Computing PageRank scores in graphs Foundations of Data Science - M1 MOSIG

The goal of this lab is to study the PageRank algorithm which aims to assess the importance of a node in a graph (see IR book, Chapter 21).

Data We will use the following representation for graphs.



For each line (but the first one) the first number corresponds to a node and the following numbers to the outgoing links of that node. You'll find examples of graphs under http://ama.liglab.fr/~gaussier/Courses/FoundationsDataScience/Lab-PageRank.

Transition probability matrix Write, in Python, a code which computes the transition probability matrix fo a graph represented as above. The transition matrix P is defined by:

$$P_{ij} = \begin{cases} \lambda \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}} + (1-\lambda)\frac{1}{N} & \text{if } \sum_{j=1}^{N} A_{ij} \neq 0\\ \frac{1}{N} & \text{otherwise,} \end{cases}$$

where $A_{ij} = 1$ if there exists a link from node *i* to node *j*, and 0 otherwise.

Questions:

- 1. First set λ to 0.85 and write a code (in Python) which computes the PageRank with the power method as given in Algorithm ??.
- 2. Chose different values for the stopping criterion (e.g., $\epsilon = \{10^{-3}, 10^{-4}, 10^{-5}\}$). What do you observe?
- 3. Add a few *hubs* (nodes with many outgoing links) and a few *authorities* (nodes with many incoming links). Which nodes have the highest PageRank?
- 4. Try and increase the PageRank of some pages. Explain your approach and validate it experimentally.
- 5. Try different values for λ . What is the behaviour of the algorithm when λ tends to 0?

If you have time Write a code (in Python) which generates a random graph with a given number of nodes N (input parameter of your code). Test your PageRank code on newly generated graphs.