

FGV EMap
João Pedro Jerônimo

Ciência de Redes

Revisão para A1

Rio de Janeiro
2025

Conteúdo

1	Grafos	3
2	Medidas de Centralidade	6

Grafos

De antemão valhe ressaltar que essa matéria, por mais que seja chamada de **Ciência de Redes**, o termo **rede** se refere a um grafo, não ao tipo específico de grafo que se é visto em **Fluxo em Redes** quando estudamos matemática discreta. Então que já fique esclarecido de antemão que, ao citarmos redes, estamos nos referindo a um grafo no geral, desde que o contrário seja explicitado

Essa sessão será apenas algumas definições que não foram passadas no curso de Matemática Discreta, então conceitos que forem citados sobre grafos e não houver definição nesse resumo, a mesma estará no recap de Matemática Discreta. Aqui segue algumas notações sobre grafos para que não fique confuso:

- $G(V, E) :=$ Grafo com conjunto de vértices V e de arestas E (edges)
- $N(v) :=$ Vizinhança do vértice v (Neighbourhood)
- $\delta(v) :=$ Grau do vértice v
- $K_n :=$ Grafo completo com n vértices
- $K_{m,n} :=$ Grafo completo bipartido com m vértices no primeiro conjunto e n vértices no segundo
- $X(G) :=$ Número cromático de G
- $X'(G) :=$ Número cromático por arestas de G

Definição 1.1 (Grau Médio): Dado um grafo não-dirigido $G(V, E)$, o grau médio de G é:

$$\delta_{\text{med}}(G) := \frac{1}{|V|} \sum_{v_i \in V} \delta(v_i) \quad (1)$$

Se G é dirigido, podemos definir os graus médios de entrada e saída

$$\delta_{\text{med}}^{\text{in}}(G) := \frac{1}{|V|} \sum_{v_i \in V} \delta^{\text{in}}(v_i) \quad \text{Entrada} \quad (2)$$

$$\delta_{\text{med}}^{\text{out}}(G) := \frac{1}{|V|} \sum_{v_i \in V} \delta^{\text{out}}(v_i) \quad \text{Saída} \quad (3)$$

Definição 1.2 (Distribuição do Grau): A distribuição do grau de um Grafo $G(V, E)$ é a distribuição da variável aleatória X , sendo X o grau do vértice que eu escolho ao acaso

Para os teoremas a seguir e daqui em diante, consideremos a matriz de incidência de forma que $A_{ij} = 1$ se a aresta j se conecta no vértice i e, 0 do contrário (-1 se G for dirigido).

Teorema 1.1: Dado um grafo $G(V, E)$ e sua matriz de incidência A , temos que:

$$\text{n}^\circ \text{ de ciclos} = |E| - \text{posto}(A) \quad (4)$$

Demonstração:

$$\begin{aligned} \text{posto}(A) + \dim(N(A)) &= |E| \\ \Leftrightarrow |E| - \text{posto}(A) &= \dim(N(A)) \end{aligned} \quad (5)$$

Porém, a dimensão do núcleo de A é a quantidade de ciclos no grafo, então eu tenho que:

$$\text{n}^\circ \text{ de ciclos} = |E| - \text{posto}(A) \quad (6)$$

□

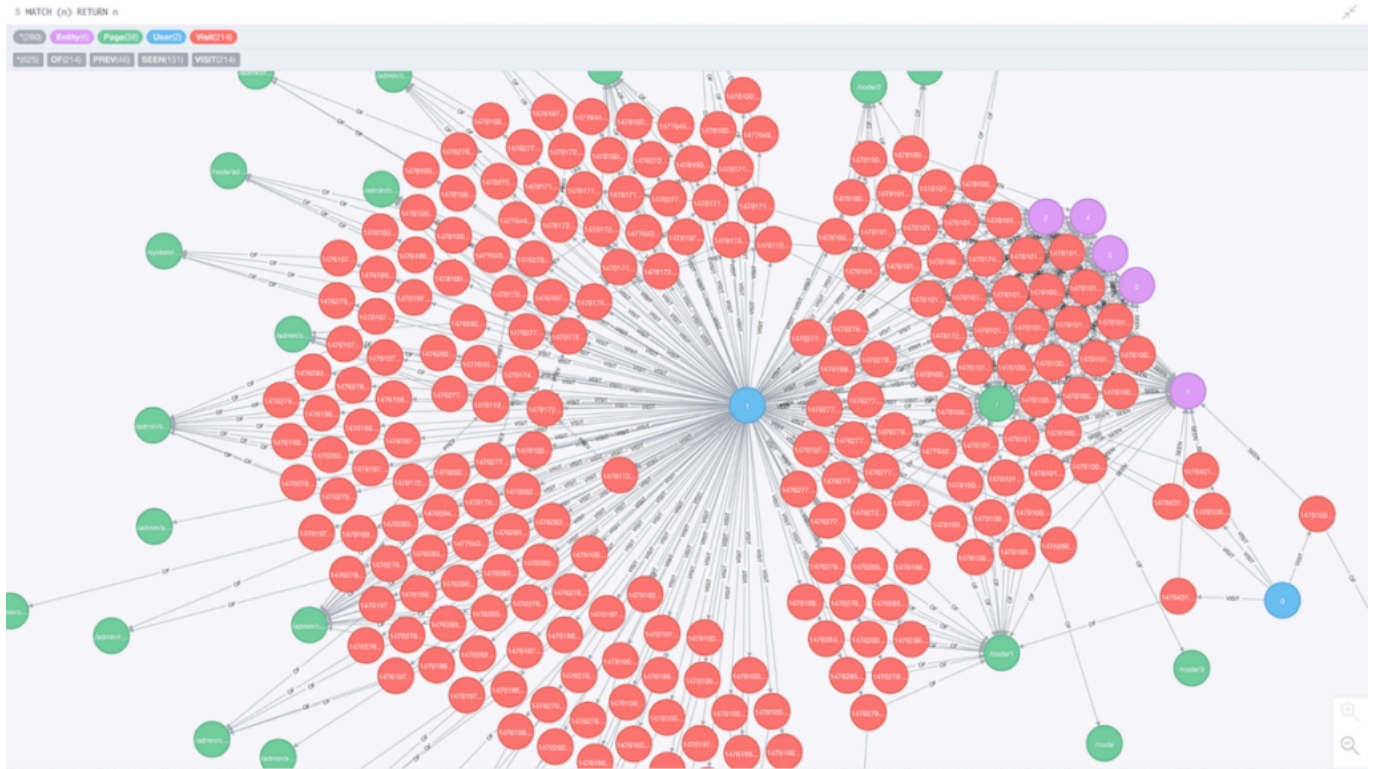
Definição 1.3 (Coeficiente de Clustering): Dado um grafo $G(V, E)$, o coeficiente de clustering de um nó $v \in V$ é:

$$C(v) := \frac{2E_v}{\delta(v)(\delta(v) - 1)} \quad (7)$$

onde E_v é a quantidade de arestas ligadas aos nós vizinhos

Medidas de Centralidade

Quando estamos vendo aplicações reais de grafos, é muito comum querermos ver o “quão importante” um nó é no contexto que estamos analisando. Por exemplo, se nosso grafo representa as conexões entre servidores que um pacote pode percorrer, faz muito sentido querermos ver qual o servidor que quase todos os pacotes percorrem



Imagine que esse é o grafo que estávamos falando (Não importa o que ele representa de verdade, só finge que é o caso que falamos), então o nó azul tem uma importância MUITO grande, mas como podemos medir isso? Nem sempre o grafo vai tá arrumadinho assim pra gente. Daí que surgem as medidas de Centralidade.

Definição 2.1 (Farness/'Lonjura'): Dado um grafo $G(V, E)$, a farness de um vértice v_i é dada por

$$L(v_i) := \sum_{v_i \neq v_j \in V} d(v_i, v_j) \quad (8)$$

onde $d(v_i, v_j)$ é o tamanho do menor caminho entre v_i e v_j

Essa medida mede o quão longe o nó está dos outros, de forma que, quanto maior essa medida é, menos importante o meu nó é (Depende do contexto analisado)

Definição 2.2 (Closeness/Proximidade): Dado um grafo $G(V, E)$, a proximidade/closeness do vértice $v_i \in V$ é dada por:

$$C(v_i) := \frac{|V|}{L(v_i)} \quad (9)$$

Por convenção, se v_i e v_j estão em componentes conexas separadas em G , então $d(v_i, v_j) = \infty$, o que torna a definição de antes inútil, então podemos redefinir como:

$$C(v_i) := \frac{1}{|V|} \sum_{v_i \neq v_j \in V} \frac{1}{d(v_i, v_j)} \quad (10)$$

Definição 2.3 (Betweenness/Intermediação): Dado um grafo $G(V, E)$ e $P(v_i, v_j)$ o conjunto de todos os menores caminhos possíveis entre v_i e v_j , então a intermediação de v_i é:

$$B(v_i) := \sum_{v_s, v_t \in V} \frac{|c \in P(v_s, v_t); v_i \in c|}{|P(v_s, v_t)|} \quad (11)$$

Saindo um pouco dessas definições, vamos tentar pensar em alguma medida mais básica e intuitiva. Uma medida bem padrão que podemos pensar logo de cara é simplesmente o grau do vértice, já que, quanto mais vértices ele se ligar, mais importante ele é! Em muitas literaturas sobre redes o grau do vértice é chamado de **Centralidade de Grau**.

Um outro pensamento que pode surgir a partir desse é: “Poxa, meu vértice tem um grau alto, então ele é importante, mas eu quero valorizar aqueles vértices que se conectam com ele, afinal, se ele é importante, os vértices que estão diretamente ligados nele também são, não é?”, e esse pensamento não está errado! É dessa ideia que surge a centralidade por autovetor. Funciona assim: Vamos inicialmente assumir que todos os nossos vértices v_i tem importância $x_i^{(0)} = 1$, o que não me é muito útil agora, porém, vamos tentar fazer uma nova estimativa baseada nos vizinhos, que tal a nova centralidade do vértice v_i ser a soma da centralidade dos vizinhos? Isso faz com que a importância do v_i se baseie no quão importante são seus vizinhos! Eu posso expressar isso com uma fórmula:

$$x_i^{(1)} = \sum_j A_{ij} x_j^{(0)} \quad (12)$$

Onde A é minha matriz de adjacência. Se meu nó v_i não é vizinho de v_j , então $A_{ij} = 0$ o que faz com que minha centralidade $x_j^{(0)}$ não seja somada. Posso reformular isso de forma matricial:

$$x^{(1)} = Ax^{(0)} \quad (13)$$

onde $x^{(k)}$ é o vetor com entradas $x_i^{(k)}$. Se fizermos esse processo várias vezes, depois de k passos, vamos ter algo do tipo:

$$x^{(k)} = A^k x^{(0)} \quad (14)$$

Tomemos a liberdade, então, de escrever $x^{(0)}$ como uma combinação linear dos autovetores w_j de A de forma que

$$x^{(0)} = \sum_{j=1}^n c_j w_j \quad (15)$$

Para alguma escolha apropriada de c_j . Então temos:

$$x^{(k)} = A^k \sum_{j=1}^n c_j w_j = \sum_{j=1}^n c_j \lambda_j^k w_j = \lambda_1^k \sum_{j=1}^n c_j \left(\frac{\lambda_j}{\lambda_1} \right)^k w_j \quad (16)$$

De forma que λ_j são os autovalores de A e λ_1 pode ser, sem perda de generalização, o maior de todos em módulo. Como $\lambda_i/\lambda_1 < 1 \forall \lambda_i$ com $i \neq j$, então:

$$\lim_{k \rightarrow \infty} \sum_{j=1}^n c_j \lambda_j^k w_j = c_1 \lambda_1 w_1 \quad (17)$$

Ou seja, o vetor de centralidades que limita as centralidades que eu fiz antes é proporcional ao autovetor associado ao maior autovalor de A , que é equivalente a dizer que o vetor de centralidades x satisfaz:

$$Ax = \lambda_1 x \quad (18)$$

Definição 2.4 (Centralidade Autovalor): Seja r um vetor com as centralidades dos vértices v_i de uma rede G de forma que r_i = centralidade de v_i , então:

$$Ar = \lambda_1 r \quad (19)$$

onde λ_1 é o maior autovalor de A

Agora temos outro problema. Quando temos um grafo dirigido, essa medida de centralidade autovalor já não funciona, já que se um nó não tem nenhuma aresta apontando para ele (Apenas saem arestas dele), ele não terá sequer uma centralidade, e isso afeta não só esse vértice como os vértices que ele aponta, que não terão nenhuma “pontuação” adicionada por serem apontados por esse vértice, e isso não pode ocorrer, já que não faz muito sentido na maioria das aplicações práticas. O que podemos fazer para contornar isso? Então entra a solução a seguir:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta \quad (20)$$

Onde α e β são constantes positivas. O primeiro termo é a centralidade autovetor que vimos antes, porém o termo β garante que os nós que comentei anteriormente (Sem grau de entrada) possuam uma pontuação e possam contribuir para a pontuação dos nós que eles apontam. Essa medida é interessante por conta do termo α que balanceia o termo constante e a medida de centralidade autovetor. Podemos expressar isso de forma matricial:

$$x = \alpha Ax + \beta \mathbf{1} \quad (21)$$

Onde $\mathbf{1} = (1, \dots, 1)$. Se rearranjarmos para x , obtemos:

$$x = \beta(I - \alpha A)^{-1} \mathbf{1} \quad (22)$$

Normalmente colocamos $\beta = 1$ pois não estamos interessados em saber o valor exato das centralidades, mas saber quais vértices são ou não mais ou menos centrais.

$$x = -\alpha \left(A - \frac{1}{\alpha} I \right)^{-1} \quad (23)$$

Perceba que eu quero que $A - \frac{1}{\alpha} I$ seja invertível, e isso acontece quando $\frac{1}{\alpha} \neq \lambda_j$ onde λ_j são os autovalores de A . Ou seja, o meu α não é completamente arbitrário, eu vou ter que analisar

o contexto da minha aplicação. Porém, muito comumente, se é utilizado $\alpha = \frac{1}{\lambda_1}$ com λ_1 sendo o maior autovalor

Definição 2.5 (Centralidade de Katz): Dado uma rede $G(V, E)$ e duas constantes $\alpha, \beta > 0$, a centralidade de katz do nó v_i é:

$$K(v_i) = \alpha \sum_j A_{ij} v_j + \beta \quad (24)$$

Onde A é a matriz de adjacência de G

Um outro tipo de medida surge quando queremos responder a questão: “Se eu estou navegando entre meus nós, ao longo prazo, qual é o nó que eu mais vou percorrer/parar nele?”. Um exemplo são páginas na internet que referenciam entre si, daí surge o nome da medida: **PageRank**. O que fazemos essencialmente é transformar a rede em uma cadeia de markov. Por exemplo:

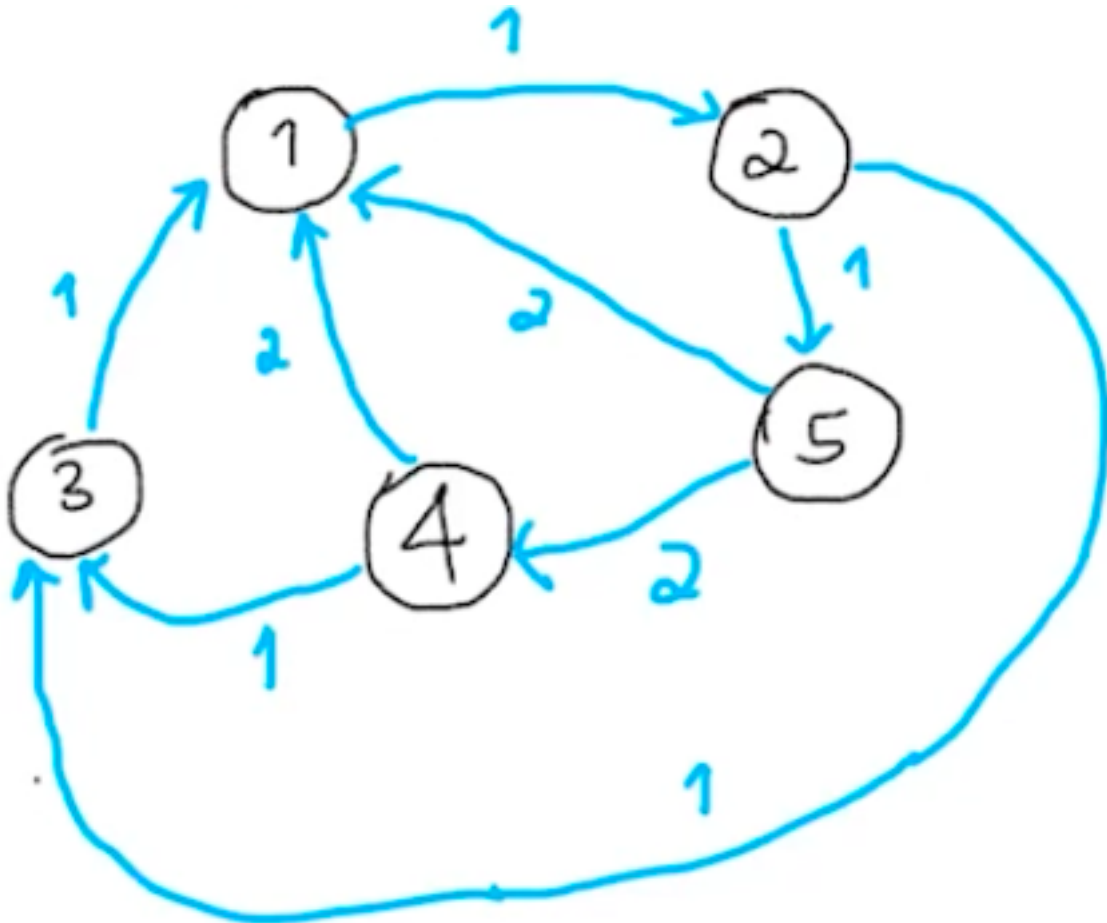


Figura 1: Grafo de Exemplo 1

Vamos supor que estamos no nó 4 e queremos escolher aleatoriamente entre os nós 3 e 1 para irmos, como podemos ver na distribuição dos pesos (Nesse exemplo, isso indica que a página 4 tem 2 links referenciando a página 1 e apenas 1 link referenciando a página 3), então teríamos:

$$\begin{aligned} \mathbb{P}(4 \rightarrow 3) &= \frac{1}{3} \\ \mathbb{P}(4 \rightarrow 1) &= \frac{2}{3} \end{aligned} \quad (25)$$

E fazemos isso definindo uma matriz estocástica H de tal forma que:

$$H_{ij} = \frac{A_{ij}}{\sum_k^n A_{ik}} \quad (26)$$

Com A sendo a matriz de adjacência. De forma que a soma de todos os elementos de uma coluna dê 1. Agora que vem o truque interessante. Dado um vetor $p \in \mathbb{R}^n$ de forma que cada entrada de p_i representa a chance de eu ir do nó que eu estou para o nó v_i (Ou seja, p tem que ser alguma coluna de H), ao fazer a operação:

$$Hp \quad (27)$$

Eu estou ponderando as probabilidades de p com os seus respectivos nós, ou seja, $(Hp)_k$ representa a probabilidade esperada de que, ao sair do nó v_i , eu vá para o nó v_k . Se isso é verdade e, como eu defini antes, eu quero saber qual nó é mais visitado conforme se passa o tempo, faz sentido eu refazer esse processo inúmeras vezes, então eu tenho uma centralidade do vértice v_i :

$$r = \lim_{t \rightarrow \infty} H^t p \quad (28)$$

Com p sendo a i -ésima coluna de H . Porém isso ainda nos trás um problema, veja essa outra rede:

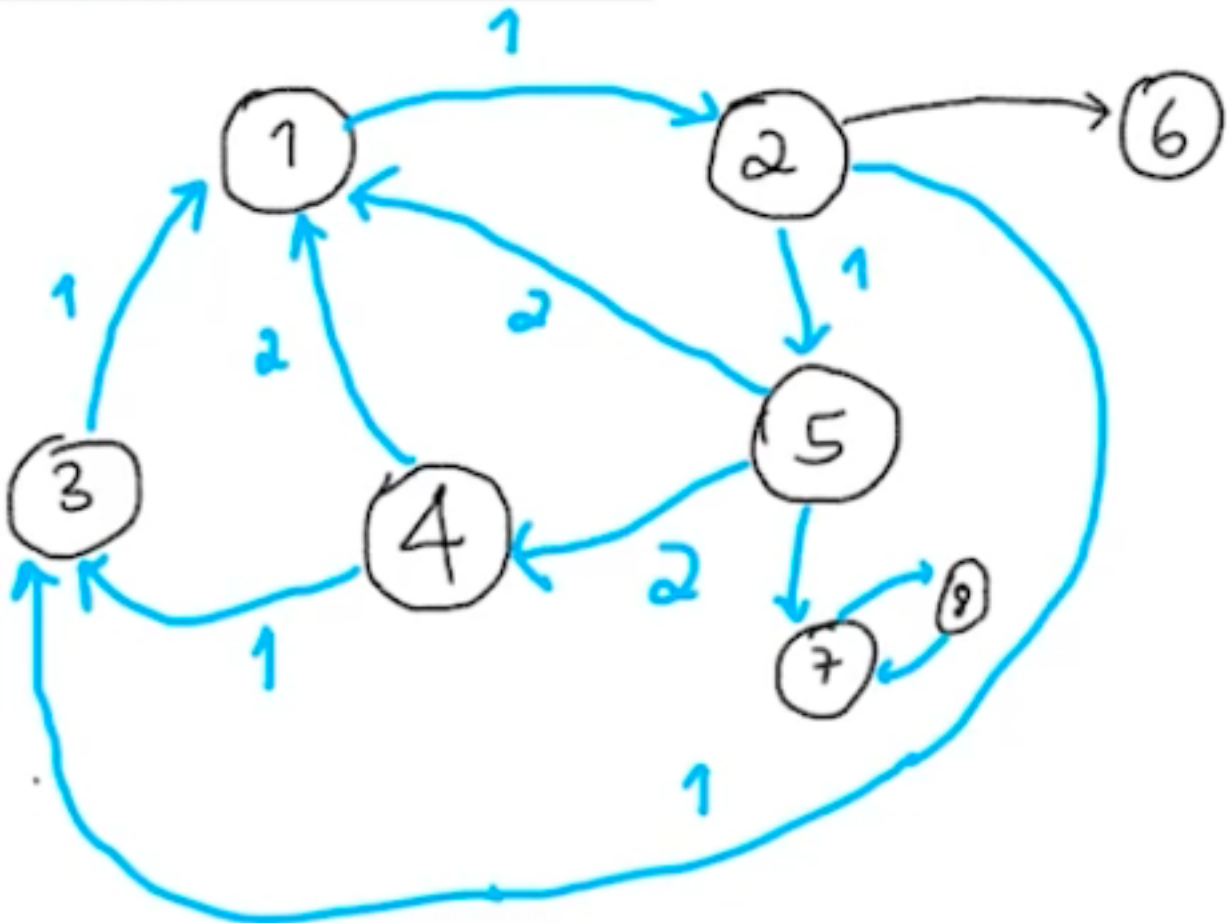


Figura 2: Grafo de Exemplo 2

Veja que, por conta do nó 6, eu não posso transformar meu esquema em uma cadeia de markov, pois eu teria uma coluna de 0, e no caso dos nós 7 e 8 eu teria um problema por conta que eles sempre vão um para o outro. Como podemos resolver isso? O PageRank vem para resolver isso. Vamos pensar no caso da internet, você navegador aleatório, uma hora, pode se cansar de estar onde estar, e visitar uma página aleatoriamente dentro da sua rede, e é nessa ideia que trabalhamos em cima.

Definimos um $\alpha \in (0, 1)$, onde podemos interpretar α como a chance do meu navegador permanecer no meu nó. Definimos então nossa nova matriz de chances da seguinte forma:

$$\mathbb{G} = \alpha H + (1 - \alpha)C \quad (29)$$

De forma que C é uma matriz $n \times n$ com todas as entradas iguais a $1/n$ para representar um dirigido onde todos os nós apontam para todos os outros nós (Representando a ideia de que eu posso ir para o nó que eu quiser). Porém, há uma propriedade que, se eu tenho uma combinação convexa entre duas matrizes estocásticas/markovianas, então o resultado é uma matriz markoviana. Ou seja, eu ainda posso aplicar a mesma ideia de antes do vetor p_0 inicial e aplicar o limite, assim, eu vou obter meu vetor de centralidades r , de tal forma que

$$\lim_{t \rightarrow \infty} \mathbb{G}^t p_0 = r \quad (30)$$

Definição 2.6 (PageRank): Sejam a matriz \mathbb{G} como definida anteriormente e o vetor inicial p_i sendo a i -ésima coluna de \mathbb{G} , então o vetor de centralidades PageRank r onde a k -ésima entrada é a centralidade de v_k , então:

$$r = \lim_{t \rightarrow \infty} \mathbb{G}^t p_0 \quad (31)$$