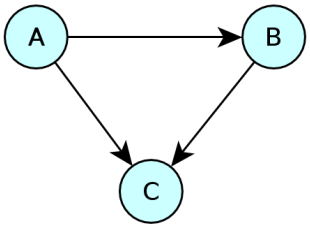
Bielefeld University
Faculty of Technology

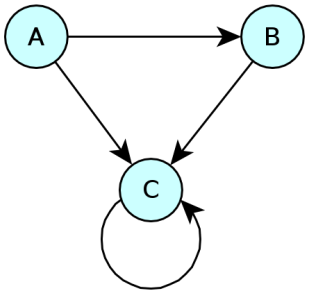
Big Data Analytics

Summer 2023

Number 04, Submission Deadline: May 29 2023, 11:59 PM

1. **Page Rank:**





(a) Create transition matrices M_A and M_B for the graphs shown above. (2 P)

(b) Perform 10 iterations (per graph) of iterative page rank as defined in the lecture. (2 P)

(c) What can you observe in the results? (2 P)

(d) Change your previous solution to:

- eliminate dead ends recursively (remember to also delete the edges connecting to them) (3 P)
- include the concept of taxation in order to counter-act spider traps. (3 P)

How do the results for the graphs above differ with one or both mechanisms enabled? (2 P)

2. **Loading a real graph:**
   It is now time to apply the techniques to a larger graph from real data which you can find in the provided folder *material*[1]

   (a) Create a transition matrix given the links in the datasets (ignore the value column for this). (2 P)

   (b) Calculate the PageRank for all nodes in the graph using the code from Task 1. (2 P)

3. **Topic-sensitive PageRank:**
   Topic-sensitive (or topic-specific) PageRank is often used to compute personalized PageRank.

   (a) Implement a solution that calculates the topic-sensitive PageRank for a given node in the graph and apply it to the graph you created in Task 2. (4 P)

   (b) Output the topic-sensitive PageRank ($TSPR$) for each of the following nodes: (2 P)
   - $TSPR('css')['angularjs']$ (meaning the TSPR value for 'angularjs' in the topic 'css')
   - $TSPR('angularjs')['css']$
   - $TSPR('jquery')['bootstrap']$
   - $TSPR('bash')['linux']$

   (c) For each of the topics above (css, angularjs, jquery, bash): output the top 5 nodes and their respective TS-PR value. (2 P)

Important:
Please submit your group solution via LernraumPlus. You are free to hand solutions in as PDFs or Jupyter Notebooks.

[1] Or download the files: `https://www.kaggle.com/stackoverflow/stack-overflow-tag-network`