The network, as a metaphor, is ubiquitous in historical research. The formal analysis and use of network methods will be an incredible boon to historians, adding nuance and precision to an already important metaphor. We turn to network analysis and visualization in our final two chapters, in the hopes that we can foster *better* network analyses and visualizations in the humanities.

Chapter 6

Network Analysis

Networks and network visualizations are becoming increasingly important in digital humanities and digital history research. There are, however, very few resources meant for historians and the particular challenges that network analysis poses for us. In this chapter, we cover the fundamentals of network analysis so that the approach can be used in your own research — and you will know how to critically read others' use of networks in theirs.

Formal networks are mathematical instantiations of the idea that entities and connections between them exist in consort. They embody the idea that connectivity is key in understanding how the world works, both at an individual and a global scale. Graph theory, social network analysis, network science, and related fields have a history dating back to the early 18th century, cropping up in bursts several times since then. We are currently enjoying one such resurgence, not incidentally co-developing along with the popularity of the Internet, a network backbone connecting much of the world to one system.

The idea that relationships are essential to understanding the world around us is, of course, ancient. The use of formal network methods for historical research, however, is much more recent, with only a few exceptions dating back beyond 30 years. Marten Düring has aggregated a thorough multilingual bibliography at http://historicalnetworkresearch.org. This chapter will go over a few examples of how historians have used networks, in what situations you might or might not want to use them, and the details of how networks work mathematically and technically.

In the 1960s, Eugene Garfield created the "historiography," a technique to visualize the history of scientific fields using a network of citations or

historical narratives laid out temporally from top to bottom.[1] Garfield developed a method of creating historiographs algorithmically, and his contemporaries hoped the diagram would eventually be used frequently by historians. The idea was that historians could use these visuals to quickly get a grasp of the history of a discipline's research trajectories, either for research purposes or as a quick summary in a publication.

A citation analysis by White and McCann looking at an 18th century chemistry controversy took into account the hierarchical structure of scientific specialties.[2] The authors began with an assumption that, if two authors both contributed to a field, the less prominent author would always get cited alongside the more prominent author, while the more prominent author would frequently be cited alone. One scientist is linked to another if they tend to be subordinate to (only cited alongside of) that other author. The resulting networks, called entailograms, proved particularly useful in showing the solidification of a chemical "paradigm" over a period of 35 years. Lavoisier begins as a lesser figure in 1760 and eventually becomes the most prominent chemist by 1795; by that time, most chemists who were cited at all were cited alongside Lavoisier. Following the entailogram over time reveals conflict and eventual resolution.

Citation analysis, whose practitioners are also called bibliometricians or scientometricians, is a rapidly growing discipline. Despite the hopes of its founders and though many of its practitioners conduct historical research, historians rarely engage with the field.

Historical sociologists, anthropologists, economists, and other social scientists have been using formal network methods for some time and tend to exchange ideas with historians more readily. One such early sociological work, by Peter Harris, also employed citation analysis, but of a different sort. Harris analyzed citations among state supreme courts, looking at the interstate communication of precedent from 1870–1970.[3] By exploring who relied on whom for legal precedent over the century, Harris showed

[1] Eugene Garfield (1973), "Historiographs, Librarianship, and the History of Science," in *Toward a Theory of Librarianship: Papers in Honor of Jesse Hauk Shera*, Conrad H. Rawski (ed), Metuchon, NJ: Scarecrow Press, pp. 380–402.

[2] Douglas R. White and H. Gilman McCann (1988), "Cites and Fights: Material Entailment Analysis of the Eighteenth-Century Chemical Revolution," in *Social Structures: A Network Approach*, Barry Wellman and Stephen D. Berkowitz (eds), Cambridge, UK: Cambridge University Press, pp. 380–400.

[3] Peter Harris (1982), "Structural Change in the Communication of Precedent among State Supreme Courts, 1870–1970," *Social Networks*, **4**(3), 201–212.

that authority was originally centralized in a few Eastern courts but slowly became more diffuse across a wide swath of the United States. Network approaches can be particularly useful at disentangling the balance of power, either in a single period or over time. A network, however, is only as useful as its data are relevant or complete. We need to be extremely careful when analyzing networks not to read power relationships into data that may simply be imbalanced.

In a study on 19th century women's reform in New York, Rosenthal *et al.* revealed three distinct periods of reform activity through an analysis of organizational affiliations of 202 women reform leaders.[4] These 202 women were, together, members of over a thousand organizations, and the researchers linked two organizations together based on how many women belonged to them both. The result was a network of organizations connected by the overlap in their member lists and a clear view of the structure of women's rights movements of the period, including which organizations were the most central. The study concludes, importantly, by comparing network-driven results to historians' own hypotheses, comparing its strengths and weaknesses with theirs. For research on organizations, network analysis can provide insight on large-scale community structure that would normally take years of careful study to understand. As much as networks reveal communities, they also obscure more complex connections that exist outside of the immediate data being analyzed.

The study of correspondence and communication networks among historians dates back centuries, but its more formal analysis is much more recent. The *Annales* historian Robert Mandrou[5] and the historian of science Robert A. Hatch[6] both performed quantitative analyses of the Early Modern Republic of Letters, exploring the geographic and social diversity of scholars, but neither used formal network methods. In a formal network

[4] Naomi Rosenthal, Meryl Fingrutd, Michele Ethier, Roberta Karant, and David McDonald (1985), "Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State," *American Journal of Sociology*, **90**(5), 1022–1054.

[5] Robert Mandrou (1978), *From Humanism to Science 1480-1700*, 2nd edn, Atlantic Highlands, NJ: Humanities Press.

[6] Robert Alan Hatch (1998), "Between Erudition & Science: The Archive & Correspondence Network of Ismaël Boulliau," in *Archives of the Scientific Revolution: The Formation and Exchange of Ideas in Seventeenth-Century Europe*, Michael Cyril William Hunter (ed), Woodbridge: Boydell & Brewer.

study of Cicero's correspondence, Alexander and Danowski[7] make the point that large-scale analyses allow the historian to question not whether something exists at all, but whether it exists frequently. In short, it allows the historian to abstract beyond individual instances to general trends. Their study looks in 280 letters written by Cicero; the network generated was not that of whom Cicero corresponded with, but of information generated from reading the letters themselves. Every time two people were mentioned as interacting with one another, a connection was made between them. Ultimately the authors derived 1,914 connections between 524 individuals. It was a representation of the social world as seen by Cicero. By categorizing all individuals into social roles, the authors were able to show that, contrary to earlier historians' claims (but more in line with later historians), knights and senators occupied similar social and structural roles in Cicero's time. This is an example of a paper that uses networks as quantitative support for a prevailing historical hypothesis regarding the structural position of a social group. Studies of this sort pave the way for more exploratory network analyses; if the analysis corroborates the consensus, then it is more likely to be trustworthy in situations where there is not yet a consensus.

In what is now a classic study (perhaps the only study in this set relatively well-known beyond its home discipline), Padgett and Ansell used networks deftly and subtly to build a historical hypothesis about how the Medici family rose to power in Florence.[8] The authors connected nearly 100 15th century Florentine elite families via nine types of relations, including family ties, economic partnerships, patronage relationships, and friendships. Their analyses reveals that, although the oligarch families were densely interconnected with one another, the Medici family — partially by design and partially through happy accident — managed to isolate the Florentine families from one another in order to act as the vital connective tissue between them. The Medici family harnessed the power of the economic, social, and political network to their advantage, creating structural holes and becoming the link between communities. Their place in the network made the family a swing vote in almost every situation, giving them a power that eventually gave rise to a 300 year dynasty.

[7]Michael C. Alexander and James A. Danowski (1990), "Analysis of an Ancient Network: Personal Communication and the Study of Social Structure in a Past Society," *Social Networks*, **12**(4), 313–335.

[8]John F. Padgett and Christopher K. Ansell (1993), "Robust Action and the Rise of the Medici, 1400–1434," *American Journal of Sociology*, **98**(6), 1259–1319.

Before Facebook and MySpace, the first network of people to come to mind would probably be kinship or genealogical networks with linkages between family members. Historiographic studies of these networks stem from early prosopographical (or collective biography) methods, but explorations using social network analysis are more recent. Looking at a large town in southwest Germany in the early 19th century, Lipp explored whether and how the addition of an electoral system affected the system of kinship networks that previously guided the structure of power in the community.[9] Surprisingly, in an area to become known for its democratic reforms, Lipp showed that a half century of elections had not reduced the power of kinship in the community — in fact, kinship power only became stronger. Lipp also used the network to reveal the prominent actors of local political factions and how they connected individuals together. In this case, networks were the subject of study rather than used as evidence, in an effort to see the effects of political change on power structures.

Trade networks are particularly popular among economists but have also had their share of historical studies. Using the records of nearly 5,000 voyages taken by traders of the East India Company between 1601 and 1833, Erikson and Bearman show how a globalized economy formed out of ship captains seeking profit out of the malfeasance of private trade.[10] Captains profited by using company resources to perform off-schedule trades in the East, inadvertently changing the market from a dyadic East–West route to an integrated and complex global system of trading. The authors used a network as evidence, in this case the 26,000 links between ports each time two were connected along a trading route. Over 200 years, as more ports became connected to one another, the East India Company lost control in a swath of local port-to-port connections. The authors show that the moments at which private trade was at its peak were also critical moments in the creation of more complex trade routes. While network analysis is particularly powerful in these many-century longitudinal studies, they also must be taken with a grain of salt. Without at least a second dataset of a different variety that is connected to the first, it is difficult to disentangle what effects were caused by the change in network structure, and what effects

[9]Carola Lipp (2005), "Kinship Networks, Local Government, and Elections in a Town in Southwest Germany, 1800–1850," *Journal of Family History*, **30**(4), 347–365.

[10]Emily Erikson and Peter Bearman (2006), "Malfeasance and the Foundations for Global Trade: The Structure of English Trade in the East Indies, 1601–1833," *American Journal of Sociology*, **112**(1), 195–230.

were merely external and changed both the network and the effect being measured. Global networks also tend to entangle geographic and relational distances, a fact which should not be glossed over when trying to understand the lived experiences of historical actors, which may diverge greatly from a network representation.

Folklorists have a long tradition of classifying folktales based on types, motifs, and various indices in order to make finding, relating, and situating those tales easier on the scholars balancing thousands upon thousands of tales. These schemes are often inadequate to represent the multidimensional nature of folktales, such as a tale that is classified as being about manor lords, but also happens to include ghosts and devils as well. Tangherlini and colleagues[11] came up with a solution by situating a collection of 19th century Danish folktales in a network that tied tales to subjects, authors, places, keywords, and the original classification schemes, resulting in a network connecting 3,000 entities together by 50,000 ties and made them easily browsable in an online interface. The interface made it significantly easier for folklorists to find the tales they were looking for. It also aided in serendipitous discovery, allowing scholars to browse many dimensions of relatedness when they were looking at particular tales or people or places.

Lineage studies with networks are not limited to those of kinship. Sigrist and Widmer[12] used 1,000 18th century botanists, tracing a network between masters and disciples, to show how botany both grew autonomous from medical training and more territorial in character over a period of 130 years. The authors culled their group of botanists from various dictionaries and catalogues of scientific biography, and found by connecting masters to disciples, they saw botanists from different countries had very different training practices, and the number of botanists who traveled abroad to study decreased over time. The study juxtaposes a history of change in training practices and scientific communities against traditional large scientific narratives as a succession of discoveries and theories.

As is clear, historical network analysis can be used in a variety of situations and for a variety of reasons. The entities being connected can be articles, people, social groups, political parties, archaeological artefacts, stories,

[11] James Abello, Peter Broadwell and Timothy R. Tangherlini (2012), "Computational Folkloristics," *Communications of the ACM*, **55**(7), 60.

[12] René Sigrist and Eric D. Widmer (2012), "Training Links and Transmission of Knowledge in 18th Century Botany: A Social Network Analysis," *Redes: Revista Hispana Para El Análisis de Redes Sociales*, **21**, 347–387.

and cities; citations, friendships, people, affiliations, locations, keywords, and ship's routes can connect them. The results of a network study can be used as an illustration, a research aid, evidence, a narrative, a classification scheme, and a tool for navigation or understanding.

The possibilities are many, but so too are the limitations. Networks can be dangerous allies; their visualizations, called graphs, tend to be overused and little understood. Ben Fry, a leading voice in information visualization, aptly writes:

> There is a tendency when using graphs to become smitten with one's own data. Even though a graph of a few hundred nodes quickly becomes unreadable, it is often satisfying for the creator because the resulting figure is elegant and complex and may be subjectively beautiful, and the notion that the creator's data is "complex" fits just fine with the creator's own interpretation of it. Graphs have a tendency of making a data set look sophisticated and important, without having solved the problem of enlightening the viewer.[13]

It is easy to become hypnotized by the complexity of a network, to succumb to the desire of connecting everything and, in so doing, learning nothing. The following chapter, beyond teaching the basics of what networks are and how to use them, will also cover some of the many situations where networks are completely inappropriate solutions to a problem. In the end, the best defense against over — or improperly — using a network is knowledge; if you know the ins and outs of networks, you can judge how best to use them in your research.

Network Analysis Fundamentals

As we showed in the previous section, networks are versatile tools for exploring our connected world. Unfortunately, the rhetorical utility of networks can occasionally overshadow their formal nature. It is not uncommon to read scholarly articles that invoke the vocabulary of networks without acknowledging any underlying formal definitions, rendering a metaphorically rich argument devoid of any real substance. While network analysis is frequently not the most appropriate method of analysis, when it *is* invoked appropriately, it should be treated with a healthy respect for the many years

[13] Ben Fry, *Visualizing Data: Exploring and Explaining Data with the Processing Environment* (Sebastopol, CA: O'Reilly Media, 2007).

of research that have gone into building its mathematical and conceptual framework.

The first part of this section introduces that framework, beginning first with basic vocabulary, progressing through mathematical means of measurement at both the small and large scale. Part two describes how to put the concepts in action, including what network data look like and how to manipulate or visualize them. Part three discusses when network analysis is appropriate and when it is not.

Basic Concepts and Network Varieties

Nodes, edges, and attributes

Networks, also called graphs, are the mathematical answer to the assumption that many things in this world worth studying are *interdependent*, and need to be treated as such. Despite their name, networks are *dichotomous*, in that they only give you two sorts of things to work with: entities and relationships. Entities are called **nodes** (Fig. 6.1) and the relationships between them are called **edges** (Fig. 6.2). Everything about a network pivots on these two building blocks.

Nodes can be all sorts of things, from ideas to people to bricks. Because networks have cropped up in many different disciplines over the years, all with their own vocabularies, you may see nodes referred to by many names including: *vertices*, *actors*, *agents*, or *points*. They can be used interchangeably.

Edges connect nodes (Fig. 6.2). They may represent the amount of words two books share, a friendship between two people, or a similar

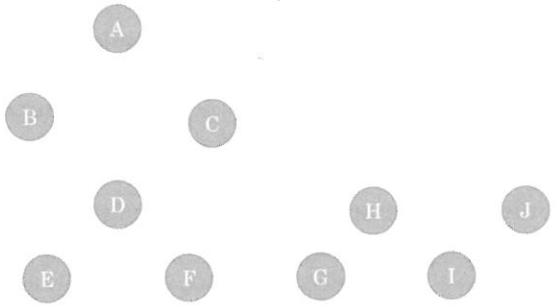

Fig. 6.1 Nodes.

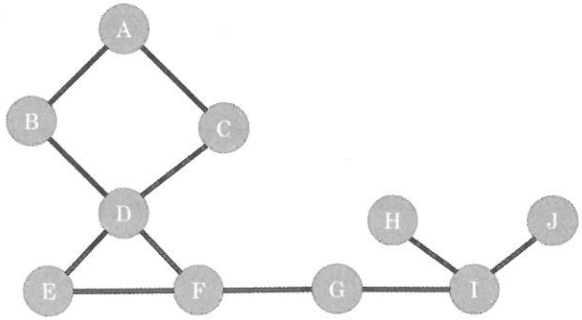

Fig. 6.2 Edges connecting nodes.

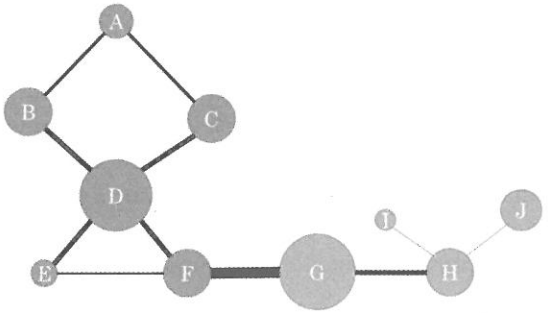

Fig. 6.3 Nodes and edges with attributes (depicted by the size of the node and the weight of the edge).

originating location connecting two museum objects. With very few exceptions, formal network representations allow for edges between two, and only two, nodes. If siblings Alice, Bob, and Carol share a kinship connection, three edges will exist (connecting Alice to Bob, Bob to Carol, and Alice to Carol), rather than one edge connecting all three. You may see edges referred to as *arcs*, *links*, *ties*, or *relations*.

Over the years, new methods have allowed researchers to use more than simple node and edge representations in analyses. Often, edges and nodes can take on any number of individual **attributes** (Fig. 6.3). In a network where nodes are cities and edges are the travel routes between them, each city can contain attributes like its population or primary spoken language, and each edge can have attributes like travel time or mode of transportation.

Static and dynamic networks

Many networks one encounters appear to be **static** snapshots of a moment in time: one large visualization of nodes connected by edges, whether it be about people or books or ideas. In reality, the underlying networks often represent a **dynamic** process that evolves and changes over time. Networks of ports based on shipping routes necessarily change with economic and weather conditions, and a static snapshot could only lead to a picture of dozens of ports with docked ships, and occasional ships at sea, but all unconnected to one another.

Network analysts generally deal with these issues in one of three ways:

1. Aggregate all of the data into one giant network representing the entire span of time, whether it is a day, a year, or a century. The network is static.
2. Slowly build the network over time, creating snapshots that include the present moment and all of the past. Each successive snapshot includes more and more data, and represents each moment of time as an aggregate of everything that led up to it. The network continues to grow over time.
3. Create a sliding window of time, e.g. a week or five years, and analyze successive snapshots over that period. Each snapshot then only contains data from that time window. The network changes drastically in form over time.

Each method has its benefits and drawbacks, and historians should think carefully about which would best suit their method of inquiry.

Isolates, dyads, and triads

The smallest unit of meaningful analysis in a network is a **dyad**, or a pair of nodes connected by an edge between them (Fig. 6.4). Without any dyads, there would be no network, only a series of disconnected **isolates**. Dyads are, unsurprisingly, generally discussed in terms of the nature of

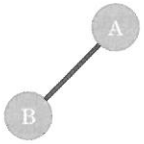

Fig. 6.4 A dyad.

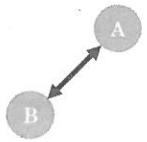

Fig. 6.5 A dyad engaged in a reciprocal relationship.

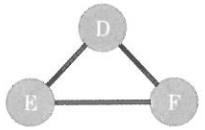

Fig. 6.6 A triad.

the relationship between two nodes. The dyad of two medieval manuscripts may be strong or weak depending on the similarity of their content, for example. Studies of dyads in large networks most often revolve around one of two concepts: **reciprocity**, whether a connection in one direction is reciprocated in the other direction, or **assortativity**, whether similar nodes tend to have edges between them.

A dyad of two Twitter users may be considered reciprocal if both parties follow one another, and an entire Twitter network may have a higher or lower **reciprocity** overall depending how many dyadic relationships tend to be reciprocal (Fig 6.5).

Assortativity, also called **homophily**, is the measure of how much like attracts like among dyads in a network. On the Web, for instance, websites (nodes) tend to link (via edges) to topically similar sites. When dyads connect assortatively, a network is considered **assortatively mixed**. Networks can also experience **disassortative mixing**, for example when people from isolated communities with strong family ties seek dissimilarity in sexual partners.

A **triad** is a group of three dyads, and another frequent unit of network analysis. The various possible configurations of three nodes are invoked surprisingly often in network analyses, even when looking at networks with millions of nodes, and is more versatile than one might expect (Fig 6.6).

For example, if brick type A and brick type B are often found at the same archaeological site, as are brick types A and C, how often are brick types B and C found at the same site (thus connecting the A–B–C triangle)? Asking that question of all the archaeological sites found in a several hundred mile

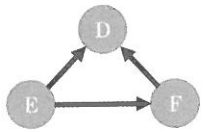

Fig. 6.7 A triad demonstrating transivity.

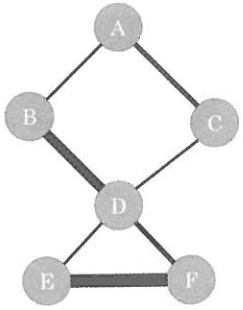

Fig. 6.8 A weighted network.

radius might yield the degree to which a trade economy was tightly knit. **Transitivity** is the concept that when A is connected to B and C, B and C will also be connected (Fig 6.7). Some networks, like those between friends, feature a high degree of transitivity; others do not.

In an evolving network of correspondents, if Alice writes to Bob, and Bob to Carol, we can ask what the likelihood is that Alice will eventually write to Carol (thus again closing the triangle). This tendency, called **triadic closure**, can help measure the importance of introductions and knowing the right people in a letter-writing community.

Weighted versus unweighted

One attribute sometimes associated with edges is **weight**. Weight is a numeric value quantifying the strength of a connection between two nodes. Most often, the higher the weight of the edge between them, the more similar or closely connected the two nodes; occasionally, though, edge weight can be used to quantify dissimilarity rather than similarity. Weight is usually connoted by line-thickness in visualizations.

Weighted networks, as they are called when they include weighted edges (Fig. 6.8), can take many forms. A network of people connected by

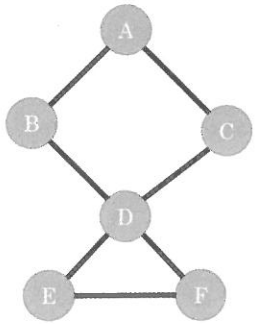

Fig. 6.9 An unweighted network.

whom they call on the phone can be weighted by the total number or length of phone calls. A network of scientific articles can be weighted by the total number of words each shares with the other. A network of activist organizations can be connected based on how many members they share. **Unweighted networks** can include the network of websites that link to each other, the US electrical grid, or a kinship network in a family tree (Fig. 6.9).

Most common algorithms used to measure things about networks (finding the most central or important nodes, calculating network density, etc.) don't play well with attributes of nodes or edges. Attributes like type or category or name are a very difficult fit into algorithmic formalisms. Edge weight is an exception to this rule. Many network measurement algorithms can take edge weight into account when performing their calculations. The calculation finding the most central person in a phone-call network should potentially change if two people occupy the same structural place in the network (keep in contact with the same people) but one makes calls vastly more frequently than the other.

Directed versus undirected

The second-most common attribute of edges is directedness. A **directed edge** is one that is part of an asymmetrical relationship, and an easy way of thinking about them is by imagining arrow tips on the edges (Fig 6.10).

Citation networks, for example, can make **directed networks**. A dozen papers written after 2005 can cite Mark Newman's famous 2004 article on

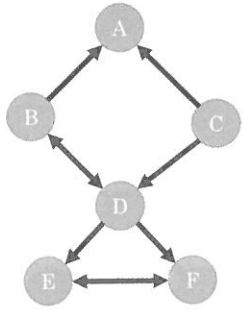

Fig. 6.10 A network with directed edges.

finding communities in networks,[14] but that does not imply that Newman's paper cites them back. Twitter follower and followee relationships form a similarly asymmetrical network; some follows are reciprocated, but others are not. Trade networks may sometimes be asymmetrical, with goods not always traveling in both directions.

Similarity networks are generally symmetrical and thus **undirected**; *Brave New World* is just as similar to *1984* as *1984* is to *Brave New World*. Facebook is also undirected, as in order for you to choose someone as a friend, they too must choose you as a friend. Highway networks are undirected, as they go in both directions, but local city road networks may be directed, as there can be a mix between one- and two-way streets.

Unlike with weighted networks, on which you can often safely use an algorithm made for unweighted networks and still get methodologically plausible results, you must be extra careful when analyzing directed networks. The directionality of an edge can have huge repercussions to a network's structure, and so algorithms made for undirected networks to find local communities or node importance might produce very unlikely results on a directed network. Be careful that you only use algorithms made specifically for directed networks when analyzing them.

Bipartite and k-partite

Basic networks, and most network algorithms, can only support one type of node per network. You can connect books that influence other books, or people who co-author together, but it has been (until now) a methodological

[14] M. E. J. Newman (2004), "Detecting Community Structure in Networks," *The European Physical Journal B — Condensed Matter and Complex Systems*, **38**(2), 321–330.

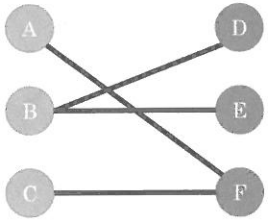

Fig. 6.11 A bipartite network.

faux pas to connect authors and books in the same network. You cannot connect apples and oranges. This is not an immediately obvious limitation, but because of the assumptions underlying most network analyses, adding multiple types of nodes in the same network can severely limit the use of the dataset. This is not a theoretical limitation, but a practical one; in general, network scientists have not yet created many algorithms to deal with these **multimodal networks**.

However, there are situations — especially when dealing with complex humanities datasets — when this multimodality can be useful. The most basic example is a **bipartite network**, also called a *bimodal* or *2-mode* network (Fig. 6.11). As the name implies, these networks support two types of nodes. For mathematical reasons to be described later, in bimodal networks, you must be careful to only allow edges *between* types and not allow edges *within* types. Taking the most recent example, this means you can draw edges from books to authors, but not between authors or between books.

Generalizing from there, some graphs can be **multimodal** or ***k*-partite**. An italic k is just a network scientist's way of saying your-favorite-number-goes-here. Like with bipartite networks, k-partite networks feature nodes of various types that can only link between — not among — each other. These networks might represent vastly complex trade networks, with nodes as books, authors, publishing houses, cities, and so forth. There are even fewer algorithms out there for dealing with k-partite networks than there are for bipartite ones, so although these varieties of networks are very useful for the initial planning and preparation of your data, you will need to reduce the complexity of the network before running any of the analyses included with most network software packages.

Directed or bipartite network transformations

The previous sections suggested that analyses on directed or bimodal networks require slightly more forethought and a more nuanced understanding

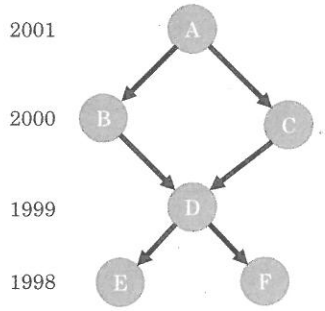

Fig. 6.12 A citation network as a directed network.

of the underlying algorithms. Although there are workflows for working with these more complicated network types, there are also methods of transforming them into simpler networks that are easier to work with, and in certain circumstances, more useful.

A citation network is an example of a directed network that, for various reasons, tends to be difficult to analyze (Fig. 6.12). For example, articles always cite earlier work (the directed arrows pointing toward the past). Because citations go to previous work rather than contemporary similar research, it winds up being difficult for algorithms to find communities of practice in the network.

Bibliometricians, people who study citations, have developed two reciprocal methods of transforming a directed, past-facing citation network into an undirected network of either current communities of practice or articles from the past that should be categorized together.

Bibliographic coupling networks are networks of scholarly articles that are connected if their bibliographies are similar. When two articles reference a common third work in their bibliographies, they get an edge drawn between them; if those two articles share ten references between them, the edge gets a stronger weight of ten. A bibliographic coupling network, then, connects articles that reference similar material and provides a snapshot of a community of practice that emerges from the decisions authors make about whom to cite. It does not change over time: articles do not change their bibliographies as time goes by.

This method is generalizable to any sort of directed network, not just citations. In practice, any time you have a directed network, you can connect two nodes if their edges point to the same third node.

Co-citation networks are the functional opposite of bibliographic coupling networks. In this case, two scholarly works are connected if they appear together in a bibliography, and the more times they are cited together, the stronger the edge between them. As opposed to bibliographic coupling networks, co-citation networks connect articles not by the choices their authors make, but by the choices future authors make about them. Also unlike bibliographic coupling networks, co-citation networks change over time: two articles may live long, separate lives before some author decides to cite them both in the same third article. In this way, co-citation networks are an evolving picture of various scholarly communities' disciplinary canons; not a view of the past as it saw itself, but a view of the past as the present decides to connect it together.

Co-citation networks are also generalizable well beyond the realm of scholarly articles; in a directed network, if two edges point from one node to two others, those two nodes get an undirected edge drawn between them. For example, in a family tree where directed edges point from parents to children, performing a co-citation analysis would produce an undirected network that connects siblings together.

Bipartite networks, those with two types of edges, can similarly be transformed into **unipartite networks**, with only one type of node, in a method quite similar to co-citation or bibliographic coupling. If you have a bipartite network of scholarly societies as one type of node, connected by edges to their members, the other type of node, you can collapse the network into a unipartite one by connecting people who are members of the same organization, or connecting societies if they share members. This collapse is useful, for example, if you want to explore which societies were more likely to influence each other via member transference and how information might have spread from one society to the next. The resulting collapsed network in this case includes only society nodes, and edges between them are stronger if the societies share more members, but the opposite network can also be generated (Fig. 6.13).

Multigraphs, hypergraphs, multiplex networks, and self-loops

The four concepts listed here are brought up mainly as ones to avoid if at all possible; common network software packages not only often do not support these features but may actually stop working or perform math incorrectly if they are included. The concepts all relate to more complex varieties of

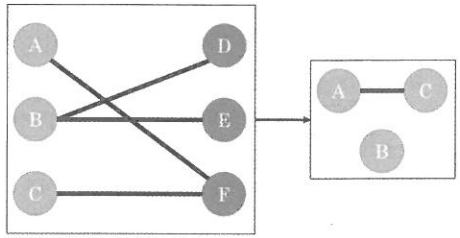

Fig. 6.13 Collapsing a bipartite network into a unipartite one (for instance, a network of members *and* scholarly societies to a network of scholarly societies *where members are the edges*).

edge relationships, and although they are legitimate ways of representing complex networks, they are not yet standard enough in network science to be explored by the historian first dipping her toes into network science.

Multigraphs are those networks that are permitted to have **parallel edges** between a pair of nodes. It means what it sounds like: duplicate edges between nodes, as in a trade network with a new edge every time some exchange is registered between the same two cities. While in theory these networks are just another reasonable representation of a series of networked events, in practice, software tends to break around them. A stopgap solution is to code a single edge with attributes to replace the parallel edges; for example, giving the trade network a weight attribute that increases as the number of exchanges go up. It is worth pointing out that in a directed network, as with Twitter follows, an edge directed from A to B (A follows B) is not considered parallel to an edge directed from B to A (B follows A). These are fine, and most software supports this use.

It is unfortunate that **hypergraphs** are rarely supported by either software packages or network data structures, as they can introduce a lot of nuance that is usually lost. Hypergraphs allow **hyperedges**, single edges that connect multiple nodes. Hypergraphs can be considered functionally equivalent to bipartite graphs; as with the society/membership example before, a group of people can be connected with one hyperedge if they all are members of the same society. In this case, the nodes are the people and the hyperedges are the societies. If you want to use hyperedges, you can create a bipartite network instead.

We saw **multiplex networks** in the previous section on the various ways network analysis has been used to study history, when discussing the Medici's relationships with other Florentine families. A multiplex network is one that allows multiple types of edges, similar to k-partite networks that

allow multiple node types. Often multiplex networks also include parallel edges, when there can be many types of edges between the same two nodes. One example would be a national network of transportation between cities, where there can be air, sea, road, or train edges between any two cities, and often more than one of those options for each pair. As will become clear, the standard network data structures make coding networks like this difficult, but you can use the same shortcut you use for multigraphs: create only one edge between nodes, but give it multiple attributes to identify its various types. In the example above, an edge between New York City and Liverpool might have four different attributes, boat (true), road (false), train (false), plane (true), rather than four separate edges.

The last variety of odd edge covered here is the **self-loop**. More innocuous than the rest, a self-loop is simply an edge that goes from and to the same node. A structural network that aggregates people into single nodes representing their entire social class, and edges connecting them based on interactions between classes, would presumably have self-loops on every node because each social class has people that talk to each other. An email network, with nodes as people and edges as the volume of emails sent between them, would have self-loops for people who email themselves. Self-loops are supported by all software packages, but can sometimes subtly affect algorithms in ways that will change the results but not produce errors. For this reason, it is good to avoid self-loops if you can, unless you know specifically how they will affect the algorithms you plan on using to measure your network.

Explicit and natural versus implicit and derived

Although in practice the networks produced look and act the same, the various ways a network is created have repercussions on how it should be interpreted and analyzed. This introduction will not go beyond pointing out the difference, because interpretations can vary depending on particular use cases.

Historians will want to note when their networks are **explicit/physically instantiated** and when they are **implicit/derived**. An explicit network could be created from letters between correspondents or roads that physically exist between cities. A derived network might be that of the subjectively-defined similarity between museum artefacts or the bibliographic coupling network connecting articles together if they reference similar sources.

The difference between explicit and implicit networks is not a hard binary; it can be difficult or impossible to determine (or not a reasonable question) whether the network you are analyzing is one that has its roots in some physical, objective system of connections. It is still important to keep these in mind, as metrics that might imply historical agency in some cases (e.g. a community connected by letters they wrote) do not imply agency in others (e.g. a community connected by appearing in the same list of references).

Local Metrics — Close Reading/The Trees

One of the benefits of networks is their ability to offer several perspectives of analysis, both across scales and subsections. A node occupies a particular neighborhood, which in turn has its own neighbors, all of which fit together in the global network. This allows the historian to see not only how historical actors interact with their immediate relations, but how standard or unique those interactions are and how they affect or are influenced by the larger whole.

There are many formal ways to explore these interactions, and they are often classified according to their scale. Sometimes you want to look at your networks under a macroscope to clarify general patterns and structure, but other times you may want to focus in with your microscope to discover local variations or intricacies. For example, you may want to pinpoint the most well-connected political lobbyists or the most well-traveled trading routes. The advantage of applying network analysis to these questions is that local effects are always contextualized or driven by global interactions. The following section shows some basic metrics used to approach networks at the node, dyad, or triad level, the smallest of scales.

Paths — path length, shortest path, average path length

A **path** is a fairly intuitive concept: it is the series of edges between two nodes. The interpretation of a path changes from dataset to dataset, but its formal definition remains the same.

The ORBIS project, a map of the ancient Roman world, allows you to calculate paths between any two cities. In most cases, there are multiple routes between any two cities and those routes can vary in size. The number of roads one needs to traverse to get from Rome to Constantinople is the **path length**; a direct route would take you across 23 roads, but you could also take a wide detour through modern day Spain, requiring you to traverse

across another 30 roads. The path between two nodes that requires the fewest steps possible is the **shortest path**.

Both historians and historical actors are often interested in optimizing their path according to some criteria, whether it is shortest distance, least expensive, or fastest travel time. These are often not the same. The shortest path between two cities, for example, may go through a treacherous mountain path. There may also be multiple shortest paths, such as two routes that diverge but require the same amount of steps.

You can get an idea of how well connected your network is by looking at *all* shortest paths between every possible pair of nodes and averaging them together. This **average path length** is a measurement of the connectedness of a network. In networks with a fairly high average path length, like ORBIS, one can expect that, if they picked a node at random, it would take them quite a few steps to get to where they need to go. Other networks have a fairly low average path length, like the network of movie stars who act in movies together. This low average path length is what allows the movie trivia "Six Degrees of Kevin Bacon" game to exist. It is played by picking an actor at random and naming actors they co-starred with until you reach Kevin Bacon. For example, the path length between Edward Norton and Kevin Bacon is three: Edward Norton starred in *Fight Club* with Brad Pitt; Brad Pitt starred in *Ocean's Eleven* with Julia Roberts; Julia Roberts starred in *Flatliners* with Kevin Bacon (Fig. 6.14).

It is important to adapt your interpretation of path length based on the dataset and historical question being asked. Paths between cities are fairly easy to interpret and can be useful for historians who want to infer, for example, the most likely route someone traveled. Historically, the distance between two cities is a meaningful measurement, one that enables trade and travel, and one that continues to be meaningful the farther away one travels. Paths between two family members on a genealogical network are a little different. Very close paths are inherently meaningful but unless the question being asked deals with genetics, inbreeding, or royalty, it is very

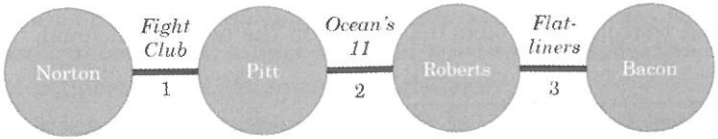

Fig. 6.14 Path length between Edward Norton and Kevin Bacon, via co-starring in movies.

difficult to distinguish the meaningful difference between a path length of five or 15. Average path lengths at a population level, however, once again become meaningful. For example, the average genealogical path length of Ashkenazi Jews is quite low, suggesting an insulated and often isolated population compared to other ethnic groups of similar size.[15]

Remember that average path length is a property of a network, whereas shortest path and path length are properties of a pair of nodes, described by a particular series of edges.

Centrality and prestige — degree, closeness, and betweenness

Historians are often interested in locating the most active or well-connected actors in a network, whether that means the early modern thinkers whose ideas were the most influential, or the cities that were most often traveled through due to their central position on trading routes. The metrics used to explore these topics are **centrality**, when the directionality of edges is not relevant to the question, or **prestige**, when directionality is relevant. There are many ways to measure both centrality and prestige, and how they are interpreted can change depending on the method and the data at hand. While these metrics can become mathematically complex, often the simplest solutions are the best ones.

Degree centrality is generally the first network metric taught; it is both simple and powerful, and it is an attribute that every node in a network has. A node's **degree** is, simply, how many edges it is connected to. This generally correlates with how many **neighbors** a node has, where a node's neighborhood is those other nodes connected directly to it by a single edge. In Fig. 6.15, each node is labeled by its degree.

In a network of Facebook friendships, the higher someone's node degree, the more friends they have. In a network of non-fiction books connected to one another by shared topics, an encyclopedia would have an extremely high degree because its topical diversity would connect it to most other books in the network. High degree centrality corresponds to well-connectedness, and what that implies differs depending on the network at hand.

[15]Shai Carmi, Ken Y. Hui, *et al.* (2014), "Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins," *Nature Communications*, **5**, 4835. http://www.nature.com/ncomms/2014/140909/ncomms5835/full/ncomms5835.html.

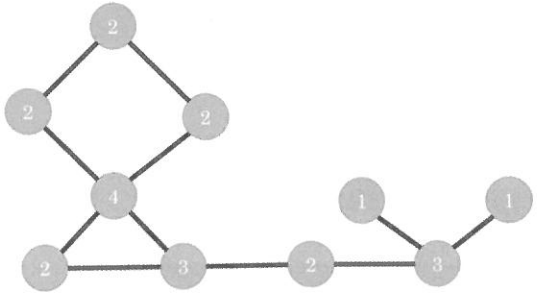

Fig. 6.15 Degree centrality.

In fairly small networks, up to a few hundred nodes, degree centrality will be a fairly good proxy for the importance of a node in relation to a network. If we have a network of 100 authors, connected to one another when they co-author a work, those with the highest degree are those who co-author with the most other people in the network. It reasonably follows, if the data collection was thorough, that these high degree authors are the ones responsible for connecting this network together; they are a force of cohesion in the network.

This use of degree centrality also works on larger networks but one must be more careful when interpreting importance. If we have a network of 10,000 co-authors, drawn from articles published in late-20th century physics, a few extremely high-degree nodes will crop up that are of little importance to network cohesiveness.

A few more recent works in high-energy physics have hundreds or thousands of authors on a single article.[16] Each of the authors on one of these articles would by definition have hundreds of co-authors, and thus they would all have an extremely high degree centrality. This would be the case even if none of the authors had co-authored works with people outside this one article, resulting in hundreds of authors with high degree centrality who may have very little importance to the cohesiveness of the overall network. A historian viewing the network as simply an ordered-list of the highest degree nodes would have difficulty discerning the truly central figures in the network, whereas even a brief glance at a network visualization would reveal the peripherality of the so-called high-degree nodes.

[16]See, for instance, Atlas Collaboration (2012), "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Physics Letters B*, **716(1)**, 1–29 http://dx.doi.org/10.1016/j.physletb.2012.08.020.

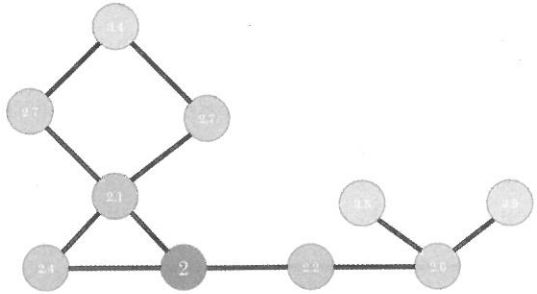

Fig. 6.16 Closeness centrality.

Network scientists have created more complex metrics to deal with these and other potentially confounding effects. Two such metrics are **betweenness centrality** and **closeness centrality**, both of which are attributes that can be calculated for each individual node, and both of which attempt to contextualize the local centrality measurement against the larger network.

A node's **closeness centrality** is a measure of how close it is to every other node in the network (Fig. 6.16). In the network of travel routes and cities of the ancient Roman Empire modeled by ORBIS, the city with the highest closeness centrality is the one that would be the fastest to travel to from any given city in the network. A city that has many direct routes to other cities, but is still at the periphery of the network (e.g. Emerita Augusta, present-day Mérida, Spain), would have a high degree centrality but still would not be ranked high on closeness centrality. Rome would likely have a high closeness centrality, as it is both directly connected to many cities and geographically central enough that it would take very few roads to get to any other city on the map.

There are multiple ways to calculate closeness centrality, but the simplest one is to take the average distance from a node to every other node on the network. In the Roman city example, if there are 100 cities in the network (including Rome), the closeness centrality is found by finding the shortest paths from Rome to all 99 other cities, adding their distances together, and dividing by 99. The resulting number is the average path length from Rome to any other city, and the lower it is, the closer Rome is to the rest of the world. This calculation can then be repeated for every other city in order to find which cities have the lowest closeness centrality. Note that, unlike most other centrality metrics, closeness is ranked by the lowest number rather than the highest.

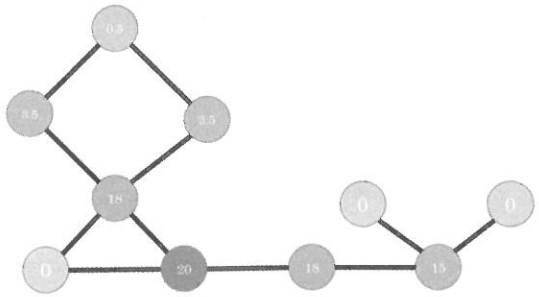

Fig. 6.17 Betweenness centrality.

Another centrality measurement that mitigates the issue of high-degree nodes on the periphery is **betweenness centrality** (Fig 6.17). In the Roman city network example, the city with the highest betweenness centrality is the one that a traveler is most likely to travel through when going from any one destination to any other. Formally, it is the node that sits on the shortest paths between all node-pairs in the network.

In the United States, the road system in the state of Indiana is such that, on almost any road trip from one corner of the state to another, the fastest route is one that goes through Indianapolis. The city is the hub of Indiana from which the highway spokes emanate; to get from Bloomington to Terra Haute, one must travel through Indianapolis; to get from Muncie to Columbus, one must travel through Indianapolis. As the city sits in the middle of the shortest paths between most pairs of cities in the state, it has the highest betweenness centrality.

Calculating betweenness and closeness centralities requires finding the shortest path between every pair of nodes in a network, which can be computationally time-consuming depending on the size of the network.

PageRank

PageRank is an algorithm developed by Google to improve search results. It gives high scores to popular pages. PageRank is a measurement of **prestige**, which means it measures the centrality of a node and takes edges leading to or from it into account. Its calculation is more mathematically complex than will be addressed in this book, however as long as you understand the concept behind it, you can accurately employ it in your historical research.

If you were a robot crawling the World Wide Web, following hyperlinks you come across at random and occasionally jumping to another site entirely even if there is no hyperlink to it, you would be acting out one method of calculating PageRank. You would, undoubtedly, find yourself ending up on some sites very frequently and others very infrequently. Wikipedia, for example, is linked to all across the web; odds are if you keep clicking links at random, you will end up on Wikipedia. On the other hand, you probably would not frequently end up at a *Lord of the Rings* fan page from 1997. The frequency with which you land on a site is equivalent to that site's PageRank; Wikipedia, BBC, and Amazon all have high PageRanks because the probability of ending up there by clicking links is quite high.

Another way of conceptualizing PageRank is by pretending each website has a certain amount of votes it can give to other sites and it signifies those votes by hyperlinking to the sites it votes for. The more votes a site has gotten, the more it is allowed to give. Also, every single site has to use *all* its votes, which means if BBC collects a lot of votes (a lot of sites link to it), but it itself only links to one other site (e.g. Wikipedia), all of the votes in BBC's possession go to that one site. Thus, even if BBC were the *only* site to link to Wikipedia, Wikipedia would still have a huge PageRank, because it collected all of BBC's votes.

This type of algorithm was originally developed for science citation networks, where an article replaces a website and a citation replaces a hyperlink. It is a surprisingly powerful algorithm for finding important nodes in directed networks of any sort. A historian calculating the PageRanks of letter-writers in early modern Europe would find they correspond quite well with the figures everyone remembers. PageRank can also be deceptive, however. A famous head of state and the recipient of letters from powerful people, for example, may have written the majority of her letters to a secretary. Even if this secretary played a fairly small role in early modern politics, her PageRank would have been quite high because she received the lion's share of her employer's votes.

Local clustering coefficient

The **local clustering coefficient** (LCC) measures the extent to which a node's neighbors are neighbors of one another. Every node has its own LCC; nodes with low LCCs connect separate communities and nodes with high LCCs do not (Fig. 6.18). Formally, the LCC of a node is calculated by dividing the number of edges between neighbors of that node divided

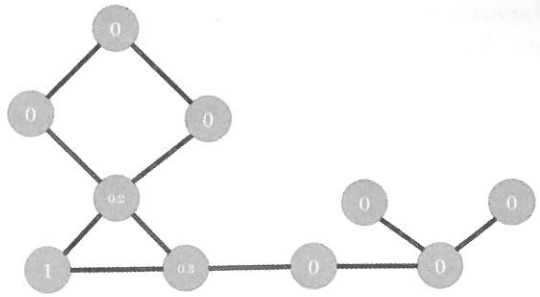

Fig. 6.18 Local clustering coefficient.

by the number of possible edges between its neighbors. In your friendship network, this manifests as the number of your friends who are friends with one another, divided by the total possible number of friendships between pairs of your friends.

This can be a useful measurement for many varieties of historical networks. In social or communication networks, a node with a high LCC may be in a brokerage position. As a short connection between otherwise disconnected people, this person can act as a conduit for information or can play a part in introducing communities together. In an electrical grid, a few high LCC nodes might be a sign of a weak network — if one high LLC node goes offline, sections of the grid may become disconnected. In a network of languages connected by their linguistic similarities, a high LLC node may be a language which two geographically distant languages contribute to, or adapt from.

Global Metrics — Distant Reading/The Forest

An aggregate of individual components is not a network. A network is what happens when the relationships between individual components come together and every entity therein becomes part of a coherent whole. Network analysis at the global level reveals subtle trends that may not be apparent when exploring individual nodes or edges, and oftentimes it can lead to shifts in the types of questions historians ask.

Global network measurements also betray one of the shortcomings of network analysis for historical research: homogeneity. Certain fundamental properties of how people and things self-organize tend to create networks that look remarkably similar when zoomed out. At this scale, it can be

difficult to tell apart a network of Italian families from a network of genetic similarity in rats. A danger historians need to watch out for when reporting these measurements is that they often do not imply any sort of novelty about the network under study. These global metrics are most useful when measured *in comparison to* other networks; early modern and present day social networks both exhibit scale-free properties, but the useful information is in how those properties differ from one another.

Density

The most basic global network measurement is of its **density**. Every network's density ranges between zero and one, and is defined by the number of edges it has divided by the number of edges it potentially could have. The social network of a small 13th century village would have incredibly high density, as almost everyone would know almost everyone else. The density measurement for a network connecting American cities by interstate highways would be quite low, on the other hand, because the farther two cities are from one another, the less likely they are to have a highway traveling directly between them. Distance would be much less of a limiting factor if this were a network of commercial flights.

Historians can find a use for density in comparisons across networks. In trade networks of iron, cotton, and tobacco, for example, each traded good may travel to just as many outposts and cities, but the trading networks might have incredibly different densities. This might suggest differences in demand or of transportation expense across certain routes. An increasing density over time among a group of authors connected by co-authorships and citations would reveal the solidification of a community of scholars, or perhaps even a new discipline.

Generally speaking, the densest networks are also the least analytically fruitful. With every node connected to every other, the basic measurements like centrality or reciprocity have very little meaning. If you find yourself with an extremely dense network, it is usually a sign that the data need to be prepared differently or a different technique should be used.

Diameter

A network's **diameter** is a function of its shortest path lengths. Formally, it is a network's longest shortest path, or a measurement of how many

edges one needs to traverse to travel between the two most disparate nodes on a network. Even networks with a short average path length may have a high diameter. In the movie trivia game "Six Degrees of Kevin Bacon," for example, an actor who played one minor role in an early 20th century silent film might be 15 steps away from a voice actor of a cartoon movie produced last year. This would be an example of a network with a short average path length but a large diameter. In fact, the diameter of a movie co-acting network is likely undefined because there are bound to be some actors that can never reach the rest of the network.

For the historian, it is unlikely that the diameter will be more useful than the average path length on most occasions. Most of our networks are incomplete; they are subsets of networks from the past that we have constructed or reconstructed, but which do not represent the entirety of what we wish to study. It may be that the data are lost or not completely digitized, or that we wish to constrain our problem to a smaller scale. Whatever the case, nodes at the periphery of a historical network often only appear so because that is where the data end. In those cases, a network's diameter is more a measurement of a historian's reasoning when creating a dataset than a useful metric for understanding the past. While this holds particularly true for a network's diameter, it is in general important to think through how the process of collecting data may affect any algorithm that is applied to it.

Reciprocity and transitivity

Reciprocity and transitivity are global network metrics that have implications for the way in which networks evolve. As they both measure actual edges against expected edges, they are related to, but more specific than, measurements of density.

In a directed network, **reciprocity** measures the extent to which dyads form reciprocal relationships. In a network with reciprocity of one, every directed edge is reciprocated. On Facebook, for example, you cannot list someone as a friend without them also listing you as a friend; thus, the Facebook friendship network has reciprocity of one. Reciprocity is equal to the number of reciprocal edges divided by the number of total edges in a network. Networks with no reciprocated edges have a reciprocity of zero. One example is a legal precedent citation network, as the edges must point toward previously decided cases and two cases cannot cite each other as precedents.

Social networks tend to become more reciprocal over time. On Twitter, as in life, one-way transactions generally become two-way relatively quickly. Unsurprisingly, in the early modern Republic of Letters, we see this as well. The probability that a person will correspond with another in the network increases if they have a history of previous contact. Reciprocity is then not only a measurement of a network but also a prediction of its future. The metric is an extremely valuable one in most circumstances. Historians might measure the reciprocity of military engagements among nations of opposing alliances, giving them insight into the balance of power in a war.

Transitivity is a global measurement of how frequently triads are completely connected. It is similar to reciprocity in that it divides actual connections by potential ones. If A connects to B, and B to C, a transitive edge is one that connects A and C. Transitivity is essentially the percent of connected dyads with transitive edges between them. Like with reciprocity, many evolving networks tend toward a higher transitivity, and the mechanism is fairly straightforward. In social networks, you often meet new people through pre-existing friends. Once you are connected to your friend's friend, you have created a closed triangle where all three nodes are connected to one another.

Networks with high transitivity often appear clumpy, with many small communities of densely connected individuals. A historian may wish to see the evolution of transitivity across a social network to find the relative importance of introductions in forming social bonds.

Degree distribution, preferential attachment, and scale-free networks

All the global network measurements covered so far produce a simple number that is easy to compare across networks. Not all useful measurements can collapse into one number, however, and one example is a network's **degree distribution** (Fig 6.19). Recall that a node's degree is defined by the number of edges connected to it. A network's degree distribution is an ordered list of the degrees of every node in that network, usually visually represented in a chart. It is used to determine how skewed node degree tends to be across a network.

It is easy to think that most networks would have a "normal" distribution of node degrees; that is, some nodes have a tiny amount of edges connected to them, some nodes have quite a few, but most sit squarely in the middle. This is generally not the case. Social networks, article citation

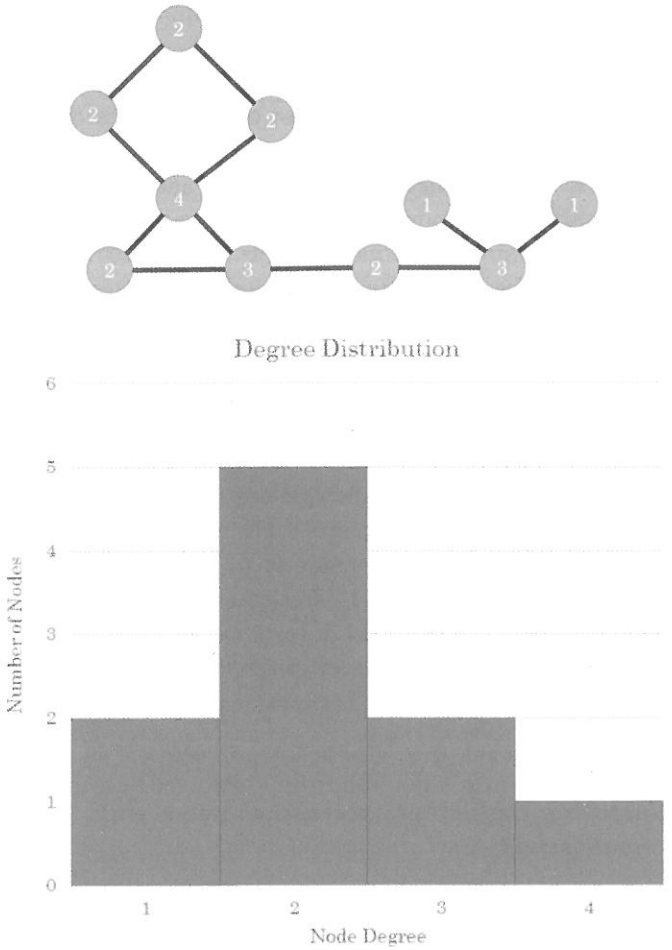

Fig. 6.19 Degree distribution of the network displayed above.

networks, airline travel networks, and many others feature a significantly more skewed degree distribution. A few hub nodes have huge numbers of edges, a handful more have a decent amount of edges but significantly fewer, and most nodes are connected by very few edges.

In network parlance, this long-tailed degree distribution is usually called **scale free** or **power law**, although the networks described as such often do not meet the formal mathematical definitions of those terms. Scale-free networks have a handful of major hubs that keep the rest of the nodes connected to one another. In transportation networks, these are the cities

one has to go through to reach anywhere; in correspondence networks, these are the people without whom information would spread much more slowly.

A number of mechanisms have been suggested to explain this similarity across so many networks, but the most convincing historical argument revolves around **preferential attachment**, or rich-get-richer. In a friendship network, this means that those who have a lot of friends are at a higher risk of making even more friends; they have a larger circle that can, in turn, introduce them to even more people. This positive feedback loop continues until a few people dominate the landscape in terms of connectivity. This is also true of citation networks, in that articles that already are highly cited are more likely to be seen by other scholars and cited again in turn.

This fairly ubiquitous feature of networks can provide some insight for historians but not necessarily by its very existence. If a historical network exhibits a long-tail distribution, with very prominent hub nodes, the structure itself is not particularly noteworthy. What is worthwhile is figuring out which nodes made it to the top and why. Why do all roads lead to Rome? How did Mersenne and Hartlib develop such widespread correspondence networks in early modern Europe? The answers to these questions can cut to the heart of the circumstances of a historical period, and their formalization in networks can help guide us toward an answer.

Global clustering coefficient, average shortest path, and the small world

The term small world is thrown around rather frequently as a catchall for any network that seems surprisingly close-knit. Although this is not exactly wrong, the term has a formal definition, and mistaking the folk definition for the formal one can lead to misunderstandings or improper conclusions. A **small world** network is one in which the shortest path between any two random nodes grows proportionally to the logarithm of the number of nodes in the network. This means that, even if the network has many, many nodes, the **average shortest path** between them is quite small. These networks also have a high **global clustering coefficient**, which is the average LCC across all nodes.

In short, a small world network is one in which groups of nodes tend to cluster together into small communities, and the path between any two nodes is usually quite short. The "Kevin Bacon" network is one such small world, as are certain transportation and communication networks. It is not uncommon for small world networks to share many qualities with scale-free networks, and the two can overlap.

The further back in history one goes, the less the globe looks like a small world network. This is because travel and distance constraints prevented short connections between disparate areas. However, if one defines the network as only those connections within a sufficiently small region for the time period, small world properties tend to reappear.

Historians can draw a number of inferences from small world networks, including the time it might have taken for materials to circulate within and across communities and the relative importance of individual actors in shaping the past. The existence of these networks can also play a role in thinking through counterfactual history. Small world networks tend to be resilient against small contingent events upsetting the network structure, which means information, diseases, and so forth will tend to spread widely and quickly regardless of small historical perturbations.

Connecting the Global/Distant with the Local/Close

The true benefit of networks for historians lies not in the local (that is what we have always been good at), nor the global (we have some experience with that, too), but in the way it connects the two. Local features are embedded seamlessly into the whole unit, and global properties are very clearly made up of the aggregate of individual parts. This makes network analysis a promising methodology for connecting the historiographic traditions of microhistories and the Braudelian *longue durée*.

Hubs and bridges

Networks help us pinpoint those channels or individuals that are essential for connecting disparate communities. These can take the form of individual nodes, which are often called hubs, or an edge between a node pair, called a bridge. Each of these terms has multiple formal definitions, but the ideas are the same: a **hub** is a node without which the path between its neighbors would be much larger, and a **bridge** is an edge which connects two otherwise unconnected communities.

Hubs relate to many of the concepts already covered. Scale-free networks usually have a few very central hubs, which themselves tend to have high betweenness centralities. That is, they sit on the shortest path between

many pairs of nodes. These nodes are vulnerable points in the network; without them, information takes longer to spread or travel takes quite a bit longer. A hub in a network of books connected by how similar their content is to one another would have a different meaning entirely: the most central node would likely be an encyclopedia, because it covers such a wide range of subjects. The meaning of these terms always changes based on the dataset at hand.

Bridges also relate to previous concepts. In small world networks, bridges are those edges that connect two dense but isolated clusters together, thus allowing the network to retain a surprisingly small average path length alongside such close-knit communities. The simplest historical example of a bridge is a bridge itself, connecting two landmasses across a body of water. If that bridge collapses, people would need to go well out of their way to get from one side to the other.

Hubs and bridges help connect the local with the global because they are individual metrics that are defined by how they interact with the rest of the network. On its own, a single marriage between two families might seem unremarkable, but if these are royal families marrying into some until-then disconnected foreign power, or two people marrying across faiths in a deeply religious community, one simple bridge takes on new meaning. Network analysis aids in finding these unusually connective entities en masse and with great speed, leaving the historian with more time to explore the meaning behind this connection.

Cliques, communities, and connected components

What are hubs or bridges connecting? There are a slew of terms that describe different organizational patterns of networks but the three most basic are cliques, communities, and connected components. A **clique** is a section of a network that is maximally dense; that is, every node is connected to every other node. Individual nodes in cliques may be connected to nodes outside of the clique. In a network of authors citing other authors, a clique is a subset of authors who all cite one another. An individual author may cite others outside of the clique as well but those external authors are not part of the clique unless they cite and are cited by every other member. All closed triads are by definition cliques, and more cliques lead to a higher clustering coefficient. By definition, any group of connected nodes, each with an LCC of one, will form a clique because, for each, every neighbor is a neighbor to every other.

Sometimes cliques are too strict to be useful when looking for groups of nodes in networks. There may be a network of corporations that are connected to one another if they share people on the board of directors. Not every corporation shares directors with the same other companies, but it is still clear that the same group of people are connecting these companies together, even if they do not strictly form a clique. These well-connected groups of nodes, though not maximally connected, form a **community**. There are many different ways to define a community, and communities may or may not be overlapping. One popular definition of a community is a group of nodes that have more internal edges between them than external edges connecting to other communities. It is a metric of relative density.

Like most network concepts, a community is not well defined. Multiple algorithms produce different results and group the network in different ways. Some algorithms group edges, some group nodes; others produce hierarchically nested communities, or communities that overlap. The most frequently used community detection method, called modularity, places every node in a network into a community based on how well-connected its neighbors are. Modularity is successful when there's a high ratio of edges connecting nodes *within* any given community compared to edges linking out to other communities. The best community detection approach is the one that works best for your network and your question; there is no right or wrong, only more or less relevant. A good "sanity check" is to see if the algorithm you're using puts the people or entities you know *should* be grouped together in the same community and leaves out the ones that do not belong. If it works on the parts of the data you know, you can be more certain it works on the parts of the data you do not.

The most inclusive term for a group of nodes is a connected component. A **connected component** is simply a group of nodes that can possibly reach one another via a path of any number of edges. Undirected networks are occasionally classified by the number of connected components they contain; that is, the number of different groups of nodes which are completely disconnected from one another.

Although this class of node-group seems so large as to be effectively meaningless, connected components can be very useful to understand, for example, who could possibly have heard a particular rumor. They define the scope of a connected world. The one caveat for historical networks is that we often lack the full story, and nodes that appear disconnected might be otherwise if more evidence presented itself. This is why, in historical network analysis, it is often better to rely on the edges and nodes that do

exist than those that do not. Measurements of connectivity should be seen as lower bounds: your network is *at least as connected as* this measurement shows; the shortest path *is at most this long.*

The strength of weak ties and structural holes

The premise of this section is that networks reveal the interplay between global properties and local activities, and this discussion on weak ties and structural holes is a prime example of just such a benefit. Many networks form as small worlds. They are densely packed communities with many links between communities, which allows any community to reach another in only a few steps.

This is the view from the macroscope; under the microscope, the focus shifts. Two cliques, adjacent to one another, happen to have a single edge connecting them together. This single edge, connecting two otherwise insular and disconnected clusters, is called a **weak tie**, as the connection between these two communities could break easily. A downed power-line, for example, could disconnect two sections of an electrical grid, or a single estranged cousin may break the connection between two sides of a family. Careful readers will have noted that the definition of a weak tie is curiously similar to that of a bridge. This dichotomy, the weakness of a connection alongside the importance of a bridge, has profound effects on network dynamics.

In social networks, weak ties may be old college friends whom you barely keep in touch with; in early trade networks, a weak tie may be that single treacherous trade route between two disconnected worlds, offering great reward at great peril. In practice, weak ties tend to be weak not only because their breaking can fragment the network but also because the connections themselves tend to be tenuous. On weighted networks, these edges generally have a very low weight. This is the case because of the way networks tend to evolve. If you have a particularly close friend from another community, you are more likely to introduce him to your other close friends, at which point the connection between the two communities immediately multiplies. The tie is no longer weak.

These weak ties, however, are extremely important.[17] We will once again use the example of social networks, however this property holds across many

[17]The canonical study is Mark Granovetter (1973), "The Strength of Weak Ties," *The American Journal of Sociology*, **78**(6), 1360–1380.

other varieties as well. When an individual is looking for something, for example a new job, it is unlikely she will look to her closest friends for advice. She and her best friends know the same people, they have access to the same resources and information, they are interested in the same sort of things, and as such are particularly poor sources for news. Instead, she will reach out to her cousin whom she knows works for an organization that is hiring, or an old teacher, or someone else with whom she has a weak connection. Empirically, network studies have shown connections to outside communities tend to be the most fruitful.

The strength of weak ties has been hypothesized as a driving force behind the flourishing of science in 17th century Europe.[18] Political exile, religious diaspora, and the habit of young scholars to travel extensively, combined with a relatively inexpensive and fast postal system, created an environment where every local community had weak ties extending widely across political, religious, and intellectual boundaries. This put each community, and every individual, at higher risk for encountering just the right serendipitous idea or bit of data they needed to set them on their way. Weak ties are what make the small community part of the global network.

The study of weak ties is the study of the connections between separated communities, but the beneficial effects of spanning communities also accrue in nodes. Someone or something sitting at the intersection of two or more otherwise disconnected communities is said to occupy a **structural hole**[19] in the network, and that position accrues certain advantages on both the node and the network. As weak ties correspond to bridges, structural holes are the negative space that become filled by hubs (in sociology, sometimes known as **brokers**). People who fill structural holes in corporations have empirically been shown to find more career success. In the first section of this chapter, the Medici family was said to have strategically created structural holes in the network of Florentine families. From this position, the Medici were the only family who could reap the benefits of every community in the network; they were able to play the communities against one another, and they could act as network gatekeepers whenever they saw fit. In early modern Europe, individuals like Henry Oldenburg, the Secretary of the Royal Society, sat at the center of many networks and were able to funnel

[18]Davis Lux and Harold Cook (1998), "Closed circles or open networks? Communicating at a distance during the scientific revolution," *History of Science*, **36**, 179–211.

[19]Ronald S. Burt (2004), "Structural Holes and Good Ideas," *American Journal of Sociology*, **110**, 349–399, doi:10.1086/421787.

the right information to those who needed it and keep it from those who should not have had access.

Networks and Time

Network science is a young science; the study of temporal networks, though extensive, is still in its infancy. As important a subject as the evolution of networks over time is for historiography, its undeveloped state in the literature will make this section necessarily brief.

There is no consensus on how temporal networks should be represented. It is clear, for example, that the network of trade between cities in and around the Roman Empire changed drastically over time, however we cannot say with certainty when particular routes were in use or when they were not. Even if we could peer directly into the past, it is not easy to define the point at which a route's use changed from sporadic to regular.

Even in modern networks with full data availability and clear-cut events, it is not clear how to model network change. The network of phone calls in the United States creates an ever-changing network, where each phone call is an additional edge. You cannot take a snapshot of a single moment, however, and say it represents the call network of the United States, because what about all the people who call each other at different points throughout the day?

One solution is to aggregate phone data into **time slices**, or chunks of time at a particular resolution (hour, day, month, year) for which all phone data are combined. In this approach, each time slice is a snapshot of the network, and the study of how that snapshot changes at each slice is the study of the evolving network. This approach is the most common, but because small perturbations can lead to drastic reconfigurations of networks, it is sometimes difficult to compare time slices against one another.

The other obvious solution, modeling every bit of network data at every moment, may make for a more pleasantly continuous network but is notoriously difficult to represent in a meaningful way. It is also difficult to write algorithms that map to our current conceptual framework of network analysis under these dynamic approaches.

We hope that a renewed interest among historians in network analysis, and a willingness to collaborate with network scientists, will pave the way for more nuanced and meaningful approaches to understanding the evolution of networks over time.

Further Reading and Conclusion

There are many resources available to those interested in delving further into network analysis. *Networks, Crowds, and Markets* by Easley and Kleinberg[20] is an approachable introduction and most similar to what this section would look like if it had been expanded into its own book. It is particularly noteworthy for its ability to connect abstract network concepts with real, modern social phenomena. Wasserman and Faust's *Social Network Analysis*[21] is a canonical textbook that focuses on smaller scale network analysis, and the associated issues of data collection and preparation. It is more mathematical than the former, and recommended for the historian who wants to go from understanding network analysis to applying it, especially on small-to-medium scale datasets. Newman's *Networks: An Introduction*[22] is *the* network textbook for modern, large-scale network analysis of any sort, from metabolic networks to social. It is highly formal and mathematical, and is recommended for those who want to seriously engage with their network science colleagues, and who want to help develop new algorithms, tools, and methods for historical network analysis.

Textbooks also exist for the various network analysis software packages available, which encompass both concepts and tool use. For those already familiar with UCINET, or who are particularly interested in matrix manipulations, *Analyzing Social Networks* by Borgatti, Everett, and Johnson[23] is the book to read. Those wishing to do network analysis in Excel using NodeXL should find *Analyzing Social Media Networks with NodeXL* by Hansen, Shneiderman, and Smith.[24] Pajek, one of the most feature-rich network analysis programs, can be learned from *Exploratory Social Network Analysis with Pajek* by De Nooy, Mrvar, and Batagelj.[25] The Network

[20] David Easley and Jon M. Kleinberg (2010), *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge, UK: Cambridge University Press.
[21] Stanley Wasserman and Katherine Faust (2010), *Social Network Analysis: Methods and Applications*, 1st edn, Cambridge, UK: Cambridge University Press.
[22] Mark E. J. Newman (2010), *Networks: An Introduction*, 1st edn, New York, NY: Oxford University Press.
[23] Stephen Borgatti, Martin G. Everett and Jeffrey C. Johnson (2013), *Analyzing Social Networks*, 1st edn, Thousand Oaks, CA; London: SAGE Publications Ltd.
[24] Derek Hansen, Ben Shneiderman and Marc A. Smith (2010), *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, 1st edn, Burlington, MA: Morgan Kaufmann.
[25] Wouter de Nooy, Andrej Mrvar and Vladimir Batagelj (2005), *Exploratory Social Network Analysis with Pajek*, Cambridge, UK: Cambridge University Press.

Workbench (NWB) and the Sci^2 Tool both have extensive user manuals linked from their homepages.[26] Unfortunately, the software we recommend for historians beginning in network analysis, Gephi, does not yet have an extensive centralized learning guide.[27]

[26] Information for NWB can be found at http://nwb.cns.iu.edu/Docs/NWBTool-Manual.pdf; Sci^2 http://sci2.wiki.cns.iu.edu/.

[27] That said, Clement Levallois has a suite of excellent tutorials at http://clementlevallois.net/gephi.html.

Chapter 7

Networks in Practice

This section shows how to take the conceptual framework of network analysis and apply it in practice. A historical network analysis will often require a similar series of steps:

1. *Deciding on a dataset,*
2. *Encoding, collecting, or cleaning the data,*
3. *Importing the data into a network analysis package,*
4. *Analyzing the data,*
5. *Visualizing the data,*
6. *Interpreting the results,*
7. *Drawing conclusions.*

The framework is not universal, and the process usually requires a lot of repetition and some steps may be omitted or added depending on circumstance. Steps four and five, while covered in a small section in this book, are extremely open-ended, and historians who wish to learn more are encouraged to delve into the tool of their choice using the Further Reading section of this chapter. The steps we take the most time explaining usually take the least time in practice; less than 5% of the time spent on a project will be time spent analyzing and visualizing data. Most time will be spent on collecting, cleaning, and interpreting.

Picking a Dataset

How can you know whether the data you have will be amenable for network analysis? The answer depends on the project, of course, but unfortunately it is often difficult to tell at the outset of a project which data will be most useful, if any, for a network analysis. Many network analyses lead to dead ends, and the more experience you have, the earlier you will begin to