Chapter 5

The life cycle of a network study

How do we obtain new knowledge? A common distinction is between "*hypothesis*-driven" and "*data*-driven" research. Hypothesis-driven research (or "normal science") begins with a research question or a hypothesis that is built on existing knowledge or "paradigms." Then, driven by the question, researchers seek measurements or data that can answer the question. Einstein's theory of relativity is a good example of such hypothesis-driven research. The idea and mathematical theory was built on several hypotheses about how space and light behaves. The theory then produced several testable hypotheses that were later tested empirically to confirm[1] the theory.

By contrast, in data-driven research, the measurements, observations, or simply *data* are the driver of the research. Researchers may not necessarily have a concrete question in mind at the beginning of the research, but will identify interesting patterns from the data through an iterative, exploratory analysis. Kepler's laws are a good example because Kepler discovered these laws by carefully examining the data that had been collected without any concrete idea or hypothesis about universal laws in mind.

These two approaches are also tied to the nature of data that they deal with. Robert Groves, expert in survey methods and 23rd director of the US Census Bureau, classified research data into *designed* data and *organic* data [194]. (Sociologist Matthew Salganik referred to the latter as *found* data [412].) In traditional social science research, the researcher usually design surveys or data collection methods, which are then used to collect the actual data. Yet in many areas of data science, this model is flipped—usually, the data has already been collected as a byproduct (e.g., a social media company's user logs) without any research questions in mind. Researchers then *find* the usage of such data. Often, the researchers do not form the research question before examining the data. We will talk more about this distinction later in this chapter.

In practice, many research projects contain both aspects. The research questions and hypotheses are often iteratively sharpened by a better understanding of the data obtained through exploratory analyses. Sharper questions can also dictate additional data collection or inspire novel ways to dissect and combine existing datasets.

Although data-driven research is sometimes criticized due to its risk of "researcher-

[1] Or, more properly, failed to rule out the theory.

https://doi.org/10.1017/9781009212601.008 Published online by Cambridge University Press

degrees-of-freedom" and "p-hacking," it is important to understand that *hypothesis-free observation* is an integral part of the scientific process and plays a foundational role in science. At the same time, it is also crucial to understand the importance of research design and the risks of *p-hacking* and related issues.

In this chapter, we follow the sequence of steps that hypothesis-driven research would follow. This does not mean that this is the only way research can be done. We chose this sequence because the process of hypothesis-driven research is often overlooked in data science, and this process is followed at least during parts of even heavily data-driven research projects. We discuss specific tools and techniques underlying both approaches in later chapters, particularly Ch. 11.

5.1 Network questions

A network study often begins with a *network question*, meaning a question that needs to take connections (edges) or networks into account. Network questions can be questions that are about the structure and dynamics of networks themselves (e.g., "*what are the topological characteristics of this network?*") or questions about the impact of network structure on network elements ("*can we predict the clinical outcome of an individual by knowing the size and characteristics of their social network?*").

Note that the problem itself does not necessarily dictate whether it is a network question or not—some problems can either be studied as a network question or as a non-network question. For instance, understanding the spread of an infectious disease like COVID-19 can be studied either by considering the contact network between people and the spread of disease through this network or by ignoring the network. In fact, *network epidemiology*—considering network structure as critical elements to understanding epidemics—is relatively new and still not in the epidemiological mainstream.

5.1.1 Types of network questions

The potential questions can range from micro- ("*can we identify social influence between a pair of students?*") to macro-level ("*what is the degree distribution of this network?*"), and from descriptive ("*what are the communities in this network?*") to prescriptive ("*what kinds of interventions can we implement to enhance communication within a company?*"). One data scientist may just be curious about the properties of a new network that no one has seen before. Another may need to create the best link prediction algorithm to minimize user *churn* on their online social media service. The questions can also be defined only loosely and emerge from the exploratory analysis of a network.

In any case, the question will eventually dictate the type of network data that is needed as well as the potential methods of analysis. Even if the analysis began with hypothesis-free exploration, the question should be clarified as quickly as possible to avoid critical mismatch between the question and the data, which can lead to a lot of wasted effort and flawed analysis.

https://doi.org/10.1017/9781009212601.008 Published online by Cambridge University Press

5.1.2 Questions guide operationalization

It often goes under-appreciated that numerous networks can be defined (or "operationalized") from a *single* dataset. For instance, there is no unique "***Twitter network***"; numerous entities can be considered nodes, each of which will result in a different network. Do we consider each user account as a node? How about hashtags, tweets, words, or groups of users? How about the edges? Do we consider the "following" relationship as a directed edge? Should we only consider an edge to be present when the relationship is reciprocated (both i follows j and j follows i), leading to an undirected network? Or, maybe we can construct a bipartite network between users and hashtags? We can keep asking these questions, in social media data and in fact in nearly all data we expect to encounter.

The point is that there are often numerous choices one has to make even to *define* a network, especially when the raw dataset is rich and complex. Consider a ***protein–protein interaction network***. Even if we ignore all the complexities about defining nodes (proteins), several networks can be defined because there exist multiple types of edges. One network can be defined by the edges that have been discovered and studied in the *literature*; another network can be defined by the edges discovered through a systematic *pair-wise* interaction probing (e.g., based on the Y2H (Yeast Two-Hybrid) assay); another can be defined by the AP–MS (Affinity Purification–Mass Spectrometry) method that tends to discover *cliques*[2] rather than binary edges. Depending on which types of edges you include in your network, the network will exhibit very different structural properties as well as biases. Overlooking these biases and characteristics will ensure flawed analyses and results.

Nevertheless, it is your *question* that clarifies what could be the *right* or *ideal* operationalization of your network. For instance, although there are numerous networks that we can identify from Twitter data, once we have a concrete question in mind, it becomes clearer which network fits the question. Do you want to study how social media users are exposed to misinformation? This probably requires you to think about the network structure through which the misinformation can flow, the "following" network on Twitter. Now we have a clue what network to use. Or do you want to study how Twitter users engage in political debates? Then it will be critical to examine the network of "mentions" and "replies." Although these examples are probably still too simplistic, the point is that, in many cases, there are numerous *networks* that coexist in the same dataset and it is crucial to clarify the question and the corresponding, reasonable operationalization of the network.

More often than not, it may be impossible to collect or work with the ideal dataset, but even so, it is still important to clarify what *could* be the ideal network data because it lets us identify the assumptions, compromises, and biases that we introduce when choosing the final network data. For instance, let's assume that we are interested in studying how conspiracy theories about vaccines spread across society.[3] It is of course impossible to work with data that capture *every* possible social communication involving

[2] A clique is a completely dense subgraph. It is a set of nodes within the network where every pair of nodes is connected.

[3] Antivaccination campaigns, often rooted in misinformation and fear, have plagued vaccination efforts almost from the very beginning of Jenner's smallpox inoculations [10].

the conspiracy theories. Yet, it is possible to narrow the question to something feasible, such as: "*how did a particular conspiracy theory spread on Twitter?*" or "*are there Facebook groups that promote vaccine conspiracy theories?*" In this process of finding a feasible study, we inevitably introduce biases and assumptions, such as the study population (those who use a particular social media vs. everyone) and the types of communication. It is critical to explicitly identify these biases and assumptions.

Especially when working with existing data, it is easy to overlook these simple, yet fundamental questions. The network data that is currently available may not be right to address the question and it is essential to collect new data or transform the current data. So, *what is your question*?

5.2 Collecting, constructing, and cleaning network data

Once we have a question, we can concretely think about the ideal network that we would like to analyze. Sometimes we can and should collect the network data ourselves; sometimes we need to clarify how to operationalize the network from an existing dataset; in other cases, we should find an existing dataset.

5.2.1 Designed data vs. organic data

As discussed earlier in this chapter, one important distinction about data is whether the data is *designed* or *organic* (or "found") [194, 412]. Designed data refers to datasets that are collected to answer specific research questions. Organic data, on the other hand, refers to datasets that were collected for purposes other than the questions being asked. They may be collected by a company to answer their own questions about products, users, etc., but are later repurposed for research. Designed data tends to be small (because they are collected specifically for the research in question) but high-quality (same reason); organic data tends to be big but may not necessarily contain the exact information that can best answer the question.

Let us talk about an example. If we ask a question about the spread of vaccine conspiracy theories on Twitter, we can either design the data or use organic data. Designed data would most likely mean that we perform a survey that asks people on Twitter about the exact things that we want to know about them. The organic data route would mean that we simply take whatever the Twitter data already contain and try to find with it, as best we can, reasonable operationalizations and proxies.

When we design our datasets, because we are already asking the exact questions that we want to get answers to, the usual barrier is the ability and resources to collect enough data. On the other hand, when we work with existing data, the most critical question is often about finding the best proxies for the certain quantities that we really want to estimate and dealing with all kinds of biases present in the data.

Returning to the misinformation example, Twitter is (presumably) not *designed* to gather vaccine misinformation data, so we will necessarily have organic data. When we use the organic data that Twitter is already collecting, we would have to take data from Twitter and build a network from the data. This process again involves lots of choices. Usually it is not possible to access Twitter's full database (unless you work *at* Twitter)

https://doi.org/10.1017/9781009212601.008 Published online by Cambridge University Press

meaning we would need to collect what Twitter makes public. There can be multiple ways to do so (e.g., using the "Streaming API" or the "Search API"[4]) and different data sources may introduce different types of biases. The operationalization of network edges also requires choices. Should we collect who follows whom? Should we simply collect retweets? How about mentions? Shall we combine all of these link types or just pick one of them? These are all tricky questions that need to be carefully evaluated based on the research question and the characteristics of the edge types.

Even after defining a network, questions remain. There are other types of data we can collect. For instance, we need to decide how much information we want to collect about nodes and links. Do we want to collect the profiles of the users? They may change, so do we want to monitor them over time? We can go even further. Do we want to contact those users directly and perform surveys? Directly asking people can lead to invaluable data that lets us peek into their social communication outside Twitter.[5] We can also choose to collect more information about individual edges. Say we want to use only the retweets. Then, should we collect various information *about* those individual retweets, such as the number of retweets, those who retweeted a particular tweet, timestamps of the retweets, and so on? There are numerous metadata (Ch. 9) that can be potentially helpful later. All these decisions affect how the results can be interpreted and how much the study can reveal.

Unfortunately, the importance of these questions is often overlooked. Many studies will simply choose the most convenient or already walked paths, which can create a mismatch between what they are asking and what the data represents.

5.2.2 Exploratory and confirmatory network analysis

Broadly speaking, given an extracted network, analysis can take two paths, exploratory or confirmatory. Both paths can rely on statistical, computational, or mathematical tools, tools which form the bread-and-butter of network science research. And they are not exclusive: the tangled path of real science often follows both.

Exploratory network analysis Many network questions can be directly answered or illuminated greatly by performing exploratory network analysis. Exploratory network analysis usually involves network visualizations as well as measurements of network statistics. Computational methods, such as community detection or graph embedding, can also be employed for this analysis as well. The goal of exploratory analysis is to understand the overall network structure and to guide further analysis.

Confirmatory network analysis On the other hand, confirmatory analysis aims to test a concrete hypothesis by employing statistical models. For instance, consider a protein–protein interaction network derived from the Affinity Purification–Mass Spectrometry (AP–MS) technique discussed earlier. AP–MS discovers interactions using a

[4] An API or Application Programming Interface is a specification that allows computer processes to communicate with one another and send and receive data. In this case, the API is a specification created by Twitter that says, in essence, "Here is how you can ask us for data and this is what the data will look like if you are allowed to see it and ask for it correctly."

[5] Surveys should always be done following appropriate ethical guidelines; Ch. 3.

https://doi.org/10.1017/9781009212601.008 Published online by Cambridge University Press

"bait protein" designed to capture an entire interacting complex of associated "prey" proteins that can then be identified by subsequent analysis. As discussed above, from the network's point of view, this technique does not sample links independently but instead gathers entire dense subgraphs, groups of nodes called cliques which are completely connected. A network derived from AP–MS will not be representative of the entire structure—for example, we expect triangles, cliques of size 3, to be overrepresented, perhaps heavily so. We can test this hypothesis by building a null model that captures the overall density of our AP–MS network but randomizes away the complexes, then we can compare the observed quantity of triangles in the real data with what we would observe if the null model held true.

We dive deeper into network exploration and confirmatory analysis in Ch. 11; Ch. 13 focuses on visualizing networks, which are often used for presentation but are also useful for network data exploration.

5.3 Iterating on the cycle

Of course, the story does not end after collecting, cleaning, and analyzing the network data. The results of a study are not open and shut. Instead, as the current study addressed your original questions, new questions will emerge. The data were limited to a single time period; does the result hold at other moments? The network only consisted of links from a given layer; what about other layers? The network came from an API that aggregated all activities; do the results hold under a different aggregation? Or when the API changes?

The net result of this is that the outcome of one study will inform the starting point of the next. A new network dataset can be collected and now, armed with the knowledge gleaned from the previous study, we can learn better than before about the system of interest.

5.4 Summary

Network studies follow an explicit form, from framing questions and gathering data, to processing those data and drawing conclusions. And data processing leads to new questions, leading to new data and so forth. Network studies follow a repeating life cycle. Yet along the way, many different choices will confront the researcher, who must be mindful of the choices they are making with their data and the choices of tools and techniques they are using to study their data.

Bibliographic remarks

The tension between open-ended exploration and hypothesis-driven confirmation has been at the foundations of science since the enterprise began. Yanai and Lercher [500] discuss some of the disadvantages of having a hypothesis, that it may lead researchers to fixate their attention on the preexisting question, causing them to miss out on other, critical features within the data. Yet being completely free of hypotheses is also risky,

https://doi.org/10.1017/9781009212601.008 Published online by Cambridge University Press

especially if statistical tests intended for confirmatory work are used in an exploratory fashion. Head et al. [207] discuss the prevalence and effects of such practices on scientific research. The long-standing tension between these competing views can only mean that they must be synthesized. As Tukey writes in his seminal book on exploratory data analysis, "[E]xploratory and confirmatory can—and should—proceed side-by-side" [465].

Readers interested in learning more about the dichotomy between designed and organic data are encouraged to consult Groves [194], which includes interesting examples of approaches that use both types of data. Salganik [412] discusses the dichotomy further, in particular in regards to the "big data" era and its influence on social science research.

Readers interested in learning more about using online social network and social media data may wish to consult Russell [409]. Kumar et al. [260] cover further the various approaches and nuances to using Twitter in particular, going beyond the examples we discussed here. Keep in mind that the world of social media is a fast-moving place, and it is likely that these sources will be outmoded in some respects despite still serving as pertinent introductions to the area.

Exercises

5.1 Suppose you have been hired by a large college campus to study how college students spend time together. You realize that college students often spend a lot of time on their computers and smartphones! If you have access to data from the campus wireless network, you should be able to track students, especially if there are a large number of wireless access points.

Write a brief study proposal to give to the college to convince the IT department, which runs the wireless network, to give you access to data that can help you answer your question. In your proposal, map out how a network where nodes are students will be constructed from the data, what it tells you about how students spend time together, what it *doesn't* tell you about how students spend time together, and what kind of next steps may be involved after your study is complete.

5.2 What ethical concerns are there with the data you wish to collect from the previous question? How can the study be modified to address these concerns?

5.3 (**Focal network**) Consider the flavor network. This network was derived from a reference text used by food chemists to give specific flavors to new foods. Suppose a new edition of this book is released. Briefly, describe a study to check the existing flavor network and see if it needs to be updated. Assuming it does need to be updated, describe how to go about doing so.