Social Network Analysis

Aya ben hriz

INTRODUCTION

- We will work with a database on recruitment juries constituted by two students of the class during their internship from 2017 to 2020.
- **The goal** of this learning unit is to realize data visualizations in order to get a better understanding of the social networks that exist within higher education.

Methodology

- **Tools:**
  - NetworkD3
  - Igraph
  - Ggplo2
  - Gganimate
  - Circlize

- **Methods:**
  - Multipartite graph
  - Edge list
  - Adjacency matrix
  - Sankey diagram
  - Chord plot
  - Market Basket Analysis

Results

Results I

Figure 1. Multi-partite external jury members graph for 2017.

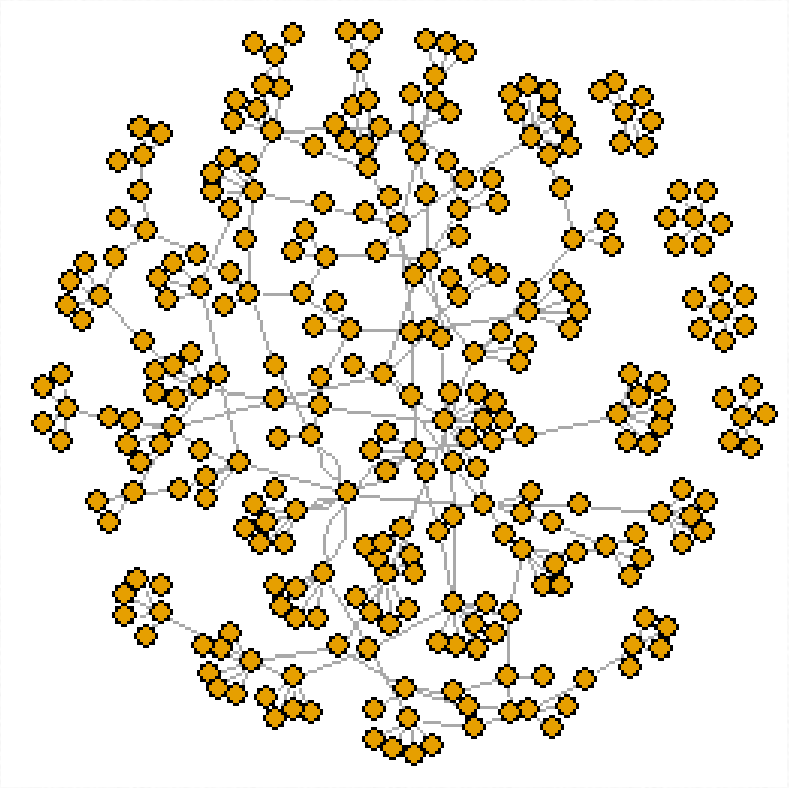
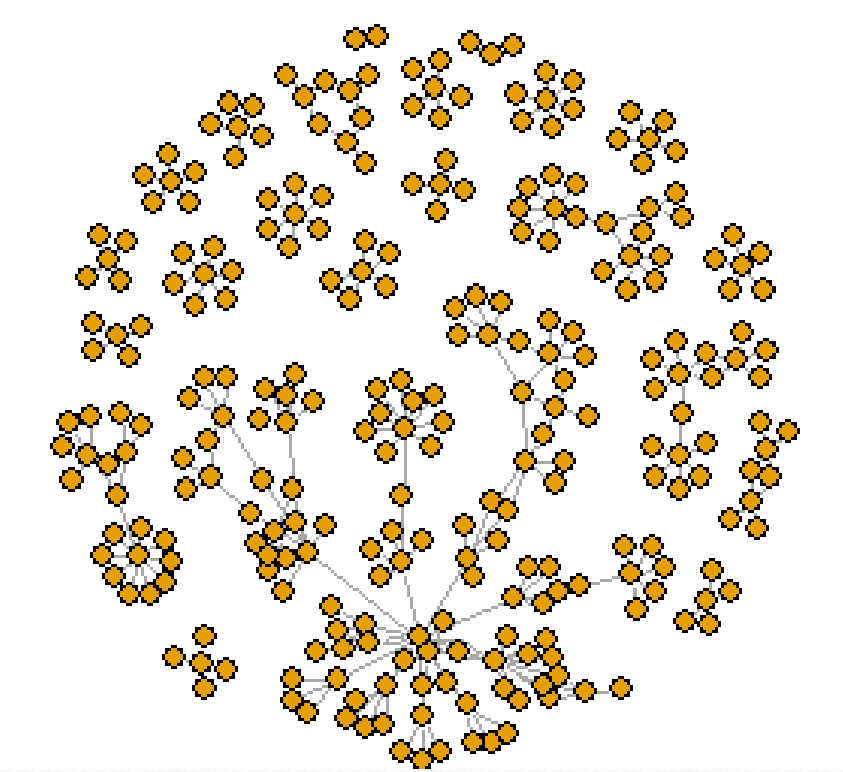


Figure 2. Multi-partite internal jury members graph for 2017.



Results II

Figure 3. Multi-partite external jury members graph for 2017changing betweenness.

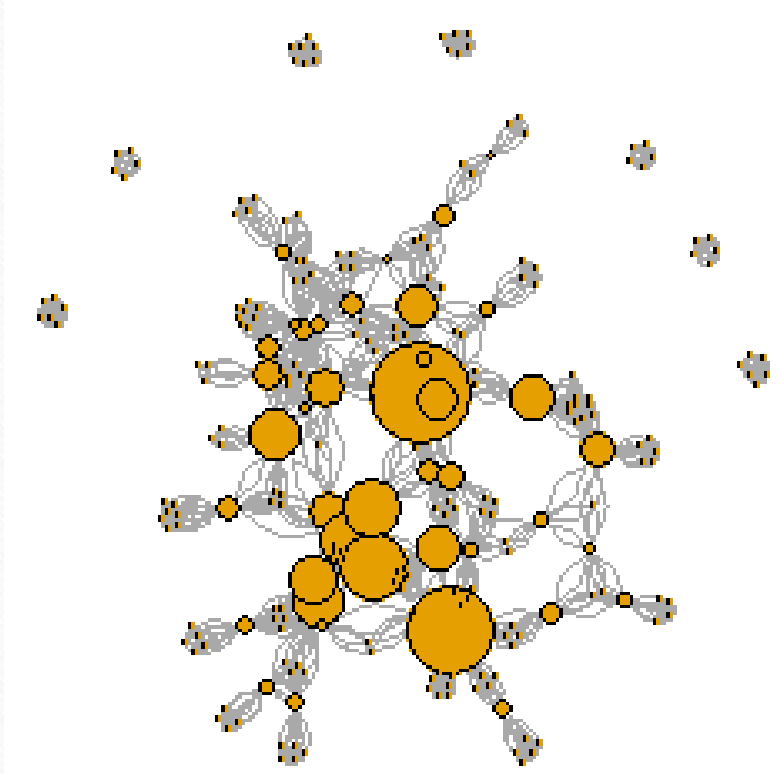
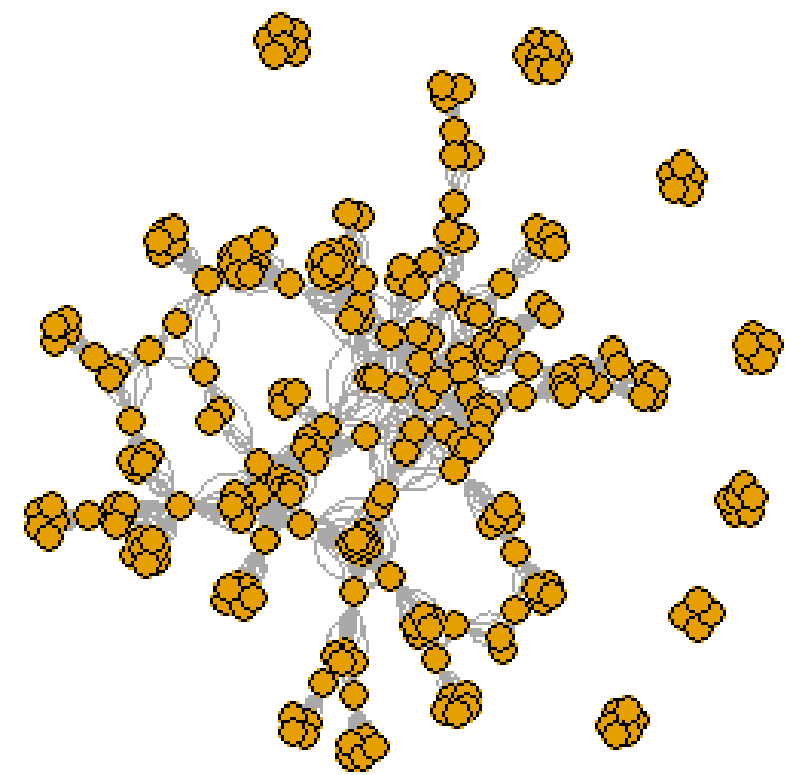


Figure 4. Multi-partite external jury members graph for 2017changing strength.



Market Basket Analysis

Figure 5. Heatmap of the the title of the field of the national council of universities.

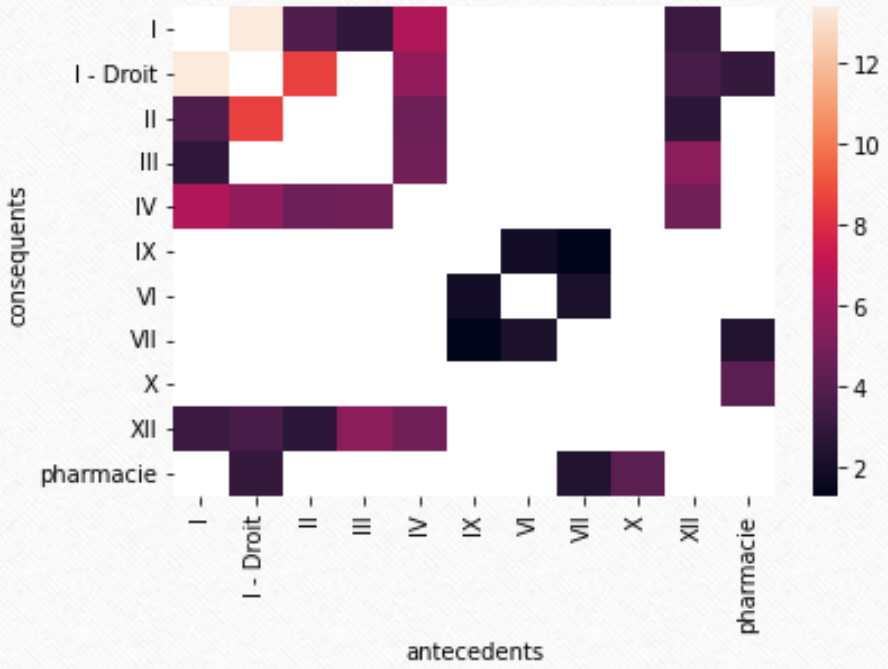
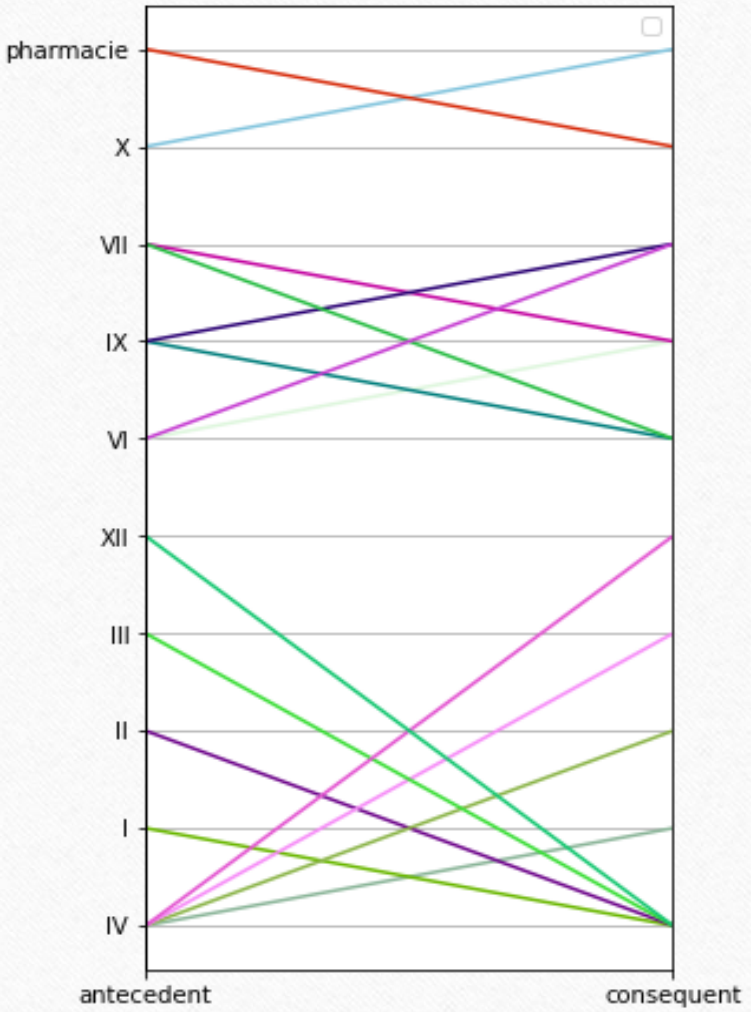


Figure 6. Parallel coordinates plot of the the title of the field of the national council of universities.



Discussion:

- Fig 1. shows a lot of connections between the nodes that is explained by the fact that some juries are the same in different universities.
- Fig 2. shows centrality in the lower side of the graph only and that can be explained by the fact that there is one big institution that is divided into small ones and therefore they are connected by the same jury.
- Fig 3. and Fig 4. show that some nodes are thicker than other ones and that can be explained by the fact that some juries are more active than others and they have more connections with other juries based on betweeenness and strength.

Discussion:

- Fig 5. Let's recall that a heatmap visualizes the intensity of the relationships between pairs of objects.
  If we pick a cell in the matrix, such as the I* as row and III* as column we can see that color is black indicating a low intensity , notice that the columns are antecedents, and the rows are consequent. Lighter colors indicate higher support.

- Fig 6. The parallel coordinates plot will allow us to visualize whether a rule exists between an antecedent and consequent.
  We can see that certain fields such as IV* are antecedents for many others. IV* and I* seem to be strongly associated. This is also true for VI* and VII* and many other cases.

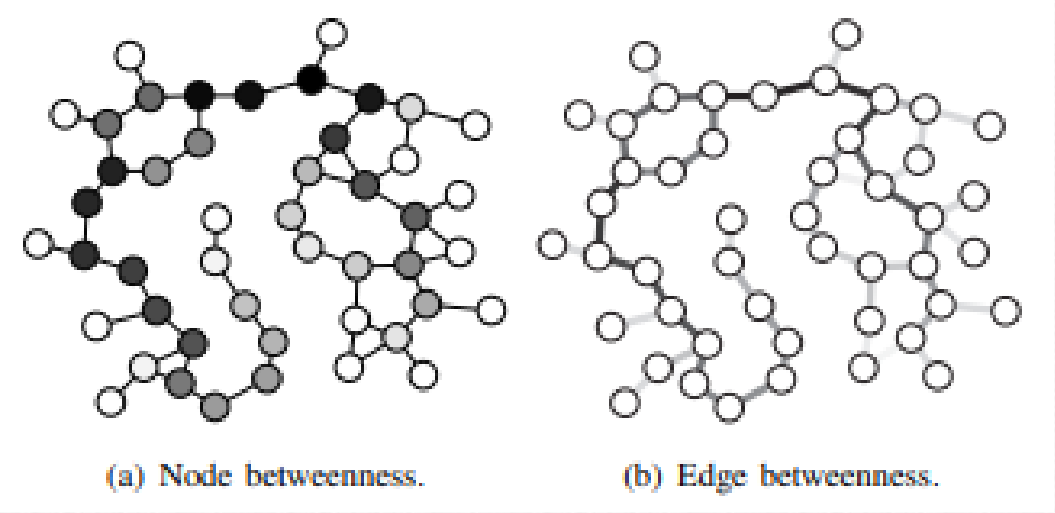
* Link reference for I,II etc meaning:
https://www.galaxie.enseignementsup-recherche.gouv.fr/ensup/pdf/qualification/sections.pdf

Concept of lift

$$\frac{Support(X\&Y)}{Support(X)Support(Y)}$$

- Lift metric evaluates the relationship between items.It’s a way to improve support.
- Numerator gives us the proportion of transactions that contain both X and Y
- Denominator tell us what that proportion would be if X and Y were randomly and independently assigned to transactions.
- A lift value of >1 tells us that 2 items occur in transactions together more often than we could expect based on their individual support values.
- This means the relationship is unlikely to be explained by random chance.

Difference between node and edge betweenness.



(a) Node betweenness. (b) Edge betweenness.

Node betweenness:

$$B_i \triangleq \sum_{\substack{k \in \mathcal{V} \\ k \neq i}} \sum_{\substack{\ell \in \mathcal{V} \\ \ell \neq i,k}} \frac{\sigma(k, \ell, i)}{\sigma(k, \ell)},$$

Bi is the number of

times node i lies on the shortest path between any pair of nodes k and ℓ. In the case where there are multiple shortest paths between nodes k and ℓ, the fraction of them that go through node i is counted. It follows that Bi attempts to measure how "strategically located" node i is within the graph G.

Figure (a) illustrates the notion of node betweenness, where the darker a node, the higher its betweenness..

Edge betweenness:

$$B_{\{i,j\}} \triangleq \sum_{k \in \mathcal{V}} \sum_{\substack{\ell \in \mathcal{V} \\ \ell \neq k}} \frac{\sigma(k, \ell, \{i, j\})}{\sigma(k, \ell)},$$

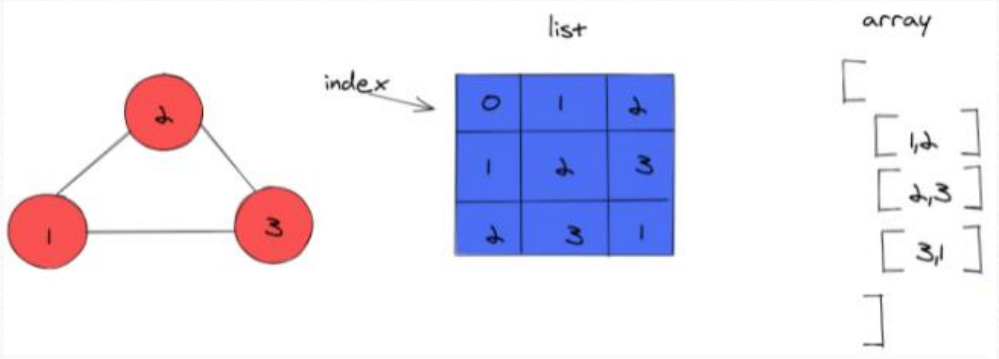
B{i,j} is the fraction of

times edge {i, j} lies on the shortest paths between any pair of nodes k and ℓ. It therefore characterizes how strategically located edge {i, j} is within the graph G,.

Figure (b) depicts the notion of edge betweenness, showing that it parallels that of node betweenness.

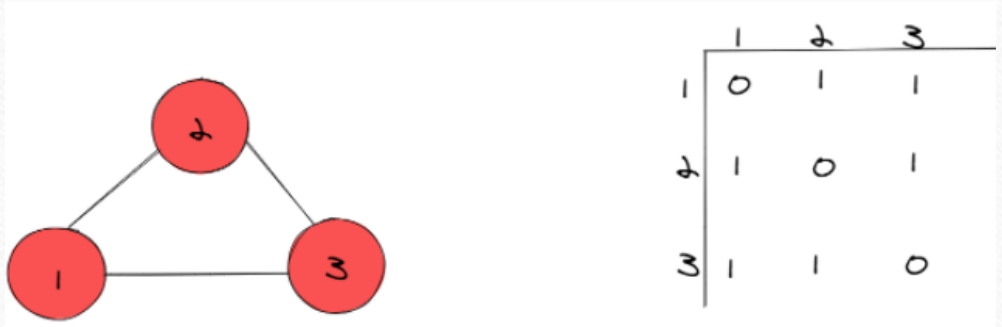
Difference between an edge list and an adjacency matrix.

Edge list



An edge list is a list or array of all the edges in a graph. Edge lists are one of the easier representations of a graph. In this implementation, the underlying data structure for keeping track of all the nodes and edges is a single list of pairs. Each pair represents a single edge and is comprised of the two unique IDs of the nodes involved. Each line/edge in the graph gets an entry in the edge list, and that single data structure then encodes all nodes and relationships.

Adjacency matrix



While an edge list won't end up being the most efficient choice, we can move beyond a list and implement a matrix. For many, a matrix is a significantly better kinesthetic representation for a graph. An adjacency matrix is a matrix that represents exactly which vertices/nodes in a graph have edges between them. It serves as a lookup table, where a value of 1 represents an edge that exists and a 0 represents an edge that does not exist. The indices of the matrix model the nodes.

Difference between a bipartite or a multipartite graph.

Bipartite

- A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V. Vertex sets U and V are usually called the parts of the graph.

Multipartite

- A k-partite graph is a graph whose vertices are or can be partitioned into k different independent sets. Equivalently, it is a graph that can be colored with k colors, so that no two endpoints of an edge have the same color.

Thank you