UNIVERSITÄT DES SAARLANDES
MAX-PLANCK-INSTITUT INFORMATIK

Dr.-Ing. Ralf Schenkel
Dr. rer. nat. Marc Spaniol

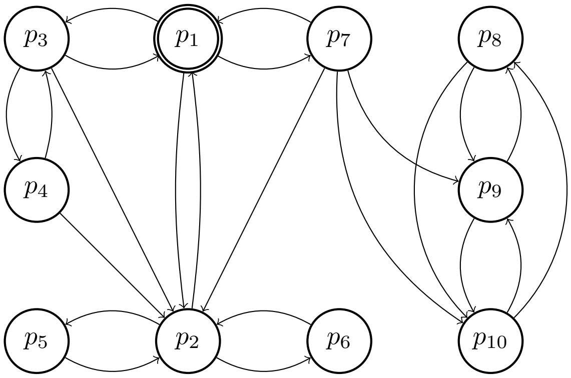
# Web Dynamics (SS 10)
# Assignment 4

Handout on: May 27, 2010

## Due on: June 7, 2010
## Hand in to: Dr. Marc Spaniol, room 425, building E1.4 (MPII)
## or AG5 Secretary, room 402, building E1.4 (MPII)

## Exercise 4.1: Archiving Strategies

Consider the graph of a Web site, the change probabilities ($\lambda_i$) per page ($p_i$) as well as their first time-point of change ($\mu_i$) in $[t_1; t_{10}]$ given below:



$$\lambda_1 = \frac{8}{10} \quad \lambda_2 = \frac{4}{10} \quad \lambda_3 = \frac{2}{10} \quad \lambda_4 = \frac{1}{10} \quad \lambda_5 = \frac{6}{10}$$
$$\lambda_6 = \frac{99}{100} \quad \lambda_7 = \frac{3}{10} \quad \lambda_8 = \frac{9}{10} \quad \lambda_9 = \frac{7}{10} \quad \lambda_{10} = \frac{5}{10}$$

$$\mu_1 = t_4 \quad \mu_2 = t_9 \quad \mu_3 = t_7 \quad \mu_4 \in \emptyset \quad \mu_5 = t_6$$
$$\mu_6 = t_2 \quad \mu_7 \in \emptyset \quad \mu_8 = t_3 \quad \mu_9 = t_5 \quad \mu_{10} = t_8$$

a) An archiving crawler is allowed to download one page per time unit $t \in [t_1; t_{10}]$. Write down the schedules for crawls starting at $t_1$ when applying the topological archiving strategies BFS (breadth-first-search) and DFS (depth-first-search) given the fixed seed $p_1$ as well as the non-topological measurable coherence strategy introduced in the lecture (slides 39 ff., configured with an "ignorant" risk threshold of $\eta = 1$).

b) What is the resulting measurable coherence ($C(c)$) relative to the start of crawl at $t_1$ for each of the three strategies ($c_{BFS}$, $c_{DFS}$, and $c_{\eta=1}$) resulting from the schedules of part a)?

## Exercise 4.2: Incoherence

Incoherence occurs if at least one change of a page $p_i$ has happened between start of crawl at $t_s = t_1$ and its time of download at $t(p_i)$.

a) Compute for a measurable coherence crawl the probability of incoherence of a page $p_i$ (with change probability $\lambda_i$) to be downloaded at $t_k$ relative to start of crawl at $t_s = t_1$.

b) Assume a measurable coherence crawl where the probability of incoherence of page $p_x$ at the fifth download position ($t_5$) is $\kappa(p_x) = 0,35$. What is its corresponding change probability $\lambda_x$?

c) Consider the fifth download position ($t_5$) of a measurable coherence crawl (configured with a "moderate" risk threshold of $\eta \sim 0,5$) where one of the three pages $p_a$, $p_b$, $p_c$ can be chosen for download next. Assume the probability of incoherence for the three pages currently is $\kappa(p_a) = 0,35$, $\kappa(p_b) = 0,45$, and $\kappa(p_c) = 0,9$. What would be the best choice? Explain!

## Exercise 4.3: Shingling

A shingling is a set of unique "shingles"-contiguous subsequences of $n$ tokens ($n$-grams) in a document - that can be used to gauge the similarity of two documents. For a given shingle size, the degree to which two documents A and B resemble each other can be expressed as the ratio of the magnitudes of their shinglings' intersection and union, or:

$$r(A, B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|},$$

where $S(A)$ and $S(B)$ are the sets of $n$-grams for the documents $A$ and $B$. This definition is identical with the Jaccard coefficient describing similarity and diversity of sample sets.

Let $A$ be the sentence "This product is not bad, it is actually quite good." and $B$ be "This product is not good, it is actually quite bad.". Compute the similarity between $A$ and $B$ using unigrams (1-grams) and using 4-grams on word level, i.e. a unigram consists of one word, a digram (2-gram) consists of two neighbouring words and so on. Which approach gives better result? When is it appropriate to use $n$-grams with smaller size and when to use $n$-grams with bigger size?