Machine Learning for Graphs and Sequential Data Exercise Sheet 5

Graphs: Ranking

1 PageRank

Problem 1: Consider a directed graph $G = (V, E)$ with $V = \{1, 2, 3, 4, 5\}$, and $E = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 4), (3, 5), (4, 5), (5, 4)\}$.

a) Set up the equations to compute PageRank for G, where the teleport probability is 0.2.

b) Set up the equations for topic-sensitive PageRank for the same graph, with teleport set $\{1, 2\}$. Solve the equations and compute the ranking vector.

c) Give examples of pairs (S, v), where $S \subseteq V$ and $v \in V$, such that the topic-sensitive PageRank of v for the teleport set S is equal to 0. Explain why these values are equal to 0.

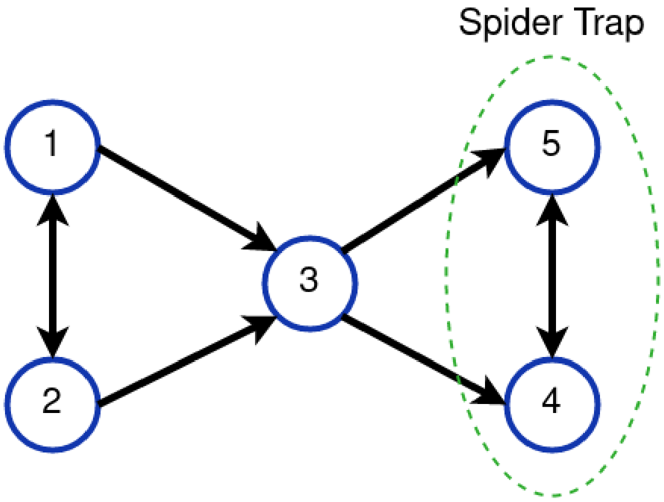


Figure 1: Graph from Problem 1

a) We want a teleport probability of 0.2, so $\beta = 0.8$. There are $N = |V| = 5$ nodes and their out degrees are $d = \begin{pmatrix}2 & 2 & 2 & 1 & 1\end{pmatrix}$. From here, setting up the equations is straightforward.

$$r_1 = \beta\frac{r_2}{2} + \frac{1-\beta}{5} \qquad r_2 = \beta\frac{r_1}{2} + \frac{1-\beta}{5} \qquad r_3 = \beta\left(\frac{r_1}{2} + \frac{r_2}{2}\right) + \frac{1-\beta}{5}$$

$$r_4 = \beta\left(\frac{r_3}{2} + r_5\right) + \frac{1-\beta}{5} \qquad r_5 = \beta\left(\frac{r_3}{2} + r_4\right) + \frac{1-\beta}{5}$$

b) In topic-sensitive PageRank we are only teleporting to specific nodes, thereby boosting their PageRank compared to nodes that are not in the teleport set. This means that only the second term of each equation changes and we can adapt them directly from part a.

$$r_1 = \beta\frac{r_2}{2} + \frac{1-\beta}{2} = \frac{2}{5}r_2 + \frac{1}{10} \qquad r_2 = \beta\frac{r_1}{2} + \frac{1-\beta}{2} = \frac{2}{5}r_1 + \frac{1}{10}$$

$$r_3 = \beta\left(\frac{r_1}{2} + \frac{r_2}{2}\right) = \frac{2}{5}(r_1 + r_2)$$

$$r_4 = \beta\left(\frac{r_3}{2} + r_5\right) = \frac{2}{5}r_3 + \frac{4}{5}r_5 \qquad r_5 = \beta\left(\frac{r_3}{2} + r_4\right) = \frac{2}{5}r_3 + \frac{4}{5}r_4$$

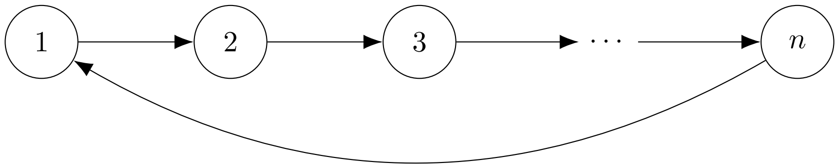
Now we solve them for $\boldsymbol{r}$.

$$\begin{aligned}
r_1 = \frac{2}{5}r_2 + \frac{1}{10} = \frac{2}{5}\left(\frac{2}{5}r_1 + \frac{1}{10}\right) + \frac{1}{10} = \frac{4}{25}r_1 + \frac{7}{50} \Rightarrow r_1 = \frac{25}{21}\frac{7}{50} = \frac{1}{6} \\
r_2 = \frac{2}{5}r_1 + \frac{1}{10} = \frac{2}{30} + \frac{1}{10} = \frac{1}{6} \\
r_3 = \frac{2}{5}(r_1 + r_2) = \frac{2}{5}\frac{1}{3} = \frac{2}{15} \\
r_4 = \frac{2}{5}r_3 + \frac{4}{5}r_5 = \frac{2}{5}r_3 + \frac{4}{5}\left(\frac{2}{5}r_3 + \frac{4}{5}r_4\right) = \frac{18}{25}\frac{2}{15} + \frac{16}{25}r_4 \Rightarrow r_4 = \frac{25}{9}\frac{18}{25}\frac{2}{15} = \frac{4}{15} \\
r_5 = \frac{2}{5}r_3 + \frac{4}{5}r_4 = \frac{4}{75} + \frac{16}{75} = \frac{4}{15}
\end{aligned}$$

c) The question can alternatively be asked as: Are there subgraphs that are disconnected or have no out-going edges? Think of the random surfer analogy. Without random teleports a random surfer would always end up in the component that they started in or one of the subgraphs from which there is no escape. So if we identify them, we can choose S as a any subset of their nodes and v from the remaining nodes that cannot be reached.

Look at the graph as drawn in Figure 1. We see that nodes 4 and 5 form a spider-trap, a closed subgraph with no outward edges, but also nodes 3, 4 and 5 together. So examples would be and $S \subseteq \{3, 4, 5\}$ and and $v \in \{1, 2, 3\} \setminus S$.

Problem 2: The PageRank algorithm is applied to the cycle graph of n nodes shown below. The teleport probability is $1 - \beta$ and the teleport set S consists of all the nodes, that is $S = \{1, \ldots, n\}$.



What is the final PageRank score of each node i as a function of n and β as computed by the algorithm?

Let $p(i)$ be the predecessor of node i in the cycle. Then the PageRank equation of node i is

$$r_i = \beta \frac{r_{p(i)}}{1} + \frac{1-\beta}{n} = \beta r_{p(i)} + \frac{1-\beta}{n}.$$

Now we can iteratively plug in the equations for the predecessors.

$$\begin{aligned} r_i &= \beta \left(\beta r_{p(p(i))} + \frac{1-\beta}{n} \right) + \frac{1-\beta}{n} = \beta^2 r_{p(p(i))} + \beta \frac{1-\beta}{n} + \frac{1-\beta}{n} \\ &= \beta^2 \left(\beta r_{p(p(p(i)))} + \frac{1-\beta}{n} \right) + \beta \frac{1-\beta}{n} + \frac{1-\beta}{n} \\ &= \beta^3 r_{p(p(p(i)))} + \beta^2 \frac{1-\beta}{n} + \beta \frac{1-\beta}{n} + \frac{1-\beta}{n} \\ &= \dots \end{aligned}$$

Let's repeat this process $k \in \mathbb{N}$ times where p^k is iterative function application.

$$= \beta^k r_{p^k(i)} + \frac{1-\beta}{n} \sum_{i=0}^{k-1} \beta^i$$

We recognize the geometric series and simplify with the partial sum formula.

$$= \beta^k r_{p^k(i)} + \frac{1-\beta}{n} \frac{1-\beta^k}{1-\beta} = \beta^k r_{p^k(i)} + \frac{1-\beta^k}{n}$$

Node i is its own n-th predecessor, so for $k = n$ we get

$$r_i = \beta^n r_i + \frac{1-\beta^n}{n}.$$

Solving for r_i and simplifying leads us to

$$(1-\beta^n) r_i = \frac{1-\beta^n}{n} \;\Leftrightarrow\; r_i = \frac{1}{n}$$

which makes intuitive sense because all nodes are exactly the same.

2 Spam farms

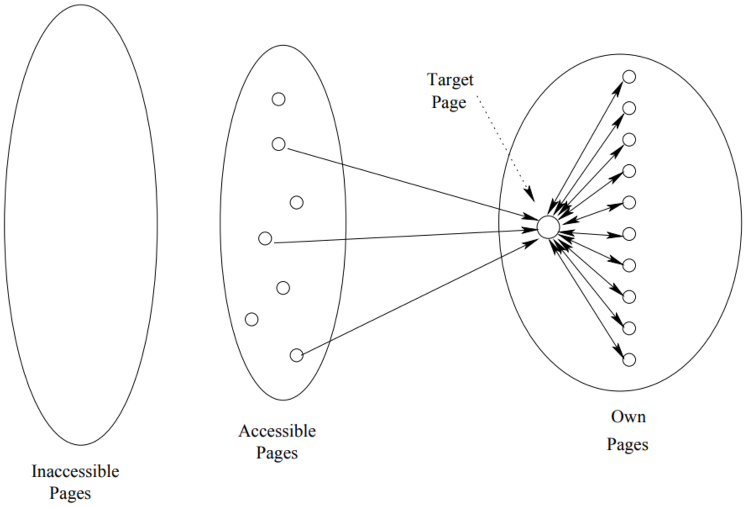
A collection of pages whose purpose is to increase the PageRank of a certain page is called a spam farm.

Problem 3: From the spammer's point of view the Web is divided into three parts:

1. Inaccessible pages: the pages that the spammer cannot affect. Most of the Web is in this part.
2. Accessible pages: those pages that, while they are not controlled by the spammer, can be affected by the spammer. For example the spammer can leave comments on blog posts that point to any desired page creating a link between them.

3. Own pages: the pages that the spammer owns and controls.

The spam farm consists of the spammer's own pages, organized in a special way as seen below, and some links from the accessible pages point to the spammer's pages.



Let there be n pages on the Web in total, and let some of them be a spam farm with a target page t and k supporting pages. Let x be the amount of PageRank contributed by the accessible pages. That is, $x = \sum_{p \in \{p \in V \mid p \in S_{\text{acc}},\, (p,t) \in E\}} \beta r_p / d_p$ where S_{acc} is the set of accessible pages, r_p is the PageRank score of a page p, d_p is the degree of a page p and $1 - \beta$ is the teleport probability. Determine the PageRank of the target page as a function of x, n, k and β ignoring interdependencies between the variables. Can you identify the multiplier effect of link farms? How does the size of the link farm influence the PageRank or the target page?

The PageRank equation for any node o in the set of own pages S_{own} is given by

$$r_o = \beta \frac{r_t}{k} + \frac{1-\beta}{n}.$$

We can set up the equation for the target node

$$r_t = x + \beta \sum_{p \in S_{own}} \frac{r_p}{d_p} + \frac{1-\beta}{n}$$

and plug in the rank of the owned pages.

$$= x + \beta \sum_{p \in S_{own}} \left(\beta \frac{r_t}{k} + \frac{1-\beta}{n} \right) + \frac{1-\beta}{n}$$

They are identical for all owned pages, so we can simplify

$$= x + \beta k \left(\beta \frac{r_t}{k} + \frac{1-\beta}{n} \right) + \frac{1-\beta}{n}$$

and simplify some more to isolate r_t.

$$= x + \beta^2 r_t + \frac{k}{n}\beta(1-\beta) + \frac{1-\beta}{n}$$

Solving for r_t gets us

$$r_t = \frac{1}{1-\beta^2}\left(x + \frac{k}{n}\beta(1-\beta) + \frac{1-\beta}{n}\right)$$

Remember $a^2 - b^2 = (a-b)(a+b)$.

$$\begin{aligned} &= \frac{x}{1-\beta^2} + \frac{k}{n}\frac{\beta(1-\beta)}{(1-\beta)(1+\beta)} + \frac{1-\beta}{n(1-\beta)(1+\beta)} \\ &= \frac{x}{1-\beta^2} + \frac{k}{n}\frac{\beta}{1+\beta} + \frac{1}{n(1+\beta)} \end{aligned}$$

First, we notice that the contributions x that the spammer placed on legitimate sites are multiplied by $\frac{1}{1-\beta^2}$. This is called the multiplier effect. For the common teleport probability of 0.15, or $\beta = 0.85$, we get a multiplier of $(1 - 0.85^2)^{-1} \approx 3.6$. This boost happens even with a single page in the link farm and stems from the 2-cycle that the farm page forms with the target page. The incoming rank contribution is caught in that cycle and reflected back onto the target node thereby exceeding its original magnitude.

Second, the PageRank of the target node is linear in the size k of the link farm if we ignore interactions between k and x or n. In practice these are negligible because $k \ll n$.