Michael Kerber, Robert Tichy **WS 2024/25**

Numerical Linear Algebra for Data Science

Projects Last update: December 5, 2024

Project 3
This project will investigate the page rank in some more detail. We will first construct two random models to simulate the internet.

a) For a fixed n (number of website) and some $p \in (0, 1)$, put a link from website i to website j with probablility p independently (for $i \neq j$). How do you have to choose p so that the expected number of links per page is 3? We call this the *uniform model*.

b) Alternatively, consider an incremental process where website $1, \ldots, k-1$ are already constructed (with links), and website k is added. Again, we chose a probability p_k (which can depend on k now) and put a link to website j with $j < k$ with probability p_k independently. Again, how do we need to choose p_k such that the expected number of links is 3? We call this the *incremental model*

c) Implement the page rank algorithm (with $p = 0.2$, initial vector $(1/n, \ldots, 1/n)$, and 50 iterations). How do the page ranks of both models differ? Give an explanation for that behavior.

Your implementation of page rank should be able to give results with around 1000-5000 websites within seconds. We now try to manipulate the page rank. Fix $n = 1000$ from now on.

d) Find the website with the lowest page rank in your model. Imagine that you own this website and you would like to boost its page rank. You can add 10 additional pages to the internet, with an arbitrary number of links. Devise and implement a strategy to improve the page rank of the formerly lowest ranked page. How far can you boost it? Try your strategy with both the uniform and the incremental model and discuss the differences.

e) You are willing to invest more money (=create more website) to make sure that your website is surely top-ranked. Try to increase the number of additional pages. How many do you need to guarantee that in the uniform model, no matter how the concrete instance looks like?

f) (Bonus) There are many other random models thinkable to simulate the internet. Try to think of a different model, and apply the manipulation also in that model.