Information Retrieval and Data Mining (IRDM) SS 2015
Prof. Dr.-Ing. Sebastian Michel
MSc. Koninika Pal
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
Sheet 4: Handout 03.06.2015, Presentation 16.06.2015

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN
http://dbis.informatik.uni-kl.de

Hints. *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. Note that still you have to be able to demonstrate your solution in the exercise session.*

Assignment 1: Markov Chains (1 P.)

Consider the Markov chain described by the following transition probability matrix

$$P = \begin{bmatrix} 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \end{bmatrix}$$

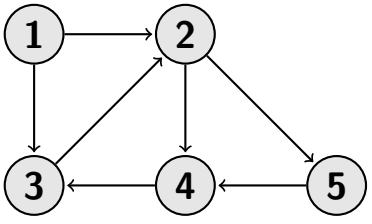
(a) Draw the Markov chain described by P.

(b) Determine its stationary state probabilities by solving a system of linear equations.

(c) 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.)

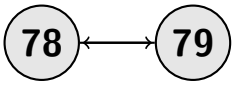
(a) Consider the following directed graph $G(V, E)$



- Determine the transition probability matrix P of the Matrix chain induced by PageRank when setting $\epsilon = 0.2$. 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 $\epsilon = 0$ 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.
- What is the effect of setting $\epsilon = 1.0$.

(c) Assume that a graph consisting of $|V| = 100$ vertices (all of which with non-zero outdegree) has the following isolated component consisting of two vertices:



What is the PageRank score of the two vertices assuming $\epsilon = 0.2$?

Assignment 3: HITS (1 P.)

Consider again the graph from Problem 2.

(a) Determine the matrices A, A^TA, and AA^T

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

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

(d) Interpret the hub vector $h^{(1)}$ and authority vector $a^{(1)}$ obtained after the first iteration. What do the components correspond to?