Exercise 9: PageRank - solution

1 Problem 1

The goal is to find how many pages of the k pages we need to add and how we should interlink them to maximize the PageRank of rolexx.com. Let’s assume that we have some pages already interlinked by the spammer and see what happens if the random surfer is walking around. The random surfer can be modelled as having three states:

- R - visiting the rolexx.com website
- K - visiting one of the \( l < k \) websites added by the spammer
- O - visiting the other (legitimate) websites

Since the crawler has no memory, we can model this as a three-state Markov chain (Figure 1).

Maximizing the PageRank of rolexx.com is equivalent to maximizing the probability that rolexx.com is visited i.e. the probability of being in state R \( (p_R) \). We have \( p_R = p_Kp_{KR} + p_Op_{OR} + p_Rp_{RR} \), which transforms to \( p_R = \frac{p_Kp_{KR} + p_Op_{OR}}{1-p_{RR}} \) and this is the expression that we want to maximize. \( 1 - p_{RR} \) is a constant so we ignore it and only need to maximize \( p_R = p_Kp_{KR} + p_Op_{OR} \)

Case a.) To maximize \( p_{KR} \) we need to link from all the \( l \) pages to rolexx.com and put no links among the \( l \) pages \( (p_{KK} \) minimized) and from the \( l \) pages to other pages \( (p_{KO} \) minimized). We also set \( l = k \) to maximize \( p_K \). If we don’t put any links from rolexx.com to anywhere then it becomes a sink and the random

![Figure 1: The Markov state transition diagram for the surfer.](image-url)
walker has no link to take next and will always select an arbitrary page and jump to it. To prevent that we link from rolexx.com to one of the \( k \) pages. The final interlinking that maximizes the PageRank of rolexx.com in the a.) case is on Figure 2.

**Case b.**) We keep the link structure from a.) but now we can additionally control the \( p_{OR} \). To maximize this probability we link from all the legitimate websites to rolexx.com. We can also increase the chances of visiting rolexx.com by linking via one of the \( k \) websites, then rolexx.com can be reached in two hops by the surfer. We create a link from all legitimate websites to the \( k \) websites. The final interlinking is on Figure 3.

The spamming in case a.) is feasible to do in the context of Google, the spammer can easily create an arbitrary number of new websites and link between them. In case b.) it is a bit hard to convince the whole World Wide Web to link to the \( k + 1 \) pages. Please have a look at the Wikipedia entry for "link farms" for a discussion.
2 Problem 2

Let’s assume that the weight of the link from the page $p_i$ to $p_j$ is $w_{ij}$. The weight equals zero if there is no link. The only modification that needs to be made is to set the values of the transition matrix $R_{ij} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}$, which is simply the weights normalized by their sum to have transition probabilities proportional to the weights.