Data to Network: Scientometrics

Scientometrics (https://en.wikipedia.org/wiki/Scientometrics) is the field of study which concerns itself with measuring and analysing scholarly literature. Network science is commonly used to do scientometrics. In most cases, the original data is a scientific respository such as https://dblp.org, https://arxiv.org, https://hal.archives-ouvertes.fr, etc.
A natural way to create networks from publications is to use articles as nodes and citations as (directed) edges. There are 2 limits to this approach: 1)Data for citations is much less available than title, authors, abstract and affiliation data, and 2)The created graph has very peculiar properties: it's called a DAG (Directed Acyclic Graph), and many notions such as betweenness or communities have non-conventional interpretations.
We will adopt another approach, which consists in creating co-occurence datasets. The best known is the co-authoring network (nodes are authors, a link represent collaboration captured as co-authoring a paper). However, we can create many variants of this, in which nodes can be articles, institutions, journals, keywords, etc. and edges represent some form of direct or undirect co-occurrence.

1. Obtaining publication data
   (a) We will use HAL, a French online repository on which French researchers are required to upload references of their publications. Follow the tutorial at https://colab.research.google.com/github/Yquetzal/Teaching_notebooks/blob/main/Networks/HAL_class.ipynb to learn how to make requests to this repository using python.
   (b) Practice a little bit, by trying to write your own request. You can get inspiration from a get started notebook: https://colab.research.google.com/github/Yquetzal/Teaching_notebooks/blob/main/Networks/HALquickStart.ipynb
   (c) Find a topic of interest, for instance all papers written about COVID in Lyon, or any other topic of interest to you, that is not too foreign to you (CNN, DNA, Pluto, Soccer, ...). In the following, you will create a network from this data and interpret it. Be careful with ambiguous terms, you might need to add constraints: forbidden words, search in full abstract or keywords, add additional words, etc.
2. Drawing, cleaning data
   (a) A first step is often to draw your network to see how it looks like. You can use your favorite solution: networkx, pyvis, Gephi... (If you don't know what to use, I recommend Gephi)
   (b) Explore the data, and search for unexpected, unsatisfactory aspects: maybe you have too many nodes to draw, or too few. Maybe you will spot some publications/authors/lab/word that should not be there. In any of these situations, try to refine your request
   (c) Another aspect to play with is when postprocessing the data: typically, you will remove nodes that appear too few times, and similarly with edges. This postprocess is not *cheating*, very much the contrary: real data automatically extracted from the real word is always noisy, unreliable, and generally of poor quality (sorry! Welcome to the real world :)). Removing elements for which you have a single observation is usually a good practice. But removing more than that is often interesting too. It shifts the meaning of the edges, for instance, from "two institutions have at least one work in common" to "two institutions have a regular, sustained collaboration".
3. Analysis
   (a) Run with your favorite tool a basic analysis: identification of most important nodes, etc.
4. Feedback
   (a) Share your insights with other students

5. Going Further
   (a) An interesting possibility with this data is to analyze evolution along time. You can plot the activity in a field, or of a researcher, institutions, etc. through several years and observe the evolution
   (b) Networks of coocurences of words in title, abstracts, etc. were not covered in the example. They might be very useful to understand the organisation of a field: starting with COVID for instance, you might find a network composed of all the different research aspects of the topic: medical, social, economic, etc.
   (c) Another way to access to scientometrics data is to download the full dataset of a repository, for instance ArXiV https://arxiv.org/help/bulk_data or Plos (https://plos.org/text-and-data-mining/). Of course, managing such large quantity of data has its challenges...
   (d) Google Scholar is probably the most complete scientific database, unfortunately it does not offer an API or open datasets. Some people have built custom tools to get information from it. It's slower than an official API, but it works. See for instance (https://scholarly.readthedocs.io/en/latest/?badge=latest)