Lab project : Analysis of Real Networks

Complex Systems

M1 Computer Science - Fall Semester 2023-2024

Université Côte d'Azur

Christophe Crespelle
christophe.crespelle@univ-cotedazur.fr

The goal of this project is to analyze two complex networks derived from real data: one small (between 3K and 12K nodes) and one larger (between 35K and 105K nodes). You will find the two networks assigned to you, as well as a link to download the raw data, on the Moodle space of the course.

Work to be done.

The first thing you need to do is filter the raw data from these two networks and calculate basic statistics (as done in Lab 1). Then, you will choose one or two directions for further analyses. These analyses should aim to answer one or more questions that you choose to investigate about the structure of these two networks. Below, I suggest several possible themes. You can choose one among them, mix and cross questions from several suggested themes, or, even better, choose a direction you have imagined yourself.

Submission.

Your submission will consist of a report that includes the results of your analyses. This report must be in PDF format, with a maximum of 10 pages (including title, references, etc.), using a font size of at least 11pt. It may contain appendices that will be read at the discretion of the evaluator. My advice is to select the analyses that yield the most interesting results from those you have conducted and place the others in an appendix. **You must explicitly and very clearly formulate the questions your analyses attempt to answer** and give an objective description of what your plots contain. Pay special attention to the readability of the axis labels. You must also provide a critical analysis of the results highlighted by these plots. Finally, for each statistic you calculate, you must provide in the appendix the tools, command lines, or code you used to produce these statistics. If you have written code, send the source files in addition to the report.

1 Getting Started and Filtering Data

The first step of your project will be to filter the raw data of your two networks and format it as described in Lab 1. You will perform all the requested analyses from Lab

1: calculate the four fundamental properties of your networks and compare them to those of the Erdős-Rényi and configuration models with the same parameters. If you wish, you may skip calculating the distribution of the local clustering coefficient per node (Question 9). If you encounter difficulties calculating the average distance, remember that you can simply do it by sampling (`https://www.youtube.com/watch?v=E0-_aSMkwGw`).

Below, you will find several suggestions for possible further analyses. You must explore at least one or two topics. However, you don't to choose them from those proposed here. Instead, you can mix the questions from different proposed directions or propose another direction yourself. Do not hesitate to do so : the originality of the analyses conducted will be highly valued in the grading.

2 Distances

This section proposes to deepen the metric properties of the network, that is, distances and paths within the network. You may, for example, investigate the following questions.

Question 1. Compute the distribution of distances in the network.

Definition 1 (Eccentricity)**.** *The eccentricity of a vertex u is the maximum distance between u and each of the other vertices in the graph:* $exc(u) = \max_{v \in V \setminus \{u\}} \{dist(u, v)\}$.

Question 2. Compute the distribution of the eccentricity of vertices.

Question 3. What is the minimum eccentricity of a vertex, and how many vertices achieve this minimum?

Question 4. How many pairs of vertices realize the diameter? How many vertices are involved in such a pair?

Question 5. What is the minimum number of vertices to remove from the graph in order to decrease the diameter by 1? By k?

3 Centrality

One possibility is to study and compare different centrality metrics in the network.

Question 6. Compute at least two notions of node centrality and study the correlations between their scores.

Question 7. Study the correlations of the rankings these different centrality metrics provide for the nodes.

Question 8. On which type(s) of nodes do they agree? On which type do they disagree?

Question 9. For metrics defined also on the edges (betweenness, eigenvector, PageRank), study the correlation between the centrality of an edge and those of its two incident nodes.

4 Communities

Another possible direction is the study of the community structure of the network.

Question 10. Compare the results of different community partitioning methods. Are they similar? Where do they differ?

Question 11. Study the stability of the Louvain method when changing the order in which the nodes are processed. Are the different partitions obtained similar? Where are they identical? Where do they differ?

Question 12. Examine the structure of the network between communities. How many edges are between distinct communities? How many are within some community? What is the distribution of the number of edges between two communities? What is the degree distribution of a community?

Question 13. Study the frontiers between communities. What are the nodes at the frontiers of multiple communities? Are there many of them? What are their characteristics?

Question 14. Analyze the community structure within the largest community.

5 Core/periphery of the Network

The question pursued here is to uncover the structure of the core (k-core) and the periphery of the network and to understand how these structures differ from that of the entire network.

Definition 2 (see `https://en.wikipedia.org/wiki/Degeneracy_(graph_theory)`)**.** *The k-core of a network is the set of nodes that remain after iteratively removing all nodes of degree at most $k-1$, also removing those that become of degree at most $k-1$ during this process. The core of the network is the k-core obtained for the maximum k such that the k-core is non-empty.*

Question 15. Determine the k-cores of the network. How does the number of nodes in the k-core evolve as a function of k?

Question 16. Where are the edges of the network located in relation to its different layers of k-core? For example, one could:

- plot the distribution of the difference between the layers the two endpoints of an edge,
- plot the evolution of the number of edges in the k-core as a function of k,
- plot the evolution of the number of edges incident to a node in the k-core as a function of k.

The *rich club* coefficient (see `https://en.wikipedia.org/wiki/Rich-club_coefficient`) measures how well-connected high-degree nodes are to each other.

Question 17. Determine the *rich club* coefficient of the network.

Question 18. Is there a relationship between the core and the rich club?

Question 19. Do the core or the rich club have the same structure as the entire network?

Question 20. What are the properties of the periphery of the network, that is, the part that is not in the core?

6 Link Prediction

One can also be interested in the structure of the most predictable links in the network and the possible relationships with other structural properties of the network, such as its communities, its core, the centrality of its nodes, etc.

Question 21. Use a link prediction method to identify the pairs of nodes most likely to form a new edge. Where are these pairs located in relation to the communities and the core of the network?

Question 22. Rank the existing links in the network by decreasing predictability. To do this, remove one link from the network and compute the score assigned to it by the prediction method.

Question 23. How do the properties of the network evolve when we add its links one by one in decreasing order of predictability? Does the network of the most predictable links have the same structure as the entire network?

Question 24. Is there a relationship between the predictability of the links in the network and the importance of the nodes to which they are attached?