

Google Maths: PageRank & Markov Chain

Florent ONANA ASSOUGA

Under the Supervision of

Prof. Philip Anthony Knight

Strathclyde University, Glasgow, UK



AIMS

African Institute for
Mathematical Sciences
CAMEROON

February 3, 2024



Table of contents

- 1 Motivation-Objective
- 2 History of Search Engines
 - Google and Googleplex
 - History
- 3 PageRank Algorithm
 - A web as a Directed Network
 - Importance Score
 - Computing the Importance Score Eigenvector
- 4 Markov Chain and Google's PageRank Algorithm
 - Stochastic interpretation of PageRank
 - Applications
- 5 Conclusion



Motivation-Objective

Problem Statement

Suppose we enter “complex network” into Google’s search engine. Google responds by telling us there are **24.9 million** web pages containing those terms. On the first page, however, there are links to ten web pages that Google judges to have the highest quality and therefore the ones we are most likely to be interested in. **How does Google assess the quality of web pages?**

- Brief History of search engines
- Explain one of the core ideas (eigenvalue, Markov chain) behind PageRank on how Google calculates web page rankings.



Google and Googleplex



Figure: Larry Page and Sergey Brin

- Founded by [Larry Page](#) and [Sergei Brin](#)(Stanford University 1996 as part of a research project).
- [PageRank\(PR\)](#) is an algorithm used by Google search to rank web pages in their search engine results.
- It is not the only algorithm used by Google to order search result but it the first algorithm that was used by the company and it is the best known.



History of Eigenvalue problem behind PageRank

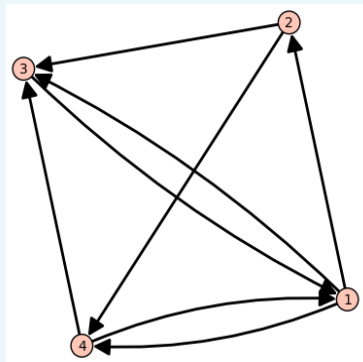
The eigenvalue problem behind PageRank's algorithm was used in many scoring problem:

Year	Author	Objective
1895	Edmund Landrau	Determining the winner of a chess tournament
1976	Gabriel Pinski and Francis Narin	scientometrics ranking scientific journal
1977	Thomas Saaty	Analytical hierarchy process
1995	Bradley Love and Steven Sloman	Cognitive model for concepts



Web as a Directed Network

- Think of web as a collection of vertices and edges.
- Vertices (V) are world wide web pages.
- Edges (E) are links or hyperlinks.



- If u and v are two websites, the edge (u, v) indicates that website u contains a hyperlink to website v .

Figure: Example of a web with 4 pages



Naive approach to rank

- We are looking for quantitative rating of a web page's importance.
- Suppose the web contains n pages, each page indexed by an integer k , $1 \leq k \leq n$. We will use x_k to denote the **importance score of page k** in the web. $x_k \geq 0$ and $x_j > x_k$ means that the page j is more important than page k .

A Naive Approach

A naive approach is to rank the sites by the number of incoming hyperlinks: it is the **in degree**.

Criticism

A link from an “important” page should carry more weight than a link from some random blog!

Second approach: A democracy of the web

- If page j contains n_j links, we will boost page k 's score by $\frac{x_j}{n_j}$. We denote by $L_k \subset \{1, 2, \dots, n\}$ the set of pages with a link to page k .
- We require

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}. \quad (1)$$

- Equation (1) can be written as

$$Ax = x, \quad \text{with } x = (x_1, x_2, \dots, x_n)^T \quad (2)$$

and

$$A_{ij} = \begin{cases} \frac{1}{n_j} & \text{if page } j \text{ links to page } i \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$



Second approach: A democracy of the web ctd

- Equation (2) transforms the web ranking problem into the “standard” problem of finding an eigenvector with eigenvalue 1 for the square matrix A (called **link matrix**): A is a stochastic matrix.

Definition

A **Markov matrix** (or **stochastic matrix**) is a square matrix A whose all of its entries are nonnegative and the entries in each column sum to one.

Proposition

Every stochastic matrix has 1 as an eigenvalue.

Criticism

The eigenspace for the eigenvalue 1 is not always 1: this leads to **Non-Unique Rankings**.

Google approach: Modification to the link matrix A

- Let denote by R an $n \times n$ matrix with all entries $\frac{1}{n}$ and $\alpha \in [0, 1]$.

Definition (Google matrix)

The following matrix M is called the **Google matrix**:

$$M = (1 - \alpha)A + \alpha R. \quad (4)$$

Theorem (Perron–Frobenius Theorem (circa 1910))

If M is a Markov matrix with all positive entries, then M has a unique steady-state vector, ie, there exists a unique x such that: $Mx = x$.



Computing the Importance Score: The power method

- We start with a typical vector x_0 .
- We generate the sequence $x_{k+1} = Mx_k$ or $x_k = M^k x_0$ and let $k \rightarrow \infty$.
- To solve the problem of magnitude, the sequence x_k can be defined as

$$x_k = \frac{Mx_{k-1}}{|Mx_{k-1}|}. \quad (5)$$

Theorem

The Google matrix M for a web with no dangling nodes has a unique vector q with positive components such that $Mq = q$ and

$$q = \lim_{k \rightarrow \infty} M^k x_0 \quad \forall x_0 > 0. \quad (6)$$

Markov Chain from a Random Walk

- What does this have to do with Markov chains?
- Brin and Page considered web surfing as a stochastic process:

Quote

PageRank can be thought of as a model of user behavior. We assume there is a “random surfer” who is given a web page at random and keeps clicking on links, never hitting “back” but eventually gets bored and starts on another random page.

- Surfer clicks on a link on the current page with probability $1 - \alpha$; opens up a random page with probability α .
- A page's rank is the probability the random user will end up on that page,



Example 1

- The value of α originally used by Google is reportedly 0.15.

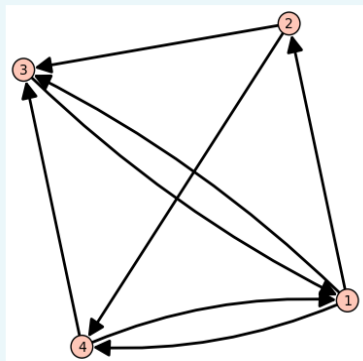
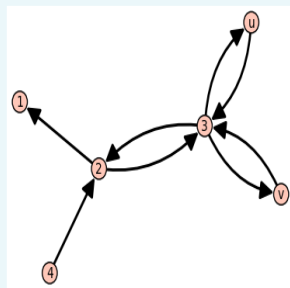
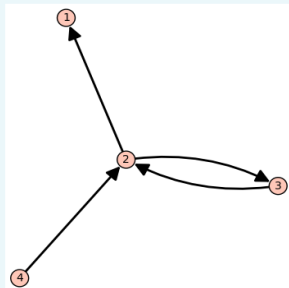
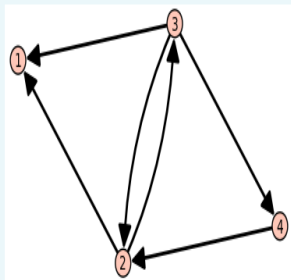


Figure: Example of a web with 4 pages

- Naive approach:** The importance vector score is $(x_1, x_2, x_3, x_4) = (2, 1, 3, 2)$. So the page **3** would be the most important, then **1,4,2**.
- Second approach:** The importance score is $(x_1, x_2, x_3, x_4) = (12, 4, 9, 6)$. The most important page is **1**, then **3,4,2**. **Surprising, but why?**
- Google approach:** The vector score is $(x_1, x_2, x_3, x_4) = (0.368, 0.142, 0.288, 0.202)$. This yields the same ranking as the earlier



Example 2: Sybil attack



- Website 3 wants to improve its PageRank, which is ≈ 0.23 in the initial graph on the left.
- First all outgoing links to website that do not link back are removed: the PageRank improves to ≈ 0.27 .
- The owner of website 3 create fake websites u and v whose purpose is to exchange links with 3: The PageRank becomes 0.41.





Conclusion

- There are other link-based ranking algorithms for web pages: [HITS\(Hyperlink-Induced Topic Search\)](#) algorithm by [Jon Kleinberg](#), [IBM CLEVER](#) project, [TrustRank](#), [SALSA](#) algorithm ...
- The initial version of PageRank algorithm can be fooled, for example in a [Sybil attack](#): The owner of a website u creates fake websites whose purpose is to exchange links with u .
- It is unknown how exactly Google ranks websites today, and specifically how the engineers at Google mitigate the effects of attacks.



References

-  Kurt Bryan, Tanya Leise *The \$25, 000, 000, 000 eigenvector The linear algebra behind Google*
-  Ian Rogers, *The Google Pagerank algorithm and how it works*, [http : //www.iprcom.com/papers/pagerank/](http://www.iprcom.com/papers/pagerank/) (accessed August 1, 2005).



**THANK YOU FOR YOUR KIND
ATTENTION**

