Machine Learning for Graphs and Sequential Data Exercise Sheet 5

Graphs: Ranking

1 PageRank

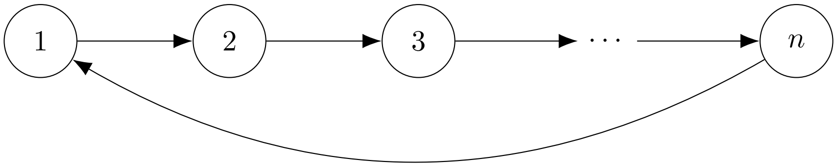
Problem 1: Consider a directed graph $G = (V, E)$ with $V = \{1, 2, 3, 4, 5\}$, and $E = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 4), (3, 5), (4, 5), (5, 4)\}$.

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.

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 $S = \{1, \dots, n\}$.



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

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.

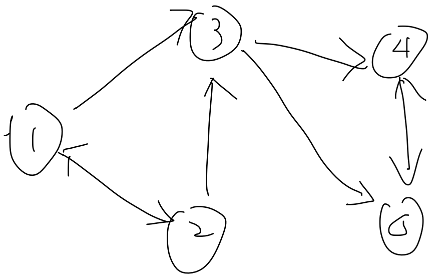
Let there be n pages on the Web in total, and let some of them be a spam farm with a target page t and k supporting pages. Let x be the amount of PageRank contributed by the accessible pages. That is,

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.



$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1-\beta)\frac{1}{N}$$

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

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

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

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

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

$r = \beta M r + (1-\beta)\pi$ where $\pi_i = \begin{cases} \frac{1}{|S|} & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases}$

$\downarrow$

$$r_1 = \beta \frac{r_2}{2} + (1-\beta) \cdot \frac{1}{2} \qquad (1-\beta) = 0.2$$

$$r_2 = \beta \frac{r_1}{2} + (1-\beta) \cdot \frac{1}{2}$$

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

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

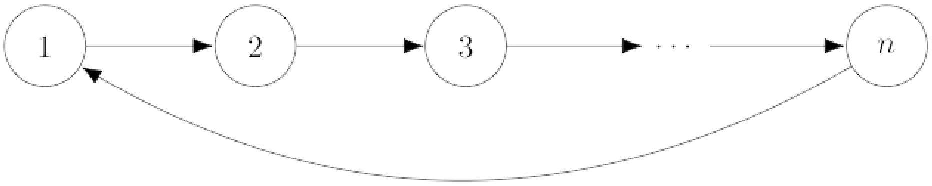
$$r_5 = \beta\left(\frac{r_3}{2} + r_4\right)$$

3) teleporte in disconned Graph

$V = 1\ 2\ 3 \setminus S$

$S = 3\ 4\ 5$

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?

$$r_1 = \beta \cdot r_n + (1-\beta) \cdot \frac{1}{n}$$

$$r_2 = \beta r_1 + (1-\beta)\frac{1}{n}$$

$$- \quad -$$

$$r_n = \beta r_{n-1} + (1-\beta)\frac{1}{n}$$

$$\downarrow$$

$$r_2 = \beta^2 \cdot r_n + (1-\beta)\beta \cdot \frac{1}{n} + (1-\beta)\cdot\frac{1}{n}$$

$$r_n = \beta^n r_n + (1-\beta)\frac{1}{n}\left(\beta^{n-1} + \beta^{n-2} \cdots + \beta^0\right)$$

$$r_n = \frac{(1-\beta)\frac{1}{n}\sum_{i=0}^{n-1}\beta^i}{1-\beta^n}$$

$$\sum_{i=0}^{n-1}\beta^i = \frac{1-\beta^n}{1-\beta}$$

$$\frac{\frac{1}{n}\cancel{(1-\beta^n)}}{\cancel{1-\beta^n}} = \frac{1}{n}$$

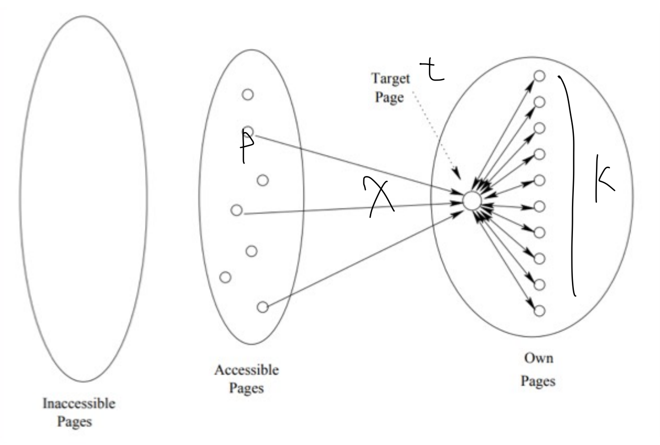
A collection of pages whose purpose is to increase the PageRank of a certain page is called a spam farm.

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,



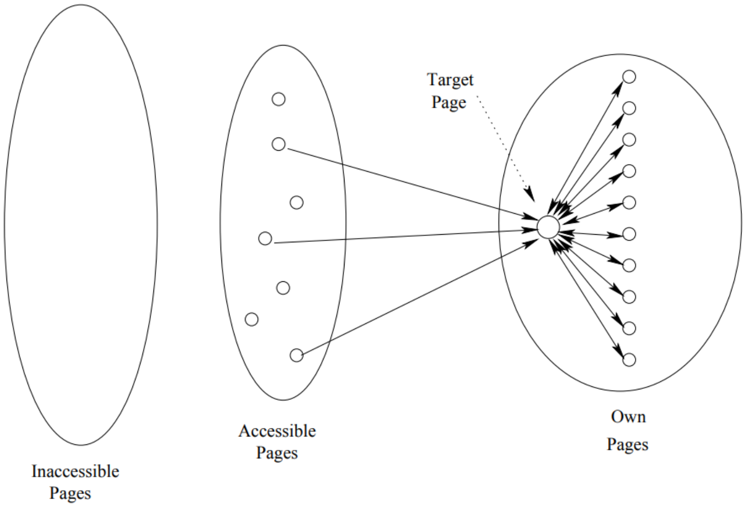
$$r_k = \beta \frac{r_t}{k} + (1-\beta)\cdot\frac{1}{n}$$

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

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

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

$x = \sum_{p\in\{p\in V\,|\,p\in S_{\text{acc}},\,(p,t)\in E\}} \beta r_p/d_p$ 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?



$x = \sum_{p \in \{p \in V \,|\, p \in S_{\text{acc}},\, (p,t) \in E\}} \beta r_p / d_p$ 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?