INSTITUTO FEDERAL DA PARAÍBA
CAMPUS CAJAZEIRAS
CURSO DE LICENCIATURA EM MATEMÁTICA

JOSÉ RUFINO RODRIGUES FILHO

A MATEMÁTICA POR TRÁS DO ALGORITMO *PAGERANK* DO *GOOGLE*

CAJAZEIRAS
2022

JOSÉ RUFINO RODRIGUES FILHO

A MATEMÁTICA POR TRÁS DO ALGORITMO *PAGERANK* DO *GOOGLE*

Monografia apresentada junto ao **Curso de Licenciatura em Matemática** do **Instituto Federal da Paraíba**, como requisito à obtenção do título de **Licenciado em Matemática**.

Orientador:

Prof. Dr. Vinicius Martins Teodosio Rocha.

Cajazeiras

2022

JOSÉ RUFINO RODRIGUES FILHO

A MATEMÁTICA POR TRÁS DO ALGORITMO *PAGERANK* DO *GOOGLE*

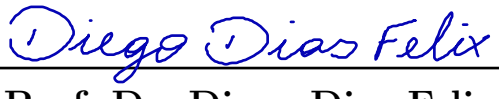
Monografia apresentada ao programa de **Curso de Licenciatura em Matemática** do **Instituto Federal da Paraíba**, como requisito à obtenção do título de **Licenciado em Matemática**.

Data de aprovação: 05/05/2022

Banca Examinadora:

Prof. Dr. Vinicius Martins Teodosio Rocha
Instituto Federal da Paraíba - IFPB

Profa. Dra. Taciana Araújo de Souza
Instituto Federal da Paraíba - IFPB



Prof. Dr. Diego Dias Felix
Instituto Federal da Paraíba - IFPB

IFPB / Campus Cajazeiras
Coordenação de Biblioteca
Biblioteca Prof. Ribamar da Silva
Catalogação na fonte: Suellen Conceição Ribeiro CRB-2218

R696m Rodrigues Filho, José Rufino

A matemática por trás do algoritmo *Pagerank* do *Google* / Jéssica Santos Silva. – Cajazeiras/PB: IFPB, 2022.

49f.:il.

Trabalho de Conclusão de Curso (Graduação em Matemática) - Instituto Federal de Educação, Ciência e Tecnologia da Paraíba-IFPB, Campus Cajazeiras. Cajazeiras, 2022.

Orientador(a): Prof. Dr. Vinicius Martins Teodosio Rocha..

1. Matemática. 2. Algoritmo *Pegerank*. 3. *Google*. 4. Álgebra Linear.

I. Rodrigues Filho, José Rufino. II. Titulo.

CDU: 51 R696m

Dedico à minha Vó, Elina Maria de Queiroz (in memoriam), e aos meus sobrinhos Guilherme Miguel Rufino e Sofia Gabrielle Rufino.

AGRADECIMENTOS

Diversas pessoas ajudaram de forma direta e indireta no desenvolvimento deste trabalho, dentre as quais eu agradeço:

Aos meus familiares, em especial ao meu irmão Rafael Rufino Rodrigues e à minha irmã Marineis Rufino Rodrigues Macena, pois sem eles a conclusão deste curso não seria possível.

Ao meu orientador, Prof. Dr. Vinicius Martins Teodosio Rocha, por ter me acompanho, com dedicação e amizade, durante todo o desenvolvimento do presente trabalho.

Aos membros da banca examinadora, Profa. Dra. Taciana Araújo de Souza e Prof. Dr. Diego Dias Felix, pela disponibilidade de participar e pelas contribuições.

Por fim, agradeço aos meus amigos e colegas de curso, em especial à Larissa Soares, ao José Jorge, à Maria Elizabete, ao Francisco Felipe, ao Antonniel Lourenço, ao Luan Ramalho, ao Francisco Alan, ao Carlos Miguel, ao Caio Henrique, ao Fabrício Rodrigues, à Maria Lavínia, à Paloma Alves, ao Derek Allan e ao Francisco André.

"Não há maior riqueza do que o conhecimento,
nem maior pobreza do que a ignorância"

Ali ibne Abi Talibe

RESUMO

Neste trabalho exploramos um pouco da matemática por trás de um dos algoritmos que fizeram do *Google* uma das empresas mais bem sucedidas do mundo. Estamos nos referindo ao algoritmo *PageRank*. Ele é responsável por atribuir a cada página da *Web* uma pontuação de importância, que é utilizada, conjuntamente com outras variáveis, para ranqueá-las. Almejamos saber quais são e como são aplicados os conceitos matemáticos pelo algoritmo *PageRank* do *Google* para se atribuir tal pontuação de importância a cada página da *Web*. Concomitantemente, também buscamos expor uma noção intuitiva do algoritmo. Para isso, optamos por uma pesquisa bibliográfica, com uma abordagem qualitativa e natureza básica pura. A qual nos possibilitou encontrar, por trás do algoritmo, uma das mais interessantes aplicações da Álgebra Linear e da Teoria da Probabilidade.

Palavras-chave: *PageRank*; Cadeias de Márkov; Álgebra Linear; *Google*.

ABSTRACT

This work is dedicated to explore a part of the mathematics behind one of the algorithms that made Google one of the world's most successful companies, the PageRank algorithm. The PageRank algorithm assigns an importance rating to each web page that is used, together with other variables, to define a ranking between them. We aim to explore which mathematical concepts are used to define this rating and how are they applied. We also seek to expose an intuitive notion of the algorithm. For this purpose, we opted for bibliographical research with a qualitative approach. Our research allowed us to find, through such algorithm, one of the most interesting applications of Linear Algebra and the Theory of Probabilities.

Keywords: PageRank; Markov Chains; Linear Algebra; Google.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Embora tenha surgido nos anos sessenta como resultado de um esforço do sistema de defesa dos Estados Unidos de tornar a comunidade acadêmica e militar partes de uma rede de computadores capaz de sobreviver a um ataque nuclear, até os anos de 1990 a Internet ainda era restrita a esse público. No entanto, foi a criação de uma nova tecnologia, baseada na Internet, que mudou essa concepção.

A *World Wide Web* (WWW) revolucionou o compartilhamento de informações. A iniciativa, concebida pelo pesquisador Tim Berners-Lee quando ainda participava da Organização Europeia para a Pesquisa Nuclear (CERN), tinha objetivo duplo. Primeiro, criar uma interface de usuário única e fácil para todos os tipos de informação, para que todos pudessem acessá-la; segundo, tornar tão fácil adicionar novas informações que a quantidade e a qualidade da informação *on-line* aumentariam (BERNERS-LEE, 1992). Logo, não demoraria muito para que o uso da WWW "explodisse". Ao leitor interessado em um estudo mais profundo sobre a *World Wide Web*, além de aspectos históricos, recomendamos (BERNERS-LEE, 1992; STOLFI, 2010).

O fato é que o número de páginas/informações na *Web* começou a crescer de forma exponencial, tornando-a grande e heterogênea. Recuperar as informações passou a ser um desafio (PAGE *et al.*, 1999). Surgia a necessidade de meios eficientes para essa tarefa. Assim, começava a entrar em cena as ferramentas de busca, isto é, sites especializados em localizar outros sites (MORAIS; AMBRÓSIO, 2007).

Os diretórios foram as primeiras ferramentas desenvolvidas para localizar os recursos na *Web*, vindo a preceder os mecanismos de busca (CENDÓN, 2001). Para localizar as informações o usuário teria que navegar pelas categorias e subcategorias que o diretório disponibilizava, criadas, em sua maioria, manualmente, e torcer para que ele possuísse alguma página relacionada com o assunto que estava sendo buscado pelo usuário. Esse trabalho manual para criar as categorias de páginas foi possível no início, no entanto, conforme o número de páginas na *Web* tomava grandes proporções em um curto espaço de tempo, isso acabava se tornando impossível. Então, começaram a surgir os motores de busca.

Segundo (BRYAN; LEISE, 2006), um motor de busca, tal como o *Google*, faz três coisas: primeiro, ele localiza, com auxílio de robôs, conhecidos como *spiders*, *bots* ou *web crawlers*, todas as páginas da *Web* com acesso público; segundo, ele indexa todas as páginas que visitou na primeira etapa, a partir de palavras-chave de cada uma delas;

terceiro, ranqueia todas as páginas em seu banco de dados, para que, quando um usuário arbitrário realizar uma pesquisa, as páginas mais importantes, dentre o subconjunto de páginas que possui relação com a pesquisa do usuário, sejam mostradas primeiro. E é essa última etapa que exploramos neste trabalho.

O *Google*, em especial a sua eficiência e precisão em fornecer resultados para as buscas dos usuários, nos despertou a curiosisade. Em pesquisas preliminares descobrimos que, atualmente, para que esse site de busca consiga sempre fornecer resultados precisos, ele utiliza diversas variáveis e algoritmos, muitos dos quais sob segredo, afinal eles são a chave do sucesso dos motores de busca. No entanto, um algoritmo foi primordial para que o *Google* se tornasse a potência que é atualmente. Estamos falando do algoritmo *PageRank*.

Um algoritmo nada mais é do que uma sequência finita de passos. Mas quais são esses passos? Isso depende do problema que queremos resolver. Os dois colegas de doutorado da Universidade de Stanford, Larry Page e Sergey Brin, queriam resolver algo literalmente grande: *ranquear as páginas da Web em termos da sua "importância"* (PAGE *et al.*, 1999). E eles conseguiram. Em 1996 Page e Brin lançaram o algoritmo *PageRank*, capaz de ranquear as páginas da *Web* em termos da sua importância. Na verdade, isso foi fruto do problema de tese de Page: *entender a estrutura matemática subjacente a internet*, ideia essa que havia sido proposta pelo seu orientador Terry Wilnograd: *estudar a estrutura de links inerente a internet* (SANTIAGO, 2021).

O algoritmo *PageRank* foi um marco. Embora já existissem mecanismos de busca tais como Yahoo! e MSN, eles funcionavam, em sua essencia, como contadores de palavras (SANTIAGO, 2021). Ou seja, o buscador iria fornecer como resultado para uma busca as páginas da *Web* que possuíssem o maior número de palavras ou frases que haviam sido pesquisadas. E qual o problema desse método? Um, bem claro, é que ele poderia ser facilmente enganado. Por exemplo, se criássemos uma página na *Web* e nela colocássemos apenas a sigla IFPB, milhares de vezes, ela certamente seria apresentada em primeiro lugar, ao invés da página oficial do IFPB.

Page e Brin resolveram combinar o ranqueamento de páginas da *Web*, proporcionado pelo algoritmo *PageRank*, com a contagem de palavras em um novo buscador; nascia, então, o *Google* (SANTIAGO, 2021).

Motivados pelo que discorremos até aqui, este trabalho tem como proposta responder a seguinte pergunta: *quais são e como são aplicados os conceitos matemáticos pelo algoritmo PageRank do Google para se atribuir um valor de importância a cada página da Web?* Para tal, estabelecemos os seguintes objetivos específicos: apresentar um problema da Olimpíada Brasileira de Matemática (OBM) e explorar as relações com o algoritmo

PageRank; estudar algumas noções da teoria dos grafos no estudo da *Web*; investigar os conceitos matemáticos que o algoritmo *PageRank* aplica e justificar como o algoritmo aplica tais conceitos para se atribuir um valor de importância a cada página da *Web*.

Para alcançar os objetivos que propusemos, recorremos à uma pesquisa qualitativa, pois almeja a descrição minuciosa dos fenômenos e dos elementos que a envolvem (AUGUSTO *et al.*, 2013). Os dados analisados foram obtidos, em sua maioria, a partir de livros e artigos científicos. Desse modo, se trata de uma pesquisa bibliográfica. Além disso, é de natureza básica pura, pois visa, unicamente, à ampliação do conhecimento (GIL *et al.*, 2017).

Sobre a estrutura deste trabalho, iniciamos explorando um dos problemas da OBM. O mesmo nos permitirá introduzir, de forma didática, as principais noções matemáticas necessárias ao entendimento do algoritmo *PageRank*, além de sua noção intuitiva. Em seguida, no Capítulo 2, objetivou-se modelar o algoritmo, além de expor as devidas justificativas. Por fim, são apresentadas as considerações finais acerca do presente trabalho.

1 A NOÇÃO INTUITIVA DO ALGORITMO PAGERANK

Neste capítulo apresentamos uma das noções intuitivas do algoritmo *PageRank*, além do estudo de algumas noções preliminares que nos ajudarão no capítulo 2. Isso será feito a medida que solucionamos um dos problemas da Olimpíada Brasileira de Matemática (OBM). Ademais, o desenvolvimento deste capítulo foi baseado nos trabalhos (PAGE *et al.*, 1999), (BOLDRINI *et al.*, 1980), (MALAJOVICH, 2021), (SILVA; JÚNIOR, 2011) e (POOLE, 2014).

Page *et al.* (1999), Boldrini *et al.* (1980), Malajovich (2021), Silva e Júnior (2011) e Poole (2014).

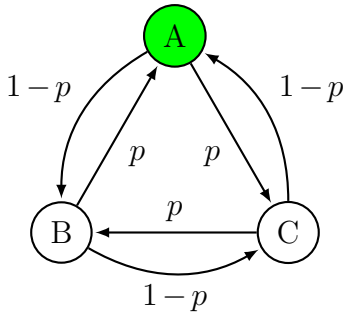
1.1 A RÃ DÕ E O ALGORITMO *PAGERANK*

A XXXI OBM, nível universitário, em um dos seus problemas nos apresenta a rã Dõ, que descansa no vértice A de um triângulo equilátero ABC. A cada minuto, Dõ salta para um dos vértices adjacentes, com probabilidade p de o salto ser no sentido horário e $1-p$ de ser no sentido anti-horário, com $p \in (0,1)$. Então, em seu item a, nos pede para mostrar que, conforme o número de saltos tende ao infinito, a probabilidade da rã estar no vértice A é de $1/3$.

Embora não pareça, esse problema possui uma certa relação com o algoritmo *PageRank*. Mas, antes de apresentá-la mostraremos o que se pede.

Primeiramente podemos tentar visualizar a situação descrita. A Figura 1.1 representa os vértices A, B e C do nosso triângulo equilátero. As probabilidades dos saltos são indicadas pelas setas. Além disso, destacamos com a cor verde o vértice em que Dõ se encontra inicialmente, isto é, sua posição inicial.

Figura 1.1 – Representação do problema da rã Dõ



Fonte: Autoria própria, 2022.

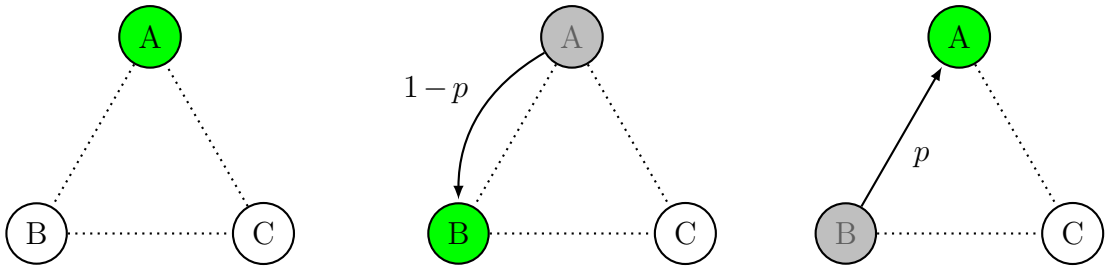
Consideremos que a_n (respectivamente b_n e c_n) indica a probabilidade da rã Dõ

estar, após n saltos, no vértice A (respectivamente B e C). Como queremos mostrar que $a_n \to 1/3$ quando $n \to \infty$, seria interessante encontrar uma forma de generalizar o valor de a_n para um $n \in \mathbb{N}$ qualquer. Com isso em mente, vamos calcular a_n para os valores iniciais de n e ver o que acontece.

Quando $n = 0$ estamos nos referindo a posição inicial da rã Dõ, que é fornecida no problema. Assim, $a_0 = 1$ e $b_0 = c_0 = 0$. Isto é, para $n = 0$ a probabilidade da rã estar no vértice A é de 100%, de estar no vértice B é de 0% e de estar no vértice C é de 0%. Quando $n = 1$ temos $a_1 = 0$, $b_1 = 1 - p$ e $c_1 = p$. Quando $n = 2$ devemos fazer uma análise mais cautelosa.

Primeiramente vamos calcular a_2. Queremos descobrir a probabilidade da rã estar no vértice A após dois saltos. Uma primeira alternativa seria saltar para o vértice B e, em seguida, saltar de volta para o vértice A. Veja a Figura 1.2.

Figura 1.2 – Primeira alternativa de saltos até o vértice A



(a) Posição inicial **(b) Salto 1** **(c) Salto 2**

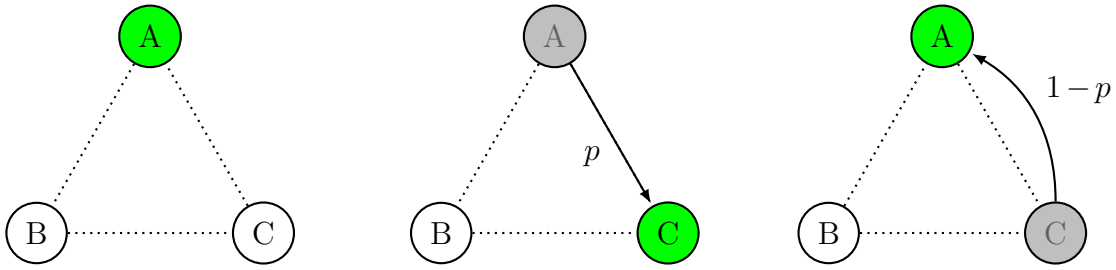
Fonte: Autoria própria, 2022.

Se P é a probabilidade de Dõ seguir esse caminho, então $P = (1 - p) \cdot p$. Note que $b_1 = 1 - p$, assim

$$P = b_1 \cdot p.$$

Uma segunda alternativa para se chegar ao vértice A, após dois saltos, seria seguir o caminho indicado na Figura 1.3.

Figura 1.3 – Segunda alternativa de saltos até o vértice A



(a) Posição inicial **(b) Salto 1** **(c) Salto 2**

Fonte: Autoria própria, 2022.

Novamente, se Q é a probabilidade de Dõ seguir esse caminho até o vértice A, então $Q = p \cdot (1-p)$. Ou ainda, como $c_1 = p$,

$$Q = c_1 \cdot (1-p).$$

Veja que as Figuras 1.2 e 1.3 representam as únicas alternativas para Dõ ir até o vértice A após dois saltos. Logo, a_2 será dado pela soma de P e Q:

$$a_2 = b_1 \cdot p + c_1 \cdot (1-p).$$

De forma análoga, tem-se

$$b_2 = a_1 \cdot (1-p) + c_1 \cdot p \quad \text{e} \quad c_2 = a_1 \cdot p + b_1 \cdot (1-p).$$

Perceba que $a_1 = 0$, o que nos levaria a $b_2 = c_1 \cdot p$ e $c_2 = b_1 \cdot (1-p)$. Mas, o nosso objetivo é fazer o leitor perceber que podemos determinar as probabilidades futuras a_2, b_2 e c_2 apenas conhecendo o vértice em que a rã se encontra no momento atual, isto é, após o salto 1. Para deixar essa ideia mais clara vamos calcular a_3, b_3 e c_3.

Agora Dõ dará três saltos. Dessa vez não precisamos de nenhum diagrama. Começaremos calculando a_3:

- Se o estado atual da rã após o segundo salto é o vértice B, cuja probabilidade é b_2, então, para se chegar ao vértice A, o terceiro salto deve ser no sentido horário (p).
- Se o estado atual da rã após o segundo salto é o vértice C, cuja probabilidade é c_2, então, para se chegar ao vértice A, o terceiro salto deve ser no sentido anti-horário $(1-p)$.

Portanto,

$$a_3 = b_2 \cdot p + c_2 \cdot (1-p).$$

Analogamente,

$$b_3 = a_2 \cdot (1-p) + c_2 \cdot p \quad \text{e} \quad c_3 = a_2 \cdot p + b_2 \cdot (1-p).$$

Não importa a quantidade n de saltos, as probabilidades futuras a_n, b_n e c_n dependem apenas do vértice em que a rã se encontra no salto imediatamente anterior,

$n-1$. Isso nos permite escrever, como queríamos, a seguinte generalização para a_n, b_n e c_n:

$$\begin{aligned} a_n &= p \cdot b_{n-1} + (1-p) \cdot c_{n-1} \\ b_n &= p \cdot c_{n-1} + (1-p) \cdot a_{n-1} \\ c_n &= p \cdot a_{n-1} + (1-p) \cdot b_{n-1}. \end{aligned}$$

Ou ainda, em forma de equação matricial:

$$\begin{bmatrix} a_n \\ b_n \\ c_n \end{bmatrix} = \begin{bmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{bmatrix} \begin{bmatrix} a_{n-1} \\ b_{n-1} \\ c_{n-1} \end{bmatrix}. \tag{1}$$

Então, como devemos continuar? Essa equação matricial parece bem complicada de resolver. No entanto, a matriz dos coeficientes possui algumas propriedades que podem nos ajudar a chegar à uma solução. Mas, para que consigamos compreendê-las, precisamos estudar algumas noções de um novo conceito, a saber, cadeias de Markov.

1.1.1 Cadeias de Markov

Um *processo estocástico* é definido como um conjunto de variáveis aleatórias (X_t) indexadas por um índice t pertencente a um conjunto T. Geralmente T é um conjunto de inteiros não-negativos. Mas, o que seria uma variável aleatória? Uma *variável aleatória* nada mais é do que uma função que descreve os resultados de um experimento através de valores numéricos. Por exemplo, considere a seguinte situação, onde a variável aleatória X_t representa o estado de uma máquina no tempo t, em minutos.

$$X_t = \begin{cases} 0, & \text{se a máquina estiver ligada} \\ 1, & \text{se a máquina estiver desligada} \end{cases}, \text{ com } t \in \mathbb{N}. \tag{2}$$

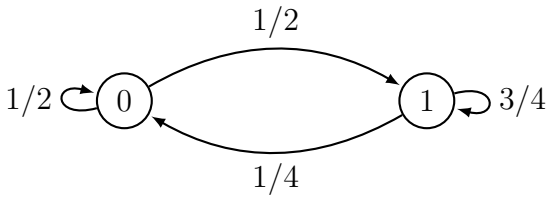
O conjunto E de valores que a variável X_t pode assumir é chamado de *espaço de estados*. Nesse caso $E = \{0, 1\}$.

A cada minuto t a máquina pode estar ligada $(X_t = 0)$ ou desligada $(X_t = 1)$ e a variável aleatória X_1 representa o estado da máquina no minuto 1, X_2 representa o estado da máquina no minuto 2 e assim por diante. Ou seja, para cada $t \in \mathbb{N}$, X_t é uma variável aleatória. Portanto, nesse exemplo, o conjunto $\{X_t, t \in T = \mathbb{N}\}$ é o que definimos como processo estocástico. Além disso, esse processo estocástico é classificado como: estado discreto (o conjunto E é enumerável) e tempo discreto (o conjunto T é enumerável).

Uma forma de representar um processo estocástico graficamente é através de diagramas. Do exemplo anterior suponha que: quando a máquina estiver no estado 0 a

probabilidade de permanecer no estado 0 seja de 1/2 e de ir para o estado 1 seja 1/2, quando a máquina estiver no estado 1 a probabilidade de permanecer no estado 1 seja de 3/4 e de ir para o estado 0 seja de 1/4. A Figura 1.4 representa essa situação.

Figura 1.4 – Representação de um processo estocástico



Fonte: Autoria própria, 2022.

Um processo estocástico será um *processo de Markov* (ou markoviano) quando a ocorrência de um estado futuro depender apenas do estado atual. Isto é, a probabilidade dos eventos futuros não dependem dos eventos passados, vindo a depender apenas do estado atual. O processo estocástico da Figura 1.4 é um dos mais simples casos de processo de Markov. Veja que, se a máquina estiver no estado 1 (desligada), então a probabilidade de transição do estado 1 para o estado 0 depende apenas do estado atual (estado 1), independendo do estado em que a máquina estava anteriormente. Ou seja, o estado atual é suficiente para determinar a probabilidade de transição para qualquer estado futuro. Matematicamente, isso equivale a escrever a seguinte igualdade entre probabilidades condicionais:

$$P(X_t = x_t \mid X_{t-1} = x_{t-1}, \ldots, X_1 = x_1, X_0 = x_0) = P(X_t = x_t \mid X_{t-1} = x_{t-1}). \qquad (3)$$

Ou seja, a probabilidade de estar no estado x_t no instante t depende somente do estado no instante $t-1$, independendo de todos os estados anteriores $(X_{t-2}, X_{t-3}, \ldots, X_1, X_0)$. Essa propriedade chama-se propriedade de Markov. E qualquer processo estocástico em tempo discreto e estado dicreto com a propriedade de Markov é definido como *cadeia de Markov*.

Perceba, então, que a Figura 1.1 é, também, uma representação de uma cadeia de Markov. Nela temos os estados A, B e C, que podem ser associados aos valores 1, 2 e 3, respectivamente. Veja que, se a rã Dõ estiver no estado 2 (vértice B) e quiser atingir o estado 3 (vértice C), a probabilidade transição do estado 2 para o estado 3 depende apenas de ela está no estado 2.

Assim sendo, vamos continuar explorando outras características a partir do nosso problema inicial. Porém, primeiro denominaremos de P a matriz dos coeficientes da equação matricial (1) e vamos nomear os vetores da seguinte forma:

$$\boldsymbol{x_n} = \begin{bmatrix} a_n \\ b_n \\ c_n \end{bmatrix}.$$

Assim, podemos escrever:

$$\boldsymbol{x_n} = P \cdot \boldsymbol{x_{n-1}}. \tag{4}$$

Desse modo, se quisermos calcular $\boldsymbol{x_1}$, devemos fazer o produto entre P e o vetor $\boldsymbol{x_0}$. Como a rã se encontra inicialmente no vértice A, temos que $\boldsymbol{x_0} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$. Ou seja,

$$\boldsymbol{x_1} = P \cdot \boldsymbol{x_0} = \begin{bmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1-p \\ p \end{bmatrix}.$$

Os vetores $\boldsymbol{x_n}$ são chamados de *vetores de estado* de uma cadeia de Markov. A matriz P é chamada de *matriz de transição*. Perceba que uma cadeia de Markov satisfaz a relação (4) para qualquer $n \in \mathbb{N}$. Aliás, com essa relação conseguimos obter, de forma iterativa, qualquer $\boldsymbol{x_n}$ desde que se conheça $\boldsymbol{x_0}$ e P. Ou seja, uma cadeia de Markov fica completamente definida quando se conhece as probabilidades de transição e o estado inicial.

Além disso, observando os vetores $\boldsymbol{x_0}$ e $\boldsymbol{x_1}$ conseguimos notar duas coisas: primeiro, as entradas desses vetores são não negativas; segundo, a soma de suas entradas é igual a 1. Vetores com essas propriedades são chamados de *vetores de probabilidade*. Consegue notar algo semelhante na matriz P? Isso mesmo, as colunas dessa matriz são vetores de probabilidade. Qualquer matriz quadrada com essas propriedades é chamada de *matriz estocástica*. Em outros termos:

Definição 1 (Matriz Estocástica)**.** *Uma matriz quadrada M, de ordem $n \times n$ e $M_{ij} \geq 0$, é chamada de estocástica se, para toda coluna j, $\sum_i M_{ij} = 1$.*

A partir do fato de que uma cadeia de Markov fica completamente definida conhecendo-se o seu estado inicial e as probabilidades de transição, seria interessante escrevermos a relação (4) apenas em função delas. E isso é algo bem simples. Veja como podemos calcular $\boldsymbol{x_2}$:

$$\boldsymbol{x_2} = P \cdot \boldsymbol{x_1} = P \cdot (P \cdot \boldsymbol{x_0}) = P^2 \cdot \boldsymbol{x_0}. \tag{5}$$

E, em geral,

$$\boldsymbol{x_n} = P^n \cdot \boldsymbol{x_0}, \quad \text{com} \quad n = 0, 1, 2, \ldots \tag{6}$$

Isso nos leva a estudar as potências de uma matriz de transição. Do nosso problema vamos calcular P^2. No entanto, com o intuito de tornar os cálculos mais claros, atribuiremos um valor a p. Como $0 < p < 1$, consideraremos que $p = 0,6$. Então, ficamos com a seguinte matriz, que chamaremos de P':

$$P' = \begin{bmatrix} 0 & 0,6 & 0,4 \\ 0,4 & 0 & 0,6 \\ 0,6 & 0,4 & 0 \end{bmatrix}.$$

Assim,

$$P'^2 = \begin{bmatrix} 0 & 0,6 & 0,4 \\ 0,4 & 0 & 0,6 \\ 0,6 & 0,4 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0,6 & 0,4 \\ 0,4 & 0 & 0,6 \\ 0,6 & 0,4 & 0 \end{bmatrix} = \begin{bmatrix} 0,48 & 0,16 & 0,36 \\ 0,36 & 0,48 & 0,16 \\ 0,16 & 0,36 & 0,48 \end{bmatrix}$$

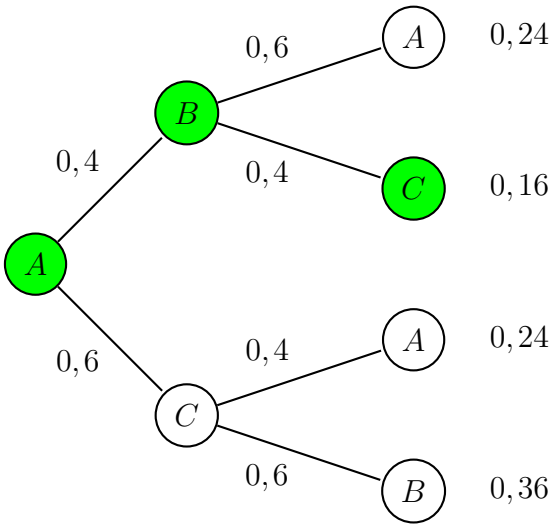
Primeiro, note que todas as entradas de P'^2 são positivas, o que motiva a seguinte definição, que será essencial no Capítulo 2.

Definição 2 (Matriz positiva)**.** *Uma matriz M é positiva se todas as suas entradas são positivas.*

Segundo, P'^2 também é uma matriz estocástica. O que nos leva a se questionar se P'^2 seria algum tipo de matriz de transição. Consideremos uma de suas entradas, por exemplo, $(P'^2)_{31} = 0,16$. O diagrama de árvore da Figura 1.5 nos mostra de onde esse valor vem.

Após dois minutos (ou dois saltos) existem quatro possibilidades de mudança de estado para a rã Dõ, e isso corresponde as quatro ramificações do nosso diagrama. Estando inicialmente no vértice A, existe apenas uma maneira de ir para o vértice C após dois minutos (destacada em verde na Figura 1.5): a rã, inicialmente no vértice A, após o primeiro minuto salta para o vértice B e, após o segundo minuto, salta para o vértice C.

Figura 1.5 – Diagrama de árvore para o estado inicial A da matriz P'^2



Fonte: Autoria própria, 2022.

O produto dessas probabilidades nos fornece o valor $0,16$, que é exatamente o valor da entrada $(P'^2)_{31}$.

Desse modo, segue que $(P'^2)_{31}$ representa a probabilidade de mover do estado 1 (vértice A) para o estado 3 (vértice C) em duas transições. Podemos generalizar esse fato, para uma matriz de transição P, da seguinte forma: $(P^n)_{ij}$ *é a probabilidade de mover do estado* j *para o estado* i *em* n *transições.*

Agora, nesse caso específico ($p = 0,6$), o que acontece com os nossos vetores de estado a longo prazo? Vamos continuar fazendo alguns cálculos.

$$
\begin{aligned}
\boldsymbol{x_0} &= \begin{bmatrix}1\\0\\0\end{bmatrix}, \boldsymbol{x_1} = \begin{bmatrix}0\\0,4\\0,6\end{bmatrix}, \boldsymbol{x_2} = P\cdot\boldsymbol{x_1} = \begin{bmatrix}0,48\\0,36\\0,16\end{bmatrix}, \boldsymbol{x_3} = P\cdot\boldsymbol{x_2} = \begin{bmatrix}0,280\\0,288\\0,432\end{bmatrix},\\
\boldsymbol{x_4} &= P\cdot\boldsymbol{x_3} = \begin{bmatrix}0,346\\0,371\\0,283\end{bmatrix}, \boldsymbol{x_5} = P\cdot\boldsymbol{x_4} = \begin{bmatrix}0,336\\0,308\\0,356\end{bmatrix}, \boldsymbol{x_6} = P\cdot\boldsymbol{x_5} = \begin{bmatrix}0,327\\0,348\\0,325\end{bmatrix},\\
\boldsymbol{x_7} &= P\cdot\boldsymbol{x_6} = \begin{bmatrix}0,329\\0,326\\0,336\end{bmatrix}, \boldsymbol{x_8} = P\cdot\boldsymbol{x_7} = \begin{bmatrix}0,330\\0,337\\0,334\end{bmatrix}, \boldsymbol{x_9} = P\cdot\boldsymbol{x_8} = \begin{bmatrix}0,335\\0,332\\0,333\end{bmatrix}.
\end{aligned}
$$

Até aqui podemos notar que o nosso vetor de estado está cada vez mais próximo do vetor $\begin{bmatrix}1/3 & 1/3 & 1/3\end{bmatrix}^T$, como se estivesse convergindo para esse valor, implicando que

a rã Dõ teria a mesma probabilidade de estar em qualquer um dos três vértices no longo prazo, ou seja, conforme $n \to \infty$, a_n, b_n e c_n tendem à 1/3. Além disso, é fácil constatar que, uma vez que essa distribuição de probabilidade for alcançada, ela nunca mudará. Então, simplesmente escrevemos

$$\begin{bmatrix} 0 & 0{,}6 & 0{,}4 \\ 0{,}4 & 0 & 0{,}6 \\ 0{,}6 & 0{,}4 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}.$$

Um vetor de probabilidade $\boldsymbol{x}$ de uma matriz estocástica P com a propriedade de que $P\boldsymbol{x} = \boldsymbol{x}$ é chamado de *vetor de estado estacionário*.

Perceba que, para calcular o vetor de estado estacionário da matriz P', fizemos várias iterações a partir do vetor de estado $\boldsymbol{x_0}$. No entanto, há uma maneira mais direta.

Do problema da rã Dõ considere que o vetor de estado estacionário que procuramos seja $\boldsymbol{x} = \begin{bmatrix} a & b & c \end{bmatrix}^T$. Além disso, vamos reescrever a equação $P\boldsymbol{x} = \boldsymbol{x}$ como $P\boldsymbol{x} = I\boldsymbol{x}$. O que implica que $(I - P)\boldsymbol{x} = \boldsymbol{0}$. Substituindo cada um dos valores, segue que:

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{bmatrix} \right) \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Ou ainda,

$$\begin{bmatrix} 1 & -p & p-1 \\ p-1 & 1 & -p \\ -p & p-1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{7}$$

Essa última igualdade nos ajuda a perceber que $(I - P)\boldsymbol{x} = \boldsymbol{0}$ nada mais é que um sistema de equações lineares homogêneo, com matriz dos coeficientes igual a $I - P$ e incógnitas a, b e c. Talvez o leitor esteja mais acostumado com a seguinte forma:

$$\begin{cases} a - pb + (p-1)c = 0 \\ (p-1)a + b - pc = 0 \\ -pa + (p-1)b + c = 0 \end{cases}.$$

Substituindo a equação 3, do sistema de equações acima, pelo resultado da soma da equação 3 com a equação 2:

$$\begin{cases} a-pb+(p-1)c=0 \\ (p-1)a+b-pc=0 \\ -a+pb+(1-p)c=0 \end{cases}. \tag{8}$$

Note que as equações 1 e 3 do sistema (8) são equivalentes. Logo, (8) equivale ao seguinte sistema:

$$\begin{cases} a-pb+(p-1)c=0 \\ (p-1)a+b-pc=0 \end{cases}. \tag{9}$$

Substituindo a equação 2 de (9) pelo resultado da soma da equação 1 com o produto da equação 2 por p:

$$\begin{cases} a-pb+(p-1)c=0 \\ (p^2-p+1)a+(-p^2+p-1)c=0 \end{cases}. \tag{10}$$

Dividindo ambos os membros da equação 2 de (10) por p^2-p+1:

$$\begin{cases} a-pb+(p-1)c=0 \\ a-c=0 \end{cases}. \tag{11}$$

Do sistema (11) é fácil concluir que $a=c$ e $c=b$. Dessa forma, a solução do nosso sistema é $a=b=c=t$, com $t\in\mathbb{R}$. Ou seja, as soluções de (7) são todos os vetores da forma:

$$t\cdot\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

No entanto, estamos procurando por um vetor de probabilidade, isto é, $a+b+c=1$. Então,

$$a+b+c=1\Rightarrow t+t+t=1\Rightarrow 3t=1\Rightarrow t=\frac{1}{3}.$$

Portanto, o vetor de de estado estacionário é igual $\begin{bmatrix}1/3 & 1/3 & 1/3\end{bmatrix}^T$ para qualquer $p\in(0,1)$. O que nos permite escrever que

$$\begin{bmatrix}0 & p & 1-p\\ 1-p & 0 & p\\ p & 1-p & 0\end{bmatrix}\begin{bmatrix}1/3\\1/3\\1/3\end{bmatrix}=\begin{bmatrix}1/3\\1/3\\1/3\end{bmatrix}.$$

Com isso solucionamos o problema da rã Dõ, isto é, mostramos que, conforme o número de saltos tende ao ifinito, a probabilidade dela estar no vértice A é de $1/3$, assim como nos demais vértices.

Analisando mais um pouco percebemos que, quando calculamos as soluções para $P\boldsymbol{x} = \boldsymbol{x}$, que nos permitiram chegar ao vetor de estado estacionário do nosso processo, estávamos calculando um autovetor da matriz P associado ao autovalor 1. De forma geral:

Definição 3. *Considere P uma matriz quadrada $n \times n$ e um vetor $\boldsymbol{v} \in \mathbb{R}^n$, não nulo. Se existe um escalar λ de maneira que $A\boldsymbol{v} = \lambda\boldsymbol{v}$, então dizemos que λ é um autovalor de P e $\boldsymbol{v}$ é um autovetor de P associado ao autovalor λ.*

Ademais, procurar por um autovetor $\boldsymbol{v}$ associado a um autovalor λ, conhecido, é um processo bem mais simples do que procurar pelos autovalores, pois, nesse caso, eles são as raízes do seguinte polinômio, chamado de polinômio característico:

$$p(\lambda) = det(P - \lambda I).$$

E, como sabemos, eles nem sempre possuem solução, além de ser um processo complicado, já que não existe uma fórmula para polinômios de grau maior do que ou igual a cinco.

1.1.2 A relação com o Algoritmo *PageRank*.

No problema da rã queríamos mostrar que, para $n \to \infty$, a probabilidade dela estar no vértice A seria de $1/3$. Verdade seja dita, o problema já fornecia a resposta final, isto é, já sabíamos que tal probabilidade seria de $1/3$. No entanto, em nenhum momento durante a nossa resolução fizemos uso desse valor. Para nós é como se a questão não houvesse informado que tal probabilidade é de $1/3$. Claro, deve existir um outro meio de resolver esse problema, talvez aplicando os conceitos de limite. Por exemplo, poderíamos trabalhar com algo do tipo: $\lim_{n\to\infty} a_n - 1/3 = 0$. O fato é que não precisamos saber "a resposta final" para aplicarmos o nosso método.

O que fizemos, desde o início, foi modelar o comportamento da rã Dõ sobre o diagrama da Figura 1.1, supondo que os saltos ocorressem de forma permanente. Ao final, encontramos como resultado um vetor de estado estacionário em que cada uma de suas entradas corresponde a probabilidade da rã está, no longo prazo, sobre o respectivo vértice. E isso é exatamente uma das definições intuitivas do algoritmo *PageRank*.

O valor de importância que esse algoritmo atribui a cada uma das páginas da *Web* é, essencialmente, a probabilidade de um usuário arbitrário visitar aquela página no

longo prazo. E como o algoritmo *PageRank* encontra essa probabilidade? Ele, basicamente, modela o comportamento desse usuário, supondo um passeio aleatório, pelo grafo da *Web*.

Mas que grafo é esse? É algo parecido com a Figura 1.1, vamos defini-lo no próximo capítulo, por enquanto imagine essa figura como uma *Web* que possui apenas três páginas, representadas por cada um dos vértices. As setas são representações dos *links* entre as páginas, mais precisamente os *forward links* e *backlinks*. Estes são, tomando a página B como exemplo, os *links* que chegam à ela; aqueles, os *links* que partem dela para as demais. Assim, a página B possui dois *backlinks* e dois *forward links*. Esses *links* permitem que, a cada clique, o usuário transite entre as páginas. Dessa forma, a entrada $(P^n)_{ij}$ representará a probabilidade de um usuário arbitrário ir da página j à página i em n cliques.

O fato é que as entradas do vetor de estado estacionário que encontraremos poderão ser utilizadas como valores de importância de cada uma das respectivas páginas. Tais valores recebem o nome de *PageRanks*.

2 RANQUEANDO AS PÁGINAS DA *WEB*

Neste capítulo vamos caminhar com o objetivo de modelar o algoritmo *PageRank*. Mas, antes apresentaremos um objeto matemático para representar a *Web*, ou melhor, para estudar a estrutura de *links* da *Web*. Além disso, no decorrer do texto serão necessários cálculos envolvendo matrizes de ordens superiores, iremos poupar o leitor de tais cálculos, mas tudo que faremos pode ser reproduzido em ferramentas como *Matrix Calculator*, que se encontra disponível em: <https://matrixcalc.org/pt/vectors.html>. Por fim, o desenvolvimento deste capítulo foi baseado nas obras (LEHMAN; LEIGHTON; MEYER, 2010), (POOLE, 2014), (BRYAN; LEISE, 2006), (SANTIAGO, 2021), (BRIN; PAGE, 1998) e (MALAJOVICH, 2021).

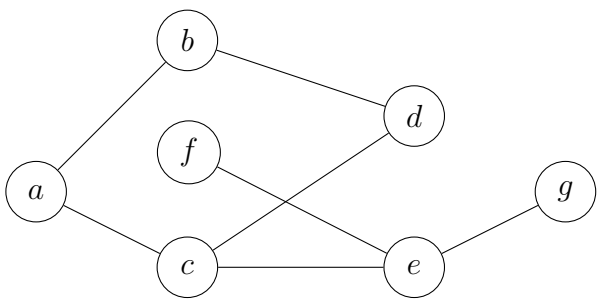
2.1 O GRAFO DA *WEB*

Nesta primeira seção nosso objetivo será estabelecer um modelo que possa ser utilizado para representar a *Web*, ou pelo menos um versão simplificada. Tal modelo, além de nos permitir "enxergá-la", nos permitirá entender algumas de suas características. Para o leitor interessado em um estudo mais profundo, recomendamos (BRODER *et al.*, 2011).

Em um primeiro momento podemos imaginar que essa tarefa é bem complexa, posto que existem páginas com tamanhos variados e conteúdos diversos. No entanto, uma vez que estamos interessados apenas na relação entre as páginas estabelecidas pelos *links*, ou seja, a estrutura de *links* da *Web*, se torna algo simples.

Os grafos são um ótimo objeto matemático que podemos utilizar, pois são uma maneira simples de representar relações entre pares de objetos. Mas, o que são grafos? De uma forma bem grosseira, são conjuntos de pontos, chamados de nós ou vértices, e um conjunto de linhas que conectam alguns desses pares de pontos, indicando alguma espécie de relação, chamadas de arestas. Um exemplo é mostrado na Figura 2.1.

Figura 2.1 – Grafo com 7 vértices e 7 arestas



Fonte: (LEHMAN; LEIGHTON; MEYER, 2010), adaptado pelo autor.

Os vértices representam objetos do nosso interesse, tais como cidades, pessoas, programas, isto é, dependem do contexto do nosso estudo. Como estamos trabalhando com a *Web*, eles representarão as páginas. Além disso, em alguns dos pares de vértices (páginas) existem arestas. Elas são utilizadas, como falamos, para indicar alguma relação entre esses vértices. Em nosso caso, estaremos relacionando as arestas com os *links*. Mas, antes de apresentarmos o grafo da *Web*, vamos formalizar algumas noções dessa teoria.

2.1.1 Grafos e Grafos direcionados

Definição 4 (Grafo)**.** *Um grafo G é uma estrutura matemática constituída por um conjunto V, não vazio, cujos elementos são chamados de vértices (ou nós) de G e um conjunto E de pares de elementos de V, cujos elementos são chamados de arestas de G e escrevemos $G = (V, E)$.*

Por exemplo, no grafo $G = (V, E)$ da Figura 2.1 temos:

$$\begin{aligned} V &= \{a, b, c, d, e, f, g\} \\ E &= \{\{a, b\}, \{a, c\}, \{b, d\}, \{c, d\}, \{c, e\}, \{e, f\}, \{e, g\}\} \end{aligned}$$

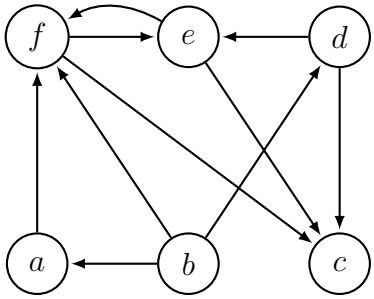
Perceba que as arestas $\{a, b\}$ e $\{b, a\}$ são as mesmas, visto que são conjuntos não ordenados. Se considerarmos que os vértices desse grafo são as representações de alguns pontos importantes de uma cidade e as arestas as vias entre esses pontos, significaria que poderíamos transitar, com algum veículo, por exemplo, do ponto a até o ponto b e do ponto b até o ponto a pela mesma via, ou seja, as vias seriam de mão dupla. No entanto, há situações onde isso não é possível, isto é, pode acontecer que algumas das vias não possuam mão dupla, permitindo que o fluxo ocorra em apenas um sentido. Aliás, isso ocorre, de fato, na *Web*. Uma página A pode possuir um *link* que leve o usuário à página B, por outro lado, a página B pode não possuir um *link* que leve o usuário à página A. Nesses casos, para que o grafo consiga representar a situação fielmente, as arestas precisam ser direcionadas, indicando os percursos permitidos, ou melhor, as relações permitidas. Grafos com essas características recebem o nome de grafo direcionado.

Um grafo com arestas direcionadas é chamado de grafo direconado ou digrafo. A definição é bem semelhante com a Definição 4, veja:

Definição 5 (Grafo direcionado)**.** *Um grafo direcionado G (ou digrafo) é uma estrutura matemática constituído por um conjunto de vértices V, não vazio, e um conjunto de arestas direcionadas E. Cada aresta direcionada de E é representada por um par ordenado de vértices $u, v \in V$.*

Nós podemos representar uma aresta direcionada como um par ordenado de vértices u, v e denotá-la por (u, v) ou $u \to v$, se a direção for de u para v.

Figura 2.2 – Grafo direcionado com 6 vértices e 10 arestas



Fonte: Autoria própria, 2022.

Por exemplo, no grafo direcionado $G = (V, E)$ da Figura 2.2, temos:

$$\begin{aligned} V &= \{a, b, c, d, e, f\} \\ E &= \{(a, f), (b, a), (b, d), (b, f), (d, c), (d, e), (e, c), (e, f), (f, c), (f, e)\} \end{aligned}$$

Em um grafo direcionado podemos definir a noção de grau de saída e grau de entrada. O grau de saída de um vértice u, denotado por $d^+(u)$, corresponde ao número de arestas distintas $(u, v_1), \ldots, (u, v_k)$. O grau de entrada do vértice u, denotado por $d^-(u)$, corresponde ao número de arestas distintas $(v_1, u), \ldots, (v_k, u)$. Por exemplo, o grau de saída do vértice f da Figura 2.2 é $d^+(f) = 2$, que corresponde as arestas (f, c) e (f, e). Já o grau de entrada de f é $d^-(f) = 3$, que corresponde as arestas (a, f), (b, f) e (e, f).

Definição 6 (Sumidouro e fonte)**.** *Quando o grau de saída de um vértice u é igual a zero, dizemos que u é um sumidouro. Por outro lado, quando o grau de entrada de um vértice u é igual a zero, dizemos que u é uma fonte.*

Na Figura 2.2 temos que o vértice b é uma fonte e o vértice c é um sumidouro.

Podemos, também, definir as noções de passeio e caminho.

Definição 7 (Passeio)**.** *Um passeio em um grafo direcionado G é uma sequência de vértices $v_0, v_1, \ldots, v_k$ e arestas $(v_0, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k)$ tais que (v_{i-1}, v_k) é uma aresta de G para todo i, onde $0 \leq i < k$.*

Por exemplo, na Figura 2.2, $(a, f), (f, e), (e, f), (f, c)$ é um passeio. Por outro lado, $(a, f), (f, e), (e, f), (f, c), (c, d)$ não é um passeio, pois (c, d) não é uma aresta desse grafo.

Definição 8 (Caminho)**.** *Um caminho em um grafo direcionado G é um passeio onde os vértices desse passeio são todos distintos.*

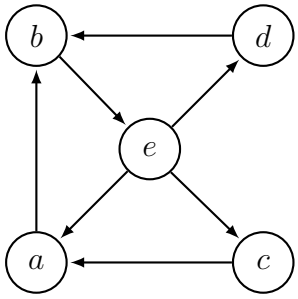
Note que, embora $(a, f), (f, e), (e, f), (f, c)$ seja um passeio, ele não é um caminho, pois o vértice f é visitado duas vezes, ou seja, nesse passeio existem vértices iguais.

Por fim, apresentamos a definição de grafo direcionado fortemente conectado.

Definição 9 (Grafo fortemente conectado)**.** *Um grafo direcionado G é chamado de fortemente conectado se, para todo par de vértices $u, v \in V$, existe um caminho direcionado de u para v em G.*

Ou seja, a partir de qualquer um dos vértices podemos visitar todos os outros. A Figura 2.3 é um exemplo de grafo fortemente conectado.

Figura 2.3 – Grafo direcionado fortemente conectado



Fonte: (LESKOVEC, 2019), adaptado pelo autor.

Podemos, então, estudar a estrutura de *links* da *Web* através de grafos direcionados. Cada vértice v de um grafo direcionado G, que agora passamos a chamar de grafo da *Web*, representa uma página. As arestas direcionadas são as relações estabelecidas pelos *links* entre os pares de páginas, mais precisamente os *forward links* e *backlinks*. Aqueles são os *links* que a página v possui e que levam o usuário a outras páginas, estes, os *links* que chegam a página v de outras páginas, e fazem com que os usuários visitem a página v. Perceba que o número de *backlinks* de uma página é igual ao seu grau de entrada e o número de *forwards links* é igual ao grau de saída.

2.2 O ALGORITMO *PAGERANK*

Nesta seção começaremos, de fato, a modelar o algoritmo *PageRank*. Para tal, apresentamos, primeiramente, uma das principais noções que norteiam toda essa busca por uma forma de se atribuir um valor de importância a cada página da Web: *links* como votos.

2.2.1 *Links* como votos

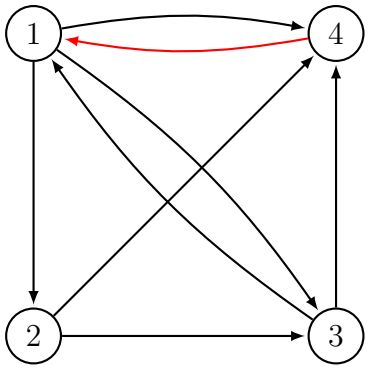
No Capítulo 1 apresentamos uma das noções intuitivas do algoritmo *PageRank*. Nele afirmamos que esse valor de importância é quantitativo, ou seja, um número. No entanto, mesmo que não soubéssemos disso seria natural pensarmos em atribuir valores numéricos a cada uma das páginas, já que eles possuem relação de ordem. Por exemplo, dados os números $0,35$ e $0,4$ sabemos que $0,4 > 0,35$. Isso implicaria que uma página com valor de importância $0,4$ seria mais importante que a página com o valor $0,35$. Portanto, se conseguíssemos encontrar uma forma de se atribuir um valor numérico a cada uma das páginas da Web poderíamos utilizá-los para ordená-las. Agora, qual deve ser o nosso ponto de partida?

Já que estamos trabalhando com a estrutura de *links* da *Web*, um bom ponto de partida é pensar que a importância de cada uma das páginas deriva dos seus *links*, mais precisamente os *backlinks*. Em verdade, essa foi a hipótese que Page e Brin levantaram quando estavam estudando a estrutura de *links* da *Web*.

Uma primeira ideia de um possível método seria atribuir a cada uma das páginas, como valor de importância, o seu número de *backlinks*, ou seja, o número de *links* que ela recebe de outras páginas. Para deixar essa ideia ainda mais intuitva, podemos pensar nos *backlinks* como votos. Então, uma página recebe alguns votos (*backlinks*) e também pode votar em outras páginas (*forward links*). A *Web* se torna uma espécie de democracia, onde páginas votam para a importância das outras (BRYAN; LEISE, 2006). Sendo assim, vamos aplicar esse método no grafo da *Web* da Figura 2.4. Mas, primeiro estabeleceremos algumas notações.

Cada uma das n páginas (vértices) do grafo da *Web* será indexada por um inteiro k, onde $1 \leq k \leq n$, tal como a Figura 2.4. Utilizaremos a notação x_k para representar o valor de importância da página k. Analogamente, d_k e b_k representarão, respectivamente,

Figura 2.4 – *Web* de quatro páginas



Fonte: (BRYAN; LEISE, 2006), adaptado pelo autor.

o número de *forward links* e *backlinks* da página k. Por exemplo, na página 2 da *Web* da Figura 2.4 temos que $d_2 = 2$ e $b_2 = 1$.

Como falamos, nossa ideia inicial é considerar x_k igual ao o número de *backlinks* da página k. Quanto mais votos uma página recebe, mais importante ela se torna. Assim, na Figura 2.4 temos que $x_1 = 2$, $x_2 = 1$, $x_3 = 2$ e $x_4 = 3$. Isso implica que a página 4 é a mais importante, as páginas 1 e 3 têm a mesma importância e, por fim, a página 2 é a menos importante. Então, um motor de busca poderia considerar a ordem em que as páginas aparecem em seus resultados de pesquisa baseada na classificação da Tabela 2.1.

Tabela 2.1 – Classificação 1

Classificação	Página	Valor de importância
1º	4	3
2º	1 e 3	2
3º	2	1

Contudo, perceba que a página 1 possui um *backlink* da página mais importante. Em outras palavras, a página mais importante vota na página 1. Isso deveria tornar a página 1 mais importante que a página 3, concorda? Estamos levando em consideração que *links* de páginas importantes têm um peso maior. Isso vai ao encontro de que o usuário não irá preferir apenas as páginas com o maior número de *backlinks*, mas, também, as páginas que possuem *backlinks* de qualidade, de páginas importantes. No entanto, o nosso método ignora essa característica. Aliás, ele seria facilmente enganado, visto que poderíamos criar novas páginas, propositalmente, e, em cada uma, adicionar um *forward link* para a página que quiséssemos aumentar a importância. Assim, votos de páginas irrelevantes teriam o mesmo peso que votos de páginas relevantes.

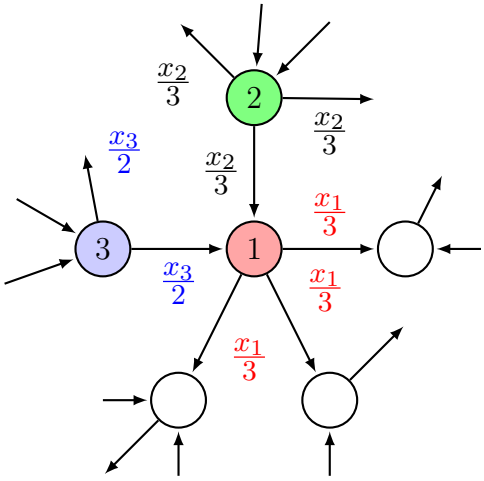
Sendo assim, esperamos que um bom método de ranqueamento não leve em consideração apenas o número de *backlinks* das páginas, mas, também, que um *backlink* de uma página importante deve ter um peso maior do que um *backlink* de uma página menos importante. Ou seja, passamos a nos preocupar, além da quantidade, com a qualidade dos *links*.

2.2.2 Aprimorando o método

Uma maneira de traduzir essa nova ideia matematicamente é considerar que cada voto tem um peso proporcional à importância da página de origem. Por exemplo, se a página k, cuja importância é x_k, possui d_k *forward links*, então cada um deles terá um peso igual a $\frac{x_k}{d_k}$. Do ponto de vista intuitivo, isso significa que a página k doará $\frac{1}{d_k}$ de sua

importância, x_k, para cada uma das páginas que ela possuir um *forward link*. A Figura 2.5 deixa essa ideia mais clara.

Figura 2.5 – Peso dos *links*



Fonte: (LESKOVEC, 2019), adaptado pelo autor.

A importância da página k ainda continuará dependendo da soma de seus *backlinks*. No entanto, agora eles possuem pesos diferentes, o que não nos permiti fazer apenas $x_k = b_k$. Então, por exemplo, o valor de importância da página 1 da Figura 2.5 será igual a $x_1 = \frac{x_2}{3} + \frac{x_3}{2}$. Generalizando,

$$x_k = \sum_{i \in F} \frac{x_i}{d_i},$$

onde F é o conjunto das páginas que possuem *forward links* para a página k, ou ainda, o conjunto das páginas que votam na página k.

Agora vamos aplicar esse novo método na *Web* da Figura 2.4 e ver o acontece:

$$\begin{aligned} x_1 &= \frac{x_3}{2} + x_4 \\ x_2 &= \frac{x_1}{3} \\ x_3 &= \frac{x_1}{3} + \frac{x_2}{2} \\ x_4 &= \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_3}{2} \end{aligned} .$$

Podemos escrever esse sistema em sua forma matricial:

$$\begin{bmatrix} 0 & 0 & 1/2 & 1 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 0 \\ 1/3 & 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} . \tag{12}$$

Essa equação parece familiar? Exatamente! Caímos no mesmo problema da rã Dõ, isto é, encontrar um vetor de estado estacionário de uma matriz estocástica, ou

ainda, encontrar o autovetor associado ao autovalor 1. O que nos leva a perguntar se o que estamos fazendo é procurar pelas probabilidades de um usuário arbitrário visitar cada uma das páginas da *Web* no longo prazo. Nesse caso, sim, podemos considerar que estamos procurando por essas probabilidades. Aliás, essa é uma das noções intuitivas do algoritmo *PageRank* que apresentamos no Capítulo 1. Neste momento o leitor pode estar se perguntando se poderíamos aplicar, diretamente, o mesmo método que utilizamos para solucionar o problema da rã Dõ. A resposta, novamente, é sim. No entanto, estaríamos escondendo a beleza do processo que é traduzir ideias em linguagem matemática. Posto isso, vamos continuar com a solução.

As componentes dos vetores que satisfazem a equação (12) são

$$x_1 = \frac{4t}{3}, x_2 = \frac{4t}{9}, x_3 = \frac{2t}{3} \quad \text{e} \quad x_4 = t, \ \text{com} \quad t \in \mathbb{R}.$$

Isto é, todos os vetores da forma

$$\begin{bmatrix} 4t/3 \\ 4t/9 \\ 2t/3 \\ t \end{bmatrix}.$$

No entanto, no problema do Capítulo 1, vimos que $x_1 + x_2 + x_3 + x_4 = 1$, isto é, estamos procurando por um vetor de probabilidade. Aliás, é algo bem intuitivo, pois podemos considerar que a soma das importâncias é igual a $100\% = 1$. Dessa forma,

$$x_1 + x_2 + x_3 + x_4 = 1 \Rightarrow \frac{4t}{3} + \frac{4t}{9} + \frac{2t}{3} + t = 1 \Rightarrow t = \frac{9}{31}.$$

O que nos leva aos seguintes valores de importância:

$$x_1 \approx 0,387, x_2 = 0,129, x_3 \approx 0,194 \quad \text{e} \quad x_4 \approx 0,290.$$

Esses valores estão ordenados na Tabela 2.2.

Tabela 2.2 – Classificação 2

Classificação	Página	Valor de importância
1º	1	0,387
2º	4	0,290
3º	3	0,194
4º	2	0,129

Quando comparamos com a Tabela 2.1 vemos que, agora, o valor de importância da página 1 é maior do que o da página 4, embora esta última tenha mais *backlinks*. Uma das explicações para isso é que a página 4 possui apenas um *forward link*, e o mesmo é direcionado para a página 1. Com isso, ao contrário das outras páginas, ela doa 100% de sua importância para a página 1. Do ponto de vista probabilístico, quando o usuário está na página 4 a chance dele ir para a página 1 é de 100%, consequentemente, a página 1 terá uma maior chance de ser visitada no longo prazo.

Com isso, parece que o método para ranquear as páginas da *Web* está mais razoável. Aliás, observando a equação (12), ele se resume a encontrar o autovetor associado ao autovalor 1 de uma matriz. Isso nos motiva a definir uma forma mais prática para encontrar a matriz a qual pretendemos calcular o autovetor associado ao autovalor 1. Pois, no exemplo anterior, tivemos que calcular, a partir de (2.2.2), cada um dos valores de importância e em seguida converter em uma equação matricial. Sendo assim, vamos denominá-la de matriz da *Web* e representá-la pela letra W. Então, definiremos a matriz da *Web* da seguinte forma:

$$W_{ij} = \begin{cases} \frac{1}{d_j}, & \text{se } j \to i \\ 0, & \text{caso contrário} \end{cases} . \tag{13}$$

Ou seja, atribuiremos o valor $1/d_j$ à entrada W_{ij} se a página j possuir um *forward link* para a página i e zero caso contrário.

Se utilizarmos (13) para encontrar a matriz da *Web* da Figura 2.4 chegaremos ao seguinte resultado:

$$W = \begin{bmatrix} 0 & 0 & 1/2 & 1 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 0 \\ 1/3 & 1/2 & 1/2 & 0 \end{bmatrix},$$

que é exatamente a matriz que encontramos quando representamos o sistema de equações (2.2.2) através de uma equação matricial. Portanto, reforçando a observação que fizemos anteriormente, o método se resume a encontrar o autovetor de W associado ao autovalor 1.

A partir daqui podemos fazer alguns questionamentos, como, por exemplo: *a matriz da Web sempre possuirá um autovetor associado ao autovalor 1?* Nota-se ainda que, no caso da *Web* da Figura 2.4, a matriz W é estocástica. O que motiva um outro questionamento: *a matriz W sempre será estocástica?*

Em geral, para os casos em que a matriz W é estocástica, sempre conseguiremos encontrar um autovetor de W associado ao autovalor 1. Para demonstrar esse fato, primeiro

demonstrarmos dois lemas:

Lema 1. *Seja* $\boldsymbol{u}$ *um vetor linha com todas as entradas iguais a* 1*. Se* P *é uma matriz estocástica, então* $\boldsymbol{u}P = \boldsymbol{u}$*.*

Demonstração. Considere a seguinte matriz estocástica P de ordem n:

$$P = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

Assim,

$$\begin{aligned} \boldsymbol{u}P &= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11}+a_{21}+\cdots+a_{n1} & a_{12}+a_{22}+\cdots+a_{n2} & \cdots & a_{1n}+a_{2n}+\cdots+a_{nn} \end{bmatrix} \\ &= \boldsymbol{u}. \end{aligned}$$

□

Lema 2. *Seja* P *uma matriz quadrada arbitrária, então* P *e* P^T *têm o mesmo polinômio característico e, dessa forma, os mesmos autovalores.*

Demonstração. Para qualquer matriz P vale que

$$\det(P) = \det(P^T).$$

Além disso, os autovalores, λ, são as soluções da seguinte equação:

$$\det(P - \lambda I) = 0.$$

Como,

$$(P - \lambda I)^T = P^T - \lambda I^T = P^T - \lambda I,$$

segue que,

$$\det(P - \lambda I) = \det(P^T - \lambda I).$$

Portanto, as matrizes P e P^T têm o mesmo polinômio característico e, consequentemente, os mesmos autovalores. □

Teorema 1. *Se P é uma matriz estocástica, então 1 é autovalor de P.*

Demonstração. Como P é uma matriz estocástica, sabemos, do *Lema 1*, que $\boldsymbol{u}P = \boldsymbol{u}$. Com isso,

$$(\boldsymbol{u}P)^T = \boldsymbol{u}^T \Rightarrow P^T\boldsymbol{u}^T = \boldsymbol{u}^T.$$

Ou seja, $\boldsymbol{u}^T$ é um autovetor de P associado ao autovalor 1. Portanto, do *Lema 2*, temos que 1 também é autovalor de P. □

Assim, mostramos que uma matriz estocástica sempre possui 1 como autovalor. Ou ainda, dada uma matriz estocástica P sempre conseguiremos encontrar um vetor de probabilidade $\boldsymbol{x}$ com a propriedade de que $P\boldsymbol{x} = \boldsymbol{x}$, definido como vetor de estado estacionário.

No entanto, há situações em que a matriz W não é estocástica. Isso ocorre quando o grafo da *Web* possue sumidouros, isto é, páginas sem *forward links*.

2.2.2.1 *Web* com sumidouros

A matriz da *Web* da Figura 2.6 é

$$W = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

Perceba que a coluna correspondente à página 3 tem todas as entradas iguais a zero. Ela é justamente o sumidouro da nossa *Web*.

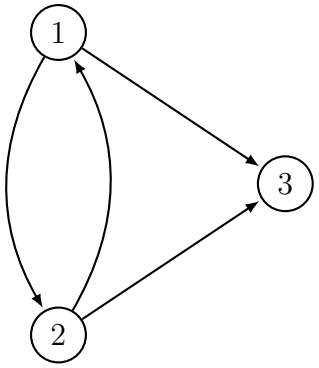
Além disso, o único vetor que satisfaz a equação matricial

$$\begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

é o vetor nulo. Isso é consequência do fato de que, nesse caso, 1 não é autovalor da matriz W.

Infelizmente esse não é o único problema. Há casos em que W, mesmo sendo estocástica, apresenta mais de um vetor de estado estacionário. Nessa situação, haveria, no mínimo, dois vetores possíveis para os valores de importância. Estamos nos referindo aos casos em que o grafo da *Web* é desconexo. Em outros termos, a *Web* é constituída de *subwebs*.

Figura 2.6 – *Web* com sumidouro



Fonte: Autoria própria, 2022.

2.2.2.2 *Web* constituída de *subwebs*.

A matriz da *Web* da Figura 2.7 é

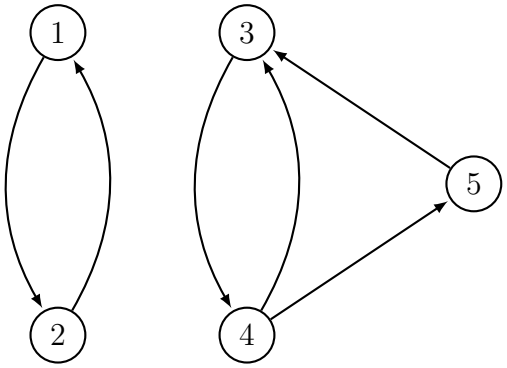
$$W = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix}.$$

Em que dois possíveis vetores de estado estacionário são:

$$\boldsymbol{u} = \begin{bmatrix} 0 \\ 0 \\ 2/5 \\ 2/5 \\ 1/5 \end{bmatrix} \quad \text{e} \quad \boldsymbol{v} = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Qual deles usar como valores de importância das páginas da *Web* da Figura 2.7? As páginas que $\boldsymbol{u}$ considera como importantes $\boldsymbol{v}$ considera com nenhuma importância e vice-versa.

Figura 2.7 – *Web* com *subwebs*



Fonte: (BRYAN; LEISE, 2006), adaptado pelo autor.

O problema da *Web* com sumidouros é fácil de resolver. Na prática, se um usuário arbitrário começa a clicar em *links* aleatórios, provavelmente ficará preso em uma página sem *forward links*. Quando isso acontece, tendemos a fazer uma nova busca ou até mesmo digitar um endereço de uma página específica, o que nos faz sair do sumidouro, é como se o usuário se teleportasse no grafo da *Web*. E isso pode ser traduzido na matriz da *Web* da seguinte forma: substituímos a coluna de zeros, que corresponde ao sumidouro, por uma nova coluna com todas as entradas iguais a $1/n$, onde n corresponde ao total de páginas. No caso da matriz W, que corresponde à *Web* da Figura 2.6, ficaríamos com a seguinte matriz W':

$$W' = \begin{bmatrix} 0 & 1/2 & 1/3 \\ 1/2 & 0 & 1/3 \\ 1/2 & 1/2 & 1/3 \end{bmatrix}.$$

Cujo vetor de estado estacionário é $\boldsymbol{x} = \begin{bmatrix} 2/7 & 2/7 & 3/7 \end{bmatrix}^T$.

Já o problema apresentado pela *Web* da Figura 2.7 é um pouco mais delicado. Embora a matriz seja estocástica, ela possui dois vetores, linearmente independentes, de estado estacionário, ou melhor, qualquer vetor $\boldsymbol{w}$ da forma

$$\begin{bmatrix} t_1 \\ t_1 \\ 2t_2 \\ 2t_2 \\ t_2 \end{bmatrix},$$

com $2t_1 + 5t_2 = 1$, pode ser considerado como um vetor de estado estacionário daquela matriz.

Perceba que, quando provamos o Teorema 1, apenas mostramos que toda matriz estocástica possui um vetor de estado estacionário. No entanto, em nenhum momento mencionamos que ele seria único. Então, em que condições uma matriz estocástica possui um único vetor de estado estacionário? Isso vai ocorrer quando o subespaço associado ao autovalor 1 da matriz W, que denotamos por $V_1(W)$, tem dimensão 1.

O subespaço $V_1(W)$ nada mais é do que o conjunto de todos os autovetores $\boldsymbol{v}$ que sastifazem $W\boldsymbol{v} = \boldsymbol{v}$. Já a dimensão desse subespaço é definida como sendo igual ao número de vetores de sua base. Mas, o que seria uma base? De uma forma bem resumida, uma base de $V_1(W)$ é um conjunto de vetores $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_n}\} \subset V_1(W)$ que satisfazem duas condições: primeiro, ele gera $V_1(W)$, ou ainda, qualquer vetor do conjunto $V_1(W)$ pode ser escrito como uma combinação linear de $\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_n}$; segundo, o conjunto $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_n}\}$

é linearmente independente, isto é, a equação $a_1\boldsymbol{v_1} + a_2\boldsymbol{v_2} + \ldots + a_n\boldsymbol{v_n} = 0$ possui como única solução $a_1 = a_2 = \ldots = a_n = 0$.

Como exemplo, podemos pegar o subespaço $V_1(W)$ da *Web* da Figura 2.4. Nesse caso, vimos que todos os vetores $\boldsymbol{v}$ que satisfazem $W\boldsymbol{v} = \boldsymbol{v}$ são da forma

$$t \cdot \begin{bmatrix} 4/3 \\ 4/9 \\ 2/3 \\ 1 \end{bmatrix}.$$

Assim, a base de $V_1(W)$ possui um único vetor, isto é, a dimensão de $V_1(W)$ é igual a 1. E, em geral, para grafos da *Web* fortemente conectados isso sempre será verdade.

Por outro lado, no exemplo da *Web* da Figura 2.7, os vetores que sastifazem $W\boldsymbol{v} = \boldsymbol{v}$ são da forma

$$\begin{bmatrix} t_1 \\ t_1 \\ 2t_2 \\ 2t_2 \\ t_2 \end{bmatrix} = t_1 \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + t_2 \cdot \begin{bmatrix} 0 \\ 0 \\ 2 \\ 2 \\ 1 \end{bmatrix}.$$

Note que $\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \end{bmatrix}^T$ e $\begin{bmatrix} 0 & 0 & 2 & 2 & 1 \end{bmatrix}^T$ são LI. Logo, a base de $V_1(W)$ possui dois vetores e, consequentemente, possui dimensão 2. O que faz com que a matriz W tenha mais do que um vetor de estado estacionário.

A *Web* possui diversos problemas, neste trabalho mencionamos apenas dois (*subwebs e sumidouros*). Na próxima seção analisaremos a solução que Larry Page e Sergey Brin encontraram para resolver essas deficiências, em especial o problema das *subwebs*.

2.2.3 Uma última modificação

O problema da *Web* com *subwebs* possui uma certa semelhança com o problema da *Web* com sumidouros. No entanto, dessa vez o usuário não ficará preso em uma página específica, mas em um conjunto de páginas. Para resolver aplica-se, novamente, a noção de teletransporte no grafo da *Web*, mas de uma forma mais elaborada.

Page e Brin inseriram uma nova constante $0 \leq p \leq 1$, que foi denominada de fator de amortecimento (*damping factor*). Então, definiram uma nova matriz G, denominada de matriz do Google, da seguinte forma:

$$G = (1-p) \cdot W + p \cdot A,$$

onde A é uma matriz com a mesma ordem de W, isto é, de ordem n, e igual a

$$A = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}.$$

Esse novo modelo indica que, em cada período de tempo, um usuário aleatório tem duas opções: com probabilidade $1-p$, ele segue um link aleatório da página em que se encontra e, com probabilidade p, ele se teletransporta para uma página aleatória. Ao fator de amortecimento p, Page e Brin atribuíram o valor $0,15$. Ou seja, a estrutura original do grafo da *Web*, W, tem um peso maior, ou ainda, o usuário escolherá se teleportar com menos frequência.

Note que, desde que a matriz W seja estocástica, G também é uma matriz estocástica. Pois,

$$\begin{aligned} \sum_{i=1}^{n} G_{ij} &= (1-p)\sum_{i=1}^{n} W_{ij} + p\sum_{i=1}^{n} A_{ij} \\ &= (1-p)+p \\ &= 1. \end{aligned}$$

E, além disso, como $W_{ij} \geq 0$ e $A_{ij} > 0$ para todo i e j, pois são matrizes estocásticas, decorre que $(1-p)W_{ij} + pA_{ij} \geq 0$, ou seja, $G_{ij} \geq 0$. Em particular, $G_{ij} > 0$ para $p \in (0,1]$, isto é, G é positiva para $p \in (0,1]$.

Por fim, a dimensão de $V_1(G)$ é igual a 1 para $p \in (0,1]$, ou seja, a matriz G possui um único vetor de estado estacionário. Para provar esse fato, primeiro, provaremos dois lemas.

Lema 3. *Se G é uma matriz positiva e estocástica, então qualquer autovetor pertencente à $V_1(G)$ tem todas as componentes positivas ou todas as componentes negativas.*

Demonstração. Considere a seguinte desigualdade triangular:

$$\left|\sum_i y_i\right| \leq \sum_i |y_i|, \text{ com } y_i \in \mathbb{R}.$$

Perceba que teremos a desigualde estrita quando os sinais de y_i forem mistos. Isto é, possui tanto números positivos quanto negativos. Agora suponha, por contradição, que $\boldsymbol{x} \in V_1(G)$, onde $\boldsymbol{x} = \begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_n \end{bmatrix}^T$, é um vetor com as componentes com sinais mistos. Da equação $G\boldsymbol{x} = \boldsymbol{x}$, segue que $x_i = \sum_{j=1}^{n} G_{ij}x_j$ e as somas $G_{ij}x_j$ são de sinais mistos, desde

que $G_{ij} > 0$. Da desigualdade trianglular (2.2.3) e sabendo que x_i possuem sinais mistos, decorre que

$$|x_i| = \left| \sum_{j=1}^{n} G_{ij} x_j \right| < \sum_{j=1}^{n} G_{ij} |x_j|.$$

Ou seja,

$$|x_i| < \sum_{j=1}^{n} G_{ij} |x_j|. \tag{14}$$

Aplicando o somatório $\sum_{i=1}^{n}$ à ambos os membros da desigualdade (14), tem-se

$$\sum_{i=1}^{n} |x_i| < \sum_{i=1}^{n} \sum_{j=1}^{n} G_{ij} |x_j|.$$

Ou ainda,

$$\sum_{i=1}^{n} |x_i| < \sum_{i=1}^{n} \sum_{j=1}^{n} G_{ij} |x_j| = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} G_{ij} \right) |x_j|.$$

Como G é estocástica, isto é, $\sum_i G_{ij} = 1$ para todo j, segue que

$$\sum_{i=1}^{n} |x_i| < \sum_{i=1}^{n} \sum_{j=1}^{n} G_{ij} |x_j| = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} G_{ij} \right) |x_j| = \sum_{j=1}^{n} |x_j|,$$

uma contradição. Assim, $\boldsymbol{x} \in V_1(G)$ possui todas as componentes positivas ou todas as componentes negativas. □

Lema 4. *Sejam $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, com $n \geq 2$, vetores linearmente independentes. Então, para algum α e β reais, não ambos iguais a zero, o vetor $\boldsymbol{x} = \alpha\boldsymbol{u} + \beta\boldsymbol{v}$ tem componentes positivas e componentes negativas.*

Demonstração. Como $\boldsymbol{u}$ e $\boldsymbol{v}$ são linearmente independentes, sabemos que $\boldsymbol{u} \neq \boldsymbol{0}$ e $\boldsymbol{v} \neq \boldsymbol{0}$, caso contrario, seriam linearmente dependentes. Agora, considere $d = \sum_{i=1}^{n} u_i$. Se $d = 0$, então $\boldsymbol{u}$ deve conter elementos de sinais mistos. Assim, fazendo $\alpha = 1$ e $\beta = 0$, podemos concluir que $\boldsymbol{x} = \alpha\boldsymbol{u} + \beta\boldsymbol{v}$ tem componentes positivas e componentes negativas. Por outro lado, se $d \neq 0$, tomamos $\alpha = -\frac{\sum_{i=1}^{n} v_i}{d}$ e $\beta = 1$, então

$$\boldsymbol{x} = -\frac{\sum_{i=1}^{n} v_i}{d} \cdot \boldsymbol{u} + \boldsymbol{v}.$$

Logo,

$$\sum_{i=1}^{n} x_i = -\frac{\sum_{i=1}^{n} v_i}{d} \cdot \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i = -\frac{\sum_{i=1}^{n} v_i}{d} \cdot d + \sum_{i=1}^{n} v_i = 0.$$

Ou seja, as somas das componentes de $\boldsymbol{x}$ é igual a zero. Como $\boldsymbol{u}$ e $\boldsymbol{v}$ são linearmente independentes, sabemos que $\boldsymbol{x} \neq \boldsymbol{0}$. Portanto, $\boldsymbol{x}$ possui componentes positivas e componentes negativas. □

A partir desses dois lemas podemos provar que a dimensão de $V_1(G)$ é igual a 1.

Teorema 2. *Se a matriz G é positiva e estocástica, então a dimensão de $V_1(G)$ é 1.*

Demonstração. Do Teorema 1 sabemos que 1 é autovalor de G, logo $\dim(V_1(G)) \geq 1$. Agora suponha, por contradição, que existem $\boldsymbol{u}$ e $\boldsymbol{v}$, linearmente independentes, pertencentes a $V_1(G)$. Sabemos que, para qualquer α e β reais, não ambos iguais a zero, o vetor $\boldsymbol{x} = \alpha\boldsymbol{u} + \beta\boldsymbol{v}$ deve pertencer a $V_1(G)$ e suas componentes são, pelo Lema 3, ou todas positivas, ou todas negativas. No entanto, pelo Lema 4, podemos encontrar α e β tais que o vetor $\boldsymbol{x}$ tenha componentes de sinais mistos, o que é uma contradição. Logo, $\dim(V_1(G)) < 2$. Implicando $\dim(V_1(G)) = 1$. □

De forma geral, para a matriz G vale o seguinte teorema:

Teorema 3 (Perron-Frobenius, caso Markoviano)**.** *Seja G uma matriz estocástica. Então,*

(i) Se λ é autovalor de G, então $|\lambda| \leq 1$;

(ii) 1 é autovalor de G;

(iii) Todo autovalor λ de G diferente de 1 verifica $|\lambda| < 1$;

(vi) Existe um autovetor à direita, associado ao autovalor 1, cujas coordenadas são todas não-negativas;

(v) Se G é positiva, então a dimensão de $V_1(G)$ é igual a 1.

Uma prova para esse Teorema pode ser consultada, detalhadamente, em (MALAJOVICH, 2021).

Antes de aplicarmos a matriz G, tal como foi definida, para calcular os valores de importância da *Web* constituída de *subwebs* da Figura 2.7, vamos aplicá-la na *Web* de quatro páginas da Figura 2.4.

Nesse exemplo a *Web* possui quatro páginas. Então todas as entradas da matriz A são iguais a 1/4, ou seja,

$$A = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}.$$

Logo,

$$G = (1-p)\begin{bmatrix} 0 & 0 & 1/2 & 1 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 0 \\ 1/3 & 1/2 & 1/2 & 0 \end{bmatrix} + p\begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}.$$

Vamos considerar a constante de amortecimento p igual a $0,15$, tal como Page e Brin propuseram. Assim, após fazer todos os cálculos, chegamos ao seguinte resultado:

$$G = \begin{bmatrix} 3/80 & 3/80 & 37/80 & 71/80 \\ 77/240 & 3/80 & 3/80 & 3/80 \\ 77/240 & 37/80 & 3/80 & 3/80 \\ 77/240 & 37/80 & 37/80 & 3/80 \end{bmatrix}.$$

Cujo único vetor de estado estacionário possui as seguintes entradas:

$$x_1 \approx 0,368,\ x_2 \approx 0,142,\ x_3 \approx 0,202 \ \text{ e } \ x_4 \approx 0,288.$$

Esses valores estão organizados na Tablea 2.3.

Tabela 2.3 – Classificação 3

Classificação	Página	Valor de importância
1º	1	0,368
2º	4	0,288
3º	3	0,202
4º	2	0,142

Quando comparamos com a Tabela 2.2, notamos que G não alterou a classificação das páginas, os valores de importância tiveram apenas uma leve variação.

Agora vamos utilizar G para calcular os valores de importância da *Web* da Figura 2.7.

Como ela é constituída por 5 páginas, segue que

$$A = \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}.$$

Assim,

$$G = (1-p)\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{bmatrix} + p\begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}.$$

Novamente, vamos considerar $p = 0,15$. Logo, a nossa matriz do *Google* é

$$= \begin{bmatrix} 3/100 & 22/25 & 3/100 & 3/100 & 3/100 \\ 22/25 & 3/100 & 3/100 & 3/100 & 3/100 \\ 3/100 & 3/100 & 3/100 & 91/200 & 22/25 \\ 3/100 & 3/100 & 22/25 & 3/100 & 3/100 \\ 3/100 & 3/100 & 3/100 & 91/200 & 3/100 \end{bmatrix}.$$

Onde o único vetor de estado estacionário possui as entradas iguais a

$$x_1 = 0,200,\ x_2 = 0,200,\ x_3 \approx 0,238,\ x_4 \approx 0,233 \ \text{ e } \ x_5 \approx 0,129.$$

2.3 CALCULANDO O *PAGERANK* NA PRÁTICA

Até aqui calculamos os *PageRanks* de *Webs* relativamente pequenas. No entanto, no momento em que estamos escrevendo esse trabalho, segundo a página (Internet Live Stats, 2022), a *Web* possui um pouco mais de $1,9$ bilhão de páginas. Calcular o vetor de estado estacionário de uma matriz de $1,9$ bilhão por $1,9$ bilhão é uma tarefa que certamente dará trabalho até mesmo aos melhores computadores. Então, como proceder?

De forma resumida, o *PageRank* é calculado através de iterações entre a matriz do *Google* e um vetor de estado inicial, normalmente com todas as entradas iguais a $\frac{1}{n}$, onde n é o número total de páginas da *Web*. Segundo (PAGE *et al.*, 1999), foram necessárias cerca de 52 iterações para chegar a uma boa aproximação para os *PageRanks* de um conjunto de páginas com cerca de 322 milhões de *links*.

Na tabela 2.4 expomos a quantidade de iterações necessárias para chegar aos mesmos valores de importância, considerando três casas decimais, que encontramos ao calcular o vetor de estado estacionário da matriz G da *Web* com *subwebs* (Figura 2.7). Para tal, consideramos o vetor de estado inicial $\boldsymbol{x_0} = \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}^T$ e $p = 0,15$.

Tabela 2.4 – Iterações entre a matriz G e o vetor de estado inicial $\boldsymbol{x_0}$

n	$G^n \cdot \boldsymbol{x_0}$	$\boldsymbol{x_n}$
1	$G^1 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,285 \;\; 0,200 \;\; 0,115]^T$
2	$G^2 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,213 \;\; 0,272 \;\; 0,115]^T$
3	$G^3 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,243 \;\; 0,211 \;\; 0,146]^T$
4	$G^4 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,243 \;\; 0,237 \;\; 0,120]^T$
5	$G^5 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,232 \;\; 0,237 \;\; 0,131]^T$
6	$G^6 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,242 \;\; 0,228 \;\; 0,131]^T$
7	$G^7 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,238 \;\; 0,236 \;\; 0,127]^T$
8	$G^8 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,238 \;\; 0,232 \;\; 0,130]^T$
9	$G^9 \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,239 \;\; 0,232 \;\; 0,129]^T$
10	$G^{10} \cdot \boldsymbol{x_0}$	$[0,200 \;\; 0,200 \;\; 0,238 \;\; 0,233 \;\; 0,129]^T$

Portanto, o algoritmo *PageRank* consiste em aplicar a matriz G sucessivas vezes a um vetor de probabilidade inicial para obter aproximações cada vez melhores do vetor de estado estacionário (SANTIAGO, 2021).

3 CONSIDERAÇÕES FINAIS

O nosso entorno está repleto de matemática. No entanto, muitas vezes, não é algo simples de perceber. Assim, para conseguirmos enxergá-la, sobretudo compreendê-la, precisamos fazer questionamentos, conjecturar, investigar. O interesse por este trabalho surgiu a partir do momento em que se questionamos como o *Google* escolhe, com precisão e eficiência, quais páginas mostrar para o usuário que realiza uma busca através dele. A partir daí, chegamos ao Algoritmo *PageRank*. Descobrimos que ele é o responsável por atribuir um valor, quantitativo, a cada página da *Web*. Este valor corresponde a importância da página e é utilizado para ranqueá-la em ordem de importância quando o *Google* apresenta os resultados de uma busca. E o que está por trás dele, ou melhor, o que está por trás do sucesso do *Google*? A resposta é simples: a matemática. Desde então o nosso objetivo passou a ser compreender e explicar, numa linguagem clara e didática, como a matemática está envolvida no processo de ranqueamento das páginas da *Web*.

Por trás do algoritmo *PageRank* encontramos uma das mais interessantes e belas aplicações da Álgebra Linear, além de conteúdo do campo da Teoria da Probabilidade. Claro, este trabalho não é definitivo, tão pouco completo, ainda existem muitas coisas para serem exploradas com mais profundidade. No entanto, acreditamos que a leitura deste material proporcionará uma compreensão razoável do algoritmo *PageRank*, em especial da matemática que estar por "baixo dos panos".

Ademais, além deste trabalho servir como um incentivo ao estudo de assuntos do campo da Álgebra Linear e da Probabilidade, tais como matrizes, sistemas de equações, processos estocásticos, ele também evidencia, a medida que expõe como o algoritmo foi modelado, a beleza do processo que é traduzir ideias em linguagem matemática.

REFERÊNCIAS

AUGUSTO, C. A. *et al.* Pesquisa qualitativa: rigor metodológico no tratamento da teoria dos custos de transação em artigos apresentados nos congressos da sober (2007-2011). **Revista de Economia e Sociologia Rural**, SciELO Brasil, v. 51, n. 4, p. 745–764, 2013.

BERNERS-LEE, T. J. The world-wide web. **Computer networks and ISDN systems**, Elsevier, v. 25, n. 4-5, p. 454–459, 1992.

BOLDRINI, J. L. *et al.* **Álgebra Linear I**. São Paulo: Harper & Row do Brasil, 1980.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

BRODER, A. *et al.* Graph structure in the web. In: **The Structure and Dynamics of Networks**. [S.l.]: Princeton University Press, 2011. p. 183–194.

BRYAN, K.; LEISE, T. The $25,000,000,000 eigenvector: The linear algebra behind google. **SIAM review**, SIAM, v. 48, n. 3, p. 569–581, 2006.

CENDÓN, B. V. Ferramentas de busca na web. **Ciência da Informação**, SciELO Brasil, v. 30, p. 39–49, 2001.

GIL, A. C. *et al.* **Como elaborar projetos de pesquisa**. [S.l.]: Atlas São Paulo, 2017. v. 6.

Internet Live Stats. **Total number of Websites**. 2022. Disponível em: <https://www.internetlivestats.com/>. Acesso em: 08 de março de 2022.

LEHMAN, E.; LEIGHTON, T.; MEYER, A. R. **Mathematics for computer science**. Relatório Técnico, 2010. Acesso em: 18 ago. 2018.

LESKOVEC, J. **Page Ranking**: Web as a graph (stanford university 2019). 2019. Disponível em: <https://www.youtube.com/watch?v=-zq9-6RbKZc>.

MALAJOVICH, G. **Álgebra linear**. UFRJ: [s.n.], 2021.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Ferramentas de busca na internet. **Relatório Técnico: Universidade Federal de Goiás**, 2007.

PAGE, L. *et al.* **The PageRank citation ranking**: Bringing order to the web. [S.l.], 1999.

POOLE, D. **Linear algebra**: a modern introduction. Trent University: Cengage Learning, 2014.

SANTIAGO, B. A matemática por trás do nascimento do google. **Instituto de Matemática e Estatística da Universidade Federal Fluminense.[Manuscrito de minicurso].**, v. 16, 2021. Disponível em: <https://www.professores.uff.br/brunosantiago/wp-content/uploads/sites/17/2018/11/MiniCurso_VCMCOBruno-Santiago.pdf.>

SILVA, T. C. M. da; JÚNIOR, V. V. **Cadeias de Markov**: conceitos e aplicações em modelos de difusão de informação. [S.l.], 2011.

STOLFI, A. d. S. **World wide web**: forma aparente e forma oculta: webdesign da interface ao código. Tese (Doutorado) — Universidade de São Paulo, 2010.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Cajazeiras
Rua José Antônio da Silva, 300, Jardim Oásis, CEP 58.900-000, Cajazeiras (PB)
CNPJ: 10.783.898/0005-07 - Telefone: (83) 3532-4100

Documento Digitalizado Ostensivo (Público)

TCC

Assunto: TCC
Assinado por: José Filho
Tipo do Documento: Anexo
Situação: Finalizado
Nível de Acesso: Ostensivo (Público)
Tipo do Conferência: Cópia Simples

Documento assinado eletronicamente por:

- José Rufino Rodrigues Filho, ALUNO (201812020018) DE LICENCIATURA EM MATEMÁTICA - CAJAZEIRAS, em 13/05/2022 08:59:09.

Este documento foi armazenado no SUAP em 13/05/2022. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 516141
Código de Autenticação: b646d2f787