UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE MATEMÁTICA

PageRank e pseudo-PageRank: convergência do algoritmo e tratamento da rede com nodos pendentes

Adriana Cristina Farinha Matos

Mestrado em Matemática

Dissertação orientada por:
Professor Doutor Carlos Manuel Ribeiro Albuquerque

2025

Agradecimentos

Aos meus pais, por acreditarem sempre que eu consigo fazer tudo. E a mim, por tentar sempre fazer tudo para não os desiludir.

Aos meus avós, pela infância feliz a fazer contas.

Aos meus amigos, por me darem o tão especial sentido de pertença e por falarem de Matemática comigo da maneira que eu mais gosto: com entusiasmo!

Aos meus professores, aos de Matemática, aos da faculdade, por me desafiarem. E por nos ensinarem tanto que se torna impossível aprender tudo. Mas é sempre bom tentar.

Ao Professor Carlos Albuquerque, pela paciência, por ser generoso com a sua disponibilidade, com a sua partilha de conhecimentos, o seu interesse genuíno por todas as questões abordadas e a sua orientação essencial durante todo este processo. Já acabou, Professor!

À Professora Maria Manuel Torres e ao Professor Jorge Buescu por todo o apoio, pelos pontos partilhados na apresentação da minha dissertação, que me ajudaram a melhorá-la, e por me incentivarem a continuar.

Ao Miguel Caldeira, o meu chefe, por plantar na minha cabeça a ideia de que devo terminar o que comecei e por tornar isso possível. E ao Rui Serra e ao José Ferreira, por me darem o empurrão inicial na direção certa.

A todos os que, de uma forma ou outra, conspiraram para que isto acontecesse.

O meu mais sincero obrigado!

Resumo

Nesta dissertação, analisa-se detalhadamente a Matemática que está na base do método de ordenação de páginas da Web por relevância utilizado inicialmente pela Google, designado por PageRank e apresentado em [1] por S. Brin e L. Page. O método PageRank baseia-se fortemente em resultados da Teoria dos Grafos e da Teoria das Matrizes Não-Negativas, especialmente no Teorema de Perron-Frobenius, que inspira a solução ao problema inicial da falta de conexão forte da rede. Outros problemas são identificados e tratados também de forma minuciosa.

A apresentação do problema do cálculo do vetor de PageRank envolve a exposição do teorema da convergência do método da potência e de duas provas de como este método pode ser aplicado para a obtenção do vetor de ranking, sendo a primeira a que encontramos em [12] e a segunda em [14]. É ainda exposto o algoritmo construído a partir do método da potência e apresentado o ponto de vista alternativo de D. F. Gleich [14], que depende da apresentação de um teorema fundamental da Teoria das Matrizes-M.

A apresentação do problema da presença dos nodos pendentes na rede é feita com a exposição do conceito de problema de pseudo-PageRank, da adaptação do algoritmo anterior a este contexto e de várias formas possíveis de transformação deste problema num de PageRank, com base no artigo [14] e noutros artigos, tais como [21] e [25]. É ainda feita uma interpretação da solução apresentada por Brin e Page em [1].

Palavras-chave: Teoria dos Grafos, Teoria das Matrizes Não-Negativas, Teorema de Perron-Frobenius, Método da Potência, Teoria das Matrizes-M.

Abstract

In this dissertation, we analyze in detail the Mathemathics underlying the method of ordering webpages by relevance initially used by Google, called PageRank and presented in [1] by S. Brin and L. Page. The PageRank method relies heavily on results from the Graph Theory and the Theory of Nonnegative Matrices, especially the Perron-Frobenius Theorem, which inspires the solution to the initial problem of the lack of a strongly connected web graph. Other problems are identified and also thoroughly treated.

The presentation of the problem of calculating the PageRank vector involves the exposition of the power method convergence theorem and two proofs of how this method can be applied to obtain the ranking vector, the first being found in [12] and the second in [14]. It is also exposed the algorithm built from the power method and presented the alternative point of view by D. F. Gleich [14], which depends on the presentation of a fundamental theorem of the Theory of M-Matrices.

The presentation of the problem caused by the presence of dangling nodes in the web is done by exposing the concept of the pseudo-PageRank problem, the adaptation of the previous algorithm to this context and several possible ways of transforming this problem into a PageRank problem, based on article [14] as well as others, such as [21] and [25]. An interpretation of the solution presented by Brin and Page in [1] is also made.

Keywords: Graph Theory, Theory of Nonnegative Matrices, Perron-Frobenius Theorem, Power Method, Theory of M-Matrices.

Índice

Introdução

A criação da World Wide Web e a consequente generalização do acesso à Internet na década de 1990 expandiu de forma grandiosa o número de documentos da rede e permitiu o crescimento contínuo do mesmo, ficando assim o sucesso do uso em massa deste sistema dependente da resposta à seguinte pergunta: como encontrar na Web a informação que se pretende?

O fim dos anos 90 ficou marcado pela ansiedade em solucionar este problema e vários motores de busca foram criados a fim de facilitar o processo de procura de informação na Internet. O objectivo em comum dos motores seria a catalogação das páginas da Web por palavras-chave e a apresentação ordenada de cada um destes catálogos (isto é, resultados de pesquisa) da forma mais rápida, organizada e eficaz possível.

Apesar do objectivo em comum, os motores de busca apresentaram métodos de ordenação bastante distintos, sendo que um em particular ganhou destaque e continua a distinguir-se da concorrência desde então. O conhecido método PageRank, correspondente ao ainda mais conhecido motor de busca Google, tem sido frequentemente analisado e discutido desde a sua apresentação por S. Brin e L. Page em [1]. A sua clara eficácia e o acesso à formulação do método inspiraram em grande escala o estudo detalhado do mesmo e da Matemática presente, bem como permitiram o desenvolvimento e adaptação do método para vários contextos, existindo hoje em dia inúmeras exposições sobre o PageRank.

Posto isto, esta dissertação foi escrita a fim de satisfazer dois propósitos.

O primeiro é o de explicar detalhadamente a Matemática presente na fundação do PageRank, que, para além de nos dar a conhecer este método em maior profundidade, permite-nos ver a Matemática como resposta a um problema concreto e crítico no contexto em que se insere.

O segundo é o de partilhar todos os casos do PageRank com o rigor que a Matemática nos exige, o que não foi um exercício fácil, dada a falta de exposições detalhadas sobre alguns casos particulares deste método. Creio que o facto da Matemática necessária para os apresentar ser mais desafiante seja uma razão para isto acontecer.

O elevado sentido de espírito crítico que ganhei nestes anos a estudar Matemática é algo que valorizo bastante. Espero tê-lo aplicado da melhor forma na escrita desta dissertação, assim como o tenho feito profissionalmente no meu dia-a-dia enquanto software tester.

Sendo assim, a dissertação encontra-se estruturada da forma seguinte.

No primeiro capítulo, é apresentado o PageRank como um problema de centralidade do grafo construído a partir da Web. É construído o ranking dos nodos do grafo a partir do vetor próprio estocástico associado ao valor próprio 1, o qual é designado por vetor de PageRank. É identificado o problema trazido pela falta de conexão forte e a solução construída com base no Teorema de Perron-Frobenius, sendo apresentado o conceito de teletransporte e feita a definição da matriz Google. Dois outros problemas são

identificados e são os temas dos dois capítulos seguintes: o problema do cálculo do vetor de PageRank e o problema trazido pela presença de nodos pendentes na rede.

No segundo capítulo, é tomada como hipótese a inexistência de nodos pendentes na rede. É feita uma análise mais profunda ao método da potência e são apresentadas duas provas de como este método pode ser aplicado na matriz Google. A segunda prova é feita após uma apresentação do ponto de vista alternativo de D. F. Gleich que depende da Teoria das Matrizes-M. É apresentado o algoritmo de aproximação ao vetor de PageRank construído a partir da iteração de von Mises em conjunto com a hipótese tomada.

No terceiro capítulo, é tratado o caso da rede com nodos pendentes. É apresentado o conceito de problema de pseudo-PageRank, bem como o respectivo vetor de pseudo-PageRank. É feita uma adaptação do algoritmo apresentado no capítulo anterior e provado como esta se aproxima eficazmente ao vetor de pseudo-PageRank. São apresentadas formas de transformação de um problema de pseudo-PageRank num problema de PageRank e é feita uma interpretação da solução inicial dada por Brin e Page.

1 PageRank e Teorema de Perron-Frobenius

Quando utilizamos um motor de busca para procurar informação na Web, seria bastante incómodo se os resultados apresentados estivessem ordenados de forma aleatória. Como poderíamos distinguir que páginas são as mais relevantes para a nossa pesquisa numa rede crescente de milhares de milhões delas? É essa distinção que os motores de busca calculam: depois de escrevermos as palavras-chave que queremos pesquisar, os ficheiros da rede que contiverem essas palavras são apresentados por ordem de relevância. Cada motor de busca apresenta um método próprio para fazer essa ordenação, sendo que alguns destes métodos sobressaem pela sua eficácia. Um utilizador do Google poderá notar que em raras vezes necessitou de visitar a segunda página de resultados apresentados para a sua pesquisa e que em muitas vezes encontrou o que pretendia logo no primeiro resultado apresentado. O eficaz método de ordenação de páginas por relevância do Google designa-se de PageRank e é sobre ele que este capítulo é escrito.

1.1 PageRank

Em 1998, S. Brin e L. Page, estudantes do Doutoramento em Ciências da Computação na Universidade de Stanford, constroem o motor de busca Google, após colaborarem no novo e eficiente modelo PageRank, apresentado em mais detalhe no artigo [1] e mencionado no artigo de apresentação do Google [2]. De notar que, curiosamente, o nome atribuído foi inspirado no termo *googol*, que era utilizado como referência ao número 10^{100}, lembrando a enorme quantidade de páginas com que um motor de busca tem de lidar.

A eficiência do modelo de Brin e Page na ordenação desta enorme quantidade de páginas tornou-se muito clara rapidamente e deste então foram feitas inúmeras exposições sobre a história do PageRank, a Matemática presente na sua constituição e as suas possíveis aplicações alternativas. O conceito base deste método é simples e encontra-se bem explicado nas exposições feitas por R. Tanase e R. Radu em [3] e por P. F. Gallardo em [4], que serviram como base na escrita deste capítulo. Podemos até obter uma boa ideia geral de como o método funciona dando uso ao próprio e utilizando o motor de busca Google para pesquisar pela palavra-chave "PageRank". Penso que é fácil adivinhar qual o primeiro resultado que surge.

Quando pensamos no conjunto de todas as páginas da Internet, normalmente utilizamos o termo rede (ou Web), pois temos consciência que as páginas estão ligadas entre si: a página x pode fazer referência à página y, através de hiperligações. A novidade trazida pelo PageRank é a ideia de que a relevância de uma página é o reflexo da relevância das páginas que se referem a ela.

Pensando, assim, na rede, podemos naturalmente representá-la através de um grafo orientado e com pesos, cujos nodos são as várias páginas na Internet e cujas arestas são as hiperligações entre si, sendo que, se a página x fizer referência à página y, a aresta existente estará direcionada do nodo da página x

para o nodo da página y. Os pesos das arestas reflectirão a maneira como a relevância de uma página se distribui pelas suas referências.

Sendo o PageRank um método para encontrar o nodo mais relevante num grafo, este é considerado uma medida de centralidade. Centralidade é um termo usado na Teoria dos Grafos e, em particular, na análise de redes, e é o que estima a importância relativa de cada nodo numa rede. Existem várias medidas de centralidade diferentes que podem ser usadas para esta estimativa, como a de grau, a de proximidade e a de vetor próprio, sendo esta última a mais similar ao PageRank. Mais informações sobre o conceito de centralidade e sobre as suas várias medidas podem ser encontradas em [5].

Consideremos o seguinte exemplo. Suponhamos que estamos na presença de uma rede de seis páginas apenas e que estas se ligam entre si da seguinte maneira:

(a) A página 1 faz referência às páginas, 2, 4 e 5;

(b) A página 2 faz referência às páginas 3 e 5;

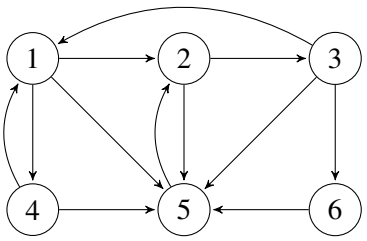
(c) A página 3 faz referência às páginas 1, 5 e 6;

(d) A página 4 faz referência às páginas 1 e 5;

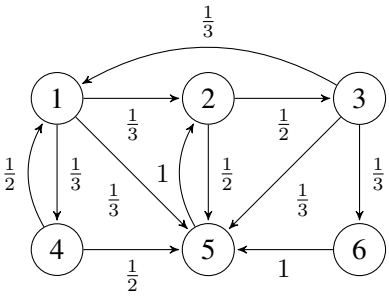
(e) A página 5 faz referência somente à página 2;

(f) A página 6 faz referência somente à página 5.

Analisando esta rede, podemos começar por construir o seguinte grafo:



Como podemos verificar, o nodo 5 é o que tem mais arestas a apontar para si, é o mais popular. Será que é o mais relevante? Se o nodo 5 for o mais relevante, em que posição ficará o nodo 2, a sua única referência? Pensemos no assunto desta forma: cada página distribui em porções iguais a sua relevância pelas páginas a que se refere. Logo, o nodo 1 distribuirá a sua relevância pelos nodos 2, 4 e 5 (dando $\frac{1}{3}$ a cada um); o nodo 2 distribuirá a sua relevância pelos nodos 3 e 5 (dando $\frac{1}{2}$ a cada um); assim por diante. Desta forma, podemos juntar pesos às arestas do grafo acima da seguinte forma:



Tenhamos em conta a definição seguinte.

Definição 1.1.1. *Numa rede de n páginas, definimos a matriz de distribuição por hiperligações, H, de dimensão $n \times n$, que representa a distribuição de relevância por hiperligações entre páginas e cujas entradas são dadas por*

$$H_{ij} = \begin{cases} 0, \textit{se a página } j \textit{ não apresentar nenhuma hiperligação para a página } i \\ \frac{1}{n_j}, \textit{caso contrário} \end{cases}$$

onde $i, j \in \{1, ..., n\}$ e, para cada j, n_j é a quantidade de páginas da rede a que a página j se refere através de hiperligações.

É fácil ver, pela definição, que uma matriz de distribuição por hiperligações é, essencialmente, a matriz transposta da matriz de adjacência de um grafo orientado e com pesos, representativo do contexto específico da Web (em que os nodos são páginas e as arestas são hiperligações).

Sendo assim, a matriz de distribuição por hiperligações do grafo acima é a matriz

$$H = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix}.$$

Podemos dizer que, neste caso, a matriz H é estocástica, tendo em conta a próxima definição.

Definição 1.1.2. *Uma matriz quadrada diz-se estocástica se for não-negativa e as entradas de cada uma das suas colunas somarem 1. Um vetor diz-se estocástico se for não negativo e as suas entradas somarem 1.*

Denotemos agora por r_1, r_2, r_3, r_4, r_5 e r_6 a relevância das 6 páginas, respectivamente.

Analisando a rede exemplificada, pretendemos obter as soluções do sistema:

$$\begin{cases} r_1 = \frac{1}{3}r_3 + \frac{1}{2}r_4 \\ r_2 = \frac{1}{3}r_1 + r_5 \\ r_3 = \frac{1}{2}r_2 \\ r_4 = \frac{1}{3}r_1 \\ r_5 = \frac{1}{3}r_1 + \frac{1}{2}r_2 + \frac{1}{3}r_3 + \frac{1}{2}r_4 + r_6 \\ r_6 = \frac{1}{3}r_3 \end{cases}$$

O que é o mesmo que procurar as soluções de

$$H \begin{bmatrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 \end{bmatrix}^T = \begin{bmatrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 \end{bmatrix}^T.$$

Ou seja, as soluções que procuramos são os vetores próprios de H associados ao valor próprio 1. Tenhamos em conta os seguintes resultados.

Teorema 1.1.1. *Seja M uma matriz não-negativa. Suponhamos que existe um vetor positivo x e um número real não-negativo λ tal que $Mx = \lambda x$ ou $x^T M = \lambda x^T$. Então, o raio espectral de M, $\rho(M)$, é igual a λ.*

Demonstração. Cf. [6], Capítulo 8, Teorema 8.3.4. □

Proposição 1.1.1. *Seja M uma matriz estocástica. Então 1 é valor próprio de M. Para além disso, $\rho(M) = 1$.*

Demonstração. Para provar a primeira parte, como qualquer matriz tem os mesmos valores próprios que a sua transposta, basta provar que 1 é valor próprio de M^T. Dado que M é estocástica, $M^T e = e$, onde e é o vetor coluna cujas entradas são todas iguais a 1. A segunda parte sai por aplicação do Teorema 1.1.1, dado que, como M é estocástica, M é não-negativa e satisfaz $e^T M = e^T$. □

A apresentação e demonstração destes resultados tem por base o segundo parágrafo do subcapítulo 8.7 do livro [6], em que, numa única frase, R. A. Horn e C. R. Johnson provam por completo a Proposição 1.1.1, *"The defining identity $Ae = e$ and $(8.3.4)$ tell us that $+1$ is not only an eigenvalue of a stochastic matrix, but also its spectral radius"*, notando que o resultado (8.3.4) mencionado corresponde ao Teorema 1.1.1.

De momento, apenas nos interessa saber que 1 é valor próprio de uma matriz estocástica. A informação relativa ao raio espectral será importante mais tarde, como veremos na secção 1.2.

Pela proposição anterior, dado que H é estocástica, 1 é valor próprio de H e as soluções procuradas existem. Neste caso, calculando os vetores próprios de H associados ao valor próprio 1, verificamos que

$$\begin{bmatrix} 6 & 30 & 15 & 2 & 28 & 5 \end{bmatrix}^T$$

é uma solução e obtemos a ordem de relevância das seis páginas: $r_2 > r_5 > r_3 > r_1 > r_6 > r_4$. Não afirmamos que $r_1 = 6, r_2 = 30, \ldots$, pois a relevância de cada página é relativa e não a podemos quantificar de forma única. De facto, e verificando facilmente que a multiplicidade geométrica do valor próprio 1 é 1, poderíamos escolher qualquer um dos vetores próprios de H associados a 1 para sabermos a ordem de relevância das páginas, visto que estes diferem apenas pela multiplicação por um escalar. Sabendo que existe um único destes vetores que é estocástico, escolhamos esse vetor v e designemo-lo por vetor de PageRank. Neste caso, o nosso vetor de PageRank é

$$v = \frac{1}{86}\begin{bmatrix} 6 & 30 & 15 & 2 & 28 & 5 \end{bmatrix}^T \approx \begin{bmatrix} 0,070 & 0,349 & 0,174 & 0,023 & 0,326 & 0,058 \end{bmatrix}^T.$$

Esclarecemos, assim, as dúvidas apresentadas anteriormente, pois verificamos que a página mais relevante é a página 2, que é a única referência da página mais popular, a página 5, ficando esta em segundo lugar neste ranking.

Esta solução foi obtida pensando no problema do ponto de vista da Álgebra Linear, mas podemos pensá-lo de outras formas.

Do ponto de vista dos Sistemas Dinâmicos, podemos começar por considerar que cada nodo tem a mesma relevância, representando por v_0 o nosso vetor de rank inicial,

$$v_0 = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}^T.$$

A partir deste momento, a relevância de cada página irá depender das distribuições de relevância contabilizadas pela matriz H. Sabemos que a nova relevância relativa da página 1 é resultado da soma entre um terço da relevância da página 3 e metade da relevância da página 4, o que se traduz, no passo 1, em $\frac{1}{3} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{6} = \frac{5}{36} \approx 0,139$. Para as outras páginas, o raciocínio é o mesmo e verifica-se que o vetor de rank do passo 1 é

$$v_1 = Hv_0 = \begin{bmatrix} \frac{5}{36} & \frac{2}{9} & \frac{1}{12} & \frac{1}{18} & \frac{4}{9} & \frac{1}{18} \end{bmatrix}^T \approx \begin{bmatrix} 0,139 & 0,222 & 0,083 & 0,056 & 0,444 & 0,056 \end{bmatrix}^T.$$

Para calcular o vetor de rank do passo 2 utilizamos o mesmo raciocínio, obtendo

$$v_2 = Hv_1 = H(Hv_0) = H^2v_0 \approx \begin{bmatrix} 0,056 & 0,491 & 0,111 & 0,046 & 0,269 & 0,028 \end{bmatrix}^T.$$

Iterando o processo, construímos a sequência de vetores de rank v_0, v_1, v_2, $\ldots$, v_k, igual a v_0, Hv_0, $H^2v_0, \ldots, H^kv_0$, e verificamos que, para $k \geq 23$, cada vetor de rank v_k tem a mesma aproximação (a 3 casas decimais), esta é estocástica e é escolhida como o nosso vetor de PageRank,

$$v = \begin{bmatrix} 0,070 & 0,349 & 0,174 & 0,023 & 0,326 & 0,058 \end{bmatrix}^T.$$

Do ponto de vista probabilístico, podemos pensar que estamos a navegar pela nossa rede seguindo sempre alguma das hiperligações que cada página tem de referência a outra página, o que dá um significado probabilístico aos pesos do nosso grafo e à relevância de cada página.

Num momento inicial, cada página tem igual probabilidade de ser visitada: $\frac{1}{6}$. No momento seguinte, a página 1, por exemplo, terá 0 probabilidade de ser visitada, caso começássemos por visitar a página 1, a página 2, a página 5 ou a página 6 (pois nenhuma das páginas tem uma hiperligação para a página 1); terá $\frac{1}{3}$ de probabilidade de ser visitada, caso começássemos por visitar a página 3 (temos hiperligações para três páginas na página 3, sendo uma delas para a página 1 e todas com igual probabilidade de serem visitadas); e terá $\frac{1}{2}$ de probabilidade de ser visitada, caso começássemos por visitar a página 4 (pela mesma razão, mas neste caso existem somente duas hiperligações).

Representando de outra forma, a probabilidade de visitarmos a página 1 no momento seguinte é $0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} = \frac{5}{36} \approx 0,139$. Continuando o mesmo raciocínio e representando as probabilidades das páginas serem visitadas após k passos num vetor de probabilidades v_k, obtemos a mesma sequência analisada no ponto de vista anterior,

$$v_0 = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}^T, v_1 = Hv_0, v_2 = H^2v_0, \ldots, v_k = H^kv_0,$$

que sabemos convergir (numa aproximação a 3 casas decimais) para o vetor

$$v = \begin{bmatrix} 0,070 & 0,349 & 0,174 & 0,023 & 0,326 & 0,058 \end{bmatrix}^T.$$

Esta distinção entre os três pontos de vista pode ser encontrada em [3].

Tenhamos agora em conta as seguintes definições.

Definição 1.1.3. *Seja $\mathscr{G}$ um grafo. Designa-se por cadeia uma sucessão $(a_1, ..., a_k)$ de arestas de $\mathscr{G}$ tal que, para cada $i = 2, 3, ..., k-1$, a aresta a_i tenha em comum com a_{i+1} uma extremidade e com a_{i-1} a outra extremidade.*

Definição 1.1.4. *Seja $\mathscr{G}$ um grafo orientado. Designa-se por caminho uma sucessão $(a_1, ..., a_k)$ de arestas de $\mathscr{G}$ tal que, para cada $i = 1, ..., k-1$, a extremidade final da aresta a_i seja a extremidade inicial da aresta a_{i+1}.*

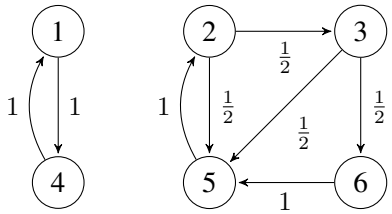
Definição 1.1.5. *Seja $\mathscr{G}$ um grafo. Dizemos que $\mathscr{G}$ é conexo se existir uma cadeia entre cada par de nodos do grafo. Caso contrário, dizemos que $\mathscr{G}$ é desconexo.*

No caso dos grafos orientados, podemos ignorar a orientação das arestas e analisar os grafos resultantes para verificar as propriedades de conexão ou desconexão, segundo a definição anterior. Os grafos orientados que são conexos são frequentemente designados na literatura por fracamente conexos. Em paralelo, define-se o conceito de grafo orientado fortemente conexo da forma seguinte:

Definição 1.1.6. *Seja $\mathscr{G}$ um grafo orientado. Dizemos que $\mathscr{G}$ é fortemente conexo se existir um caminho entre cada par de nodos do grafo.*

A forma de calcular o rank das páginas poderá não ser bem sucedida em alguns casos e até apresentar resultados ambíguos. Na verdade, a Web tem um comportamento bastante heterogéneo e, apesar do exemplo apresentado anteriormente dar origem a um grafo conexo, essa não é, como podemos esperar, a realidade do grafo formado pelas páginas da Internet. Podemos encontrar uma representação gráfica da topologia do grafo representativo da Web no artigo [4].

Vejamos o que acontece quando estamos na presença de um grafo desconexo tomando o exemplo seguinte:



As páginas 1 e 4 não têm qualquer referência às páginas 2, 4, 5 e 6; e vice-versa. Ou seja, se escolhermos uma página de forma aleatória para iniciar a nossa navegação pela Internet e continuarmos seguindo apenas as hiperligações que encontramos, começando numa componente nunca chegaremos à outra. Neste caso, a matriz de distribuição por hiperligações é

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}.$$

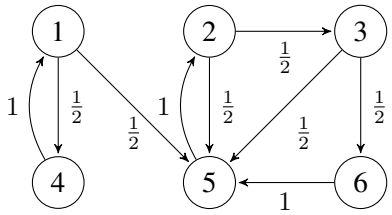
E, analisando a matriz H, verificamos que esta é estocástica, apresentando assim 1 como valor próprio, tendo este agora, no entanto, multiplicidade geométrica igual a 2.

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}^T, \begin{bmatrix} 0 & 4 & 2 & 0 & 4 & 1 \end{bmatrix}^T$$

são vetores próprios de H associados ao valor próprio 1, não são múltiplos um do outro e indicam-nos ordenações de relevância distintas. Logo, não é possível escolher um vetor único para ser o nosso vetor de PageRank, sendo necessário, portanto, refinar o método que nos leva a encontrar este vetor.

A falta de unicidade da solução não é o único problema relacionado com a topologia do grafo. Se adicionarmos uma aresta que ligue as duas partes conexas do grafo, por exemplo, o grafo torna-se conexo, mas outro problema surge.

Adicionando uma aresta do nodo 1 para o nodo 5, transformamos o grafo do exemplo anterior no grafo conexo



e obtemos a matriz de distribuição por hiperligações

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}.$$

Esta matriz também é estocástica e, portanto, apresenta 1 como valor próprio. Mas, apesar de agora este ser um valor próprio simples, o vetor próprio associado é o vetor

$$\begin{bmatrix} 0 & 4 & 2 & 0 & 4 & 1 \end{bmatrix}^T,$$

que apresenta entradas nulas. Estas atribuem relevância nula para as páginas 1 e 4 da rede. Isto acontece devido ao facto do grafo ser conexo, mas não fortemente conexo: partindo de um dos nodos 2, 3, 5 ou 6, não é possível chegar ao nodo 1 nem ao 4, ficando a relevância das páginas distribuída apenas pelos outros quatro nodos. No entanto, todas as páginas devem ter relevância e isso deve traduzir-se num vetor de PageRank com todas as entradas positivas.

Para percebermos melhor a relação entre a topologia dos grafos e as propriedades das matrizes de distribuição por hiperligações, tenhamos em conta a seguinte definição.

Definição 1.1.7. *Seja M uma matriz quadrada de dimensão n. Dizemos que M é uma matriz redutível se existir uma matriz de permutação P de dimensão n, tal que*

$$P^T M P = \begin{bmatrix} X & Z \\ 0 & Y \end{bmatrix}$$

onde X é uma matriz quadrada de dimensão s, Y é uma matriz quadrada de dimensão $n - s$ e Z é uma matriz de dimensão $s \times (n - s)$ para algum $s \in \mathbb{N}$. Caso contrário, dizemos que M é uma matriz irredutível.

Analisando a definição anterior, percebemos que a presença de uma matriz de distribuição por hiperligações redutível indica-nos a falta de conexão forte no grafo, dado que, a menos de ordenação, os s primeiros nodos não têm arestas direcionadas para os restantes $n - s$, não sendo possível, assim, a formação de caminhos entre um nodo dos primeiros s e outro nodo dos restantes $n - s$. Poderá indicar-nos até mesmo a desconexão do grafo, caso o bloco superior Z seja nulo. Poderemos, a partir daqui, também pensar na relação entre a matriz de distribuição por hiperligações irredutível e a conexão forte do grafo. Quem diz matriz de distribuição por hiperligações, diz, tal como referido anteriormente, matriz transposta da matriz de adjacência do grafo ou, até mesmo, matriz de adjacência do grafo (se uma for irredutível, a outra também será, e vice-versa). Esta ideia traduz-se no resultado seguinte.

Teorema 1.1.2. *Seja $\mathscr{G}$ um grafo orientado e seja G a sua matriz de adjacência. Então, G é irredutível se e só se $\mathscr{G}$ for fortemente conexo.*

Demonstração. Cf. [6], Capítulo 6, Teorema 6.2.24. □

Podemos encontrar de forma bastante detalhada as ideias que rodeiam esta relação entre a topologia do grafo e a propriedade de irredutibilidade da matriz de adjacência, bem como mais sobre a teoria das matrizes não-negativas irredutíveis, nos Capítulos 6 e 8, respectivamente, do livro [6]. É interessante perceber que a relação apresentada pelo teorema é tão intrínseca que se confunde por vezes com uma definição, tal como acontece no subcapítulo 5.3 do livro de A. Bonato [7].

Os problemas da solução não ser única ou positiva gerados pelos grafos desconexos ou não fortemente conexos estão presentes na exposição sobre o PageRank feita por D. Austin em [8]. Esta é uma exposição que também poderá ser considerada um bom ponto de partida para entender a base de funcionamento do PageRank. A exposição [3] de Tanasu e Radu apresenta o problema da multiplicidade trazido pela desconexão, mas não apresenta o problema da relevância nula trazido pela falta de conexão forte. No artigo [4] de Gallardo, é apresentada a questão da unicidade da solução, no entanto não existe a exigência desta ser positiva, sendo apresentada desde início como um vetor não-negativo (ao longo da sua exposição há um maior foco na propriedade da matriz ser não-negativa do que na de ser estocástica).

Podemos agora questionar se, para ultrapassar estes problemas, precisamos que o grafo seja fortemente conexo, ou seja, que a matriz de distribuição por hiperligações H seja irredutível. Obteremos a resposta na próxima secção.

1.2 Teorema de Perron-Frobenius

Oskar Perron (7 Maio 1880 – 22 Fevereiro 1975) foi um matemático alemão cujo trabalho contribuiu para diversas áreas da Matemática, como as Equações Diferenciais, as Equações Diferenciais Parciais, entre outras, sendo mais conhecido pelo seu paradoxo, o Paradoxo de Perron, que ilustra o perigo de se assumir que existe sempre solução para um problema:

Seja N o maior número natural. Se $N > 1$, então $N^2 > N$ e N não é o maior número natural. Logo, $N = 1$, o que é absurdo.

Perron foi também presidente da Sociedade Alemã de Matemática na década de 1940 e teve um papel importante no combate aos efeitos nefastos da política nazi da época, que tiveram repercussões claras ao nível académico e científico no seu país. Mais informações sobre a história de Oskar Perron e as suas contribuições para a Matemática podem ser encontradas em [9].

Em 1907, Perron formulou o teorema que dá o mote à Teoria de Perron-Frobenius e que é apresentado de seguida.

Teorema 1.2.1 (Teorema de Perron)**.** *Seja M uma matriz quadrada positiva. São verdadeiras as afirmações seguintes:*

(a) o raio espectral de M, $\rho(M)$, é um valor próprio de M algebricamente simples

(b) $\rho(M) > 0$

(c) se $\lambda \neq \rho(M)$ for valor próprio de M, então $|\lambda| < \rho(M)$

(d) existe um único vetor v tal que $Mv = \rho(M)v$ e v é um vetor estocástico; o vetor v é positivo

Demonstração. Cf. [6], Capítulo 8, Teorema 8.2.8. □

A exposição e demonstração de toda a Teoria de Perron-Frobenius pode ser encontrada no capítulo 8 do livro [6]. Neste capítulo, também é provado que o raio espectral de uma matriz estocástica é 1 (ver subcapítulo 8.7 de [6]), tal como já foi mencionado anteriormente na apresentação do Teorema 1.1.1 e da Proposição 1.1.1, sendo que é a Proposição 1.1.1 que contem este resultado. Assim, aplicando o Teorema de Perron, se M for uma matriz estocástica positiva, temos que 1 é valor próprio de M com um único vetor próprio estocástico associado, v, e v é positivo.

A aplicação do Teorema de Perron não pode ser feita nas matrizes de distribuição por hiperligações de forma geral, dado não existir a garantia destas matrizes se apresentarem nas condições do teorema, pois poderão ter entradas iguais a 0. Uma generalização ao teorema surge, para matrizes não-negativas, alguns anos mais tarde.

Georg Frobenius (26 Outubro 1849 - 3 Agosto 1917) foi um matemático alemão com contribuições importantes em diversas áreas da Matemática, tais como a Teoria dos Números, a Teoria dos Grupos, entre outras. Foi aluno, na Universidade de Berlim, de Kronecker e de Kummer e, em 1870, concluiu o seu doutoramento sob a supervisão de Weierstrass. Em 1891, com forte indicação de Weierstrass, que considerava Frobenius como um dos seus discípulos mais brilhantes, tornou-se professor na Universidade de Berlim, onde tinha como principal objectivo continuar o legado da investigação e ensino que experienciou enquanto estudante e contribuir para o conhecimento da Matemática Pura. Uma das suas contribuições mais conhecidas foi a criação do conceito de caracter de um grupo e as suas ideias à volta deste conceito, que geraram a Teoria dos Caracteres de Grupos e que contribuíram imenso para a Teoria da Representação de Grupos. Mais dados sobre a vida de Georg Frobenius e as suas contribuições para a Matemática podem ser encontrados em [10].

Entre 1908 e 1912, Frobenius formulou a generalização para o Teorema de Perron.

Teorema 1.2.2 (Teorema de Perron-Frobenius)**.** *Seja M uma matriz quadrada não-negativa e irredutível. São verdadeiras as afirmações seguintes:*

(a) $\rho(M)$ *é um valor próprio de* M *algebricamente simples*

(b) $\rho(M) > 0$

(c) existe um único vetor v *tal que* $Mv = \rho(M)v$ *e* v *é um vetor estocástico; o vetor* v *é positivo*

Demonstração. Cf. [6], Capítulo 8, Teorema 8.4.4. □

Se $M > 0$, então trivialmente $M \geq 0$ e M é irredutível, portanto verifica-se que o Teorema de Frobenius é, de facto, uma generalização do Teorema de Perron e é a este teorema que se dá o nome de Teorema de Perron-Frobenius. O valor próprio $\rho(M)$ e o vetor próprio v, caracterizados no teorema, são frequentemente designados por raíz de Perron de M e vetor de Perron de M, respectivamente.

Tendo em conta os resultados de Perron e Frobenius, na presença de uma matriz de distribuição por hiperligações estocástica que seja irredutível ou até mesmo positiva, podemos identificar o vetor de PageRank com o único, positivo e estocástico vetor de Perron da matriz. Assim, respondendo à pergunta deixada no final da secção anterior, um grafo fortemente conexo não apresentará os problemas exemplificados anteriormente.

O Teorema de Perron-Frobenius é apresentado, de formas distintas, em [3] e em [4], sendo que neste último encontramos ambos os teoremas, o de Perron e o de Perron-Frobenius.

Na secção que se segue, veremos a solução apresentada por Brin e Page para que nos encontremos nas condições de aplicar o Teorema de Perron-Frobenius.

1.3 A solução de Brin e Page

Brin e Page encontram a solução para a eficácia do seu método em redes que apresentam componentes desconexas ou falta de conexão forte interpretando a experiência de um utilizador da Web como algo que

não se limita à existência de ligações entre as páginas.

Ao navegarmos na Internet, podemos escolher a próxima página a visitar seguindo uma hiperligação existente na página em que estamos ou simplesmente voltando a considerar todas as páginas da rede e escolhendo uma delas. Esta segunda maneira de escolher a próxima página da Internet a visitar é o que distingue o PageRank dos outros métodos de busca até então formulados. Esta alternativa, geralmente designada como opção de teletransporte, é a solução para o problema que surge quando procuramos o vetor de Pagerank considerando simplesmente a matriz de distribuição por hiperligações H.

A opção de teletransporte traduz-se numa nova matriz T e a consideração desta matriz em conjunto com a matriz H faz da nossa rede uma rede fortemente conexa, pois partindo de qualquer página há sempre possibilidade de visitar no momento seguinte qualquer página da rede. A matriz T geralmente é construída considerando uma distribuição uniforme, sobre as páginas, da probabilidade de uma página ser escolhida na opção de teletransporte. Ou seja, numa rede de n páginas,

$$T = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix},$$

é geralmente a matriz escolhida para representar a possibilidade de teletransporte na rede. No entanto, a construção da matriz T poderá ser feita considerando uma distribuição mais personalizada, dadas a impossibilidade de calcular com precisão a probabilidade de o utilizador se teletransportar de uma página para outra e a vantagem que se ganha em adaptar o método de cálculo de rank a casos mais particulares.

A personalização da matriz T recairá na escolha do vetor estocástico mencionado na definição assim construída:

Definição 1.3.1. *Numa rede de n páginas, onde H é a sua matriz de distribuição por hiperligações, definimos a matriz de distribuição por teletransporte,*

$$T = te^T,$$

onde e é o vetor coluna de dimensão n cujas entradas são todas iguais a 1 e t é um vetor estocástico não-negativo de dimensão n tal que $H + te^T > 0$.

Se considerarmos que a primeira opção de seguir somente hiperligações tem uma certa probabilidade $0 < \alpha < 1$ de ser tomada, a opção de teletransporte terá probabilidade $1 - \alpha$. Mais uma vez, dadas a impossibilidade do cálculo de α e a vantagem da personalização, este torna-se no segundo fator personalizável, sendo que inicialmente, para o Google, $\alpha = 0,85$ [2].

As escolhas do vetor t e do valor α, que designaremos por vetor e parâmetro de teletransporte, respectivamente, têm bastante importância no resultado final do cálculo do vetor de rank, sendo que contextos diferentes levarão a escolhas diferentes. Estes dois fatores são as peças necessárias para a construção da matriz que representa a solução procurada por Brin e Page, cuja definição é a seguinte.

Definição 1.3.2. *Numa rede de n páginas, com H e $T = te^T$ as matrizes de distribuição por hiperligações e por teletransporte, respectivamente, onde t é o vetor de teletransporte e $0 < \alpha < 1$ o parâmetro de teletransporte, definimos a matriz Google,*

$$G = \alpha H + (1-\alpha)T = \alpha H + (1-\alpha)te^T.$$

G é uma matriz positiva, pois $H + te^T > 0$, e, sendo que T é sempre estocástica, G é estocástica quando H o é. Isto significa que, mesmo que existam componentes desconexas ou a conexão da rede seja fraca, G está nas condições do Teorema de Perron-Frobenius (e até mesmo no de Perron). Temos $\rho(G) = 1$, quando G é estocástica. Logo, neste caso, pela alínea c) do Teorema 1.2.2, o vetor estocástico v que soluciona $Gv = v$ existe, é único e é um vetor positivo.

Retomando o exemplo da rede desconexa apresentado anteriormente, escolhendo $\alpha = 0,85$ e t igual ao vetor coluna com todas as entradas iguais a $\frac{1}{6}$, obtemos

$$G = 0,85 \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \end{bmatrix} + 0,15 \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix} = \begin{bmatrix} \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{7}{8} & \frac{1}{40} & \frac{1}{40} \\ \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{7}{8} & \frac{1}{40} \\ \frac{1}{40} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{7}{8} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{1}{40} & \frac{9}{20} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} & \frac{7}{8} \\ \frac{1}{40} & \frac{1}{40} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \end{bmatrix},$$

uma matriz positiva e estocástica com vetor de Perron igual a

$$v \approx \begin{bmatrix} 0,17 & 0,22 & 0,12 & 0,17 & 0,24 & 0,08 \end{bmatrix}^T,$$

que tomamos como o vetor de PageRank deste caso e que nos dá a ordenação das 6 páginas por relevância, sendo que as páginas 1 e 4 estão empatadas, tal como seria esperado.

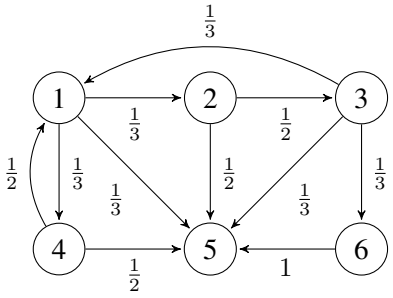
1.4 Outras preocupações

Na secção anterior, afirmo que a matriz Google G é estocástica quando a matriz de distribuição por hiperligações H o é, precisamente por H nem sempre ter esta propriedade.

Nos exemplos apresentados anteriormente, as respectivas matrizes H são estocásticas, visto cada página ter referência a pelo menos uma outra página. Porém isso nem sempre acontece, dado existirem páginas na Internet que não têm nenhuma hiperligação (páginas que não estão completamente descarregadas, por exemplo [1]). Estas páginas não distribuem a sua importância para nenhuma outra, formam nodos na rede que se designam por nodos pendentes (do inglês *dangling* [1]) e fazem com que o cálculo do vetor de PageRank precise de especial atenção.

Definição 1.4.1. *Num grafo orientado, designamos por nodo pendente qualquer nodo do grafo que não tenha arestas dirigidas a nenhum nodo.*

Se considerarmos a rede do primeiro exemplo excluindo a única hiperligação da página 5, esta nova rede é representada pelo grafo seguinte:



O nodo 5, pela definição anterior, é um nodo pendente e a matriz de distribuição por hiperligações H para este exemplo é

$$H = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix}.$$

A matriz H não é estocástica, pois todas as entradas da quinta coluna são nulas, e dado que qualquer matriz Google G construída a partir de H é da forma $G = \alpha H + (1-\alpha)T$ com T estocástica e $0 < \alpha < 1$, temos que as entradas da quinta coluna de G somam apenas $1 - \alpha < 1$. Logo, G também não será estocástica.

Tomando valores concretos, escolhendo $\alpha = 0,85$ e t igual ao vetor coluna com todas as entradas iguais a $\frac{1}{6}$, obtemos

$$G = \begin{bmatrix} \frac{1}{40} & \frac{1}{40} & \frac{37}{120} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} \\ \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{1}{40} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{37}{120} & \frac{9}{20} & \frac{37}{120} & \frac{9}{20} & \frac{1}{40} & \frac{7}{8} \\ \frac{1}{40} & \frac{1}{40} & \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \end{bmatrix},$$

e $\rho(G) = 0,643 < 1$, portanto, a equação $Gv = v$ não apresenta solução para além da solução nula, o que torna impossível a obtenção do vetor de PageRank.

Como vimos anteriormente, como G é sempre positiva, pelo Teorema de Perron-Frobenius, $Gv = \rho(G)v$ tem sempre solução, esta é única e positiva. Se G for estocástica, $\rho(G) = 1$ e a solução estocástica de $Gv = \rho(G)v$ é o vetor que procuramos. No caso contrário, em que G não é estocástica, não podemos afirmar o mesmo.

Este caso problemático, originado pelos nodos pendentes, é mencionado por Brin e Page em [1], sendo apresentada também (em poucas palavras) uma forma de tratar a sua presença na rede. Uma interpretação desta forma de lidar com os nodos pendentes será exposta no Capítulo 3.

Assim, a solução de Brin e Page para a desconexão ou a falta de conexão forte da rede através da apresentação do conceito de teletransporte e da construção da matriz Google apenas resulta como pretendido na ausência de nodos pendentes.

Este facto poderá ser difícil de compreender na leitura do artigo [3], que, apesar de apresentar um caso com nodos pendentes, afirma que a matriz Google de uma rede é sempre estocástica, não relacionando a presença destes nodos problemáticos na rede com uma matriz Google resultante não-estocástica e apresentando, consequentemente, a solução de Brin e Page como uma solução universal.

Na exposição feita em [4] é mencionado que a presença de nodos pendentes resulta numa matriz de distribuição por hiperligações não-estocástica. No entanto, a construção da matriz Google é também apresentada como solução universal, dado que o artigo se centra na propriedade da não-negatividade, não exigindo que a matriz Google nem o vetor de PageRank sejam estocásticos.

Em contraste, na exposição [8], a menção feita aos nodos pendentes e ao problema por eles gerado é acompanhada por uma forma alternativa de lidar com estes nodos.

Uma boa prática na leitura de exposições que não mencionam os nodos pendentes ou o problema gerado pela sua presença na rede poderá ser assumir que a rede não tem nodos pendentes. No entanto, é importante tratar com rigor o caso contrário. Em paralelo, existem exposições que mencionam e procuram tratar a rede com nodos pendentes. A análise de algumas dessas soluções será feita no Capítulo 3.

Agora, assumindo que estamos a trabalhar com uma rede sem nodos pendentes, resta-nos uma preocupação: o tempo de cálculo do vetor de PageRank. Calcular a solução de $Gx = x$, resolvendo o sistema de equações ou através do polinómio característico, por exemplo, será um processo praticamente interminável dada a dimensão da matriz positiva G (no momento de escrita desta dissertação, estima-se existirem 5,38 milhares de milhões de páginas indexadas na Web [11]).

Os pontos de vista dos Sistemas Dinâmicos e Probabilístico, que analisamos na primeira secção, apresentam uma alternativa de cálculo que se designa por método da potência. Veremos como este método se distingue no cálculo do vetor de PageRank por se traduzir num algoritmo passível de ser lido e computado eficazmente por um computador e cuja velocidade de aproximação da solução desejada está dentro do praticável.

Uma exposição mais detalhada sobre este método e algoritmo gerado, a sua consequente aplicação à matriz Google e a demonstração da eficácia do mesmo será feita ao longo do próximo capítulo.

2 Cálculo do vetor de PageRank

Um bom método de ordenação de páginas por relevância tem de ter em conta dois fatores: tem de ser constituído por uma fórmula eficaz de atribuição de relevância que garanta a existência e unicidade do ranking das páginas e por uma forma que garanta que o cálculo deste seja feito em tempo útil.

No capítulo anterior, verificámos que o primeiro fator é satisfeito, assumindo que a rede não tem nodos pendentes. Neste capítulo, verificaremos o segundo fator, explorando o método da potência, que será analisado de seguida.

A prova de que o método da potência pode ser aplicado para o cálculo do vetor de PageRank será apresentada de duas formas: primeiramente fazendo uma exposição baseada nos trabalhos feitos por K. Bryan e T. Leise em [12] e por A. S. Camargo e A. P. Galves em [13], e depois expondo o ponto de vista alternativo apresentado por D. F. Gleich em [14]. Ambas as formas têm como requisito a matriz Google ser estocástica, portanto assumiremos, ao longo de todo este capítulo, a não existência de nodos pendentes.

2.1 Método da Potência

Richard von Mises (19 Abril 1883 - 14 Julho 1953), não confundir com o seu conhecido irmão economista Ludwig von Mises, foi um matemático austríaco cujo trabalho contribuiu para o enriquecimento de diversas áreas da Matemática e da Matemática Aplicada, como a Teoria da Probabilidade, a Mecânica dos Fluidos, a Aerodinâmica, entre outras. Para além de matemático, também era piloto qualificado, acabando por leccionar em 1913 o primeiro curso universitário sobre Voos Motorizados. Esta qualificação fez também com que participasse na Primeira Guerra Mundial junto do exército Austro-Húngaro enquanto piloto e projetista de aeronaves. Foi professor de Collatz em Berlim, antes das políticas nazis dominarem a Alemanha. Após isso, decidiu retirar-se do país e prosseguir como professor em Istambul, primeiramente, e em Harvard, mais tarde. Mais informações sobre a história e trabalho de von Mises podem ser encontradas em [15].

Uma das suas contribuições mais conhecidas é o método da potência, conhecido também como iteração de von Mises. A apresentação do teorema da convergência do método requer a definição seguinte. Mais sobre o método da potência pode ser encontrado no livro [16].

Definição 2.1.1. *Sejam $\lambda_1, ..., \lambda_n$ os valores próprios de uma matriz quadrada M de dimensão n. Dizemos que λ_1 é o valor próprio dominante de M se $|\lambda_1| > |\lambda_i|, i = 2, ..., n$. Aos vetores próprios de M associados ao seu valor próprio dominante designamos de vetores próprios dominantes de M.*

Teorema 2.1.1. *Seja M uma matriz diagonalizável sobre $\mathbb{R}$ e com valor próprio dominante. Então existe um vetor w não nulo tal que a sequência de vetores seguinte, onde $||.||$ representa uma norma, converge para um vetor próprio dominante de M de norma igual a 1:*

(a) $w, \frac{Mw}{||Mw||}, ..., \frac{M^k w}{||M^k w||}$, *..., caso o valor próprio dominante seja positivo*

(b) $w, \frac{(-1)Mw}{||Mw||}, ..., \frac{(-1)^k M^k w}{||M^k w||}$, *..., caso contrário*

Demonstração. Como M é diagonalizável, M tem n vetores próprios linearmente independentes x_1, ..., x_n associados aos valores próprios λ_1, ..., λ_n. Como M possui valor próprio dominante, suponhamos que os valores próprios estão ordenados de forma a que λ_1 seja o valor próprio dominante de M. Como x_1, ..., x_n são linearmente independentes, estes vetores formam uma base de $\mathbb{R}^n$ e qualquer w pode ser escrito da forma $w = c_1x_1 + c_2x_2 + ... + c_nx_n$. Escolhamos w de forma a que o seu coeficiente c_1 seja diferente de 0. Assim,

$$\begin{aligned} Mw &= c_1 (Mx_1) + c_2 (Mx_2) + ... + c_n (Mx_n) \\ &= c_1 (\lambda_1 x_1) + c_2 (\lambda_2 x_2) + ... + c_n (\lambda_n x_n) \end{aligned}$$

e, para qualquer número natural k,

$$\begin{aligned} M^k w &= c_1 \left(\lambda_1^k x_1\right) + c_2 \left(\lambda_2^k x_2\right) + ... + c_n \left(\lambda_n^k x_n\right) \\ &= c_1 \lambda_1^k \left[x_1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k x_2 ... + \frac{c_n}{c_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k x_n \right]. \end{aligned}$$

Por sua vez,

$$\begin{aligned} \left|\left|M^k w\right|\right| &= \left|\left| c_1 \lambda_1^k \left[x_1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k x_2 ... + \frac{c_n}{c_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k x_n \right] \right|\right| \\ &= \left| c_1 \lambda_1^k \right| \left|\left| x_1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k x_2 ... + \frac{c_n}{c_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k x_n \right|\right|. \end{aligned}$$

Como λ_1 por definição é maior em valor absoluto que qualquer outro valor próprio, temos que cada uma das frações $\frac{\lambda_2}{\lambda_1}, ..., \frac{\lambda_n}{\lambda_1}$ é menor que 1 em valor absoluto. Portanto, cada fração $\left(\frac{\lambda_2}{\lambda_1}\right)^k, ..., \left(\frac{\lambda_n}{\lambda_1}\right)^k$ converge para 0 quando k tende para infinito. Posto isto e tomando qualquer um dos casos $\lambda_1 > 0$ ou $\lambda_1 < 0$, temos que a sequência correspondente tende para $s\frac{x_1}{||x_1||}$, um múltiplo do vetor próprio dominante x_1, onde $s = 1$ no caso de $c_1 > 0$ ou $s = -1$ no caso contrário. □

Notemos que o enunciado do teorema apresenta duas alternativas de sequências de vetores e que ambas utilizam a divisão pela norma. Isto é feito de modo a que a convergência esteja garantida no caso do valor próprio dominante ser um valor negativo e no caso do seu valor em absoluto ser maior que 1.

No caso particular em que o valor próprio dominante é positivo e igual a 1, basta trabalhar com a

sequência $w, Mw, ..., M^k w, ...$ e a convergência é garantida. Esta sequência é a que mais frequentemente é chamada de método da potência ou iteração de von Mises na literatura.

A matriz Google, apesar de apresentar um valor próprio dominante, sendo que este é igual a 1, não é necessariamente uma matriz diagonalizável. Logo, não satisfaz as condições do teorema.

Este facto é importante de notar, mas nem sempre o encontramos explicitado na literatura do PageRank. Em particular, a exposição [3] referenciada no capítulo anterior não menciona a necessidade da matriz ser diagonalizável e com valor próprio dominante para que o método da potência funcione, nem demonstra como este método pode ser aplicado à matriz Google, apesar de apresentar esta aplicação como forma mais eficaz de cálculo do vetor de PageRank. Já [4] e [8] mencionam as condições necessárias para a aplicação do método da potência, mas não indicam que a matriz Google poderá não as satisfazer. Estas duas últimas referências apresentam, no entanto, boas exposições de como o método da potência funciona e mostram que a velocidade de convergência da iteração depende do segundo maior valor próprio em módulo da matriz ($|\lambda_2|$): quanto mais próximo este estiver de 0, mais rapidamente a iteração convergirá.

Em [13], são explicitadas de forma clara as condições necessárias para o funcionamento do método da potência e também o facto da matriz Google não ser necessariamente diagonalizável. É também apresentada uma demonstração da possibilidade de aplicação da iteração de von Mises no cálculo do vetor de PageRank que não possui a necessidade da matriz Google satisfazer as condições usuais do método. Esta apresentação é uma reformulação do trabalho originalmente feito em [12].

O artigo [12] é uma das referências mais encontradas nas bibliografias dos trabalhos que compõem a literatura do PageRank. Ao lê-lo ficamos, não só com uma boa ideia geral do que é o PageRank, mas também com bastante noção dos problemas trazidos pela falta de conexão forte da rede ou pela presença de nodos pendentes. Na abordagem que faz ao cálculo do vetor de PageRank, apesar de não fazer menção à condição de matriz diagonalizável na apresentação do que é o método da potência, apresenta uma prova essencial de como é possível de facto aplicar esta iteração à matriz Google.

Apresentaremos, então, de seguida, esta prova de que é possível utilizar a iteração de von Mises para obter o vetor de PageRank, ou seja, de que é possível encontrar a solução procurada em $\lim_{k\to\infty} G^k v_0$, para um certo vetor inicial v_0.

No teorema do método da potência qualquer norma pode ser utilizada, daí a sua apresentação com uma norma genérica. A partir daqui iremos restringir o nosso foco à norma-1, que é a norma mais adequada para utilizar neste contexto, dado estarmos a trabalhar com matrizes e vetores estocásticos. Podemos relembrar a definição da norma-1 pelo seguinte.

Definição 2.1.2. *Seja v um vetor de dimensão n. Designamos por norma-1 de v, representado por $||v||_1$, o valor*

$$||v||_1 = \sum_{i=1}^{n} |v^{(i)}|.$$

Utilizando a norma-1, ter $\lim_{k\to\infty} G^k v_0 = v$ é o mesmo que ter $\lim_{k\to\infty} ||v - G^k v_0||_1 = 0$. Sendo v o vetor de PageRank, precisamos de encontrar um vetor inicial v_0 a partir do qual consigamos, aplicando

a iteração de von Mises, v_0, $v_1 = Gv_0$, $v_2 = Gv_1$, ..., $v_{k+1} = Gv_k$, ..., uma aproximação progressiva a v.

Tenhamos em conta os seguintes resultados.

Proposição 2.1.1. *Seja M uma matriz estocástica de dimensão $n \times n$ e seja v um vetor de dimensão n tal que a soma das suas entradas é nula. Então o vetor Mv é tal que a soma das suas entradas é nula.*

Demonstração. Denotemos as entradas do vetor Mv por $w^{(i)}$ para cada $i \in \{1, ..., n\}$. Queremos provar que

$$\sum_{i=1}^{n} w^{(i)} = 0.$$

Dado que

$$w^{(i)} = \sum_{j=1}^{n} \left(m^{(ij)} v^{(j)} \right)$$

para cada $i \in \{1, ..., n\}$ e que

$$\sum_{i=1}^{n} m^{(ij)} = 1$$

para cada $j \in \{1, ..., n\}$, temos

$$\begin{aligned} \sum_{i=1}^{n} w^{(i)} &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(m^{(ij)} v^{(j)} \right) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} \left(m^{(ij)} v^{(j)} \right) \\ &= \sum_{j=1}^{n} \left(v^{(j)} \sum_{i=1}^{n} m^{(ij)} \right) = \sum_{j=1}^{n} v^{(j)} = 0. \end{aligned}$$

□

Proposição 2.1.2. *Seja M uma matriz estocástica de dimensão $n \times n$ e seja v um vetor de dimensão n tal que a soma das suas entradas é nula. Então $||Mv||_1 \leq c||v||_1$, com $c = \max_{1 \leq j \leq n} |1 - 2 \min_{1 \leq i \leq n} m^{(ij)}| < 1$.*

Demonstração. Continuando com a notação introduzida na proposição anterior, temos

$$||Mv||_1 = \sum_{i=1}^{n} \left| w^{(i)} \right|.$$

Se v for o vetor nulo, então $||Mv||_1 = c\,||v||_1 = 0$, para qualquer constante c. Suponhamos agora que v não é o vetor nulo. Sem perda de generalidade, consideremos que as s primeiras entradas de Mv são positivas ou nulas e que as restantes $n - s$ são negativas. Como, pela proposição anterior, a soma das entradas de Mv é nula, existem entradas com sinais opostos, pelo que $s \neq n$. Assim,

$$\begin{aligned}
\sum_{i=1}^{n} \left| w^{(i)} \right| &= \sum_{i=1}^{s} w^{(i)} - \sum_{i=s+1}^{n} w^{(i)} \\
&= \sum_{i=1}^{s} \sum_{j=1}^{n} m^{(ij)} v^{(j)} - \sum_{i=s+1}^{n} \sum_{j=1}^{n} m^{(ij)} v^{(j)} \\
&= \sum_{j=1}^{n} \left(v^{(j)} \sum_{i=1}^{s} m^{(ij)} \right) - \sum_{j=1}^{n} \left(v^{(j)} \sum_{i=s+1}^{n} m^{(ij)} \right) \\
&= \sum_{j=1}^{n} \left(v^{(j)} \left(\sum_{i=1}^{s} m^{(ij)} - \sum_{i=s+1}^{n} m^{(ij)} \right) \right).
\end{aligned}$$

Definamos a grandeza

$$q^{(j)} = \sum_{i=1}^{s} m^{(ij)} - \sum_{i=s+1}^{n} m^{(ij)}.$$

É fácil perceber que $-1 < q^{(j)} < 1$ para cada $j \in \{1, ..., n\}$. Estudando os valores extremos de $q^{(j)}$, temos que o maior valor de $q^{(j)}$ é encontrado quando $s = n - 1$ e o menor valor é encontrado quando $s = 1$. No primeiro caso,

$$q^{(j)} = \left(1 - \min_{1 \leq i \leq n} m^{(ij)} \right) - \min_{1 \leq i \leq n} m^{(ij)} = 1 - 2 \min_{1 \leq i \leq n} m^{(ij)}.$$

No segundo caso,

$$q^{(j)} = \min_{1 \leq i \leq n} m^{(ij)} - \left(1 - \min_{1 \leq i \leq n} m^{(ij)} \right) = -1 + 2 \min_{1 \leq i \leq n} m^{(ij)}.$$

Assim,

$$-1 < -1 + 2 \min_{1 \leq i \leq n} m^{(ij)} \leq q^{(j)} \leq 1 - 2 \min_{1 \leq i \leq n} m^{(ij)} < 1 \Leftrightarrow \left| q^{(j)} \right| \leq \left| 1 - 2 \min_{1 \leq i \leq n} m^{(ij)} \right|.$$

Temos, portanto,

$$\sum_{i=1}^{n} \left| w^{(i)} \right| = \sum_{j=1}^{n} v^{(j)} q^{(j)} \leq \sum_{j=1}^{n} \left| v^{(j)} q^{(j)} \right| = \sum_{j=1}^{n} \left| v^{(j)} \right| \left| q^{(j)} \right|.$$

Considerando

$$c = \max_{1 \leq j \leq n} q^{(j)} = \max_{1 \leq j \leq n} \left| 1 - 2 \min_{1 \leq i \leq n} m^{(ij)} \right|,$$

temos

$$\sum_{j=1}^{n} \left|v^{(j)}\right| \left|q^{(j)}\right| \leq \sum_{j=1}^{n} \left|v^{(j)}\right| c = c \sum_{j=1}^{n} \left|v^{(j)}\right| = c \left|\left|v\right|\right|_1 .$$

□

Proposição 2.1.3. *Seja M uma matriz estocástica de dimensão $n \times n$ e seja v um vetor de dimensão n tal que a soma das suas entradas é nula. Então $||M^k v||_1 \leq c^k ||v||_1$, para todo o $k \in \mathbb{N}$, com $c = \max_{1 \leq j \leq n} |1 - 2 \min_{1 \leq i \leq n} m^{(ij)}| < 1$.*

Demonstração. Primeiramente, precisamos de mostrar que $M^{k-1}v$ é um vetor tal que a soma das suas entradas é nula. Tomemos $d = k - 1$. Por indução em d, para $d = 1$ é verdade pela Proposição 2.1.1. Suponhamos que é verdade para $d - 1$. Como

$$M^d v = M\left(M^{d-1}v\right),$$

pela Proposição 2.1.1 e pela hipótese de indução, $M^d v$ é um vetor tal que a soma das suas entradas é nula. Provemos agora, por indução em k, que

$$\left|\left|M^k v\right|\right|_1 \leq c^k \left|\left|v\right|\right|_1 .$$

Para $k = 1$ é verdade pela Proposição 2.1.2. Suponhamos que é verdade para $k - 1$. Assim,

$$\left|\left|M^k v\right|\right|_1 = \left|\left|M\left(M^{k-1}v\right)\right|\right|_1 \leq c \left|\left|M^{k-1}v\right|\right|_1 \leq c\left(c^{k-1}\left|\left|v\right|\right|_1\right) = c^k \left|\left|v\right|\right|_1 .$$

□

Relembremos que ao longo deste capítulo assumimos a não existência dos nodos pendentes na rede, isto é, que a matriz Google é estocástica. Assim, a matriz Google encontra-se nas condições das proposições anteriores.

Na secção 1.1 do capítulo anterior, na abordagem aos pontos de vista dos Sistemas Dinâmicos e Probabilístico, considerou-se para vetor inicial da iteração, v_0, o vetor que reflete a igual relevância das páginas da rede no estado inicial ou, por outras palavras, que reflete a igual probabilidade das páginas serem visitadas no momento inicial. Isto é, sendo n o número de nodos da rede, o v_0 considerado foi o vetor de dimensão n cujas entradas são todas iguais a $\frac{1}{n}$. Este vetor é estocástico e positivo.

O teorema seguinte mostra que qualquer vetor estocástico e positivo poderá ser escolhido como vetor inicial da iteração de forma a obtermos como limite o vetor de PageRank pretendido.

Teorema 2.1.2. *Seja G uma matriz Google de dimensão n e seja v o seu vetor de PageRank. Seja v_0 um vetor de dimensão n, estocástico e positivo. Então:*

$$v = \lim_{k\to\infty} G^k v_0.$$

Demonstração. Sendo que v e v_0 são dois vetores estocásticos e positivos, temos que

$$||v||_1 = ||v_0||_1 = 1.$$

Sendo $u = v - v_0$, temos que o vetor u satisfaz

$$\sum_{i=1}^{n} u^{(i)} = 0.$$

Pela Proposição 2.1.3,

$$0 \leq \left|\left|G^k u\right|\right|_1 \leq c^k \left|\left|u\right|\right|_1.$$

Tomando o limite em k, temos $\lim_{k\to\infty} c^k = 0$, pois $0 < c < 1$, e $\lim_{k\to\infty} 0 = 0$. Portanto, pelo Teorema das Sucessões Enquadradas,

$$\lim_{k\to\infty} \left|\left|G^k u\right|\right|_1 = 0.$$

No entanto,

$$\begin{aligned}
\lim_{k\to\infty} \left|\left|G^k u\right|\right|_1 = 0 &\Leftrightarrow \lim_{k\to\infty} \left|\left|G^k (v - v_0)\right|\right|_1 = 0 \\
&\Leftrightarrow \lim_{k\to\infty} G^k (v - v_0) = 0 \\
&\Leftrightarrow \lim_{k\to\infty} G^k v - G^k v_0 = 0 \\
&\Leftrightarrow \lim_{k\to\infty} v - G^k v_0 = 0 \\
&\Leftrightarrow \lim_{k\to\infty} v - \lim_{k\to\infty} G^k v_0 = 0 \\
&\Leftrightarrow v - \lim_{k\to\infty} G^k v_0 = 0 \Leftrightarrow v = \lim_{k\to\infty} G^k v_0.
\end{aligned}$$

□

Concluímos assim que a iteração de von Mises pode ser usada na matriz Google G para encontrar o vetor de PageRank.

No entanto, é importante notar que a matriz G é uma matriz positiva com dimensão n igual ao número de páginas indexadas na Web. O que implica que cada iteração com G envolva $O(n^2)$ multiplicações e somas, sendo que n se estima ser superior a 5,38 milhares de milhões [11]. Felizmente, é possível reduzir muito substancialmente o número de cálculos em cada iteração.

Tendo por hipótese a matriz Google $G = \alpha H + (1-\alpha)T$, com H e $T = te^T$ as matrizes de distribuição

por hiperligações e por teletransporte, respectivamente, t o vetor de teletransporte e α o parâmetro, e v um vetor estocástico de dimensão n, a iteração Gv equivale a $\alpha Hv + (1-\alpha)t$, dado que, por v ser estocástico e pelo facto do produto entre uma matriz estocástica e um vetor estocástico resultar num vetor estocástico, $Tv = t$. Esta equivalência simplifica significativamente os cálculos em cada iteração, visto que as matrizes de distribuição por hiperligações são matrizes esparsas, tendo em conta que cada página possui em média menos de 15 hiperligações [17]. Um algoritmo mais fácil de computar pode, assim, ser feito com base na iteração centrada na matriz H.

Outro fator importante num algoritmo é a apresentação de uma condição de paragem. Na condição clássica, deverá escolher-se uma tolerância ao erro, ε, e comparar-se a distância entre o vetor resultante de cada iteração e o da sua anterior. Quando a distância for menor do que ε, a iteração acaba. Esta condição permite que o rank das páginas da rede seja calculado por um computador num número finito de iterações.

Posto isto, é construído o algoritmo esperado para a obtenção de uma aproximação suficientemente satisfatória do vetor de PageRank da rede, que indicará o desejado ranking de todas as páginas.

Definição 2.1.3. *Seja H uma matriz de distribuição por hiperligações de uma rede de n páginas. Sejam t estocástico de dimensão n o vetor de teletransporte e $0 < \alpha < 1$ o parâmetro de teletransporte. Seja ε>0 a tolerância ao erro. Seja v_0 de dimensão n o vetor inicial. Definimos um método de iteração para o cálculo do ranking das páginas da rede pelo seguinte algoritmo:*

Algoritmo 1: Algoritmo de aproximação ao PageRank

$v_1 \leftarrow \alpha H v_0 + (1-\alpha)t$
$\delta \leftarrow ||v_1 - v_0||_1$
$k \leftarrow 1$
while $\delta > \varepsilon$ ***do***
  $v_{k+1} \leftarrow \alpha H v_k + (1-\alpha)t$
  $\delta \leftarrow ||v_{k+1} - v_k||_1$
  $k \leftarrow k+1$
end while

O teorema anterior demonstra que o algoritmo converge para o vetor de PageRank sempre que o vetor inicial v_0 escolhido for estocástico e positivo. Na próxima secção, demonstraremos o mesmo para dois vetores iniciais específicos.

Medindo a distância entre vetores através de uma norma, dizemos que dois vetores estão próximos quando existe uma diferença pequena entre eles entrada a entrada. Isto significa que o uso da condição de paragem presente no algoritmo, para além de limitar o número de iterações feitas pelo computador e, por sua vez, o tempo necessário para o cálculo do ranking, não desvirtuará gravemente o resultado obtido em comparação com o do vetor de PageRank, sempre que é escolhido um valor de ε apropriado.

Em 1998, na apresentação do método feita em [1], Brin e Page estimaram que a rede seria constituída por 75 milhões de páginas (sendo 24 milhões não pendentes) ligadas por 322 milhões de hiperligações,

que o cálculo de cada iteração demoraria 6 minutos, determinando o PageRank em 5 horas, e que apenas precisariam de 52 iterações no total. O controlo da quantidade de iterações e do tempo necessário para as fazer é importante também para que o cálculo do ranking seja feito com frequência de modo a que se mantenha o mais atualizado possível, dada a constante mutação da Web.

2.2 A alternativa de Gleich

No artigo [14], Gleich começa por transformar o problema de cálculo de vetores próprios associado à descoberta do vetor de PageRank num sistema linear.

Consideremos $G = \alpha H + (1-\alpha)te^T$ uma matriz Google de uma rede de n páginas, em que H é a matriz de distribuição por hiperligações, t é o vetor de teletransporte e α o parâmetro de teletransporte. Podemos transformar o problema habitual $Gx = x$ da seguinte forma, tendo em conta que $x > 0$ e $e^T x = 1$:

$$\begin{aligned} Gx = x &\Leftrightarrow \left(\alpha H + (1-\alpha)\, te^T\right) x = x \\ &\Leftrightarrow \alpha Hx + (1-\alpha)\, te^T x = x \\ &\Leftrightarrow (1-\alpha)\, te^T x = x - \alpha Hx \Leftrightarrow (1-\alpha)\, t = (I - \alpha H)\, x, \end{aligned}$$

onde I é a matriz identidade de ordem n. Ou seja, o nosso foco agora será resolver o sistema linear $(I - \alpha H)x = (1-\alpha)t$, ao qual chamaremos de sistema PageRank.

Tenhamos agora em conta a seguinte definição.

Definição 2.2.1. *Seja M uma matriz quadrada de dimensão n da forma $M = sI - N$, onde $N \geq 0$ e $s > \rho(N)$. Então M diz-se ser uma matriz-M.*

É fácil perceber que $I - \alpha H$ é uma matriz-M em que $s = 1$, $N = \alpha H \geq 0$ e $1 > \alpha = \rho(\alpha H)$. A transformação do problema no sistema linear $(I - \alpha H)x = (1-\alpha)t$, sabendo que $I - \alpha H$ é uma matriz-M, indica-nos que a solução x existe, é única e $x \geq 0$, se tivermos em conta o teorema seguinte.

Teorema 2.2.1. *Seja M uma matriz quadrada de dimensão n cujas entradas não pertencentes à diagonal principal são não-positivas, isto é, $m_{ij} \leq 0$ sempre que $i \neq j$, para $i, j \in \{1, ..., n\}$. São equivalentes:*

(a) M é uma matriz-M

(b) M é invertível e $M^{-1} \geq 0$

Demonstração. Cf. [18], Teorema 5'. □

Em 1958, K. Fan mostrou pela primeira vez a equivalência anterior [18], dando o mote à caracterização da classe das matrizes-M. Na verdade, o teorema 2.2.1 é um corolário do teorema 5' de Fan, que engloba

apenas a equivalência que necessitamos. O teorema original tem 8 alíneas e a sua demonstração recorre à apresentação e demonstração de seis outros resultados. Por esta razão, não se incluiu nesta exposição.

Duas abordagens diferentes na determinação da classe das matrizes-M são sugeridas por este teorema: por um lado, podemos pensar na classe das matrizes não-negativas invertíveis e determinar que inversas dessas matrizes são matrizes-M; por outro lado, podemos pensar na classe das matrizes com diagonal principal positiva e restantes entradas não-positivas e determinar quais destas matrizes têm inversa não-negativa.

É importante mencionar que o trabalho realizado por Perron e Frobenius no âmbito da teoria das matrizes não-negativas teve um papel essencial nos desenvolvimentos de outros matemáticos que formaram a teoria das matrizes-M. Assim, mesmo que vejamos o PageRank desta nova perspectiva trazida pelas matrizes-M, Perron-Frobenius acaba sempre por ocupar o seu lugar. Isto é mencionado por R. J. Plemmons [19], que também apresenta um conjunto de caracterizações das matrizes-M invertíveis num teorema com 40 condições. Anos mais tarde, Plemmons junta-se a A. Berman e juntos escrevem o livro [20], reunindo desta vez 50 condições equivalentes a uma matriz ser uma matriz-M invertível, sendo que a propriedade da não-singularidade com inversa não-negativa corresponde à 38ª.

Após apresentar o PageRank desta nova perspectiva, Gleich demonstra, de uma forma alternativa à apresentada na secção anterior, que o método da potência pode ser aplicado neste contexto, caracterizando no teorema seguinte os erros após k iterações do Algoritmo 1 e tomando dois vetores iniciais diferentes.

Para a demonstração do teorema, precisamos trabalhar com a norma-1 matricial. Tenhamos em conta a seguinte definição.

Definição 2.2.2. *Seja M uma matriz quadrada de dimensão n. Designamos por norma-1 de M, denotando por $||M||_1$, o valor*

$$||M||_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |m_{ij}| .$$

Esta definição é apresentada no Exemplo 5.6.4 do subcapítulo 5.6 do livro [6]. Neste exemplo, Horn e Johnson demonstram que a norma-1 assim definida corresponde à norma induzida pela norma-1 vetorial. Também demonstram, no Teorema 5.6.2 presente no mesmo subcapítulo, que qualquer norma induzida por uma norma vetorial satisfaz as condições de norma matricial. Logo, a definição anterior corresponde a uma norma matricial.

Pela definição, a norma-1 de uma matriz corresponde ao valor máximo entre as normas-1 dos vetores que compõem as colunas da matriz.

Teorema 2.2.2. *Seja H uma matriz de distribuição por hiperligações de uma rede de n páginas. Sejam t estocástico de dimensão n o vetor de teletransporte e $0 < \alpha < 1$ o parâmetro de teletransporte. Seja v o vetor de PageRank. Após k iterações do Algoritmo 1, são verdadeiras as afirmações:*

(a) se $v_0 = t$, então

$$||v - v_k||_1 \leq ||v - t||_1 \alpha^k \leq 2\alpha^k$$

(b) se $v_0 = 0$, então o vetor de erro $v - v_k \geq 0$, para qualquer k, e

$$||v - v_k||_1 = e^T (v - v_k) = \alpha^k$$

Demonstração. (a) Dado que v é solução do sistema $(I - \alpha H)\, x = (1 - \alpha)\, t$, temos que

$$v = \alpha H v + (1 - \alpha)\, t.$$

Pelo Algoritmo 1, temos

$$v_k = \alpha H v_{k-1} + (1 - \alpha)\, t.$$

Assim,

$$\begin{aligned} ||v - v_k||_1 &= ||\alpha H v + (1-\alpha)t - (\alpha H v_{k-1} + (1-\alpha)t)||_1 \\ &= ||\alpha H v - \alpha H v_{k-1}||_1 \\ &= ||\alpha H (v - v_{k-1})||_1 \\ &\leq \alpha ||H||_1 ||v - v_{k-1}||_1. \end{aligned}$$

Como $||H||_1 = 1$, obtemos a desigualdade

$$||v - v_k|| \leq \alpha \, ||v - v_{k-1}||_1 .$$

Da mesma forma obtemos as sucessivas desigualdades

$$||v - v_{k-1}||_1 \leq \alpha \, ||v - v_{k-2}||_1 , ..., ||v - v_1||_1 \leq \alpha \, ||v - v_0||_1 .$$

E, portanto,

$$||v - v_k||_1 \leq \alpha^k \, ||v - v_0||_1 .$$

Supondo que $v_0 = t$, dado que v e t são dois vetores estocásticos positivos, $||v - t||_1$ é no máximo igual a 2. Logo, provamos as desigualdades

$$||v - v_k||_1 \leq ||v - t||_1 \alpha^k \leq 2\alpha^k.$$

(b) Dado que $v = \alpha H v + (1-\alpha)\, t$, temos que v é um ponto fixo da função

$$f(x) = \alpha H x + (1-\alpha)\, t.$$

Pelo processo iterativo correspondente ao Teorema do Ponto Fixo, temos que, para cada k, $f^k(v) = v$, isto é,

$$\alpha^k H^k v + \alpha^{k-1} H^{k-1} (1-\alpha)\, t + \alpha^{k-2} H^{k-2} (1-\alpha)\, t + ... + (1-\alpha)\, t = v.$$

Por outro lado, supondo que $v_0 = 0$, pelo Algoritmo 1, obtemos

$$v_k = \alpha^{k-1} H^{k-1} (1-\alpha)\, t + \alpha^{k-2} H^{k-2} (1-\alpha)\, t + ... + (1-\alpha)\, t.$$

Assim, $v - v_k = \alpha^k H^k v \geq 0$, para cada k, e

$$||v - v_k||_1 = \left|\left|\alpha^k H^k v\right|\right|_1 = \alpha^k \left|\left|H^k v\right|\right|_1 = \alpha^k,$$

dado que $H^k v$ é sempre um vetor estocástico e é não-negativo.

□

3 Formas de lidar com os nodos pendentes

Relembrando o exemplo da rede com um nodo pendente apresentado na secção 1.4 do Capítulo 1,

$$G = \begin{bmatrix} \frac{1}{40} & \frac{1}{40} & \frac{37}{120} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} \\ \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{1}{40} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{37}{120} & \frac{9}{20} & \frac{37}{120} & \frac{9}{20} & \frac{1}{40} & \frac{7}{8} \\ \frac{1}{40} & \frac{1}{40} & \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \end{bmatrix},$$

aplicar a iteração de von Mises usual à matriz não estocástica G, isto é, calcular $v_0, Gv_0, G^2v_0, ...$ a partir de qualquer vetor inicial v_0, resulta numa aproximação ao vetor nulo, o que é coerente com o facto de $\rho(G) < 1$, tal como mencionado anteriormente.

Se, por outro lado, utilizarmos a iteração correspondente à alínea (a) do Teorema 2.1.1, obtemos uma aproximação ao vetor estocástico que soluciona $Gx = \rho(G)x$, que será o vetor

$$x \approx \begin{bmatrix} 0,161 & 0,110 & 0,111 & 0,110 & 0,420 & 0,088 \end{bmatrix}^T.$$

No primeiro caso, não obtemos qualquer tipo de ranking das páginas da rede. No segundo, obtemos um vetor estocástico e positivo, tal como queremos que o nosso vetor de ranking seja, mas que não é solução da equação $Gx = x$.

Em todo o caso, podemos novamente afirmar a impraticabilidade do cálculo de cada iteração com a matriz Google originada pela Web, por esta ser positiva e ter uma dimensão extraordinariamente grande, tal como foi observado no caso sem nodos pendentes.

É importante, então, questionar o que deve ser feito quando existem nodos pendentes na rede e avaliar as soluções que encontramos nas exposições sobre o PageRank que tratam este caso com mais rigor, tais como as de D. F. Gleich [14] e de M. Bianchini, M. Gori e F. Scarselli [21], não esquecendo a solução trazida pelos próprios criadores do método, Brin e Page, apresentada desde logo em [1].

3.1 Pseudo-PageRank

Na literatura sobre o PageRank em que existe distinção entre o tratamento das redes com e sem nodos pendentes, é comum a utilização do termo pseudo-PageRank para designar o caso em que estes nodos estão presentes, ficando assim o uso do termo PageRank apenas para o caso em que não estão. Esta distinção facilita a leitura das exposições que apresentam o cálculo do PageRank nas redes sem nodos pendentes como uma solução global, podendo estas manter o uso do termo PageRank e ficando subentendido que o caso exposto é o mais simples.

A distinção entre termos surge com definições próprias do problema de pseudo-PageRank segundo várias perspectivas, tal como acontece nas várias apresentações do PageRank. Em paralelo com a perspectiva apresentada no capítulo anterior do problema de PageRank, Gleich também apresenta em [14] o problema do pseudo-PageRank com a construção de um sistema linear. Antes de o apresentarmos, tenhamos em conta a seguinte definição.

Definição 3.1.1. *Uma matriz quadrada diz-se sub-estocástica se for não-negativa e as entradas de cada uma das suas colunas somarem no máximo 1. Um vetor diz-se sub-estocástico se for não-negativo e as suas entradas somarem no máximo 1.*

De referir novamente que, no caso da presença de nodos pendentes na rede, a matriz de distribuição por hiperligações não será estocástica, pois possuirá (uma ou mais) colunas que somam 0. Esta matriz será, no entanto, sub-estocástica e será denotada por $\bar{H}$.

Tendo, então, uma matriz de distribuição por hiperligações sub-estocástica $\bar{H}$, Gleich apresenta o sistema linear $(I - \alpha\bar{H})\bar{x} = f$, em que $0 < \alpha < 1$ é o parâmetro de teletransporte e f é um vetor não-negativo. Chamaremos a este novo sistema linear de sistema pseudo-PageRank.

Nas primeiras definições de pseudo-PageRank, como a que encontramos em [22], é comum encontrar $(I - \alpha\bar{H})\bar{x} = (1 - \alpha)f$. Na minha exposição, manterei a igualdade a f de Gleich, mas veremos como estas definições são equivalentes.

Aqui a solução continua a ser existente e única, pois $I - \alpha\bar{H}$ é ainda uma matriz-M nas condições do Teorema 2.2.1, e a esta solução $\bar{x}$ dá-se o nome de vetor de pseudo-PageRank.

Não podendo aplicar a teoria das matrizes estocásticas, apesar do vetor de pseudo-PageRank existir e ser único, não podemos garantir ser um vetor probabilístico. No entanto, a ordenação das suas entradas poderá sempre servir como um ranking das páginas da rede, mesmo não existindo a interpretação probabilística garantida no caso do PageRank.

A forma mais eficiente de encontrar este vetor é através da aplicação de um algoritmo baseado no método da potência. Neste caso, a adaptação do Algoritmo 1, apresentado no capítulo anterior como método de aproximação ao vetor de PageRank, serve eficazmente para nos aproximarmos do vetor de pseudo-PageRank.

Definição 3.1.2. *Seja $\bar{H}$ uma matriz de distribuição por hiperligações de uma rede de n páginas (com nodos pendentes). Sejam f um vetor não-negativo e não-nulo de dimensão n e $0 < \alpha < 1$ o parâmetro de teletransporte. Seja ε>0 a tolerância ao erro. Seja v_0 de dimensão n o vetor inicial. Definimos um método de iteração para o cálculo do ranking das páginas da rede pelo seguinte algoritmo:*

Algoritmo 2: Algoritmo de aproximação ao pseudo-PageRank

$\bar{v}_1 \leftarrow \alpha\bar{H}\bar{v}_0 + f$
$\delta \leftarrow ||\bar{v}_1 - \bar{v}_0||_1$
$k \leftarrow 1$
while $\delta > \varepsilon$ ***do***
  $\bar{v}_{k+1} \leftarrow \alpha\bar{H}\bar{v}_k + f$
  $\delta \leftarrow ||\bar{v}_{k+1} - \bar{v}_k||_1$
  $k \leftarrow k + 1$
end while

Após a adaptação do Algoritmo 1 ao sistema pseudo-PageRank, basta adaptar também o Teorema 2.2.2 para provar que o Algoritmo 2 é de facto um algoritmo eficaz na aproximação da solução pretendida tomando dois vetores iniciais diferentes.

Teorema 3.1.1. *Seja $\bar{H}$ uma matriz de distribuição por hiperligações de uma rede de n páginas (com nodos pendentes). Seja f um vetor não-negativo e não-nulo de dimensão n e $0 < \alpha < 1$ o parâmetro de teletransporte. Seja $\bar{v}$ o vetor de pseudo-Pagerank. Após k iterações do Algoritmo 2, são verdadeiras as afirmações:*

(a) se $\bar{v}_0 = \frac{1}{1-\alpha} f$, então

$$||\bar{v} - \bar{v}_k||_1 \leq ||\bar{v} - \frac{1}{1-\alpha} f||_1 \alpha^k \leq \frac{2}{1-\alpha} ||f||_1 \alpha^k$$

(b) se $\bar{v}_0 = 0$, então o vetor de erro $\bar{v} - \bar{v}_k \geq 0$ para qualquer k e

$$||\bar{v} - \bar{v}_k||_1 = e^T(\bar{v} - \bar{v}_k) \leq ||\bar{v}||_1 \alpha^k$$

Demonstração. (a) Similar à demonstração da alínea (a) do Teorema 2.2.2, tendo em conta que $||\bar{H}||_1 = 1$ e que $||\bar{v} - \frac{1}{1-\alpha}f||_1$ é no máximo $\frac{2}{1-\alpha}||f||_1$.

(b) Similar à demonstração da alínea (b) do Teorema 2.2.2, tendo em conta mais uma vez que $||\bar{H}||_1 = 1$ e que $||\bar{H}^k\bar{v}||_1 \leq ||\bar{H}^k||_1||\bar{v}||_1 = ||\bar{H}||_1^k||\bar{v}||_1 = ||\bar{v}||_1$.

□

Este teorema prova a eficácia do Algoritmo 2 na aproximação ao pseudo-PageRank, dado que nos dois casos do teorema a distância entre a solução e o vetor resultante da k-ésima iteração fica limitada superiormente por um valor que tende para 0 à medida que k tende para infinito.

Tomando novamente o exemplo apresentado da rede com um nodo pendente e tomando $\alpha = 0,85$ e

$$f = (1-\alpha)\begin{bmatrix}\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\end{bmatrix}^T,$$

o vetor de pseudo-PageRank neste caso é

$$\bar{v} \approx \begin{bmatrix}0,054 & 0,040 & 0,042 & 0,040 & 0,118 & 0,037\end{bmatrix}^T,$$

que pode ser encontrado resolvendo diretamente o sistema linear de 6 equações ou após 10 iterações do Algoritmo 2 tomando como vetor inicial $v_0 = \frac{1}{1-\alpha}f$ e $\varepsilon = 10^{-4}$, por exemplo.

Ignorando o facto deste vetor não ser estocástico, podemos observar como ficaria o ranking das páginas da rede se tomássemos a ordem decrescente dos valores das entradas deste vetor. As páginas 2 e 4 ficariam empatadas, como seria de prever, e o ranking das páginas em geral não é descabido. O pseudo-PageRank já nos dá uma ordenação por relevância aparentemente justa, apenas fica a faltar a interpretação probabilística por este não ser estocástico.

A maior parte das soluções para o tratamento dos nodos pendentes passam por transformar o problema de modo a obtermos matrizes estocásticas e, como consequência, um vetor final estocástico. Uma solução deste género também é apresentada por Gleich.

Teorema 3.1.2. *Seja $\bar{H}$ uma matriz de distribuição por hiperligações de uma rede de n páginas (com nodos pendentes). Seja f um vetor não-negativo e não-nulo de dimensão n e $0 < \alpha < 1$ o parâmetro de teletransporte. Seja $\bar{v}$ o vetor de pseudo-PageRank. Seja $t = \frac{f}{||f||_1}$ e seja $v = \frac{\bar{v}}{||\bar{v}||_1}$. Então v é o vetor de PageRank do problema de PageRank que gera o sistema*

$$(I - \alpha H)x = (1-\alpha)t$$

com $H = \bar{H} + tc^T$, onde $c = e^T - e^T\bar{H} \geq 0$ é um vetor de correção que garante que H é estocástica.

Demonstração. Notemos, primeiramente, que t é um vetor estocático, dado que f é não-negativo e não-nulo. H também é estocástica, dado que

$$e^T H = e^T(\bar{H} + tc^T) = e^T,$$

pela definição de H, por t ser estocático e pela definição de $c \geq 0$. Para além disso, $0 < \alpha < 1$ por definição, logo podemos afirmar que estamos na presença de um problema de PageRank e que o sistema

$$(I - \alpha H)x = (1 - \alpha)t$$

é um sistema PageRank válido. Seja, então, v a solução deste sistema. Pela definição de H e denotando

$$\gamma = \frac{\alpha c^T v + (1 - \alpha)}{||f||_1},$$

obtemos

$$\begin{aligned}
(I - \alpha H)v = (1 - \alpha)t &\Leftrightarrow (I - \alpha\bar{H} - \alpha t c^T)v = (1 - \alpha)t \\
&\Leftrightarrow v - \alpha\bar{H}v - \alpha t c^T v = (1 - \alpha)t \\
&\Leftrightarrow v = \alpha\bar{H}v + \alpha t c^T v + (1 - \alpha)t \\
&\Leftrightarrow v = \alpha\bar{H}v + \gamma f \\
&\Leftrightarrow v - \alpha\bar{H}v = \gamma f \\
&\Leftrightarrow (I - \alpha\bar{H})v = \gamma f.
\end{aligned}$$

Dado que $(I - \alpha\bar{H})\bar{v} = f$, por $\bar{v}$ ser o vetor de pseudo-PageRank, obtemos

$$(I - \alpha\bar{H})v = \gamma f \Leftrightarrow (I - \alpha\bar{H})v = \gamma(I - \alpha\bar{H})\bar{v} \Leftrightarrow v = \gamma\bar{v}.$$

Como v é vetor de PageRank, $||v||_1 = 1$. Assim,

$$||\gamma\bar{v}||_1 = 1 \Leftrightarrow ||\gamma||_1 = ||\bar{v}||_1^{-1}.$$

Logo, $v = \frac{\bar{v}}{||\bar{v}||_1}$ e concluímos a demonstração. □

Este teorema prova que é possível transformar qualquer problema de pseudo-PageRank inicial num problema de PageRank cuja solução, isto é, o vetor de PageRank, seja a transformação do vetor de pseudo-PageRank num vetor estocástico pela forma habitual da divisão pela norma. Isto indica-nos que a ordenação obtida pelo vetor de pseudo-PageRank pode ser, de facto, tomada como ranking das páginas, bastando normaliza-lo para o transformar num vetor probabilístico solução do problema de PageRank associado.

Fazendo as contas para o caso do nosso exemplo, continuando com os valores de α e f considerados anteriormente, o vetor t e o vetor de correção serão

$$t = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}^T, c = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}^T,$$

e a matriz H será

$$H = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{6} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{6} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{6} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{6} & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & \frac{1}{6} & 1 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{6} & 0 \end{bmatrix}.$$

Neste caso, a 5ª coluna de H sugere que interpretemos os nodos pendentes como nodos com hiperligações para todos os nodos da rede (em oposição a não apresentarem nenhuma hiperligação). Esta interpretação é apresentada como solução por A. Ng et al em [23], por S. Kamvar et al em [24], por N. Eiron et al em [25] e também por D. Austin em [8] e pode ser justificada pensando que, quando um navegador se encontra numa página sem hiperligações, a sua única opção para continuar a navegação será escolher de forma aleatória uma das páginas da rede, teletransportando-se para essa escolha.

E, assim, o vetor que soluciona o problema de PageRank gerado é

$$v \approx \begin{bmatrix} 0,163 & 0,122 & 0,127 & 0,122 & 0,355 & 0,111 \end{bmatrix}^T,$$

que corresponde exatamente à normalização do vetor de pseudo-PageRank, $\frac{\bar{v}}{||\bar{v}||_1}$. Concluímos com o ranking das 6 páginas em que a 5ª corresponde a um nodo pendente: 5, 1, 3, 2 e 4, 6. A 5ª página é a mais relevante.

3.2 Outras formas

Na secção anterior, vimos uma maneira de transformar um problema de pseudo-PageRank num de PageRank, que está associada a uma reinterpretação dos nodos pendentes. Outras formas de lidar com estes nodos também passam por repensar a rede e a construção da matriz de distribuição por hiperligações de forma a que esta seja sempre estocástica. Alguns exemplos serão discutidos nesta secção.

O primeiro exemplo é a solução apresentada por Bianchini, Gori e Scarselli [21], que sugerem a adição à rede com nodos pendentes de um nodo falso que se auto-referencie e arestas falsas que se dirijam dos nodos pendentes existentes na rede para este novo nodo, de forma a que a rede se transforme numa sem nodos pendentes. Chamemos a esta nova rede a rede estendida da rede com nodos pendentes. A matriz de distribuição por hiperligações desta nova rede é uma extensão da matriz de distribuição por hiperligações da rede original e pode ser definida da forma seguinte.

Definição 3.2.1. *Seja $\bar{H}$ uma matriz de distribuição por hiperligações de uma rede de n páginas (com nodos pendentes). Definimos a matriz de distribuição por hiperligações da rede estendida, H_{+1}, de dimensão $(n+1)\times(n+1)$, dada por*

$$H_{+1} = \begin{bmatrix} \bar{H} & 0 \\ R & 1 \end{bmatrix},$$

onde R é uma matriz de dimensão $1\times(n+1)$ cujas entradas das colunas correspondentes a nodos pendentes são iguais a 1 e as restantes são nulas.

A matriz de distribuição por hiperligações da rede estendida é sempre estocástica e podemos então, a partir dela, obter o ranking das páginas da rede estendida tratando o caso como um problema de PageRank normal. O ranking das páginas da rede original (com nodos pendentes) corresponderá a este ranking ignorando apenas a posição do nodo falso adicionado.

Aplicando esta solução ao nosso exemplo de rede com um nodo pendente, obtemos a matriz

$$H_{+1} = \left[\begin{array}{cccccc|c} 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right].$$

Escolhendo $\alpha = 0,85$, t igual ao vetor coluna com todas as entradas iguais a $\frac{1}{7}$, $\varepsilon = 10^{-4}$ e $v_0 = t$, por exemplo, aplicando o Algoritmo 1, após 10 iterações obtemos

$$v \approx \begin{bmatrix} 0,046 & 0,035 & 0,036 & 0,035 & 0,101 & 0,031 & 0,715 \end{bmatrix}^T.$$

Esta aproximação ao vetor de PageRank sugere a seguinte ordenação dos nodos da rede estendida: 7, 5, 1, 3, 2 e 4, 6. Ignorando o nodo extra, obtemos o ranking das páginas da rede: 5, 1, 3, 2 e 4, 6, igual ao ranking obtido na secção anterior com a solução de Gleich.

O segundo exemplo de como podemos lidar com os nodos pendentes passa também por transformar a matriz de distribuição por hiperligações fazendo a seguinte reinterpretação da experiência de um utilizador da Web: ao navegarmos na Internet, podemos escolher a próxima página a visitar seguindo uma hiperligação existente na página, sendo que no caso de não existir nenhuma hiperligação voltamos à página que visitámos anteriormente, ou simplesmente considerando novamente todas as páginas da rede e escolhendo uma delas. Isto significa que o utilizador, ao deparar-se com uma página "nodo pendente", tem a opção de clicar no botão 'Voltar' e regressar à página anterior. E assim, é como se essa página tivesse hiperligações para todas as páginas que têm hiperligações para ela.

Transcrever isto numa nova matriz de distribuição por hiperligações é considerar a seguinte definição.

Definição 3.2.2. *Seja $\bar{H}$ uma matriz de distribuição por hiperligações de uma rede de n páginas (com nodos pendentes). Definimos a matriz de distribuição por hiperligações com opção 'Voltar', H_V, de dimensão $n \times n$, dada por*

$$H_V = \bar{H} + V,$$

onde V é uma matriz $n \times n$ cujas entradas são definidas por

$$V_{ij} = \begin{cases} \frac{1}{m_j}, \textit{se } j \textit{ for nodo pendente e a página } i \textit{ apresentar uma hiperligação para a página } j \\ 0, \textit{caso contrário} \end{cases}$$

onde $i, j \in \{1, ..., n\}$ e, para cada j, m_j é a quantidade de páginas da rede que se referem à página j através de hiperligações.

Tal como no exemplo anterior, a matriz de distribuição por hiperligações com opção 'Voltar' é sempre estocástica e, através dela, ficamos na presença de um problema de PageRank.

Aplicando esta solução ao nosso exemplo de rede com um nodo pendente, obtemos a matriz

$$H_V = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{5} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{5} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{5} & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{5} & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & 0 \end{bmatrix}.$$

Escolhendo $\alpha = 0,85$, t igual ao vetor coluna com todas as entradas iguais a $\frac{1}{6}$, $\varepsilon = 10^{-4}$ e $v_0 = t$, por exemplo, aplicando o Algoritmo 1, após 11 iterações obtemos

$$v \approx \begin{bmatrix} 0,171 & 0,128 & 0,134 & 0,128 & 0,321 & 0,117 \end{bmatrix}^T,$$

do qual deduzimos o ranking das páginas da rede: 5, 1, 3, 2 e 4, 6, exatamente igual ao ranking obtido pelas duas soluções apresentadas anteriormente.

N. Eiron, K. S. McCurley e J. A. Tomlin discutem em [25] vários exemplos de soluções de como lidar com os nodos pendentes, incluindo os dois exemplos anteriores, e o da secção 3.1, que já foi referido. A par do primeiro exemplo desta secção, Eiron et al apresentam também a alternativa de se considerar a auto-referenciação em todos os nodos da rede.

Outras soluções mais extremas presentes na literatura são a adição de hiperligações aos nodos pendentes de forma totalmente personalizável [26] [27] e a remoção completa destes nodos da rede [27]. Discutiremos esta última abordagem na próxima secção.

3.3 Uma interpretação da solução de Brin e Page

Brin e Page apontam desde logo, de uma forma explícita, mas muito breve, os nodos pendentes como um problema do PageRank [1]. No parágrafo dedicado aos nodos pendentes, explicam o que estes nodos são e uma razão da sua existência em grande escala na rede. Também explicam, em pouco detalhe, como estes nodos poderão ser tratados: *"(...) we simply remove them from the system until all the PageRanks are calculated. After all the PageRanks are calculated, they can be added back in (...)"*, mencionando posteriormente também o facto do vetor final obtido não estar normalizado.

A falta de detalhes do procedimento apresentado faz com que esta solução seja pouco mencionada nas exposições que encontramos sobre o PageRank. De todas as que menciono ao longo deste capítulo, apenas 3 fazem referência às palavras de Brin e Page sobre os nodos pendentes.

Farei a minha interpretação da solução e mostrarei a mesma através do exemplo de rede com um nodo pendente que tem sido usado neste capítulo.

Recordemos a matriz de distribuição por hiperligações do exemplo:

$$\bar{H} = \begin{bmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix}.$$

Notemos que remover simplesmente o 5º nodo da rede, o nodo pendente, não nos devolve uma matriz de distribuição por hiperligações estocástica, pois o 6º nodo torna-se um nodo pendente. A remoção de nodos pendentes terá de ser feita de forma sucessiva até que nenhum nodo se torne pendente. Neste exemplo, isto implica a remoção sucessiva dos nodos 5, 6 e 4, o que resulta nas seguintes matrizes de distribuição por hiperligações:

$$\bar{H}_1 = \begin{bmatrix} 0 & 0 & \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}, \bar{H}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}, H_3 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

A matriz H_3, sendo finalmente uma matriz estocástica, pode ser usada para calcular o ranking dos 3 nodos que sobraram. E, escolhendo t igual ao vetor coluna com todas as entradas iguais a $\frac{1}{3}$, é fácil ver que o vetor de PageRank neste caso é o vetor

$$v_3 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^T,$$

donde obtemos o ranking dos 3 nodos.

Todos estes passos para calcular o ranking dos 3 nodos finais da rede são suportados por G. M. Del

Corso, A. Gullí e F. Romani [26], que interpretam as palavras de Brin e Page precisamente como a remoção sucessiva de nodos pendentes da rede. No entanto, para Del Corso et al, o cálculo do ranking terminaria aqui, pois, apresentando uma interpretação mais drástica, indicam que os nodos removidos são ignorados completamente do ranking final das páginas. Ficando a faltar a interpretação ao passo após a remoção: *"they can be added back in"*. Este passo é mencionado em [25], onde Eiron et al sublinham a falta de detalhes do procedimento, não apresentando nenhuma tentativa de interpretação do mesmo.

A meu ver, adicionar os nodos pendentes à rede após o cálculo do PageRank dos outros 3 significa regressar à matriz não estocástica que contemplava a presença desses nodos, a matriz $\bar{H}$, e aplicar o Algoritmo 1 (ou o equivalente Algoritmo 2), tendo como vetor inicial o vetor com o ranking obtido dos nodos não pendentes, isto é, o vetor

$$v_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix}^T.$$

Escolhendo $\alpha = 0,85$, t igual ao vetor coluna com todas as entradas iguais a $\frac{1}{6}$ e $\varepsilon = 10^{-4}$, por exemplo, após 12 iterações obtemos

$$\bar{v} \approx \begin{bmatrix} 0,054 & 0,040 & 0,042 & 0,040 & 0,124 & 0,037 \end{bmatrix}^T,$$

do qual podemos deduzir o ranking das páginas da rede: 5, 1, 3, 2 e 4, 6, mais uma vez, igual ao ranking obtido usando as soluções apresentadas anteriormente. Notemos que este vetor é exatamente igual ao vetor de pseudo-PageRank calculado na secção 3.1.

Segundo a minha interpretação, a solução apresentada por Brin e Page sugere, de forma muito superficial, o que se concluiu da análise à teoria paralelamente construída sobre as redes com nodos pendentes apresentada na secção 3.1. No fundo, usando agora o conceito de pseudo-PageRank apresentado nesta teoria, percebemos que o vetor de pseudo-PageRank, mesmo não sendo probabilístico, pode ser desde logo usado para a obtenção do ranking de todas as páginas da rede e que qualquer problema de pseudo-PageRank é fundamentalmente um problema de PageRank.

Esta conclusão é, em certa medida, apoiada pela interpretação feita por Bianchini, Gori e Scarselli em [21], que, apesar de apresentada de forma diferente, sugere o mesmo. Segundo Bianchini et al, a solução apresentada por Brin e Page é a de assumir que cada nodo pendente tem hiperligações para todos os nodos da rede. E, tal como vimos na secção 3.1, essa transformação é a mesma que se obtém através do Teorema 3.1.2 e o ranking do problema de PageRank é simplesmente a normalização do vetor de ranking do problema de pseudo-PageRank.

Por último, e tal como foi mencionado no Capítulo 2, a importância de se maximizar a eficiência na efetuação dos cálculos lado a lado com a redução do tempo dos mesmos é outro fator que apoia a minha interpretação da solução de Brin e Page. Na secção que segue o parágrafo sobre os nodos pendentes [1], Brin e Page estimam que a remoção sucessiva dos nodos pendentes da rede a reduza de um total de 75 milhões de nodos para 24 milhões. Por outro lado, Brin e Page relembram que partir de um bom vetor inicial é essencial para a velocidade de convergência do algoritmo. Trabalhar inicialmente com os 24 milhões de nodos não-pendentes não só consegue reduzir a complexidade dos cálculos como nos pode fornecer um bom ponto de partida para a obtenção do ranking da rede completa.

Conclusão

Terminamos esta dissertação com as seguintes conclusões.

A construção da matriz positiva Google, $G = \alpha H + (1 - \alpha)te^T$, a partir da matriz de distribuição por hiperligações, H, e dos vetor e parâmetro de teletransporte, t e α respectivamente, garante-nos, por aplicação do Teorema de Perron-Frobenius, que $Gx = \rho(G)x$ tem solução estocástica única e que esta é positiva. Dado que pretendemos obter uma solução estocástica única e positiva do sistema $Gx = x$, esta construção devolve-nos o pretendido quando a rede não tem nodos pendentes, pois, neste caso, G é sempre estocástica e, consequentemente, $\rho(G) = 1$. Sendo, assim, no caso de não existirem nodos pendentes na rede, a construção da matriz Google resolve o problema da falta de conexão forte na rede e garante a existência e unicidade do procurado vetor de PageRank (restando como única preocupação o cálculo deste vetor).

Apesar da matriz Google não se encontrar nas condições do teorema que perfaz o método da potência, é provado, no caso da rede não ter nodos pendentes, que este método pode ser aplicado. É a partir dele e da hipótese de G ser estocástica que se constrói um algoritmo de aproximação ao vetor de PageRank, o Algoritmo 1, cujos cálculos de cada iteração envolvem quase exclusivamente a matriz esparsa H, tornando-o extremamente eficiente.

A apresentação do ponto de vista alternativo de D. F. Gleich, que transforma $Gx = x$ no sistema linear $(I - \alpha H)x = (1 - \alpha)t$, traz-nos uma garantia alternativa de que o sistema tem solução única e não-negativa, por aplicação do teorema que caracteriza a classe das matrizes-M, e mais uma prova de que o Algoritmo 1 se aproxima desta solução, partindo de dois vetores iniciais diferentes. Para além disso, é a partir deste ponto de vista que construímos o sistema pseudo-PageRank, $(I - \alpha\bar{H})\bar{x} = f$, e toda a teoria paralela à do PageRank que trata a presença de nodos pendentes na rede. Percebemos que qualquer problema de pseudo-PageRank é fundamentalmente um problema de PageRank, dado que a normalização da solução de um sistema pseudo-PageRank é solução do sistema PageRank associado. No entanto, podemos sempre transformar qualquer problema de pseudo-PageRank num de PageRank, através de diversas interpretações alternativas.

Em termos práticos, a construção da matriz Google a partir de qualquer rede e a aplicação do Algoritmo 1, começando num vetor inicial apropriado, dá-nos, de forma eficaz, uma aproximação ao vetor de ranking das páginas da rede, sendo que, no caso da rede ter nodos pendentes, este vetor não será estocástico. Neste caso, um bom vetor inicial para a aplicação do algoritmo poderá ser o que se obtém através da técnica da remoção dos nodos pendentes da rede, apresentada na última secção.

Bibliografia

[1] L. Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab, 1999.

[2] S. Brin e L. Page. «The anatomy of a large-scale hypertextual Web search engine». Em: *Computer Networks and ISDN Systems* 30 (1998), pp. 107–117.

[3] R. Tanase e R. Radu. *The Mathematics of Web Search*. URL: http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/index.html.

[4] P. F. Gallardo. «Google's secret and Linear Algebra». Em: *European Mathematical Society Newsletter* (2007).

[5] M. E. J. Newman. «Mathematics of Networks». Em: *The New Palgrave Dictionary of Economics* (2008).

[6] R. A. Horn e C. R. Johnson. *Matrix Analysis - Second Edition*. Cambridge University Press, 2013. ISBN: 978-0-521-83940-2.

[7] A. Bonato. *A Course on the Web Graph*. Vol. 89. Graduate Studies in Mathematics. American Mathematical Society, 2008. ISBN: 0821844679.

[8] D. Austin. *How Google Finds Your Needle in the Web's Haystack*. URL: http://www.ams.org/featurecolumn/archive/pagerank.html.

[9] J. J. O'Connor e E. F. Robertson. *Oskar Perron*. URL: https://mathshistory.st-andrews.ac.uk/Biographies/Perron/.

[10] J. J. O'Connor e E. F. Robertson. *Ferdinand Georg Frobenius*. URL: https://mathshistory.st-andrews.ac.uk/Biographies/Frobenius/.

[11] siteefy. *How Many Webpages Are There?* URL: https://siteefy.com/how-many-websites-are-there/#How-Many-Webpages-Are-There.

[12] K. Bryan e T. Leise. «The $25,000,000,000 Eigenvector: The Linear Algebra Behind Google». Em: *SIAM Review* 48.3 (2006), pp. 569–581.

[13] A. S. Camargo e A. P. Galves. «Abordagem matemática por trás do algoritmo PageRank». Em: *Revista Eletrónica Paulista de Matemática* 21 (2021), pp. 11–23.

[14] D. F. Gleich. «PageRank Beyond the Web». Em: *SIAM Review* 57.3 (2015), pp. 321–363.

[15] J. J. O'Connor e E. F. Robertson. *Richard von Mises*. URL: https://mathshistory.st-andrews.ac.uk/Biographies/Mises/.

[16] G. N. Golub e C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1996. ISBN: 0-8018-5414-8.

[17] J. Kinne e J. Axenbeck. «Web Mining for Innovation Ecosystem Mapping: a Framework and a Large-Scale Pilot Study». Em: *Scientometrics* 125 (2020).

[18] K. Fan. «Topological Proofs for Certain Theorems on Matrices with Non-negative Elements». Em: *Monatshefte für Mathematik* 62 (1958), pp. 219–237.

[19] R. J. Plemmons. «M-matrix characterizations.I—nonsingular M-matrices». Em: *Linear Algebra and its Applications* 18 (1977), pp. 175–188.

[20] A. Berman e R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial e Applied Mathematics, 1994. DOI: 10.1137/1.9781611971262.

[21] M. Bianchini, M. Gori e F. Scarselli. «Inside PageRank». Em: *ACM Transactions on Internet Technology* 5 (2005), pp. 92–128.

[22] P. Boldi et al. «Traps and Pitfalls of Topic-Biased PageRank». Em: *Algorithms and Models for the Web-Graph*. Springer Berlin Heidelberg, 2008, pp. 107–116.

[23] A. Ng, A. Zheng e M. Jordan. «Stable Algorithms for Link Analysis». Em: *SIGIR Forum (ACM Special Interest Group on Information Retrieval)* (2001). DOI: 10.1145/383952.384003.

[24] S. Kamvar et al. «Exploiting the Block Structure of the Web for Computing PageRank». Em: (2003).

[25] N. Eiron, K. S. McCurley e J. A. Tomlin. «Ranking the web frontier». Em: *Proceedings of the 13th International Conference on World Wide Web*. WWW '04. Association for Computing Machinery, 2004, pp. 309–318. ISBN: 1-58113-844-X. DOI: 10.1145/988672.988714.

[26] G. M. Del Corso, A. Gullí e F. Romani. «Fast PageRank Computation Via a Sparse Linear System (Extended Abstract)». Em: *Algorithms and Models for the Web-Graph*. Springer Berlin Heidelberg, 2004, pp. 118–130.

[27] T. Haveliwala. *Efficient Computation of PageRank*. Rel. téc. Stanford University, 1999.