Trabalho de Conclusão de Curso

Simulação de Algoritmos Distribuídos Aplicados ao Cálculo do PageRank

Oscar Neiva E. Neto

Orientador: D.Sc. Eduardo Krempser

Coorientador: D.Sc. Marcos Garcia Todorov

Faculdade de Educação Tecnológica do Estado do Rio de Janeiro - FAETERJ Petrópolis
Laboratório Nacional de Computação Científica - LNCC

Índice

Objetivo e Motivação

- Este trabalho consiste no estudo e simulação de sistemas sujeitos a saltos markovianos.
- A motivação para os estudos no tema parte dos problemas relacionados a simulação do algoritmo *PageRank*.

O Algoritmo PageRank

- A proposta do *PageRank*[1].
- O grau de importância das páginas.
- O Sistema de Busca e o *PageRank*.

Criadores do *PageRank* e barra de pesquisa do Google

[1] Brin, Sergey and Page, Lawrence, The anatomy of a large-scale hypertextual web search engine, 1998.

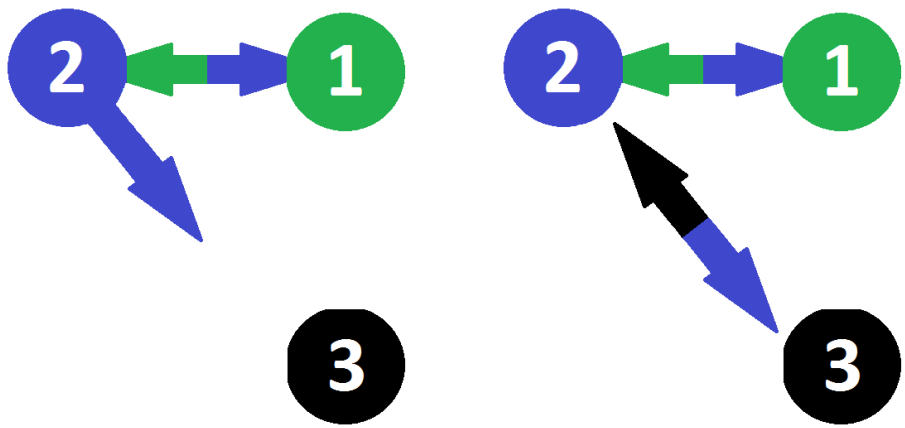
As Ferramentas de Busca

- Da *World Wide Web Worm* em 1994 até 2016.
- Nos útimos 20 anos a *Web* vem ganhando 240.000.000 páginas por ano.

Ano	Buscador	Nº de Páginas Indexadas
1994	WWWW	110 mil
1997	AltaVista	100 milhões
1998	Google	518 milhões
2016	Google	4.8 bilhões

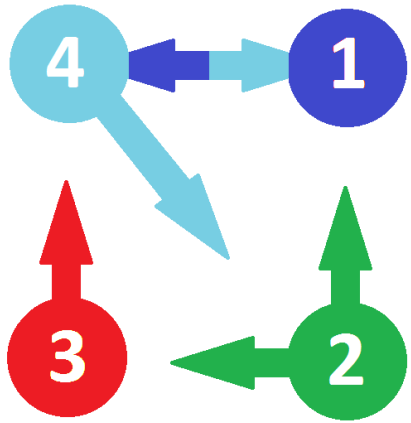
O Problema do PageRank

- A massividade da *Web* e a dificuldade do cálculo.
- A navegação entre páginas desconexas.
- A página Buraco Negro.



O Buraco Negro da *Web*

A Matriz Hiperlink



Grafo representando *links* entre páginas da *web*

$$A = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad (1)$$

A Matriz Hiperlink

- n: nós, que representam as páginas. A navegação é então representada através de uma cadeia de Markov com espaço de estados discreto de dimensão n.

- $\mathcal{E}$: arestas, que representam os *links*. Para o caso de um vértice i estar conectado a um j, tem-se que $(i, j) \in \mathcal{E}$.

$$a_{ij} = \begin{cases} \dfrac{1}{n_i}, & \text{caso } (i,j) \in \mathcal{E}, \\ 0, & \text{caso contrário.} \end{cases} \tag{2}$$

O Power Method

- Um método para obtenção do *PageRank* é o chamado *Power Method*[2].

- O *PageRank* é o ponto fixo da seguinte recursão:

$$x(k+1) = Ax(k),\ k \geq 0,\ \text{com}\ x(0) = x_0, \tag{3}$$

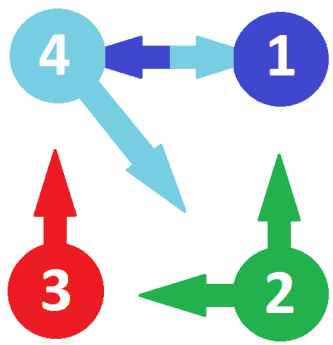
onde $x_0 \in \mathbb{R}^{n\times 1}$ é uma condição inicial positiva de soma igual a um.

[2] Ishii, Hideaki and Tempo, Roberto, The pagerank problem, multiagent consensus, and web aggregation: A systems and control viewpoint, 2014.

Questões de Convergência do Power Method

- Caso a matriz A seja irredutível, independente da condição inicial é atingida a distribuição limite.
- Diz-se que A é irredutivel se sempre existe um caminho ligando dois nós quaisquer.

$$A = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \qquad x^* = \begin{pmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{pmatrix}$$



Teleportation Model

- Embora a simplicidade do *Power Method* o torne atraente, ele pode apresentar problemas de convergência.
- O Teleportation promove exploração, sem afetar o *PageRank*.

Teleportation Model

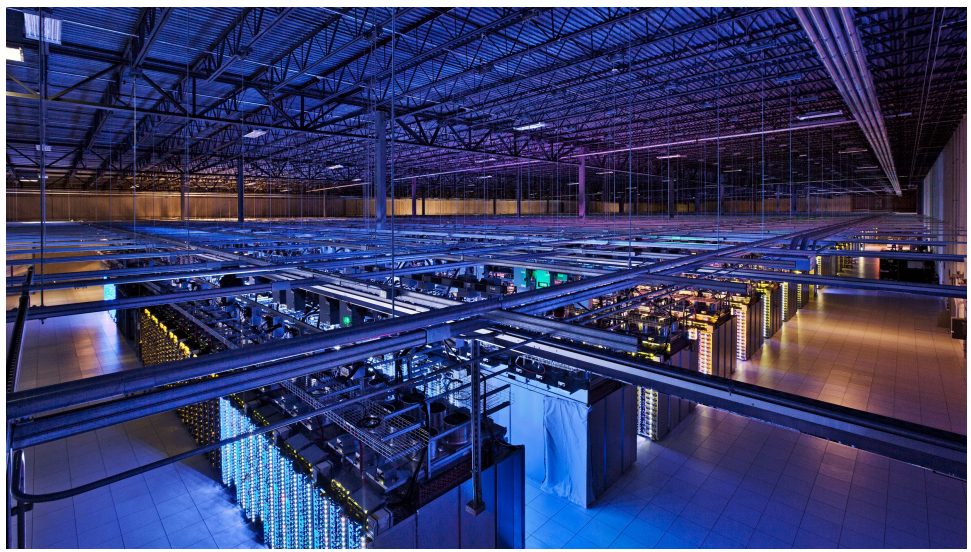
- O *Teleportation Model* é uma estratégia reconhecida para que, através de uma pequena modificação na matriz A, o método convirja globalmente.

$$M = (1 - m)A + \frac{m}{n}\mathbf{1}\mathbf{1}^T \tag{4}$$

  - $m \in (0, 1)$
  - $M \in \mathbb{R}^{n \times n}$
  - $\mathbf{1} \in \mathbb{R}^{n \times 1}$

O Modelo dos Algoritmos Distribuídos

- Tornar o cálculo menos custoso e factível.
- Explorar os recursos computacionais dos servidores.
- Emprego de algoritmos distribuídos[3].



Cluster da Google.

[3] Lei, Jianjun and Chen, Han-Fu, Distributed Randomized PageRank Algorithm Based on Stochastic Approximation, 2015.

O Modelo dos Algoritmos Distribuídos

$$A = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \tag{5}$$

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 1 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad A_4 = \begin{pmatrix} 1 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

O Modelo dos Algoritmos Distribuídos

- O modelo distribuído trata-se de um sistema dinâmico sujeito a saltos[4].

$$x(k+1) = A_{\theta(k)}x(k), \quad k \geq 0, \quad \text{com} \quad x(0) = x_0. \tag{6}$$

- Este mecanismo de seleção é modelado por $\theta = \{\theta(k), k = 0, 1, \ldots\}$ é uma cadeia de Markov[5], regida pela propriedade:

$$P\big(\theta(k+1) = j \mid \theta(k) = i_k, \theta(k-1) = i_{k-1}, \ldots, \theta(0) = i_0\big) = P\big(\theta(k+1) = j \mid \theta(k) = i_k\big). \tag{7}$$

[4] O. L. V. Costa and M. D. Fragoso and M. G. Todorov, Continuous-Time Markov Jump Linear Systems, 2013.

[5] P. Brémaud, Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, 1999.

O Modelo dos Algoritmos Distribuídos

- Adaptando ao *Teleportation Model*:

$$x(k+1) = (1-\hat{m})A_{\theta(k)}x(k) + \frac{\hat{m}}{n}\mathbf{1}\mathbf{1}^T, \tag{8}$$

  - $k \geq 0$,
  - com $x(0) = x_0$,
  - onde $\hat{m} = \frac{2m}{n-m(n-2)}$,
  - $m = 0,15$ [6].

- Problemas de convergência.

[6] Zaki, Nazar and Berengueres, Jose and Efimov, Dmitry, Detection of protein complexes using a protein ranking algorithm, 2012.

Questões de Convergência do Modelo Distribuído

- $y(k)$ é a média do conjunto de amostras $x(0), ..., x(k)$,

$$y(k) = \frac{1}{k+1} \sum_{l=0}^{k} x(l). \tag{9}$$

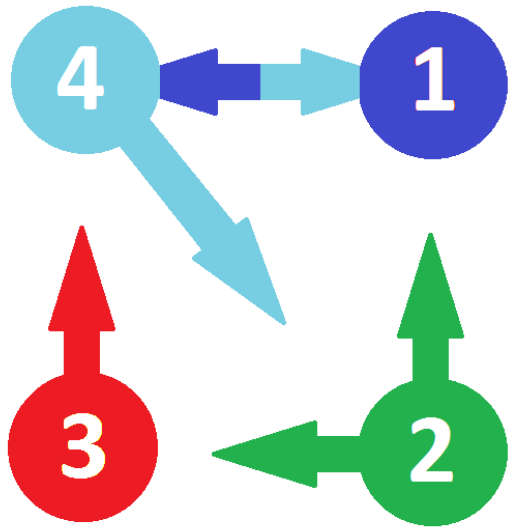
- O algoritmo converge no sentido da média quadrática:

$$\lim_{k \to \infty} \mathbb{E}[\| \, y(k) - x^* \, \|^2] = 0. \tag{10}$$

- $y(k+1)$ é a média recursiva do conjunto de amostras $x(0), ..., x(k+1)$,

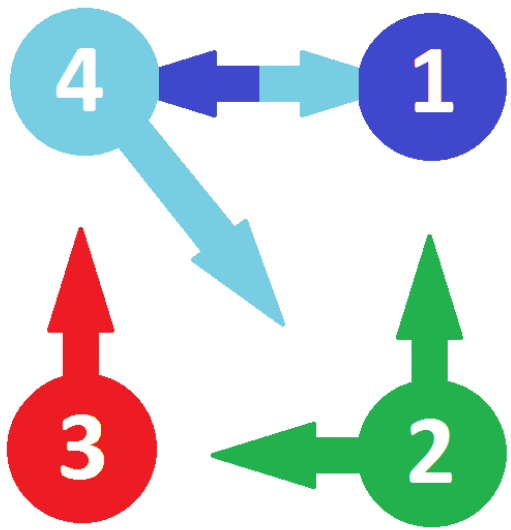
$$y(k+1) = \frac{(k+1)}{(k+2)} y(k) + \frac{1}{(k+2)} x(k+1). \tag{11}$$

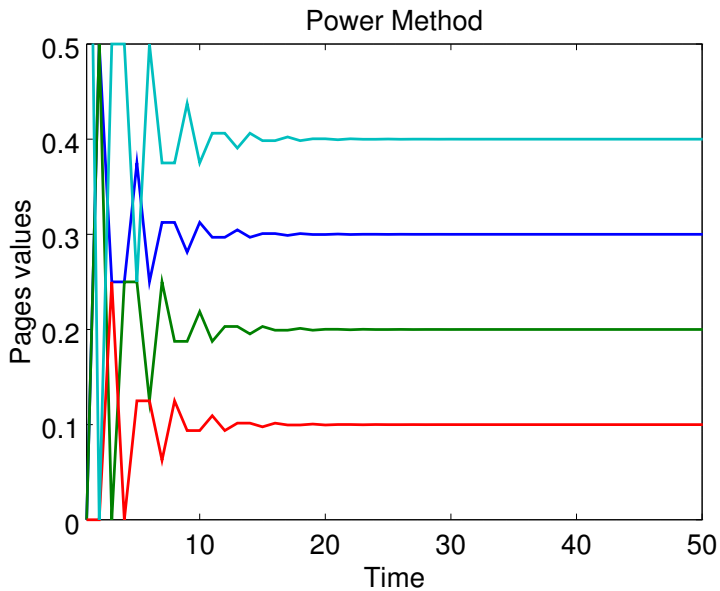
Simulações



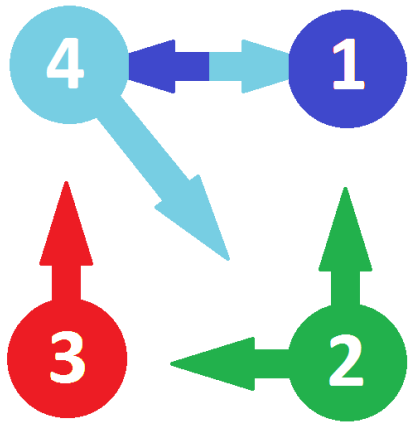
$$x_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{12}$$

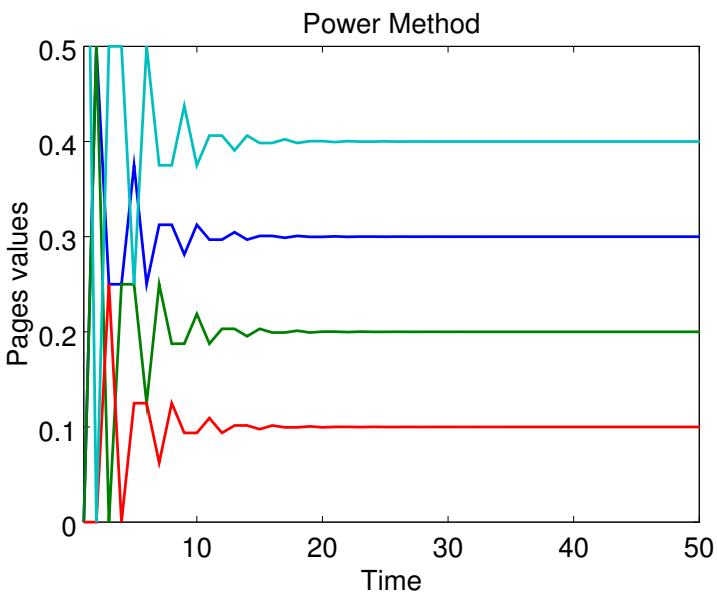
Simulação do Power Method

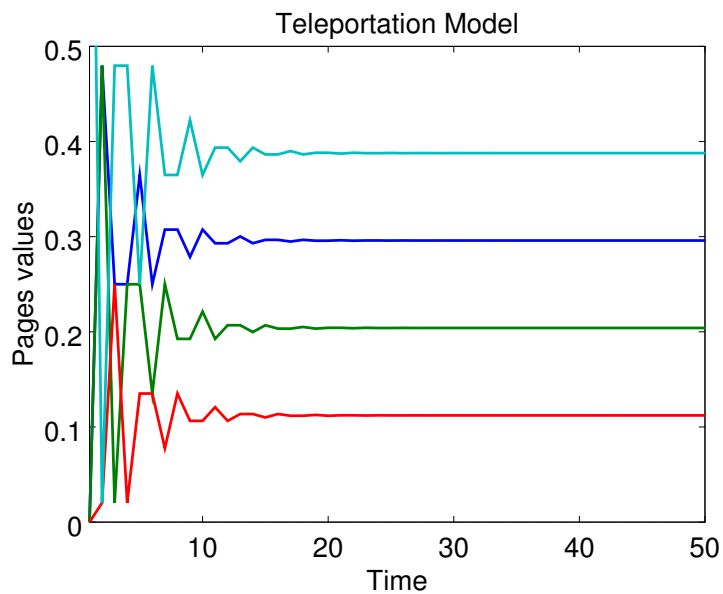




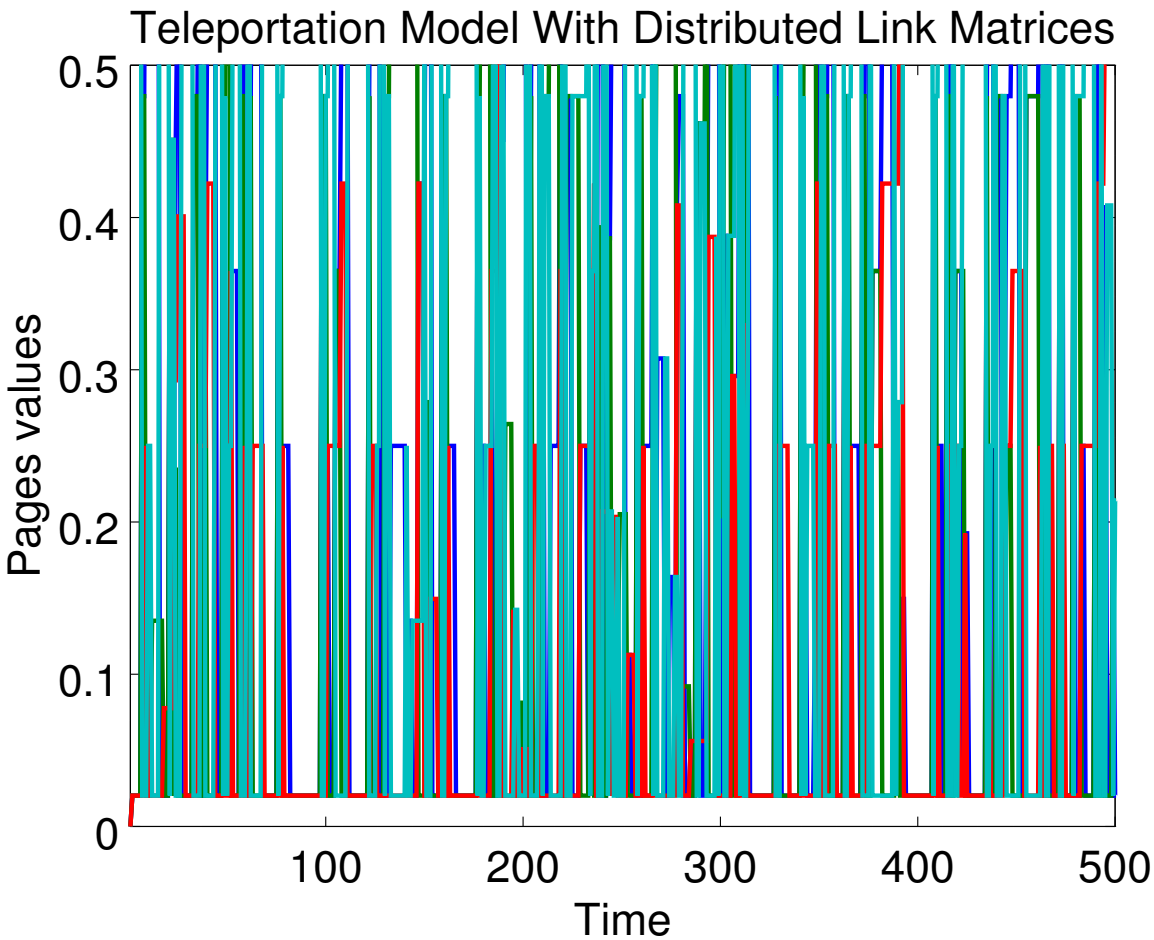
Simulação do *Teleportation Model*





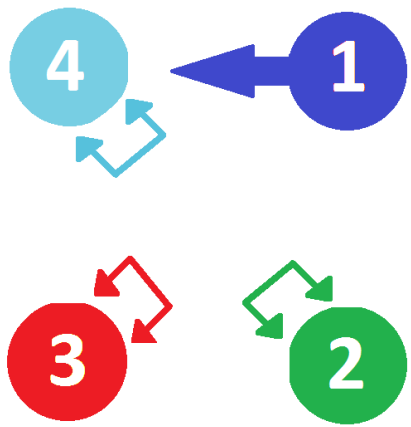


Simulação do Teleportation Model Distribuído



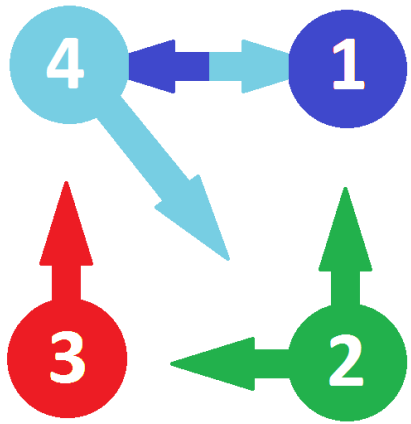
Simulação do Teleportation Model Distribuído

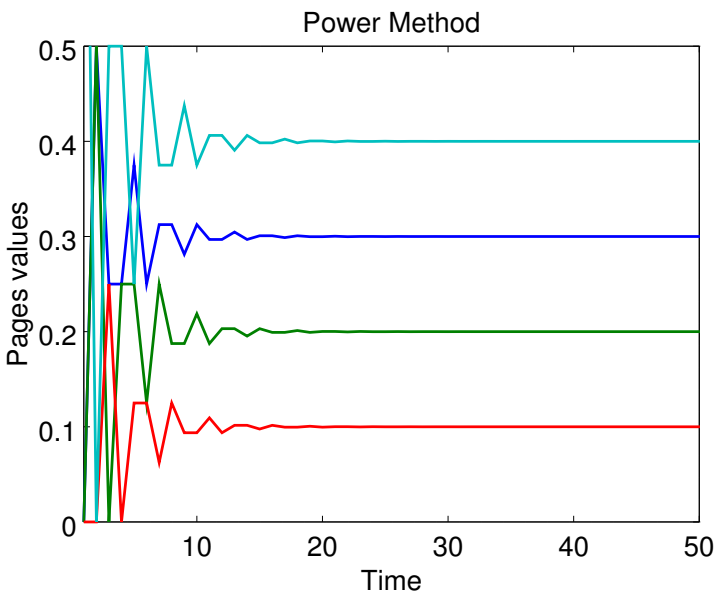
- A falta de convergência é tipica de simulações estocásticas.

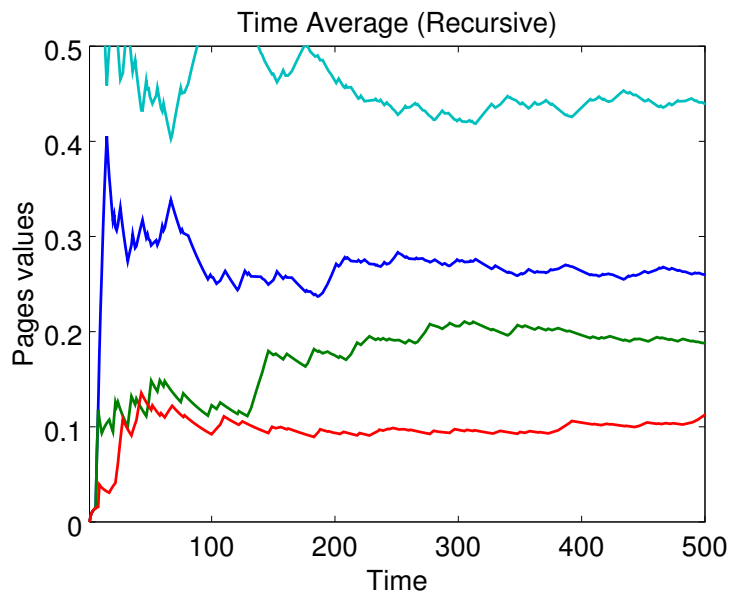


$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

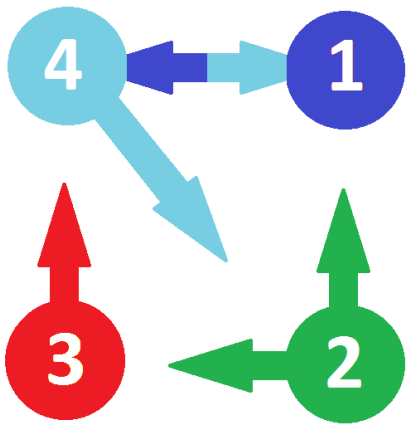
O Modelo Recursivo da Média no Tempo Aplicada a Simulação do Modelo Distribuído

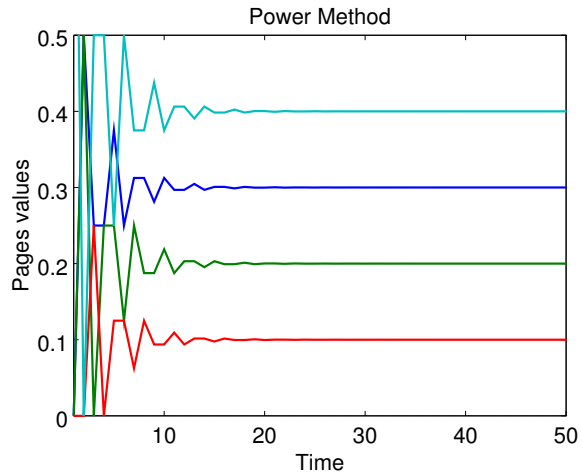


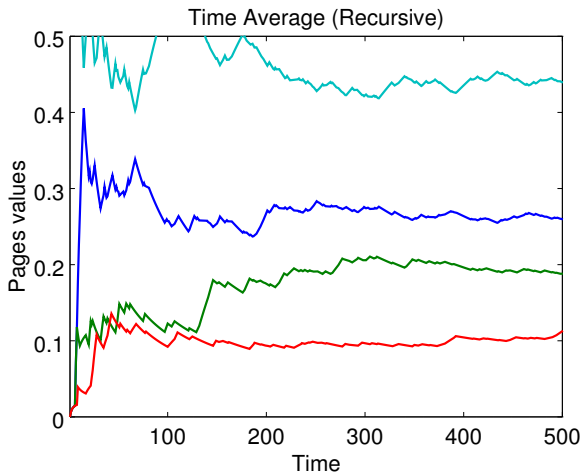


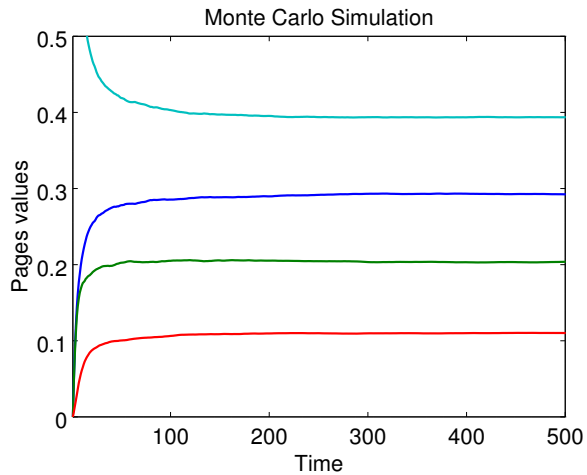


Método de Monte Carlo Aplicado após Modelo Recursivo da Média









Considerações Finais

- O sucesso dos sistemas de busca.
- Utilizar outros métodos, válidos na simulação do *PageRank*.
- Implementação com *links* já coletados por um *Web Crawling*[7].
  - Uso de outras linguagens de programação.
  - Computação Distribuída.

[7] U.K. New Zealand Univ., Statistical Cybermetrics Research Group, 2006.