Newman *et al.*, 2006, pp. 38–61).

Freeman, Linton C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, British Columbia: Empirical Press.

Gilbert, E. N. (1959). "Random Graphs." *Annals of Mathematical Statistics*, **30**: 1141–1144. doi:10.1214/aoms/1177706098.

McCulloch, Warren S. (1965). *Embodiments of Mind*. Cambridge, Massachusetts: MIT Press.

McCulloch, Warren S. and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." *Bulletin of Mathematical Biophysics*, **5**: 115–133. Reprinted in (McCulloch, 1965, pp. 19–39).

Newman, Mark, Albert-László Barabási and Duncan J. Watts (eds.) (2006). *The Structure and Dynamics of Networks*, Princeton, New Jersey. Princeton University Press.

Rauchway, Eric (January 2008). "Vast Right-Wing Conspiracy (with picture)." URL `https://edgeofthewest.wordpress.com/2008/01/28/vast-right-wing-conspiracy-with-picture/`.

Sherrington, Charles (1906). *The Integrative Action of the Nervous System*. New Haven, Connecticut: Yale University Press.

Solomonoff, Ray and Anatol Rapoport (1951). "Connectivity of Random Nets." *Bulletin of Mathematical Biophysics*, **13**: 107–117. Reprinted (Newman *et al.*, 2006, pp. 27–37).

Tilly, Charles (2005). *Trust and Rule*. Cambridge, England: Cambridge University Press.

Watts, Duncan J. and Steven H. Strogatz (1998). "Collective Dynamics of "Small-World" Networks." *Nature*, **393**: 440–442. doi:10.1038/30918.