Motif-based PageRank

Leo Cunningham | Supervisor: Andrew Wade

Durham University

Web search engines and the notion of 'importance'

Web search engines pre-1998 operated using a mixture of Boolean, probabilistic model and vector space model retrieval methods. Each came with their flaws, and the rapidly growing size of the web made meaningful results increasingly difficult to achieve.

In 1998, Brin and Page, the founders of Google, developed the idea of using a page's 'importance' as a main factor in returning results to queries. The intrinsic hyperlink structure enables the web to be represented as a directed graph, with arcs indicating outwards hyperlinks. The PageRank algorithm assigns each web page an importance ranking based on one premise: a page is important (has a higher ranking) if it is linked to by other important web pages.

The abstraction of the web into graphical form means this centrality metric can be used for any population with a notion of 'inlinking' and 'outlinking', with applications from social networks to transportation network analysis.

The PageRank Algorithm

A system of web pages $P_1, ..., P_N$ is represented as a *directed graph* $G = (V, A)$, where $V = \{P_1, ..., P_n\}$ 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. The premise of importance as given above is algebraicized by the equation:

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|F_{P_j}|} \tag{1}$$

where $F_{P_j} \neq \emptyset$ is the set of pages linked to from page P_j, and B_{P_i} is the sets of all pages linking to page P_i. To allow simultaneous computation of ranks, we use matrices:

Definition (Hyperlink matrix)**.** For a graph $G = (V, A)$, we define the *Hyperlink matrix* as

$$\mathbf{H}_{ij} = \begin{cases} \frac{1}{|F_{P_i}|} & \text{if } (P_i, P_j) \in A \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

If a page has no outlinks, its row in $\mathbf{H}$ is $\mathbf{0}^T$.

We would like to find a PageRank vector using an iterative process $\mathbf{v}^{(k+1)T} = \mathbf{v}^{(k)T}\mathbf{H}$, which is the power method applied to a Markov chain with transition matrix $\mathbf{H}$. However, the rows of $\mathbf{H}$ are *sub-stochastic*; they either add to 1 for the pages which have outlinks, or are all 0 for the pages without any (called *dangling nodes*). We fix this by replacing the $\mathbf{0}^T$ rows with $1/n\mathbf{e}^T$, and we call this new stochastic matrix $\mathbf{S}$. One particular interpretation for the PageRank process here is that of a 'random surfer' travelling the web, who has an equal probability of going to each of the pages linked to by the current page. The final PageRank vector (a stationary distribution) gives the probability of being on any one page on the web.

We can now use Markov chain theory to handle the convergence of the power method to a meaningful PageRank vector. To obtain a unique, positive PageRank vector, regardless of the starting vector, $\mathbf{S}$ needs to be irreducible and aperiodic, which are both satisfied if $\mathbf{S}$ is *primitive*. This is satisfied by the convex combination of $\mathbf{S}$ and a normalised matrix of 1s:

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

where $\alpha \in [0, 1]$. Our 'random surfer' will now teleport to a completely random web page, ignoring the hyperlink network, with probability α at each step.

Now the iterative formula

$$\mathbf{v}^{(k+1)T} = \mathbf{v}^{(k)T}\mathbf{G}, \tag{4}$$

will converge to the same PageRank vector $\mathbf{v}$, regardless of the starting vector.

Changes to the original PageRank algorithm

There are several choices regarding the structure of the original PageRank iterative process which seem natural to reconsider:

The Hyperlink Matrix

- The original scheme has a uniform weighting for rows in $\mathbf{H}$, working with the idea that the surfer is equally likely to visit each page linked to from the current page.
- 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 teleportation matrix $1/n\mathbf{e}\mathbf{e}^T$ assigns an equal probability to each page on the web.
- *Personalised PageRank* uses the teleportation matrix $\mathbf{e}\mathbf{v}^T$, 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

Unweighted PageRank (each outlink from a node has equal weight) uses only *edge-based relations* to calculate an importance score. In social and trust networks, where information can propagate through local structures [3], it makes sense to study the connection between nodes beyond an immediate edge. Consider the following example (where an arc indicates trust):

The original PageRank algorithm 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, meaning this influence is strengthened.

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

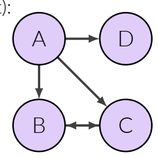


Figure 1. A 4-person social network

Motifs

The triangular structure in figure 1 is an example of a *network motif*. We use the definitions from [1] as inspiration for the following:

Definition (Network motif)**.** A network motif M is a structure, consisting of k nodes, defined by a tuple $(\mathbf{B}, \mathcal{A})$, where $\mathbf{B}$ is a $k \times k$ binary adjacency matrix (non-normalised hyperlink matrix), and $\mathcal{A} \subseteq \{1, 2, ..., k\}$ is the set of anchor nodes.

- The adjacency matrix $\mathbf{B}$ indicates the structure of the motif, and $\mathcal{A}$ determines the nodes we are interested in. If we are interested in all k nodes, M is called a simple motif. Otherwise, it is called an anchored motif. For example, if we just wanted to observe whether two nodes occur in the A and B position in the network motif above, our anchor nodes would be $\mathcal{A} = \{A, B\}$.
- In figure 1, the adjacency matrix is $\mathbf{B} = \left(\begin{smallmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right)$ for the nodes (A, B, C).

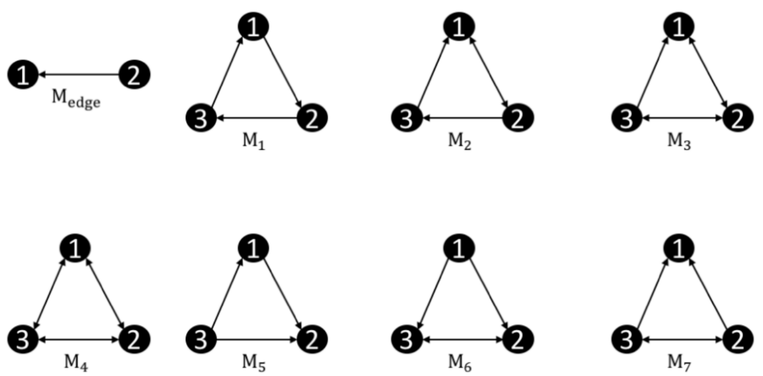
We also need a definition for the collection of all instances of a particular motif:

Definition (Motif set)**.** The motif set in an unweighted directed graph $\mathcal{G}$ with vertices V and adjacency matrix $\mathbf{W}$, denoted $\mathcal{M}(\mathbf{B}, \mathcal{A})$, is defined as

$$\mathcal{M}(\mathbf{B}, \mathcal{A}) = \{(\mathbf{v}, \chi_{\mathcal{A}}(\mathbf{v})) : \mathbf{v} \subseteq V, |\mathbf{v}| = k, \mathbf{W}_{\mathbf{v}} = \mathbf{B}\}, \tag{5}$$

where $\chi_{\mathcal{A}}(\mathbf{v}))$ selects the elements of $\mathbf{v}$ which are in the anchor set, and $\mathbf{W}_{\mathbf{v}}$ is the adjacency matrix of the subgraph induced by $\mathbf{v}$. Intuitively, this is the set of all k-subsets of V whose subgraph structure (i.e. its adjacency matrix) matches that of the motif described by B, with each subset being paired with its set of anchor nodes.

Here, we only look at simple motifs. All possible triangular simple motifs where each node is connected are given below:



Calculating the presence of motifs

The motif-based adjacency matrix, which is defined as

$$(\mathbf{W}_M)_{ij} = \sum_{(\mathbf{v}, \chi_{\mathcal{A}}(\mathbf{v})) \in \mathcal{M}} \mathbf{1}(\{i, j\} \subset \chi_{\mathcal{A}}(\mathbf{v})), \tag{6}$$

indicates how many times two nodes occur together in the anchor set of a motif M. We are able to compute $\mathbf{W}_M$ using a deconstruction of $\mathbf{W}$:

Let $\mathbf{W}$ be the adjacency matrix of the graph $\mathcal{G}$ we are analysing. Then:

- $\mathbf{B} = \mathbf{W} \odot \mathbf{W}^T$ indicates bidirectional links between nodes, where $\odot$ denotes the Hadamard (entry-wise) product, and
- $\mathbf{U} = \mathbf{W} - \mathbf{B}$ indicates unidirectional links between nodes [6].

In each ordered triangular motif, there are six ways of two nodes occurring. In motifs where there are symmetries, like M_6, these are not all unique. We give an idea of the computational method by computing the motif in figure 1, an instance of $\mathbf{W}_{M_6}$:

In M_6, consider two nodes i, j occurring in the 1 and 2 positions. Node i must have a unidirectional connection to another auxiliary node, which must have a bidirectional connection to node j. The frequency of this condition being true is given by $\mathbf{A}_{ij}$, where $\mathbf{A} = (\mathbf{U} \cdot \mathbf{B}) = \sum_k u_{ik}b_{kj}$. Node i must also have a unidirectional connection to node j, which is guaranteed with $\mathbf{A} \odot \mathbf{U}$. The same logic can be applied to the other possibilities for two nodes in M_6, displayed in the table below.

Motif	Matrix Computation	$\mathbf{W}_{M_i}$
M_1	$\mathbf{C} = (\mathbf{U} \cdot \mathbf{U}) \odot \mathbf{U}^T$	$\mathbf{C} + \mathbf{C}^T$
M_2	$\mathbf{C} = (\mathbf{B} \cdot \mathbf{U}) \odot \mathbf{U}^T + (\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{U}^T + (\mathbf{U} \cdot \mathbf{U}) \odot \mathbf{B}$	$\mathbf{C} + \mathbf{C}^T$
M_3	$\mathbf{C} = (\mathbf{B} \cdot \mathbf{B}) \odot \mathbf{U} + (\mathbf{B} \cdot \mathbf{U}) \odot \mathbf{B} + (\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{B}$	$\mathbf{C} + \mathbf{C}^T$
M_4	$\mathbf{C} = (\mathbf{B} \cdot \mathbf{B}) \odot \mathbf{B}$	$\mathbf{C}$
M_5	$\mathbf{C} = (\mathbf{U} \cdot \mathbf{U}) \odot \mathbf{U} + (\mathbf{U} \cdot \mathbf{U}^T) \odot \mathbf{U} + (\mathbf{U}^T \cdot \mathbf{U}) \odot \mathbf{U}$	$\mathbf{C} + \mathbf{C}^T$
M_6	$\mathbf{C} = (\mathbf{U} \cdot \mathbf{B}) \odot \mathbf{U} + (\mathbf{B} \cdot \mathbf{U}^T) \odot \mathbf{U}^T + (\mathbf{U}^T \cdot \mathbf{U}) \odot \mathbf{B}$	$\mathbf{C}$
M_7	$\mathbf{C} = (\mathbf{U}^T \cdot \mathbf{B}) \odot \mathbf{U}^T + (\mathbf{B} \cdot \mathbf{U}) \odot \mathbf{U} + (\mathbf{U} \cdot \mathbf{U}^T) \odot \mathbf{B}$	$\mathbf{C}$

Table 1. Motif-based adjacency matrices for triangular motifs

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" [6]. 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 3.

In [6], the linear motif-based PageRank is shown to be 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

[1] Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, jul 2016. doi:10.1126/science.aad9029.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[3] Wei Chen, Carlos Castillo, and Laks VS Lakshmanan. *Information and influence propagation in social networks*. Springer Nature, 2022.

[4] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond*. Princeton University Press, Princeton, 2006. ISBN 9781400830329. doi:doi:10.1515/9781400830329.

[5] Tom Seymour, Dean Frantsvog, Satheesh Kumar, et al. History of search engines. *International Journal of Management & Information Systems (IJMIS)*, 15(4):47–58, 2011.

[6] 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 amp; Data Engineering*, 33(05):2179–2192, may 2021. ISSN 1558-2191. doi:10.1109/TKDE.2019.2953264.