10. Using Data and Network Analysis in Humanities Research: A Guide to Getting Started

NATHANIEL D. PORTER

Network thinking and analysis are now widely used in diverse disciplines throughout the academy. In this chapter I will offer a brief primer on network analysis, aimed specifically at understanding the methods and principles used by the authors in this volume, all of whom participated in the Viral Networks workshop. I will begin by explaining basic terminology and models commonly used in network analysis, which should be valuable to anyone thinking of using network analysis or visualization in their own work. Then I will outline a typical network analysis workflow and offer tips on getting started in network analysis as a traditional humanist, based on my observations from helping workshop participants. This chapter will be most useful to those considering using network analysis for the first time. Those looking for more information or inspiration on network analysis and what it can accomplish can find resources in the book's glossary and this chapter's references.

First, let's clarify what we mean by the terms network thinking and network analysis. Chances are, even if you have never engaged in statistical analysis or other structured, formal types of data analysis, at some point you have used network thinking. Take, for instance, surveys. Traditional surveys and vital statistics , such as measures of victims of a disease reported by physicians or hospitals, are typically used to gather and analyze data about distinct and separable individuals or groups. The gold standard is a population-representative sample that reflects, as closely as possible, the characteristics of individuals in an entire group, so that you can answer questions such as, “Who is most susceptible to a particular disease?” or “How do disparities in health outcomes compare to race, poverty, or age?” The underlying assumption is that people act somewhat independently and that a good way to understand social patterns is to look at the distribution of people with different characteristics.

In contrast to traditional surveys, *network surveys* start with the assumption that social environment (family, friends, school peers, fellow group-members, etc.) is an integral part of who people are and how they make decisions. Instead of asking, for example, “Are young people most likely to contract sexually transmitted diseases?” a network approach might ask, “Does having strong relationships with family, friends, or co-workers affect the likelihood of contracting a sexually transmitted disease?” In both ways of thinking, questions can be quite nuanced, but a traditional survey is more about individuals, regardless of any ties among them, whereas, a network survey intentionally collects and draws on information specifically about the ties between and among individuals in a given environment.

In many ways, this distinction is not new to the humanities. The clearest parallel is the distinction between case study methods and comparative methods. Scholars use case studies to understand the distinctiveness and character of a single category or entity, be that an author, national or local context, time period, etc., in as much detail as possible. A comparative study focuses principally on defining a set of characteristics that can be compared or contrasted to provide insight into how these characteristics are associated with specific historical factors or outcomes. Case studies help us understand exemplary individuals, communities, or businesses, and yet the subject of a case study (e.g. Florence Nightingale, Detroit, or IBM) is rarely isolated entirely from the influence of contextual factors. Network analysis formalizes the contextual factors and relational thinking already embedded in comparative approaches to treat those very relationships as items of interest, whether as causes, effects, or simply patterns to be studied.

Formal network analysis can, no doubt, be intimidating. Many of the authors in this volume, despite having self-selected into a workshop on historical networks, initially expressed concern at the prospect of moving from close reading of specific events, actors, and processes towards coding data and producing truly relational models. With help, however, all authors came to appreciate both how coding data can produce a disciplined form of reflection and how network analytics can enhance or complement other approaches. It was not the goal of the workshop—nor is it the goal of this volume—to transform traditional historians into network scientists or data scientists, although, frankly, both network and data scientists would benefit from more of the probing attention to detail that is inherent to humanistic inquiry. Instead, the goal for both workshop participants and readers is that they be inspired to new ways of organizing and thinking about evidence and analysis, both as producers and as consumers of knowledge. Now let's delve into basic terminology and models commonly used in network analysis.

Terminology and Models

"What is a network?" The answer to this question is more complicated than it might at first seem. In the broadest sense, a network is any group of entities (people, places, words, ideas, computers, topics, institutions, etc.) that are tied to each other in one of two ways: first, through direct relationships like friendship, partnership, genealogy, or communication; and second, possession of similar characteristics, such as attending the same event or working for the same employer, words or topics that appear in the same corpus of texts, or multiple non-exclusive treatments for the same disease. In many of these cases, a network could just as easily be considered only a collection of similar items; the difference is in the importance placed on the ties. For example, a study of word usage in the works of Shakespeare might ask how the frequency of specific words changed over time or differed between plays and sonnets (non-network questions); or, instead, such a study could look for clusters of words that tend to appear together across his works and analyze the characteristics of those clusters and/or common language that spans multiple clusters (network questions). It is important to recognize that network and non-network analysis may overlap, intersect, or appear indistinguishable because, as alluded to above, it is a rare analysis that ignores context and relationships entirely. We will return below to the question of what exactly a network is, after exploring network terminology, in order to build a more technical definition that can prepare for the transition from network thinking to network analysis, which requires a clearly-defined network and explicit specification of relationships.

Network Data and Hypotheses

Two elements are basic to any network: *nodes* and *edges*. Nodes are the entities that are connected. In social analysis, nodes are often individual people or organizations. For example, consider the question of peer influence on delinquency and substance abuse among high school students. In this case, the nodes are individual high school students and possibly other important people in their lives such as parents and teachers. Edges are any relationship that ties the nodes together. In delinquency studies, the edge is often friendship, but it could equally be liking or disliking someone, being in the same class or belonging to the same sport team, working on projects together, or sitting at the same lunch table.

Some of these edges are *symmetrical ties*, meaning that both nodes connected by an edge are connected to each other in the same way. Being in a class together is such a symmetrical tie: if person A is in class with person B, person B is also in class with person A. A symmetrical tie that consists of sharing some common characteristic, rather than a mutual relationship, is called an *affiliation tie*. Others types of edges, such as friendship, can be *asymmetrical*: person A can consider person B a friend, regardless of whether it is reciprocated from B to A. Another important type of *asymmetrical tie* is network flow: if person A gives advice to person B, the relationship between them is asymmetrical, as person B is receiving advice. Certain types of network properties and hypotheses are only relevant to asymmetrical relationships.

In addition to nodes and edges, the other fundamental type of network data are *attributes*. An attribute is simply a characteristic of a node or edge. N*ode attributes* provide more information about the members of a network: a person's race or age, a place's population, mortality rates, or climate. E*dge attributes* provide information specifically about a tie: strength of friendship, frequency of communication, how commonly words occur together, the date of a connecting event. Many types of edges possess both *sign* (positive or negative, such as like/dislike) and *weight*, which is a special type of edge attribute often used in *network statistics* that represents the strength of a relationship (best friends vs. casual acquaintances).

Network analysts consider a variety of different types of properties, each of which has its own ensemble of language used to describe it. I attempt here to introduce some of the most important network properties pertaining to both whole networks and individual nodes, as well as a few typical types of arguments and the language commonly used to make them. That said, network analysis terminology varies substantially between disciplines, and it may be necessary to consult introductory or reference works within an individual discipline or subdiscipline to understand the specific language you encounter there. This is particularly true for those moving between STEM fields and the humanities and social sciences. Each property will be illustrated with example network visualizations. In general, nodes are represented by points on network visualizations and edges by lines, although there are some variants that will be discussed below.

Properties of Networks and Nodes

The first type of properties to consider are those that apply to the whole network (also called the *graph*). Figure 10.1 shows a sexual contact network of early U.S. patients diagnosed with AIDS. Node labels reflect both the state or city where the diagnosis took place and the order of AIDS diagnosis within a location, which is not identical to the likely order of HIV transmission. Edges represent sexual contact (symmetric), with arrows indicating potential transmission vectors (asymmetric) for the disease. P0 is the person believed to be the initial point of entry for the HIV virus into this contact network. Node color represents the condition(s) with which a person was diagnosed.

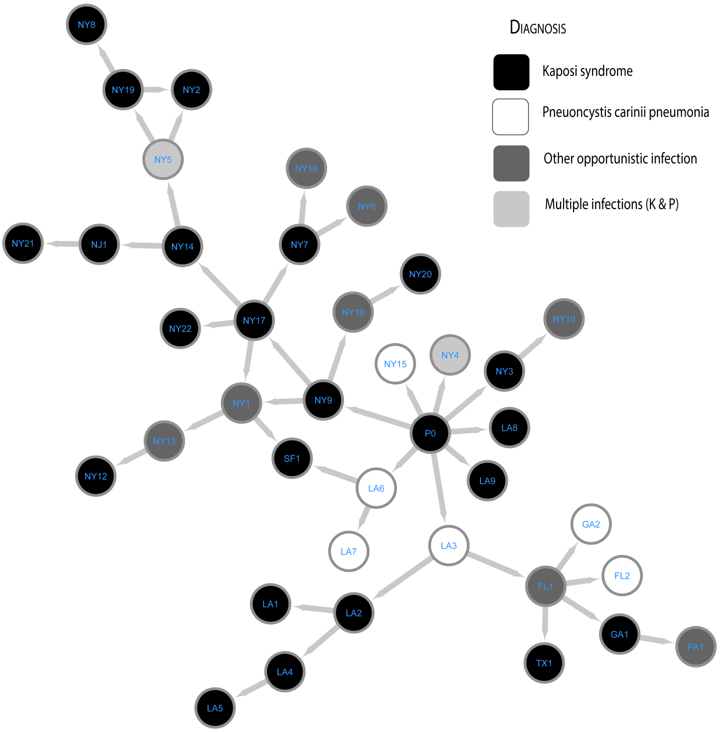

Figure 10.1: Sexual Network of Early Individuals Diagnosed with AIDS

At the most basic level, *density* measures the proportion of possible ties in the network. At one extreme, a fully connected (density = 1) network means that every node has a relationship (edge) with every other node, like a small group of close friends or collaborators. The subnetwork of NY2, NY5 and NY19 near the top of figure 10.1 has density 1. In most cases, however, graphs are *sparse* (density close to 0), particularly larger networks like collaboration across an entire discipline, friendship across a school, or partnerships between physicians licensed to practice in a state. The network in figure 10.1 has a density of 0.053. Each *isolate* (node with no adjacent edges) or *disconnected subgroup* is called a network *component*. *Centralization* measures the extent to which a small group of highly-connected nodes accounts for many of the paths between other nodes, while *clustering* measures the extent to which network components are broken into distinct, loosely connected subgroups.

Specific combinations of these network properties are tied to distinct types of network structures. The most basic structure is a *random network*. Random networks are often used for examples, simulations, or comparison standards, and occur when each edge has a similar or identical probability of being active. They are empirically rare because very few circumstances arise when context or shared characteristics have no relationship to the probability of a tie existing. *Scale-free networks* provide a closer idealized network structure, where the number of nodes with at least X edges follows a *power-law (exponential) distribution*. That is, most nodes have a small number of ties, and the proportion of nodes with at least X ties shrinks rapidly as X grows. Most empirical networks consist of a number of relatively highly-connected subgroups with a few individual nodes *bridging* subgroups to each other. Often, these bridge nodes are of high theoretical importance, for example, as key transmission vectors in the spread of disease or choke points in the diffusion of information. Cohesive subgroups or communities within a network can be distinguished by specific technical variations. The most restrictive type of subgroup is a *clique*, in which every group member shares an edge with every other; the least restrictive is a *component*, in which every member need only be reachable by tracing edges from every other.

Like networks, individual nodes can be evaluated and scored on a variety of network characteristics. Many are forms of centrality, the importance, however defined, of a given node within the network. The most basic type of *node centrality* is *degree*; that is, the total number of edges it shares with other nodes. O*ut-degree* and *in-degree* provide analogues to total degree for *asymmetric* or *directed networks*. In figure 10.1, P0 has a degree (and outdegree) of eight but an indegree of zero. The geodesic distance between two nodes is the minimum number of edges that it takes to connect them. For example P0 had contact with NY9 and NY9 had contact with NY1; the geodesic distance from P0 to NY1 is therefore two.

An individual node has high *closeness centrality* if the average distance to other nodes in its network component is low. However, in many cases, such as diffusion networks, *closeness* is less important than *betweenness*—the proportion of shortest paths (geodesics) a node is on. A node connecting two otherwise separated subgroups is sometimes called a *cutpoint* because if it weren't in the network, the components would be disconnected. Cutpoints have high betweenness. To understand the importance of cutpoints in medicine and epidemiology, consider NY17 in figure 10.1. Without NY17, transmission of HIV from NY9 and NY1 to the top section of the graph could not have occured, at least through this network. A final major concept of node centrality, *prestige centrality*, applies mainly to asymmetric networks. There are many types of *prestige centrality* measures, but all take into account the centrality of nodes tied to each node, rather than simply degree or geodesic distances, in assigning centrality scores.

Networks, nodes, and edges can have many more distinguishable properties. Often they are specific to particular disciplines or substantive research areas. Now let's consider how to assess if network analysis might be useful in your research and, if the answer is yes, how to design the early stages of a network study.

What is My Network?

Every participant in the Viral Networks Workshop was fortunate to have entered with a research project that was in some way "network" oriented. Perhaps it is surprising, then, that the most challenging question that I, as the data and visualization consultant, posed to many of them was, "What is the network you are studying?" It is encouraging, by the same token, that many participants remarked that being forced to answer this question up front was one of the most valuable technical elements of the workshops.

When trying to define a network, it is important to first consider three elements: the network's nodes, edges, and research context. Each of the three, at least in relation to an analytic project, hinges on two questions: what matters and what is measurable. In practice, the step of defining the network is often an iterative process: start with general ideas, try to define a network, check what you might actually be able to do in terms of finding and analyzing data, then refine the general ideas and try again until something workable coalesces.

I often recommend to people that they start the process by thinking about a hypothetical report on their research and drafting a title for the report that incorporates all three elements–e.g. "The Network of [edge relationship] between [nodes] in [research context]." When considering possible nodes, it is important that they share some common characteristic(s). In the early stages of their projects, a number of participants struggled with this because they tended to think of networks more like flow-charts, where anything could qualify as a node and any relationship as an edge. In principle, there is no problem with this; networks can be quite complex as long as the nodes and relationships are clearly defined. However, each additional type of node tends to limit network analysis' potential to serve as more than a glorified concept map. In some cases, more complex projects may involve constructing multiple related networks that can be compared or combined. It is usually helpful, therefore, particularly in the early stages of definition, to draw a mock-up of the network or networks of interest and think about how they might be analyzed.

The situation is slightly different for affiliation networks, which have two distinct types of nodes rather than one. These nodes are often called *actors* and *events* because early affiliation networks were based on co-attendance at specific events. I often find it helpful to think of them instead as *topics* and *ties*. For example, in an affiliation network of doctors and hospitals, where an edge represents having worked in a particular hospital, a scholar might be interested in understanding how doctors (topics) are connected by hospitals (ties) over time. Or, another scholar might be interested in how hospitals (topics) are connected across locations (attribute) by doctors (ties). In other words, in an affiliation network, the node that is the topic and the node that is the tie *is entirely dependent on the research question*. Thus, one hypothetical title for research on an affiliation network of doctors and hospitals might be: "The Network of Shared Doctors Between Army Base Hospitals during World War One."

Edges are the second element to be considered when trying to define a network. The edges of a network provide the relationship(s) of interest. Like nodes, the more comparable and clearly-defined the content of an edge is, the more likely the analysis is to be meaningful and understandable. Networks of scientific researchers, for example, can be constructed in a variety of ways. Some common examples include collaboration networks (A writes with B or is co-investigator on a grant with B), citation networks (A cites B), co-citation networks (A and B cite C), supervision networks (A served on doctoral committee of both B and C), and institutional affiliation networks (A and B were both at institution D at the same time). Each of these types of relationships is likely to be important in understanding the overall structure of a particular scientific network, or of scientific progress in general, but network analysis by definition provides a more complex (and hopefully more valid) representation than case-based models. Thus, only a very limited number of models are capable of simultaneously accounting for such a variety of network types.[1]

The final element to be considered when defining a network is *research context*. In many cases, research context will be readily apparent from the analytical question, especially for historians and other humanists, for whom analyzing sources or events within a defined corpus or timeframe is standard. Network research, however, often requires narrowing the scope or context being considered in order to obtain high-quality data, that can yield insights generalizable to other related contexts.

A pragmatic approach to defining a network is to force oneself to answer the question, "Given my general research goals, what is the most readily accessible type of topic (node), relationship (edge), and context that I could potentially measure or quantify to answer some or all of my research question?" For multiple workshop participants, the most clarifying step in this process came when I asked them to make a sample dataset with a small subset of nodes and edges. This exercise illuminated situations where membership in the set of nodes or edges was poorly defined whether through overly narrow definitions, reducing the quantity of available data, or overly broad definitions, leading to unclear data. For example, many corpuses of text are publicly available through online archives (such as Project Gutenberg or the Internet Archive) and can be used with techniques such as topic modeling (see ch 6 by Cottle) or Epistemic Network Analysis (see ch 8 by Ruis). Likewise, there are standard online sources for many types of scientific networks, such as PubMed or Web of Science. Remember, though, that not all networks need to be large to be effective. Archival data gathered on a single topic can often be conceived of as a network and then productively visualized or analyzed to gain insight that might otherwise have remained hidden if relying on close reading alone (see ch 1 by Runcie, ch 2 by Smith, and ch 7 by Archambeau).

Finding colleagues who are both interested in your topic *and* data-oriented can be a vital step in this process, whether they serve in a formal role (such as digital humanities specialists or data consultants) or an informal role, say, meeting over lunch to talk about ideas. Only one workshop participant had prior analytic expertise in the method they used for analysis, but with the help of consultation from a small number of analytic specialists and conversation with others in the workshop, each participant was able either to use network analysis to produce insight into their research questions or to determine that it was a poor fit.

Applying Network Analysis

Now that we've reviewed some basic network terminology and considered how to define a network research question, let's identify the typical steps a researcher in the humanities might go through when applying network analysis.

We've already identified the first step, which is to define the network, identify the context, and settle on a research question. Once this has been done, the next step is to make a trial dataset of a few nodes and edges. Network data can be stored in a number of forms, but the most common way is to use two tables, called a *nodelist* and an *edgelist*.[2] As the names suggest, a nodelist is a list of nodes and an edgelist is a list of edges. The nodelist includes columns with a unique identifier for each node, as well as any *node attributes*, such as personal or organizational characteristics, population size, group membership or word frequency. The edgelist minimally contains two columns, representing the two nodes related by each edge. If the data are *directed*, one column is considered a *source* and one a *target*. If edges have an indicator of strength (e.g. a *valued network*), there should be another column for *edge weight*. Any other information about the edges can be included in *edge attribute* columns. Identifiers in the nodelist and edgelist should match exactly. Comma-separated (.csv) or tab-separated (.tsv) text files, which can be created in any spreadsheet program, are typically interchangeable across software, but some programs may require different formats of input files; search the documentation for your program to find out preferred formats.

In cases where there are multiple relations or affiliations, network data can be quite complex and it may be worth considering if a database (in Access or SQLite, for example) may be more flexible, allowing you to export multiple combinations or structures of the data as networks. Unlike a single table or nodelist-edgelist format, databases can have many different tables, linked by identifiers (see data in ch 1 by Runcie for a relatively simple example).

In the case of relationship data, nodes and edges are fairly straightforward. For affiliation data, however, both types of entities (actors and affiliations) are represented as nodes in a dataset. Each *tie*, then, represents an actor being associated with an event or affiliation. This is also called a *bipartite network*, because there are two sets of distinct types of nodes that can only have direct ties between (but not within) groups. When analyzing affiliation networks, there are procedures for converting the bipartite network into a single mode network in cases where ties are based on how frequently two nodes of the same type are associated with the same nodes. Doing this allows you to focus on one type as the *topic* and the other type as a relationship.

Now it's time to create the dataset. The three main ways to do this are by hand coding, machine coding, and hybrid (or augmented) coding. This first trial dataset is typically made by hand, unless you are importing data from an existing database, such as Web of Science or PubMed (see ch 9 by Phillips), already in a network format. For smaller networks and archival research, the entire network may be hand-coded using the models above, customized to reflect the types of nodes, edges, and attributes included in your data. Machine coding is useful for very large or complex datasets, as well as data that was originally digital such as citation networks, text/topic networks (see ch 6 by Cottle), and web-scraped data. The advantages over hand-coding are time and scale, but it is also easier to miss poor-quality or irrelevant data. Hybrid coding is a relatively recent development and frequently involves coding a portion of the data by hand, then using either automated tools such as machine learning or crowdsourced workers to create a larger dataset modeled on the initial cases.[3]

The first two steps, definition and data creation, are fairly structured and should be undertaken at specific, definable points in the analytic process. The next two steps, ideally, should be iterative, with the researcher moving back and forth between adjusting visuals and considering the research insight they provide. Don't hesitate to consider multiple approaches to visualization. Visualization early in a project is intended to help discover patterns in the data that might be further investigated. Nicole Archambeau (ch 7) discovered through early visualization that, although there weren't notable gender or age patterns in canonization testimony, her analysis revealed a surprising pattern of people using the first plague mortality as a time marker, rather than a significant event. As you consider your early visualizations be sure to look at some basic network and node characteristics that are calculable in nearly every network package. Each iteration of visualization should reveal important characteristics of the network as well as answers to the research question.

As you move toward a final visualization, be sure to tease out the story your research is telling, in both its layout and design features. Visualizations are, above all else, a form of communication. They should be clearly labeled and free of visual elements that do not represent data (i.e. drop shadows). Often, peripheral elements, such as node labels, isolated nodes or very weak ties, can be removed entirely to improve clarity. Creating effective visualizations, like good writing, requires multiple drafts, critical reading by colleagues, experimentation with formats, and willingness to fail. (Always save backup copies of the data and each version of the visualization.)

Practical Advice

Assisting the cohort of scholars in the Viral Networks Workshop offered me a unique vantage point from which to observe the challenges that traditionally-trained humanists face when attempting for the first time to do network-related research. The following tips come directly from this experience.

First, not all research problems benefit from network thinking and analysis—though many can. To address this challenge, think creatively and critically about what network you are interested in and how it addresses your research question. For humanists, in particular, I would encourage starting by hand-drawing a model of what the visualization product could look like at the end—and consider how this outcome will advance your research agenda. Researchers often invest substantial effort into a project thinking it will fit a particular analytic model, only to discover that they had missed something important that they otherwise would have caught had they followed these preliminary clarifying steps. Nothing is more frustrating than spending hours hand-coding data, only to have to go back and repeat it all because of a simple oversight.

Second, get to know your data and talk about your early thoughts and findings with others outside your discipline. Doing so is vital to developing and communicating network research. A number of the authors in this volume detail the development of their research as they worked in Cytoscape or other software to explore and refine their visualizations. In every case, seeing the possibilities sparked new insight for their project—connections that might never have been made without turning a traditional history project into digital data. Each participant started the workshop with his or her own, distinctive project, but by coming together and talking with each other and a small number of outsiders, they were able to clarify their questions, goals, and processes, ultimately leading to an impressive array of chapters. Collaboration is a vital aspect of creative and scientific growth, even in disciplines where the solo scholarly endeavor is normative.

Third, any researcher who can produce an article or monograph can also succeed in creating a network analysis. The very process of applying digital humanities tools and methods to one's humanities research can be a powerful analytic stimulant. None of the projects here has the broad scope of the most prominent digital humanities projects, yet all benefited from the discipline required to turn their research materials into digital data and the possibility for unexpected discovery that comes from letting others, even computers, participate in the process.

Selecting and Learning Software Tools

Workshop participants worked primarily with two software packages, Cytoscape and Epistemic Network Analysis (ENA). These tools were chosen because of their ease of use and broad range of potential applications. Two additional packages, Gephi and the Python package scikit-learn, were used for their specialized mapping and text analysis capabilities respectively. In this section, I will provide some advice for getting started in Cytoscape and ENA, followed by an overview of other options and when they might be worth considering.

Cytoscape is a free network analysis and visualization package for all operating systems. It is most commonly used in health and biological sciences, although Miriam Posner has created an excellent tutorial,[4] used by many workshop participants, on Cytoscape for humanities applications. Additionally, Cytoscape has a large and growing collection of plug-ins, including ones calculating network and node statistics, downloading citation networks from PubMed, and allowing for easy publishing of interactive visualizations to the web.

The best way to learn Cytoscape is, frankly, to try it out. Original projects for all of the Cytoscape visualizations in this volume are available in the online supplements, and can give you a good feel for the software. When you first open Cytoscape, the splash screen will present you with options for accessing an existing project or creating a new one. In Cytoscape each project is a single file corresponding to related analysis or networks to which you can add multiple datasets, layouts, and style sets. The main window has three panels. When you open a project, the first one you will probably want to look at is the visualization at the top right. You can drag or use standard zoom gestures to get a better feel for different parts of the network, and many options are available by right-clicking on nodes or edges. You can also drag nodes with your mouse or change the layout using the Layout menu. On the bottom right is the Table Panel, where you can view or edit the source data Cytoscape used to produce the visualization. The Control Panel on the left is the heart of customization for the visuals, and allows you to select from multiple networks, adjust the appearance nodes and edges, and select subsets of the data. The Style tab in particular allows you to use colors, size, shapes, labels, or even images to represent node and edge attributes and help tell your network's story.

To import your own data into Cytoscape, start a new project and choose File-Import-Network-File from the menus and import the edgelist; then repeat the process with File-Import-Table-File and the nodelist. To add extra features like auto-imports of web data or network statistics, use the Apps menu. Once your data is imported, think about your research questions and how they might be elucidated visually and then play around with options. When you're satisfied with the product, you can save the project for use in Cytoscape, save the diagram as a picture, and save the project as a web page.

ENA is a relatively new software package, available for free, both through a web interface and the rENA package for R statistical software. Unlike Cytoscape, ENA's design is based on a specific methodology and not useful for more general exploration of networks. ENA answers variations on a single question: "How do a set of concepts co-occur throughout a corpus of coded material."

For instance, it is excellent for answering questions a researcher might have about a particular word or phrase, such as its usage and meaning vary over time or between contexts.

The original intent of ENA was to analyze and compare different stages and adaptations of educational activities, but it can be applied to any collection of sources that can be coded in terms of a small number of key ideas. Source data for ENA must match a specific format, and codes must already be created and applied prior to importing. The sample data provides helpful examples of the different elements of an ENA project, and the web interface can help guide new users through selecting variables. There are a number of good resources and tutorials to help you determine if ENA is right for your project, and to get you started with the web tool.[5] The web interface provides a user-friendly way to experiment with data and produce attractive and useful visualizations. The R package, while still in a preliminary form at the time of writing, is useful in documenting your work and making it available and replicable to others, as well as providing simple data transfer for current R users and a way to share datasets exported from the web tool.

Other widely-used, standalone network packages include Gephi, Pajek, and UCINET. Gephi and Pajek are free and cross-platform; UCINET is Windows-only and is free to try with full functionality for 90 days. Gephi is similar to Cytoscape in many ways, although the controls are less intuitive for new users. It is focused on visual design of networks, provides a great deal of customizability and multi-format exports, and has some key features that Cytoscape lacks, such as geographic network visualization with map overlays. Pajek provides a mix of both visualization and statistics features, and is particularly good for working with very large networks. UCINET has a larger variety of statistics and is among the best-documented, but its visualization tool NETDRAW is less refined and works best with smaller networks. While Cytoscape and Gephi work by importing all data into a single project that stays open and accessible throughout the session, Pajek and UCINET are more modular and require combining input and output files for each step of the process, adding flexibility but increasing the learning curve.

Network modules are available for a number of more general software packages, as well. The most user-friendly of these are NodeXL, a plug-in for Microsoft Excel for working with small to medium networks, and Tableau, an interactive data visualization tool. NodeXL's greatest advantage is its integration with Excel; editing data and moving between worksheets will be familiar to many users, and there is no need to export data to another package. Tableau, available free to students and educators, is excellent at rapidly producing clear visuals without the need to code, although its drag and drop interface can limit its flexibility. R and Python both have extensive network analysis packages, although R's are more full-featured and include many statistical procedures for simulation and modeling that are not available in other packages.

A final software class to mention here is interactive html and JavaScript visualization tools. Gephi, Cytoscape, R (via plotly), and Tableau all export visualizations that web users can visit and explore themselves, changing display options or even the network itself. However, a new collection of tools, such as d3.js and node.js, have emerged in the last few years to allow embedding network data natively in web pages with extreme customizability and interactive flexibility. Their application is limited by the need for fluency in their coding language, but they remain an option for high-impact visualization for code-savvy researchers or those collaborating with programmers. These tools are not limited to network data; they are designed as full-featured data visualization tools. To get a sense of what is possible with these packages, you can browse visualization galleries such as those at d3js.org or FiveThirtyEight.[6]

Getting Started on Your Own

At this point, some people considering network analysis for the first time may feel overwhelmed by the variety of options available. So how should one get started? The best options, depending on your access to support, are, first, to take a hands-on, instructor-led workshop or course in network analysis, and, second, to find a colleague who uses network techniques. In addition to departmental colleagues, many universities have statistics or research data consultation available through the library, statistics department, or social and demographic research centers. Tapping into experience in this way can save a great deal of frustration both on learning the language and processes involved and finding the right tools.

If in-person help is not practical or available, the next-best option is to start with a user-friendly tutorial or textbook. The Cytoscape tutorial by Miriam Posner (discussed above) combines an introduction to network concepts and data with application to real data. At present, the most accessible textbook on applied network analysis is *Analyzing Social Networks*, by Borgatti et al.[7] It provides both a readable introduction to a wide variety of network concepts and a good overview of the elements of network visualization, all using UCINET software. NodeXL, Pajek and Gephi all have hands-on books to help you get the most out of your chosen software.

I wouldn’t recommend starting with software documentation for the simple reason that all of the major packages assume existing familiarity with network analysis. Once you have experience with a single tool, you may choose to stick with it or you may discover it doesn’t meet your needs and try something else. Either way, just getting started, creating and working with network data, will be invaluable regardless of the tool you choose in the end.

Another option is to find a paper that employs methods or a particular visual approach you would consider adapting for your own research. The greatest advantage here is that you can more quickly discern whether your research question is a tractable network question and what tools or techniques may be most relevant. The challenge, however, is that many network papers are written by and for people who live and breathe network analysis or statistical programming. Often the techniques they use would be difficult if not impossible for a novice, even if they are an expert in the same subject area. Still, if you see something that makes sense to integrate in your research, you can try to learn a little more about the methods or ask a colleague if they are feasible for you. Understandably, this approach is best used in combination with the others; start with an idea of where you want to end, read carefully to find out how previous researchers got there, and then use that information to help select the tools or approaches you'll need to learn to pursue your research question.

Conclusion

My goal in this chapter has been to convince readers that, if they are successful researchers in their own substantive fields, more than likely they will be able to productively use network analysis provided that they take a few basic steps. First, they need to learn to think in terms of networks and network data. Second, their research questions and data sources must be appropriate for network analysis. And third, they must be prepared to match their goals to the appropriate tools and learning resources.

Based on my experience of the Viral Networks Workshop in the capacity of data consultant, I would encourage all humanities scholars to keep talking to colleagues, keep coming back to your research question and sources materials, and keep playing. With these conditions and exhortations in mind, you should have the tools to embark in a new direction toward network research, whether your networks consist of friends, enemies, letters, places, patients, doctors, ideas, or anything else. Whether you intend to become a network or digital humanities specialist or you simply want to enhance and complement other approaches, network thinking, tools, and visualizations are useful additions to your toolbox.

Endnotes

[1] Tom A. B. Snijders, “Statistical Models for Social Networks,” *Annual Review of Sociology* 37 (2011): 131-153.

[2] Examples can be found in the supplemental data for this chapter.

[3] Matthew J. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, 2017).

[4] Miriam Posner, *Creating Network Graphs with Cytoscape* (web page), https://github.com/miriamposner/cytoscape_tutorials.

[5] David Williamson Shaffer, Wesley Collier and A.R. Ruis, “A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social and Interaction Data,” *Journal of Learning Analytics* 3 (2016): 9-45.

[6] https://github.com/d3/d3/wiki/Gallery, https://fivethirtyeight.com/tag/data-visualization/

[7] Stephen P. Borgatti, Martin G. Everett, and Jeffrey C. Johnson, *Analyzing Social Networks* (London: Sage, 2018).