

# The PageRank algorithm

Author: Leo Cunningham

Department of Mathematical Sciences

# Table of Contents

① Motivation

② Derivation

# Web search engines and the idea of importance

## Pre-1998 web search

- e.g. Yahoo! Search, AskJeeves
- Yahoo! Search catalogued the web using humans rather than crawlers (bots which automatically index web pages), resulting in search results which had been classified as meaningful
- Capacity to provide consistently good results weakened by spammers and the ever-increasing enormity of the web

## Solution

In a search, rank pages based on their relative 'importance' using the web's hyperlink structure

# The premise behind PageRank

Brin and Page (1998), developed the idea of using a page's 'importance' as a main factor in returning results to queries

**Importance:**

- The hyperlink structure of the web enables pages to 'vote' for others by linking to them
- A page is important (has a higher ranking) if it is linked to by other important web pages

# The original summation formula

*A page is important if it is linked to by other important web pages.*

## Summation formula

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|F_{P_j}|}$$

- $r(P_i)$  is the ranking of page  $P_i$
- $F_{P_j}$  is the sets of all pages linked to by page  $P_j$
- $B_{P_i}$  is the sets of all pages linking to page  $P_i$

Each page  $P_i$  has a total 'vote' worth  $r(P_i)$ , which is split evenly between the pages it links to

- How do we find  $r(P_i)$  if we don't know the ranks of the other pages?
  - Use an iterative process, beginning with a uniform rank distribution summing to 1.

# PageRank

## The Hyperlink matrix and the iterative process

A system of web pages  $P_1, \dots, P_n$  is represented as a *directed graph*  $G = (V, A)$ , where  $V = \{P_1, \dots, P_n\}$  is the set of pages, and  $A = \{(P_i, P_j)\}$  is the set of arcs, where  $(P_i, P_j) \in A$  if page  $P_i$  links to page  $P_j$ . For a graph  $G = (V, A)$ , we define the *hyperlink matrix* as

$$\mathbf{H}_{ij} = \begin{cases} \frac{1}{|F_{P_i}|} & \text{if } (P_i, P_j) \in A \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

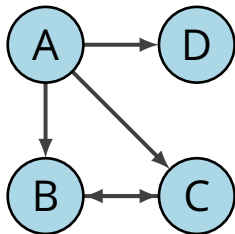
where  $F_{P_i} \neq \emptyset$  is the set of pages linked to from page  $P_i$ . If  $P_i$  has no outward links, it has a zero row in  $\mathbf{H}$ .

**Iterative process:**

$$\mathbf{v}^{(k+1)T} = \mathbf{v}^{(k)T} \mathbf{H} \quad (2)$$

# PageRank

## The Hyperlink matrix - an example



$$\mathbf{H} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- Sub-stochastic matrix -  $D$  is a dangling node
- What happens when we update the PageRank vector?
  - $A$  has no inlinks, so its rank immediately becomes zero
  - rank does not flow out of  $D$ , hence it is lost and the sums of the ranks drops to below 1, and the rank of  $D$  also becomes zero

# PageRank

## Fixing the Hyperlink matrix

Sub-stochasticity:

### Solution

Replace  $\mathbf{0}^T$  rows with  $1/n\mathbf{e}^T$ . This new matrix is denoted  $\mathbf{S}$ .

Since  $\mathbf{S}$  is stochastic, we can employ Markov chain theory.

Goal for  $\mathbf{v}$ :

- Convergence to a unique, positive PageRank vector
- Convergence regardless of starting distribution

### Solution

Transition matrix needs to be irreducible and aperiodic. This is satisfied by

$$\mathbf{G} = \alpha \mathbf{S} + (1 - \alpha) 1/n\mathbf{e}\mathbf{e}^T, \alpha \in (0, 1) \quad (3)$$



# PageRank

## The final formula and interpretations

Using the matrix

$$\mathbf{G} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{1}/n \mathbf{e}^T, \quad (4)$$

the power method

$$\mathbf{v}^{(k+1)T} = \mathbf{v}^{(k)T} \mathbf{G} \quad (5)$$

converges to a unique, positive PageRank vector regardless of initial distribution.

# PageRank

Interpretation - the random surfer:

## Markov chains:

- $\mathbf{v}$  = probability distribution of a random surfer being on a web page
- Power method, with transition matrix  $\mathbf{G}$
- At each iteration:
  - with probability  $\alpha$ : travel to a page linked to by the current page, each with equal likelihood, or
  - with probability  $1 - \alpha$ : teleport to a completely random web page, each with equal likelihood
- $\mathbf{v}$  converges to a stationary distribution representing the probabilities that the surfer will be on any one page at a particular time step.

# PageRank

## Changes to the algorithm

### The Hyperlink Matrix

- The algorithm assigns an equal probability to each node available to travel to
- The distribution of probabilities for the available nodes can be altered to reflect the strength of each connection. This change is called **weighted PageRank**.

### The Teleportation matrix

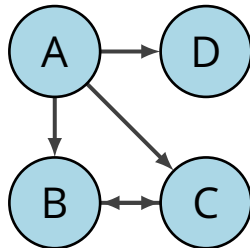
- The current teleportation matrix  $1/n\mathbf{ee}^T$  assigns an equal probability to each page on the web.
- *Personalised PageRank* uses the teleportation matrix  $\mathbf{ev}^T$ , where  $\mathbf{v}^T$  is the *personalisation vector*.  $\mathbf{v}^T$  is non-uniform and can be constructed to reflect the interests of the user, effectively resulting in a PageRank vector which prioritises pages the user has deemed more relevant.

## Higher order relations and social/trust networks

Unweighted PageRank uses only immediate *edge-based relations*, with an equal strength assigned to each arc. Here, an arc indicates trust.

*Unweighted PageRank* classifies the influence of *B*, *C*, and *D* on *A* as equal, as edges are unweighted. But considering higher order relations,

- *B* and *C* have mutual influence on each other
- they both also influence *A*



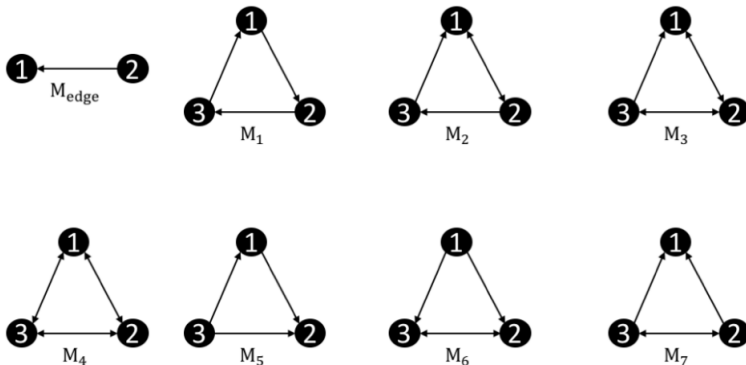
### Conclusion:

The participation in this local structure means *A* is influenced more by *B* and *C*, so their edges should be more highly weighted.

# Motifs

## Triangular motifs

The structure participated in by  $A, B, C$  is an example of a triangular motifs. Below is the set of all possible triangular motifs (adjusted for symmetries):



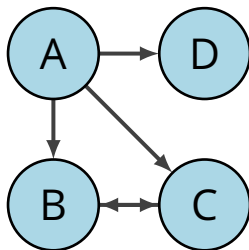
# Motif-based PageRank

- $\mathbf{W}$  is the *adjacency matrix* (an unnormalised (binary) hyperlink matrix)
- $\mathbf{W}_{M_i}$  is the *motif-based adjacency matrix*, where  $(\mathbf{W}_{M_i})_{ij}$  = the number of times node  $i$  and node  $j$  appear in motif  $M_i$  together (hence symmetric)

For the 4-user example,

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{W}_{M_6} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$



# Motif-based PageRank

*Motif-based PageRank* is based on the premise that "the weights of the links between directly connected users should be adjusted based on the local structures they participate in". This adjustment is handled by combinations of the original (edge-based) and motif-based adjacency matrices:

- **Linear Combination:**  $\mathbf{H}_{M_k} = \alpha \mathbf{W} + (1 - \alpha) \mathbf{W}_{M_k}$ ,
- **Non-Linear Combination:**  $\mathbf{H}_{M_k} = \mathbf{W}^\alpha + \mathbf{W}_{M_k}^{(1-\alpha)}$ ,

which are converted to stochastic matrices by normalisation and fixing dangling nodes, and used in place of  $\mathbf{S}$  in

$$\mathbf{G} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{1}/n \mathbf{ee}^T, \quad (6)$$





# Conclusion

## Experimental comparison

In social and trust networks, the linear motif-based PageRank has been shown to be a more effective user relevance scoring method than indegree score, betweenness score, closeness score, unweighted PageRank and weighted PageRank. Datasets used come from DBLP, an academic dataset containing publication and author records, and Epinions and Ciao, two review websites where users can rate a review's helpfulness.



# References

-  Austin R. Benson, David F. Gleich, and Jure Leskovec, *Higher-order organization of complex networks*, Science **353** (2016), no. 6295, 163–166.
-  Sergey Brin and Lawrence Page, *The anatomy of a large-scale hypertextual web search engine*, Computer networks and ISDN systems **30** (1998), no. 1-7, 107–117.
-  Amy N. Langville and Carl D. Meyer, *Google's pagerank and beyond*, Princeton University Press, Princeton, 2006.
-  H. Zhao, X. Xu, Y. Song, D. Lee, Z. Chen, and H. Gao, *Ranking users in social networks with motif-based pagerank*, IEEE Transactions on Knowledge and Data Engineering **33** (2021), no. 05, 2179–2192.

Questions?