

Chapter 7

Extracting networks from data — the “upstream task”

The goal of this chapter is to recognize that, while there are cases where it is straightforward and unambiguous to define a network given data, often a researcher must make choices in how they define the network and that those choices, preceding most of the work on analyzing the network, have outsized consequences for that subsequent analysis.

7.1 What is it?

Sitting between gathering the data and studying the network is the *upstream task*: how to define the network from the underlying or original data. Defining the network precedes all subsequent or “downstream” tasks, tasks we will focus on in later chapters. Often those tasks are the focus of network scientists who take the network as a given and focus their efforts on methods using those data.

i The *upstream task* is to define the network from the data before you begin to analyze it. Sometimes this is easier said than done: researchers often face choices, sometimes difficult choices, before they can begin to study their network.

The simplest way to visualize the upstream task is to ask yourself two questions: “*What are the nodes?*” and “*What are the links?*” By focusing on these questions, while they seem rather elementary, you can at times reveal important details about why the network was made and even if it should have been defined differently.

Consider these two questions as we discuss several examples below.

Example: social network In social network analysis, a common data source has been social media services like Twitter. Some provide a programming interface (called an API) to retrieve data on users and their activities. But there are multiple ways to generate the network from social media activity data. For instance, from Twitter, one

could construct the network using mentions, retweets, and followings. In these cases, the nodes are Twitter users but how to define links may not be clear. Which definition should we choose? Should we simply gather all of them and merge them, defining links between users when any mention, retweet, or following occurs? This can be a critical mistake depending on the research question. It has been shown that each of these connections carry different meaning [72, 115]. In political discourse¹ on Twitter, *retweets* strongly signify agreement with the original author. By contrast, *mentions* tend to cross the political aisle and are used as a channel for fighting and mockery of political opponents. In other words, the retweet network captures in-group relationships while mentions capture both.

Example: protein–protein interaction network Protein interactions are measured via high-throughput experiments (Ch. 6). When defining a network such as HuRI (Ch. 2), we need to pay attention to our experimental methods because the characteristics and biases in the methods strongly affect downstream tasks and network properties. For instance, the AP–MS and Y2H methods extract different types of links (Ch. 5), leading to potentially very different networks. But often choices need to be made, not just between methods, but within a method, to use it appropriately. To build the HuRI network, for example, the Y2H method was applied to an $N \times N$ search space of proteins nine separate times. And three different versions of Y2H were used [283]. Luck et al. varied the versions and replicated their screenings specifically to enhance the robustness of the discovered interactions. Future studies may vary their protocols further. Overall, we can see how different experimental methods, and choices when applying them, can yield very different pictures of the network being inferred.

Example: brain network Neuroscientists use imaging experiments to infer the hidden structure and functional dynamics of the living brain (Fig. 7.1). As with protein interaction networks discussed above, experimental protocols will affect the final network being extracted. The brain scanner is part of the upstream task. The field of neuroimaging has taken great pains to understand the most appropriate use of imaging studies, with sometimes great debate as to their ability to yield good descriptions of the brain. From this work has arisen a field of statistical analysis aimed at inferring the nodes and links of brain networks, the connectome. These analyses include methods to pull out signals from time series measurements of blood oxygen levels in the brain, the central focus of functional MRI (fMRI) imaging. Moving from these signals to a network requires many choices of algorithms and parameters, all of which influence the final form of the brain network. The upstream task in network neuroscience is rich and complex.

7.2 Why does it matter?

We emphasize the importance of the upstream task because everything subsequent depends on it. Perhaps you wish to study the community structure in your network but

¹ If you can call it that.

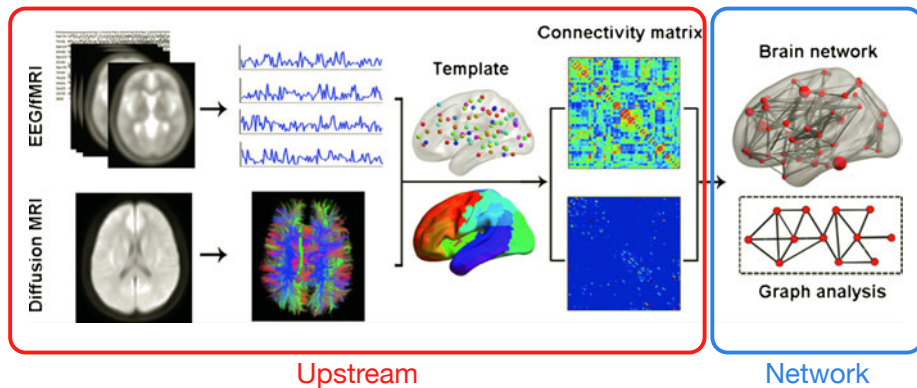


Figure 7.1 An illustration of the upstream task in network neuroscience. Here (left-to-right) functional or structural MRI data are recorded and processed into standardized time series or fiber bundles, respectively, then connectivity matrices are generated which are finally processed to make networks. Many choices are made along the way, from what algorithm to use for standardizing the data to what measure to use for comparing the time series. Network neuroscience is an example where the upstream task is highly visible and well documented due to its complexity, but many other areas feature networks drawn from comparably involved upstream tasks. Figure adapted from [94].

your definition of links depends on a data processing algorithm. If a small change to a parameter of that algorithm leads to a drastic change in the network's structure, then is your discovered community structure fundamental to the data you're investigating or is it simply due to the data algorithm?

Pay close attention to the “data generating process”!



Always think critically about the upstream task.

When *you* perform the task:

- What would change if you completed the task differently?
- Can you check whether your research results are robust to changes to the task?


When *someone else* performed the task:

- Do you have enough information to understand what they did and how?
- Did they do it in a manner appropriate to your problem?


A further effect of drawing attention to the upstream task is that it shines light on an important aspect of *data provenance*. As network data are shared, researchers can become fixated on the network and lose track of the preceding work, the extenuating circumstances, that went into creating that data.² When this happens there is a risk of

² Of course, fixating on simple messages while losing track of extenuating circumstances is by no means limited to network studies. Many scientific problems develop a “folk wisdom” where a fact or figure gets passed around, taken for granted and assumed true, all while being supported by one or a few studies or experiments that were limited in scope or even incorrect. One example is the idea that in cold weather you

unintentional misuse, you may draw a poor conclusion because of an assumption that went into the construction of the network which you did not know about.

 Take care when information on the upstream task is missing. Imagine if you start from scratch, would you be able to arrive at the same network as what you currently have? If you do not know everything about how the network was generated, you may be faced with reproducibility problems down the line.

A corollary to this holds when you are performing the upstream task yourself. Always document all aspects of this task and be prepared to replicate the task—there’s a good chance you may need to check if your later results are robust to the task by changing what you did to create the network and testing if your conclusions also change.

 Always document your upstream task thoroughly. Ensure this information stays with the network you extract. Be prepared to modify and repeat the upstream task.

7.3 Summary

This brief chapter discussed the upstream task, defining the network by creating a process to extract the network from the gathered data. Envision the upstream task by asking yourself, “*what are the nodes?*” and “*what are the links?*,” with the network following from those definitions. You will find these questions a useful guiding star as you work, and you can learn new insights by re-evaluating their answers from time to time.

Are you satisfied (currently) with the network you’ve extracted? Good, we can now turn our attention to incorporating non-network data (Ch. 9), further refining the network (Ch. 10), or exploring the network (Chs. 11–13 and beyond).

Bibliographic remarks

Little ink has been spilled on the importance of the upstream task in network science. A notable exception is the excellent perspective piece by Butts [88], which asks such simple—but foundational questions—as “when is a node a node?” and “when is an edge an edge?”

Exercises

- 7.1 Describe a network where there is one answer to the question, “what is a node?” and from that answer there is really only one answer to the question, “what is an

lose most of your heat through your head, based on a single Army study and since called into question [479]. Another, more chilling example is the considerable confusion early in the COVID-19 pandemic that arose over whether the virus spread only over short distances in respiratory droplets or was “airborne” in particles smaller than 5 microns, which spread much farther. But this 5-micron cutoff, well supported by policy, is not well supported by research and a better cutoff for farther spread may in fact be 100 microns [482]!

edge?” Describe another network where, given what nodes represent, there are many answers to the question, “what is an edge?”

- 7.2 (**Focal network**) Consider the data generating process of HuRI. What biases could be present due to it?
- 7.3 (**Focal network**) Consider the developer collaboration network. Nodes represent GitHub users, links exist between developers who coedit files. Describe a few other meaningful definitions of links for these nodes.
- 7.4 (**Focal network**) Write a table summarizing, for each focal network, answers (in your own words) to the questions, “what are the nodes?” and “what are the links?” From these answers, do you see any similarities or differences between the focal networks?

