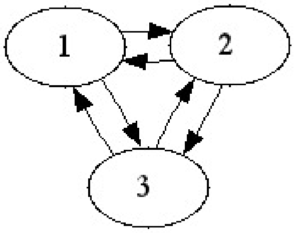
Information Retrieval and Recommeder Systems (IRRS)
Master in Data Science (MDS)

Session 5: Web Search

Exercise List, Fall 2022

Basic comprehension questions.
Check that you can answer them before proceeding. Not for credit.

1. Compute by eye the PageRank of all nodes of the following graph:

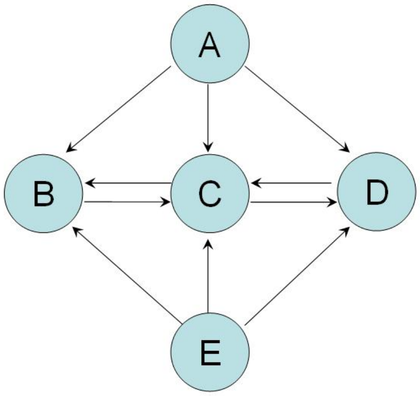


2. True or false: the pagerank of a web page depends on the query.
3. True or false: the hub and authority values of a web page depend on the query.
4. True or false: the PageRank algorithm does not take into account the content of a web page.
5. True or false: the HITS algorithm does not take into account the content of a web page.

Exercises for credit. Solving three of these exercises (not solved by the instructors in class) suffice for full credit for this assignment.

Exercise 1

Consider the following miniature web:



1. Tell the PageRank values of A and E as a function of the damping factor.

2. Justify that B and D have the same PageRank, no matter the damping factor.

3. Fix the damping factor to 0.9. Give the the Google matrix and the associated PageRank equations. Then compute the PageRank of each node.

Exercise 2

Consider a small web with three pages, A, B, and C, where A has links to B and C, B has a link to C, and C a link to B.

1. Give the initial equations for this system (no damping), the associated transition matrix, and the resulting node PageRank values.

2. Now give the Google matrix using damping factor 0.85, the associated system of equations for PageRank, and the resulting node PageRank values.

3. Give HITS' equations for hub and authority values. Solve the equations, either using the iterative method or using a numerical computation package.

Exercise 3

Give an example of a strongly connected graph with three nodes such that 1) each node has exactly two edges arriving into it 2) not all three nodes have the same page rank.

Set up the page equations for the graph you give, solve the system, and check by direct substitution that the solution you give satisfies the equations.

Exercise 4

PageRank and HITS were inspired by studies of how scientists obtain reputation from the citations to their papers from other papers. The more citations you get, and the higher the reputation of the people who cite you, the more reputation you get yourself. Also, knowing who cites who is useful for determining research communities, research trends within one area, etc.

Consider six scientists *A*lan, *K*im, *M*aria, *P*eter, *R*on, and *T*rinity, who have cited each other as follows in their bibliographies:

$$\begin{array}{rcl} A & : & K, P, R, T \\ K & : & M, P \\ M & : & K, P \\ P & : & K, M \\ R & : & A, T \\ T & : & A, R \end{array}$$

For example, Peter has cited papers by Kim and by Maria in his own papers.

1. Compute the *citation matrix* $A = (a_{ij})$ such that

$$a_{i,j} = \begin{cases} 1 & \text{if author } i \text{ cites author } j \\ 0 & \text{otherwise} \end{cases}$$

2. Compute the *co-citation* matrix for these authors. We say that two authors are co-cited if a third author cites both of them. This matrix tells which pairs of authors i and j are co-cited, and by how many third authors.

3. Compute the *bibliographic coupling* matrix. Bibliographic coupling occurs when two authors reference a common third author in their bibliographies. This matrix tells which pairs of authors i and j are bibliographically coupled, and how many times.

4. More in general, define formally the co-citation and bibliographic coupling. Then show that they can be expressed (computed) as simple functions of the citation matrix.

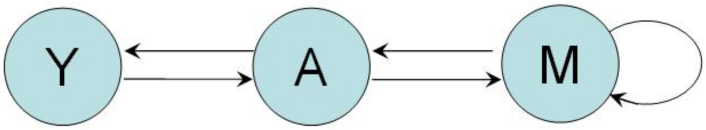
5. Consider the definition: "A number of authors constitute a related group if each member of the group has at least one coupling to every other member of the group". Give the maximal related groups (that cannot be enlarged) in the previous bibliography.

Exercise 5

Let G be the Google matrix of a web. We know that the vector of page ranks of the nodes is the one that satisfies $G^T P = P$, where ^{T} denotes the transpose of a matrix. Argue that if we do the computation without transposing G, that is, if we compute the vector S such that $GS = S$, there is always a trivial solution, independent from the web graph. Of course, we only consider vectors whose components' sum is 1. Suggestion: start working out a simple example, without damping factor, to see what goes on.

Exercise 6

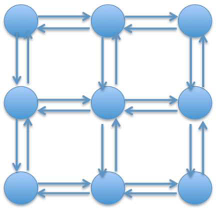
Consider the following miniature web:



Compute the PageRank equations with no damping, and the PageRank of each node. Repeat with damping factor 0.85.

Exercise 7

Consider a graph with 9 nodes aligned in a grid, where each node has links to its 4 closest neighbors, with the exception of nodes at the extremes and the corners, which each have links to their 3 and 2 closest neighbors, respectively. Compute the Pagerank of each node using a damping factor of $\lambda = 1$ first, and then generalize this to damping factors $0 < \lambda < 1$.



.