# Networks Analysis and Mining

Sorbonne Université
Master d'Informatique spécialité Réseaux

## Final Exam

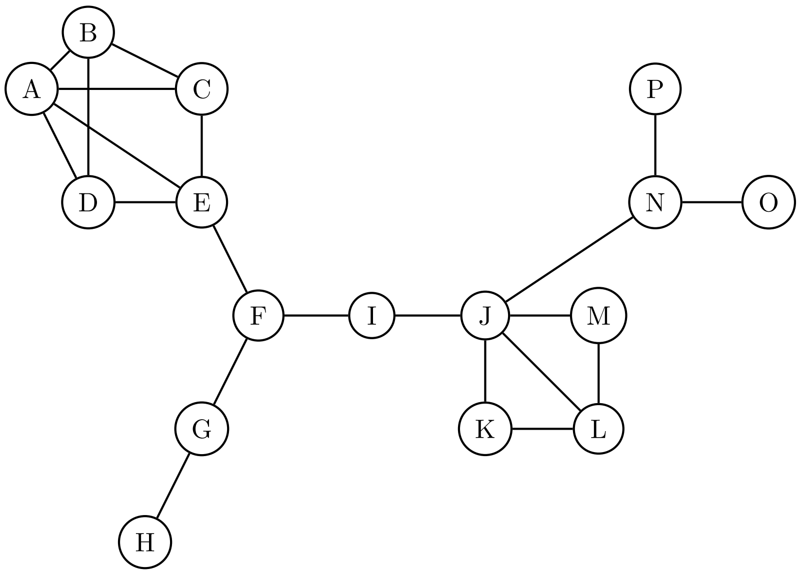
**February $16^{th}$ 2021**

Maximilien Danisch and Lionel Tabourier

The exam is 2hrs long. All documents are authorized. Points are only indicative. Thank you to answer on a separate paper. Good luck!

**Exercice 1 — *Measures on a small graph (6pts)***

Consider the following graph:




Q1. Plot the inverse cumulative degree distribution (ICDD) of the graph. *Reminder:* the ICDD shows on the Y-axis the number of nodes with a degree larger or equal to the values on the X-axis.

Q2. List all the triangles of the graph.

Q3. Compute the clustering coefficient of nodes $C$ and $J$.

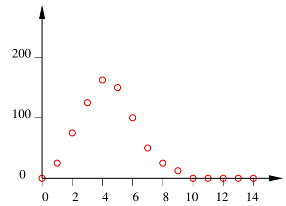
Q4. Compute the core value of all nodes in the graph.

Q5. Give the name of one node with high betweenness centrality but low degree and one node with low betweenness centrality but high degree. Justify in one sentence.

**Exercice 2 — *Graph models (4pts)***

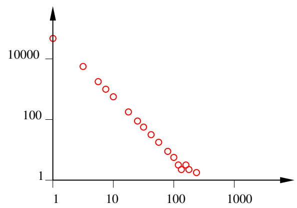
We give below measures realized on graphs. Tell in each case if the model proposed is consistent with the measures observed and if not, explain why.

Q1. Graph $G_1$:
density = 0.0001, average clustering coefficient: 0.12
degree distribution:



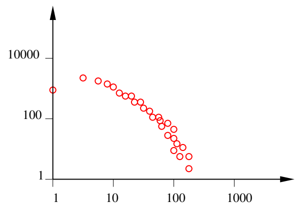

Is this graph generated with an Erdős-Rényi model?

Q2. Graph $G_2$:
density = 0.0001, average clustering coefficient: 0.004
fraction of nodes in the largest connected component: 0.38
degree distribution:



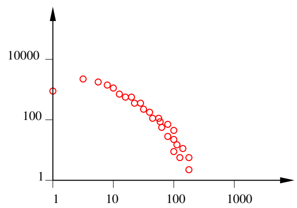

Is this graph generated with a Barabasi-Albert (preferential attachment) model?

Q3. Graph $G_3$:
density = 0.0001, average clustering coefficient: 0.001
degree distribution:




Is this graph generated with a configuration model?

Q4. Graph $G_4$:
density = 0.0001, average clustering coefficient: 0.21
degree distribution:




Is this graph generated with a Watts-Strogatz (*small world*) model?

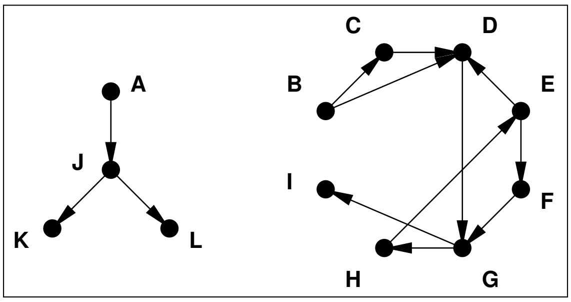
**Exercice 3** — ***About PageRank (4pts)***

Q1. In the graph below, which nodes are

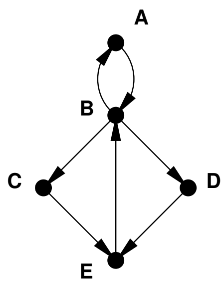
a) in the largest strongly connected component?

b) upstream of the largest strongly connected component?

c) downstream of the largest strongly connected component?




Q2. Compute the value of the PageRank of node $B$ in the (strongly connected) graph below, supposing there is no evaporation.




*Reminder:* without evaporation, the PageRank satisfies for all $x$:

$$PR(x) = \sum_{y \in Pred(x)} \frac{PR(y)}{d_{out}(y)}$$

where $Pred(x)$ are the predecessors of $x$ and $d_{out}(y)$ is the outdegree of $y$.

**Exercice 4** — ***Course understanding (4pts)***

In this exercise, you are asked to answer in one or two simple sentences.

Q1. In complex networks analysis, why do we nearly always favor using the table of adjacency lists format over the adjacency matrix format when representing graphs?

Q2. Why do we say that weak links have an essential role in a social network?

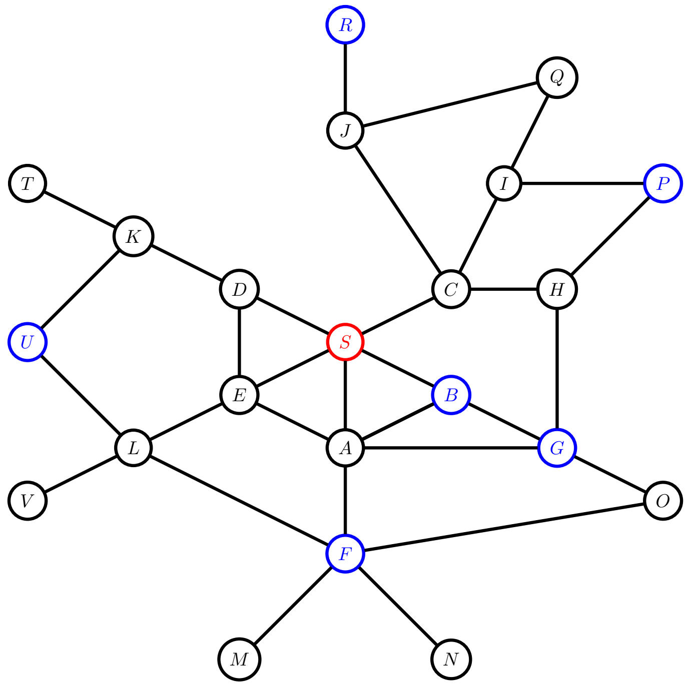
Q3. What is the initial state of every agglomerative community detection algorithm? Give an example of such an algorithm.

Q4. Give an advantage and a disadvantage of a recommendation system based on collaborative filtering in comparison with a content-based recommendation system.

**Exercice 5 — *Simulating traceroute (8pts)***

In this exercise, we simulate traceroute measurements from $S$ to several destinations using a restricted BFS algorithm, following the principle described in the *Internet topology metrology* course.

The structure of the graph $G$ under examination is given below:




Q1. a) Compute a BFS from source $S$ (to all nodes). You can present your answer as a list of nodes or a table of fathers, but it is recommended to draw the entire BFS tree.

b) Deduce the restricted BFS tree when restricting the BFS from source $S$ to destinations: $\{B, P, R, U, F, G\}$. You can present your answer as a table of fathers (with $-1$ for the nodes without father) or draw a restricted BFS tree.

Q2. a) What is the degree distribution of the original graph $G$?

b) The *sample graph* is the subgraph containing only the links and nodes observed in the restricted BFS. What is the degree distribution of the sample graph?

c) Can you briefly compare these two distributions?

Q3. a) For each link of the graph, indicate their distance to the source $S$ (for example $S - A$ is 1 step away from $S$, $A - F$ is 2 steps away from $S$, etc).

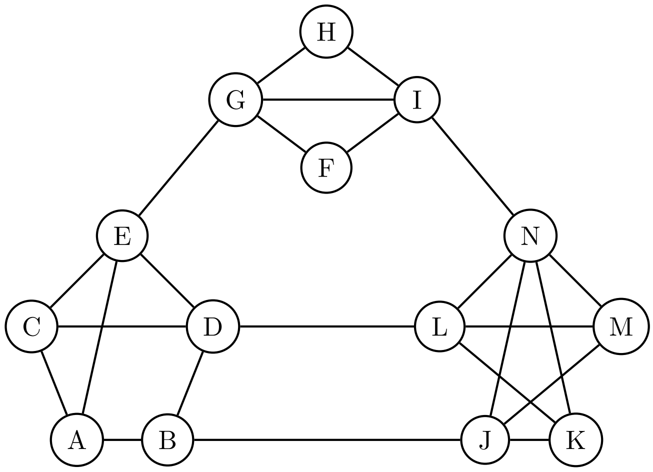
b) Plot the curve of link visibility depending on the distance of a link from the source, according to the same principle as the plot seen at the end of the course *Internet topology metrology*, that is:

- $X$-axis: number of steps from the source
- $Y$-axis: fraction of edges observed (at $X$ steps from the source) in the sample graph

Can you briefly comment this plot?

**Exercice 6 — *Community detection using conductance (8pts)***

In this exercise, we consider the following graph $G = (V, E)$




**Notion of conductance**

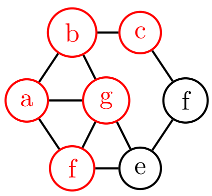
If $S$ is a set of nodes, we remind that $\overline{S}$ is the complementary set (meaning $\overline{S} = V \setminus S$).
The *conductance* $\phi(S)$ *of a set* $S$ is defined in the following way:

$$\phi(S) = \frac{\sum_{i \in S, j \in \overline{S}} \; a_{ij}}{min(a(S), a(\overline{S}))}$$

$$\text{where} \begin{cases} a_{ij} = & 1 \text{ if } (i,j) \text{ is an edge, otherwise } 0 \\ a(S) = & \sum_{i \in S} d_i \;\; \text{where } d_i \text{ is the degree of node } i \\ a(\overline{S}) = & \sum_{i \in \overline{S}} d_i \end{cases}$$

So the conductance is a ratio between the number of edges going from the nodes in $S$ to the nodes not in $S$ and the minimum between the sum of the degrees of the nodes in $S$ and the sum of the degrees of the nodes not in $S$.

For example in the following graph, we compute $\phi(\{a, b, c, f, g\})$:




- $\sum_{i \in S, j \in \overline{S}} \; a_{ij} = 3$ , as the only edges to go from $S$ to $\overline{S}$ are $(c, f)$, $(g, e)$ and $(f, e)$
- $a(S) = 15$ as there are 6 edges with both ends in $S$ and 3 edges with one end in $S$
- $a(\overline{S}) = 5$ as there is one edge in $\overline{S}$ and 3 edges with one end in $\overline{S}$
- so $\phi(\{a, b, c, f, g\}) = \frac{3}{5}$

**Q1.** Compute $\phi(\{H\})$ the conductance of the set containing only $H$ in the graph above.
**Q2.** Compute the conductance $\phi(\{G, H, I, F\})$ of $S = \{G, H, I, F\}$ in the graph above.

Now, we introduce the conductance of a partition $\Pi = \{S_1, S_2, \dots, S_k\}$ as being the maximum of the conductances of each set in the partition:

$$\phi(\Pi) = max_i\{S_i\}$$

$\Pi$ is a partition so by definition $S_1 \cup S_2 \cup \dots S_k = V$ and $S_i \cap S_j = \emptyset$ if $i \neq j$.

**Q3.** Compute the conductance $\phi(\Pi_{iso})$, which is the conductance of the partition $\Pi_{iso}$ where all nodes are isolated:

$$\Pi_{iso} = \{\{A\}, \{B\}, \ldots, \{M\}, \{N\}\}$$

**Q4.** Compute the conductance $\phi(\Pi_{all})$, which is the conductance of the partition $\Pi_{all}$ where all nodes are in a same set:

$$\Pi_{all} = \{\{A, B, \ldots, M, N\}\}$$

**Q5.** Compute the conductance $\phi(\Pi_{good})$, which is the conductance of the partition $\Pi_{good}$ defined in this way:

$$\Pi_{good} = \{\{A, B, C, D, E\}, \{F, G, H, I\}, \{J, K, L, M, N\}\}$$

**Q6.** Compute the conductance $\phi(\Pi_{bad})$, which is the conductance of the partition $\Pi_{bad}$ defined in this way:

$$\Pi_{bad} = \{\{A, C, D, E\}, \{F, G, H, I\}, \{B, J, K, L, M, N\}\}$$

**Q7.** The previous questions suggest to create a community detection algorithm based on the minimization of the conductance. What do you think is more appropriate: a) minimizing the partition conductance or b) minimizing the average conductance of the sets in a partition? Explain in one or two sentences.