Information Retrieval and Data Mining SS 2017
Prof. Dr.-Ing. Sebastian Michel
MSc. Koninika Pal
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
http://dbis.informatik.uni-kl.de
Exercise Sheet 4: Hand out 29.05.2017, Due 13.06.2017 - 16:00

In this sheet, please feel free to use R or any other language/system of your choice to perform "heavy" computations such as the power iteration method, but include the script / code in your submission.

Assignment 1: Markov Chains (1 P.)

Consider the Markov chain described by the following transition probability matrix

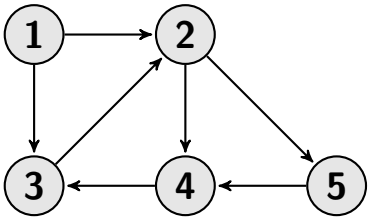
$$P = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 1 \\ 1 & 0.0 & 0.0 \end{bmatrix}$$

(a) Draw the Markov chain described by P and determine its stationary state probabilities by solving a system of linear equations.

(b) Modify the Markov chain such that it is no longer *aperiodic* and explain why this is no longer the case. Do stationary state probabilities exist for your modified Markov chain? If so, please specify them. Otherwise, explain why they do not exist.

Assignment 2: Page Rank (1 P.)

(a) Consider the following directed graph $G(V, E)$



- Determine the transition probability matrix P of the Matrix chain induced by PageRank when setting $\varepsilon = 0.2$. Compute the stationary state distribution π using the power iteration method. Begin with $\pi_i = 1/5$ as initial state probabilities and perform five iterations.
- When setting $\varepsilon = 0$ would the Markov chain induced by PageRank still be ergodic? Please justify your answer.

(b) Discuss how the special case of a complete graph (i.e., every node is connected with every other node) affects the computation of PageRank scores. Discuss further what the effect of setting $\varepsilon = 1.0$ is.

(c) Consider a link spam farm which serves to promote a useless Web page (e.g., about an online university where you obtain superb grades and titles without any exams and work) in its PageRank score.

Let us denote the page to be promoted as p. The spammer has managed to obtain a total of m links from m good Web pages that now point to p. Let us denote these pages as $g_1, \ldots, g_m$ and assume that each of them has the same PageRank γ and the same outdegree o. Furthermore, the spammer uses a spam farm of k boosting pages $b_1, \ldots, b_k$ that are artificially created in k different

Internet domains. Each of these pages b_i is designed to point to p, and p is modified to have links that point back to all of these k boosting pages; these are the only outgoing links of p. In addition, the spammer makes each page b_i to point to all other $k-1$ boosting pages.

Derive a formula for the PageRank score that p obtains this way.

Assignment 3: HITS (1 P.)

Consider again the graph from Assignment 2.

(a) Compute the hub vector $h^{(5)}$ and the authority vector $a^{(5)}$ obtained after five iterations of the algorithm from the lecture.

(b) Compare the hub vector and authority vector to the PageRank stationary state distribution determined in Assignment 2(a). Please describe your observations.

(c) Consider a graph where all edges are bi-directional. What is the characteristic of the hub vector and authority vector of such graph?