Web Search Algorithms

CAIM: Cerca i Anàlisi d'Informació Massiva

Exercise list, Fall 2025

Basic Comprehension Questions. Make sure you can answer them before proceeding.

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

$$E = \{(1,2), (2,3), (3,1), (1,3), (3,2), (2,1)\}$$

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.

Exercise 1

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

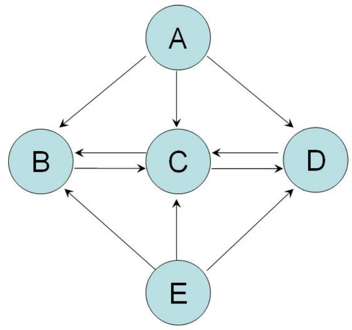
1. Give the initial PageRank equations for this system (no damping, $\lambda = 1$), the associated transition matrix (M^T), and the resulting node PageRank values.

2. Now give the **Google Matrix** (G^T) using a damping factor $\lambda = 0.85$, the associated system of equations for PageRank, and the resulting node PageRank values.

3. Give the **HITS** equations for **Hub** (**h**) and **Authority** (**a**) values. Solve the equations, possibly using a numerical computation package.

Exercise 2

Consider the following miniature web:



1. Provide the **PageRank** values of A and E as a function of the damping factor λ.
2. Justify that B and D have the same **PageRank**, regardless of the damping factor λ.
3. Fix the damping factor to $\lambda = 0.9$.
   - Give the **Google Matrix** (G^T) and the associated PageRank system of equations.
   - Compute the PageRank of each node.

Exercise 3

Give an example of a **strongly connected graph** with three nodes such that 1) each node has exactly two incoming edges (in-degree = 2) and 2) not all three nodes have the same PageRank (assume $\lambda = 1$).

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

Exercise 4

Let G be the **Google Matrix** of a web. We know the PageRank vector $\mathbf{p}$ satisfies $G^T\mathbf{p} = \mathbf{p}$. Argue that if we compute the vector $\mathbf{s}$ such that $G\mathbf{s} = \vec{s}$ (without transposing G), there is always a trivial solution, $\mathbf{s} = [1/n, \ldots, 1/n]^T$, independent of the web graph structure, where n is the number of nodes.

Exercise 5

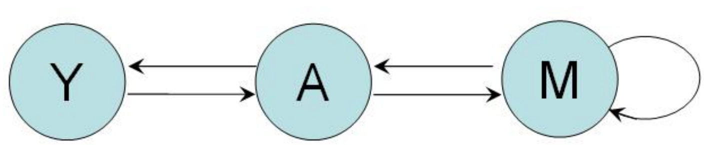
Consider six scientists (A, K, M, P, R, T), where Peter cited Kim and Maria. Their citation links are:

Author	Cites:
A	K, P, R, T
K	M, P
M	K, P
P	K, M
R	A, T
T	A, R

1. Compute the **citation matrix** $C = (c_{ij})$, where $c_{i,j} = 1$ if author i cites author j.
2. Compute the **co-citation matrix** D. (Two authors are co-cited if a third author cites both of them.)
3. Compute the **bibliographic coupling matrix** B. (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. Formally define co-citation and bibliographic coupling, and show that they can be expressed as simple matrix functions of C.
5. Based on 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 bibliography.

Exercise 6

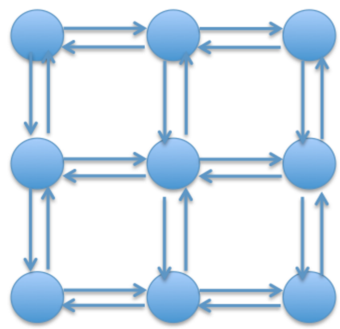
Consider the following miniature web:



1. Compute the PageRank equations with no damping ($\lambda = 1$) and the PageRank of each node.
2. Repeat the computation with a damping factor $\lambda = 0.85$.

Exercise 7

Consider a graph with 9 nodes aligned in a 3×3 grid. Each internal node links to its 4 nearest neighbors, edge nodes link to 3, and corner nodes link to 2.



1. Compute the **PageRank** of each node using a damping factor of $\lambda = 1$.
2. Generalize this result for a damping factor $0 < \lambda < 1$.

Exercise 8

Consider a simple linear graph (a "daga") with three nodes A, B, and C and two edges $A \to B$ and $B \to C$.

1. Write the transition matrix M^T for $\lambda = 1$. Using the power method (iterative multiplication), show what happens to the PageRank vector $p^{(k)}$ as $k \to \infty$.

2. Explain how **teleportation** term in the Google Matrix G^T does not fully resolve the sink problem in a practical implementation.

3. Write the full expression for G^T assuming $\lambda = 0.85$ and a modification where the dangling node C is treated as if it links to all nodes equally.

Exercise 9

Consider a **bipartite cycle graph** with four nodes A, B, C, D where the links are: $A \to B$, $B \to C$, $C \to D$, and $D \to A$.

1. Write the adjacency matrix A for this graph.

2. Using the HITS iterative update formulas ($\mathbf{a} = A^T\mathbf{h}$ and $\mathbf{h} = A\mathbf{a}$), demonstrate the problem of **score oscillation** or **instability** that can occur in HITS for graphs with high symmetry or specific cyclic structures (e.g., bipartiteness).

3. Suggest a modification to the HITS algorithm to mitigate these stability issues.