O algoritmo PageRank do Google

Miguel Frasson – ICMC/USP

2019

Resumo

Estudaremos o Algoritmo PageRank do Google para classificação de páginas

Ingredientes

Cadeias de Markov (Estatística + Álgebra Linear)

- Matrizes positivas, estocásticas e simétricas
- Autovalores e autovetores
- Teorema de Perron-Frobenius

Métodos numéricos

- Cálculo de autovalores dominantes para matrizes gigantescas (com bilhões de linhas).

Buscas na Web

No início, os usuários vinham a certos sites pela busca.
Yahoo! inventa a receita $$$ por anúncios (banners).
Para manter usuários online, viraram portais, bate-papo, e-mail etc.
A Web crescia exponencialmente mas modo de encontrar conteúdo não acompanhava.

- Yahoo! (1994)
  começou como catálogo de sites, com descrição das páginas
- Excite: primeiro indexador, software lia páginas
- Google (1998)
  primeiro algoritmo classificador, o **PageRank**.
  Sucesso absoluto, dominou as buscas desde então.
- Outros algoritmos viriam.
  Conviria estudar o **HITS**, usado hoje pelo Ask.com

Buscas até 1998

Encontrar as páginas com um termo de busca

- Técnicas liam o conteúdo da página (palavras, imagens).
- Pontuação baseada em informações contidas na página.

Problema

Baixa qualidade na ordenação (relevância) dos resultados da busca

Algoritmo PageRank

- Classificar por relevância era o principal problema da Web.
- Larry Page e Sergey Brin (Universidade de Standford) propuseram um algoritmo (PageRank)

- Implementaram nos servidores da Universidade de Standford (que detém a patente até hoje)
  $\rightarrow$ sucesso e sobrecarga nos servidores
- Tiveram que sair: fundaram o Google.

Algoritmo PageRank

Estratégia

A importância era medida por fatores externos à página.
Cada ligação (link) para uma página era um voto

- quanto mais votos, maior importância
- votos de páginas importantes têm mais peso

Algoritmo PageRank

Estratégia

A importância era medida por fatores externos à página.
Cada ligação (link) para uma página era um voto

- quanto mais votos, maior importância
- votos de páginas importantes têm mais peso

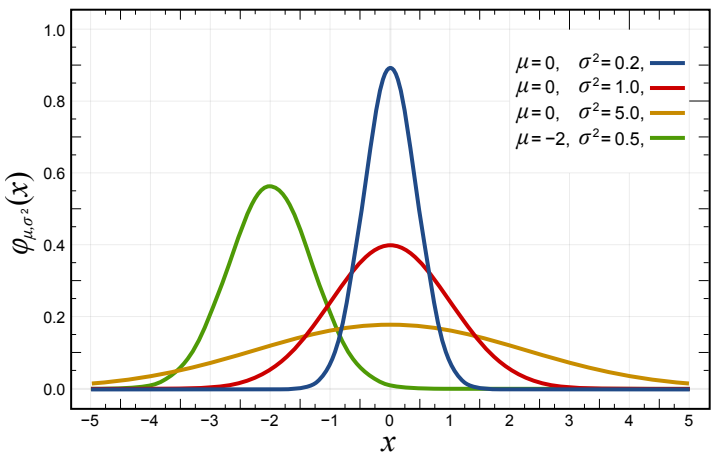
... mas a quantidade de páginas crescia exponencialmente.

A ferramenta matemática: Cadeias de Markov.

Cadeias de Markov

História das cadeias de Markov

- Desde o século 16, a teoria da probabilidade foi sendo desenvolvida.
- Por exemplo, numa longa lista do lançamentos de uma moeda ou dados, ou um sorteio com reposição, as razões dos eventos tendem a convergir (Lei dos Grandes Números)
- Por fim, até mesmo as somas de variáveis aleatórias independentes parecem convergir para a distribuição normal (Teorema Central do Limite)



- Tudo parecia pré-determinado!

História das cadeias de Markov

- Tanto *determinismo* incomodou alguns, como Nekrasov.
  Seria contra o livre-arbítrio, por exemplo:

  Independência é uma condição necessária para a Lei dos Grandes Números.

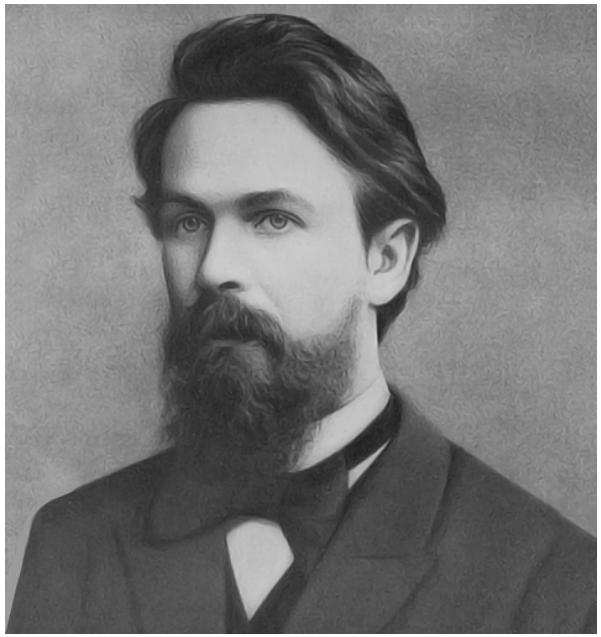
- Como a maioria dos eventos físicos são claramente dependentes de eventos passados, as leis probabilísticas se aplicariam a poucas coisas.

Pavel Nekrasov
(1853–1924)

História das cadeias de Markov

- Andrej Markov não gostou das ideias de Nekrasov:

  Esta circunstância me incita a explicar, numa série de artigos, que a Lei dos Grandes Números também pode ser aplicada a variáveis dependentes.



Andrej Markov
(1856–1922)

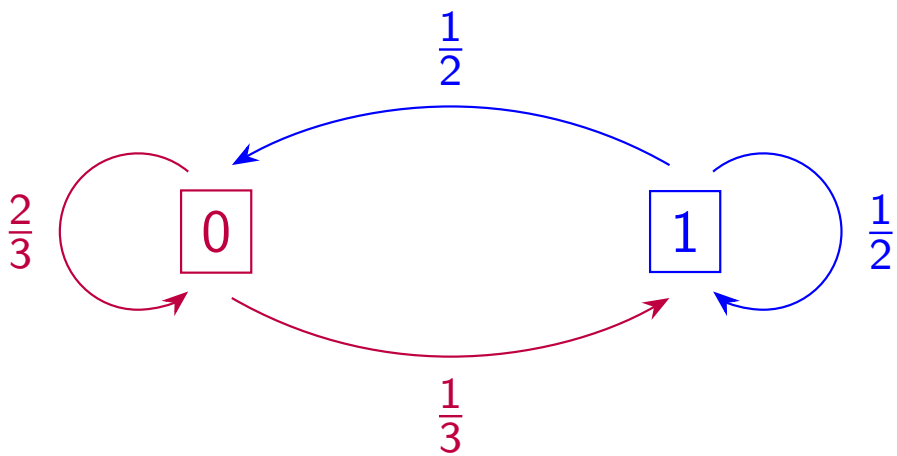
História das cadeias de Markov

- Deu o exemplo de dois "estados"

  estado 1 uma mistura uniforme de bolas brancas e pretas

  estado 0 uma mistura com mais pretas que brancas

- Começando com qualquer estado, sorteia-se (com reposição) uma bola. Se a bola sorteada for preta, o próximo sorteiro entre as bolas do estado 0. Senão, do estado 1.



História das cadeias de Markov

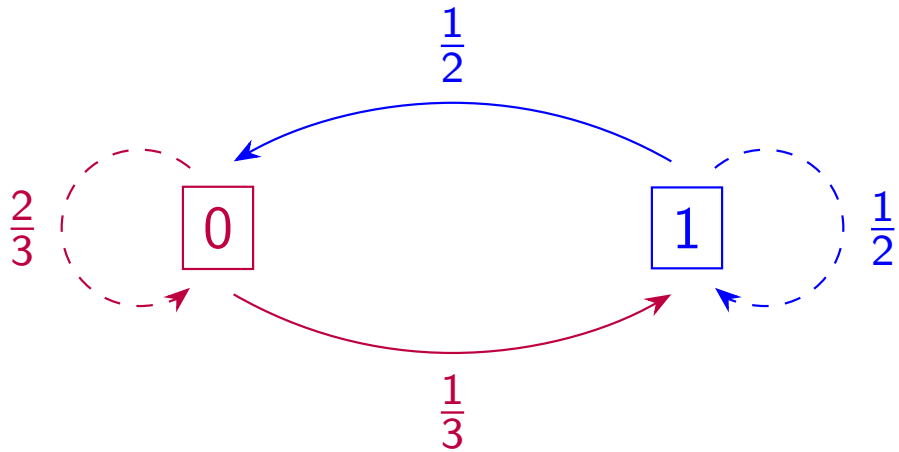
- Markov provou que,
  desde que todos os estados da máquina sejão atingíveis,
  rodando máquinas como esta um grande número de vezes,
  elas atingiriam um equilíbrio.

- Este exemplo desprovou a afirmação de Nekrasov, mostrando que mesmo eventos dependentes podem convergir para distribuições previsíveis.

Cadeias de Markov

O conceito de modelar sequências de eventos aleatórios
usando estados e transições entre estados
ficou conhecido como cadeias de Markov.

Exemplo das bolas brancas e pretas



	p_0	p_1	p_2	p_3
0	1	$\frac{2}{3}$	$\frac{2}{3}\cdot\frac{2}{3}+\frac{1}{3}\cdot\frac{1}{2}=\frac{11}{18}$	$\frac{11}{18}\cdot\frac{2}{3}+\frac{7}{18}\cdot\frac{1}{2}=\frac{65}{108}$
1	0	$\frac{1}{3}$	$\frac{2}{3}\cdot\frac{1}{3}+\frac{1}{3}\cdot\frac{1}{2}=\frac{7}{18}$	$\frac{11}{18}\cdot\frac{1}{3}+\frac{7}{18}\cdot\frac{1}{2}=\frac{43}{108}$

- As mesmas contas podem ser feitas com matrizes!

$$p_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad p_1 = \underbrace{\begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{p_0} = \begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \end{pmatrix},$$

$$p_2 = \underbrace{\begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \end{pmatrix}}_{p_1} = \begin{pmatrix} \frac{11}{18} \\ \frac{7}{18} \end{pmatrix},$$

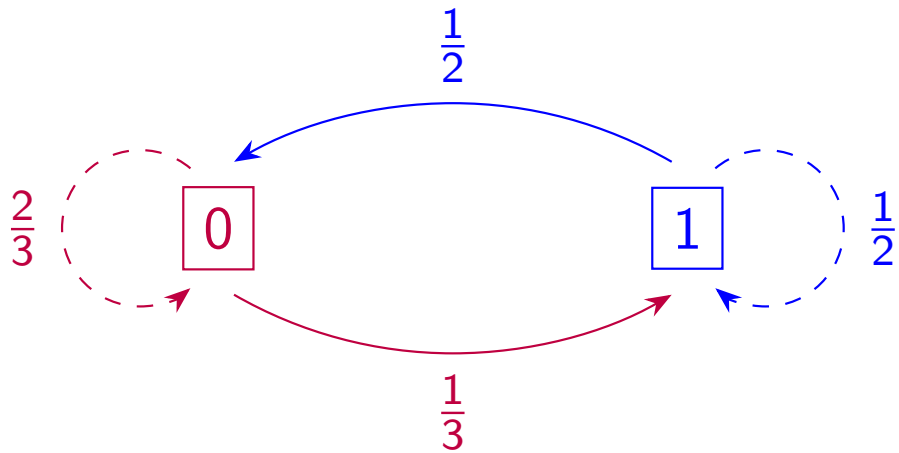
$$p_3 = \underbrace{\begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \frac{11}{18} \\ \frac{7}{18} \end{pmatrix}}_{p_2} = \begin{pmatrix} \frac{65}{108} \\ \frac{43}{108} \end{pmatrix}, \quad \cdots$$

Matrizes de transição

Matriz de transição

$$A = (a_{ij}), \qquad a_{ij} = P(X_{n+1} = i | X_n = j)$$

Exemplo (bolas brancas e pretas)



$$A = \begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}$$

Matrizes de transição

Matriz de transição

$$A = (a_{ij}), \qquad a_{ij} = P(X_{n+1} = i | X_n = j)$$

- Soma das colunas $= 1 \rightarrow$ matriz estocástica por colunas
  Motivo: soma das probabilidades do espaço todo, dado $X_n = j$.

  matriz de transição: $A = \begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}$

Matrizes de transição

Matriz de transição

$$A = (a_{ij}), \qquad a_{ij} = P(X_{n+1} = i | X_n = j)$$

- Soma das colunas $= 1 \rightarrow$ matriz estocástica por colunas
- Se p_n é o vetor com as probabilidades $P(X_n = i)$

$$\begin{aligned} p_n &= A p_{n-1} \\ &= A \underbrace{A p_{n-2}}_{p_{n-1}} = A^2 p_{n-2} \\ &\vdots \\ &= A^n p_0 \end{aligned}$$

Matrizes estocásticas por colunas

Vetor de probabilidades

O vetor coluna v é vetor de probabilidades se

- entradas $\geqslant 0$
- soma das entradas $= 1$

Matriz estocástica por colunas

- Cada coluna é vetor de probabilidades
- Se entradas positivas (> 0), matriz é positiva
  notação: $A > 0$

Teorema de Perron–Frobenius

Conjunto dos vetores de probabilidade

$$C = \{x \in \mathbb{R}^n : x \geq 0, \sum x_i = 1\}$$

Teorema (Perron–Frobenius)

Seja $A > 0$ uma matriz estocástica por colunas.

- 1 *é autovalor simples de* A.
- 1 *tem exatamente um autovetor* π *em* C: $A\pi = \pi$ *chamado regime estacionário de* A
- *Se* λ *é outro autovalor de* A, *então* $|\lambda| < 1$

$$\therefore A^n p \to \pi \text{ as } n \to \infty \quad \forall p \in C.$$

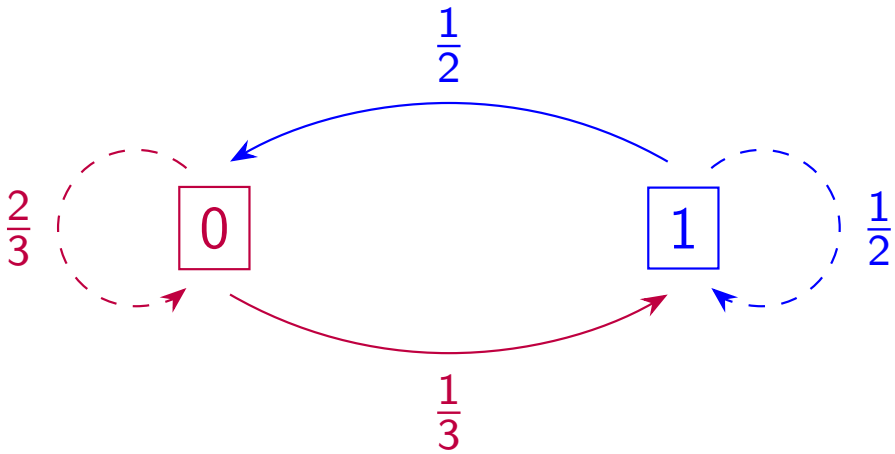
Propriedades das Cadeias de Markov

Corolário

Se $A > 0$ é a matriz de transição de uma cadeia de Markov

- 1 é um autovalor simples de A
  e todos os outros são menores (em módulo).
- Há um único estado estacionário π para o qual tendem as probabilidades dos estados
- Todas as colunas de A^n tendem a π quando $n \to \infty$

Revisitando o exemplo de Markov



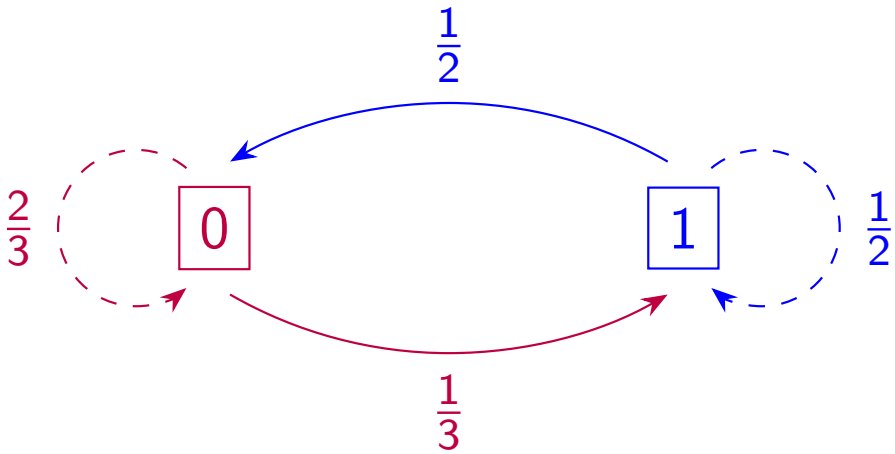
$$A = \begin{pmatrix} \frac{2}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix}$$

- Não é difícil calcular o autovetor de 1: $\pi = \begin{pmatrix} \frac{3}{5} \\ \frac{2}{5} \end{pmatrix}$
- Na força bruta, usando Octave

$$A = \begin{pmatrix} 0.66667 & 0.50000 \\ 0.33333 & 0.50000 \end{pmatrix} \qquad A^5 = \begin{pmatrix} 0.60005 & 0.59992 \\ 0.39995 & 0.40008 \end{pmatrix}$$

$$A^{10} = \begin{pmatrix} 0.60000 & 0.60000 \\ 0.40000 & 0.40000 \end{pmatrix}$$

Revisitando o exemplo de Markov



$$\pi = \begin{pmatrix} \frac{3}{5} \\ \frac{2}{5} \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$$

$n = \#$ sorteios	$b = \#$ brancas	$p_b = \frac{b}{n}$
10	5	0.50000
50	19	0.38000
100	51	0.51000
1000	412	0.41200
10000	3982	0.39820
100000	39822	0.39822

Mas se há entradas nulas?

Cuidado com A não positiva

- Se $A \geqslant 0$, não se garante a unicidade do estado estacionário.
- Por isso é importante que todo estado possa ser atingido a partir de qualquer outro estado com probabilidade não nula, como disse Markov.
- Veremos como evitaremos entradas nulas na aplicação do PageRank.

Voltando ao PageRank

O internauta imparcial

- Suponha que um internauta imparcial inicie numa página qualquer da internet
- Esta página pode ter (ou não) ligações de saída para outras n páginas.
- Ele segue uma ligação de saída com igual probabilidade $\frac{1}{n}$ (ou salta para outra página qualquer se $n = 0$)
- Ficando 1s em cada página visitada, conte o tempo gasto nela
- As páginas com mais votos tendem a ser mais visitadas
- As páginas apontadas por estas serão também mais visitadas

Modelando com cadeias de Markov

- Cada página é um estado.
- Suponha uma página com n links únicos.
- A transição segue os links de saída com igual probabilidade $\frac{1}{n}$.
- Importância da página: a entrada correspondente em π indica a média de tempo que passamos na página!

Problemas

Seja A a matriz de transição dessa cadeia de Markov.

1. $A \geqslant 0$, a imensa maioria das entradas é zero.
2. Páginas sem links de saída $\rightarrow$ coluna de zeros.
3. Vários grupos desconexos de páginas.

Problema 2: páginas sem links de saída.

Estratégias

Remover estas páginas e ignorar links para elas

- Solução inicial adotada no PageRank.
- Páginas de bom conteúdo mas sem links externos ficarão invisíveis na busca.

Saltar para uma página aleatória

- Mostrou-se que isso não altera a classificação das outras páginas.
- Substituir coluna de zeros por coluna de $\frac{1}{N}$, $N =$ número de páginas da internet.
- Assim, A torna-se matriz estocástica por colunas.

Problemas 1 e 3: $A \geqslant 0$ e blocos desconexos

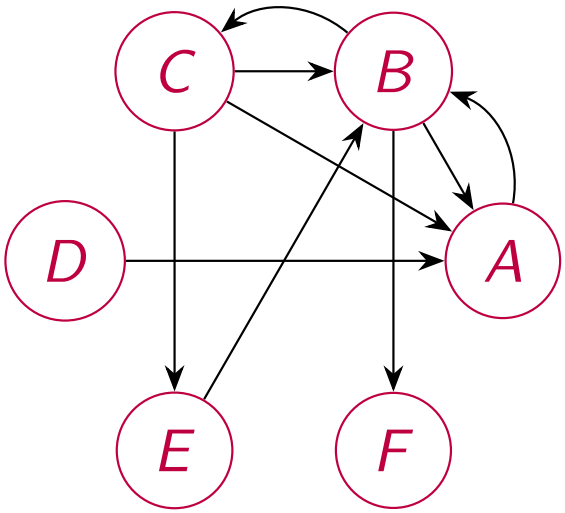
Estratégia de Page

- Conectar toda a rede, atribuindo uma chance $p \approx 0.15$ de saltar para uma página qualquer.
- $U = \left(\frac{1}{N}\right)_{N \times N}$ é matriz estocástica.
- $B = (1-p)A + pU$ é matriz estocástica
- $B > 0$!!!

Importância PageRank

- π: estado estacionário da cadeia de Markov com matriz B
- A importância da página é sua entrada correspondente em π

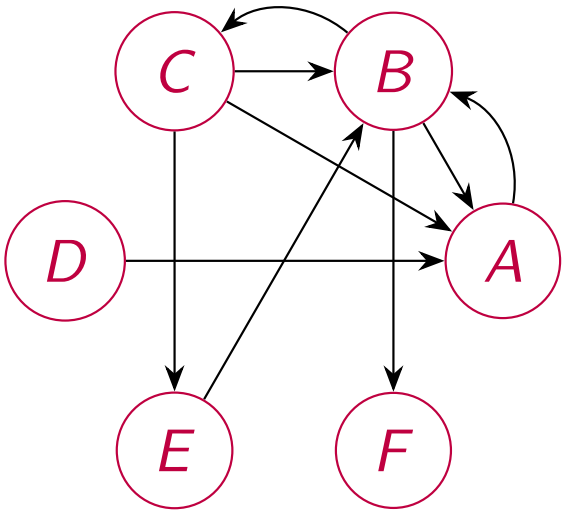
Exemplo: uma internet com 6 páginas



$$A = \begin{array}{c} \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & 1 & 0 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & 1 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}$$

Matriz de transição

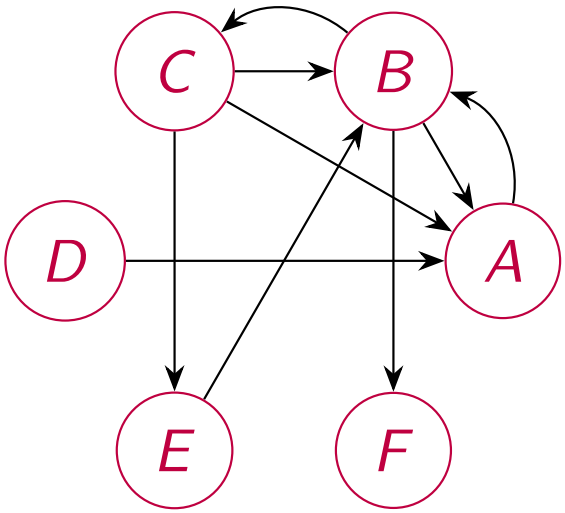
Exemplo: uma internet com 6 páginas



$$A = \begin{array}{c} \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & 1 & 0 & \frac{1}{6} \\ 1 & 0 & \frac{1}{3} & 0 & 1 & \frac{1}{6} \\ 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{6} \\ 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{6} \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}$$

Eliminando colunas de zeros

Exemplo: uma internet com 6 páginas

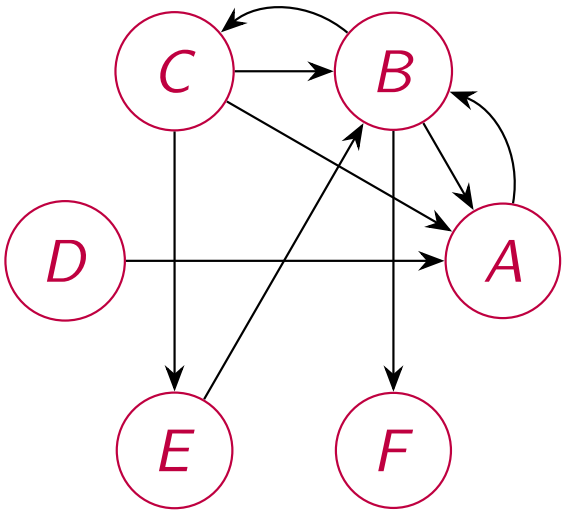


$$B = \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} \frac{17}{120} & \frac{23}{120} & \frac{23}{120} & \frac{35}{120} & \frac{17}{120} & \frac{20}{120} \\ \frac{35}{120} & \frac{17}{120} & \frac{23}{120} & \frac{17}{120} & \frac{35}{120} & \frac{20}{120} \\ \frac{17}{120} & \frac{23}{120} & \frac{17}{120} & \frac{17}{120} & \frac{17}{120} & \frac{20}{120} \\ \frac{17}{120} & \frac{17}{120} & \frac{17}{120} & \frac{17}{120} & \frac{17}{120} & \frac{20}{120} \\ \frac{17}{120} & \frac{17}{120} & \frac{23}{120} & \frac{17}{120} & \frac{17}{120} & \frac{20}{120} \\ \frac{17}{120} & \frac{23}{120} & \frac{17}{120} & \frac{17}{120} & \frac{17}{120} & \frac{20}{120} \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \end{array}$$

Eliminando zeros: $B > 0$

$$B = 0.15A + (1 - 0.15)U$$

Exemplo: uma internet com 6 páginas



$$\pi = \begin{pmatrix} \frac{67197}{362460} \\ \frac{74000}{362460} \\ \frac{56460}{362460} \\ \frac{52760}{362460} \\ \frac{55583}{362460} \\ \frac{56460}{362460} \end{pmatrix} = \begin{pmatrix} 0.225197 \\ 0.351899 \\ 0.145287 \\ 0.045582 \\ 0.086747 \\ 0.145287 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}$$

Estado estacionário π:

$$B > A > C = F > E > D$$

Calculando π no Octave

```
A = [0,  1/3,  1/3,  1,  0,  0;
     1,  0   ,  1/3,  0,  1,  0;
     0,  1/3,  0   ,  0,  0,  0;
     0,  0   ,  0   ,  0,  0,  0;
     0,  0   ,  1/3,  0,  0,  0;
     0,  1/3,  0   ,  0,  0,  0];
```

Calculando π no Octave

```
# Eliminando colunas nulas

n = size(A);

for i = 1:n
  s = sum( A(:,i) );
  if (s == 0)
    A(:,i) = (1/n)*ones(n,1);
  endif
endfor
```

Calculando π no Octave

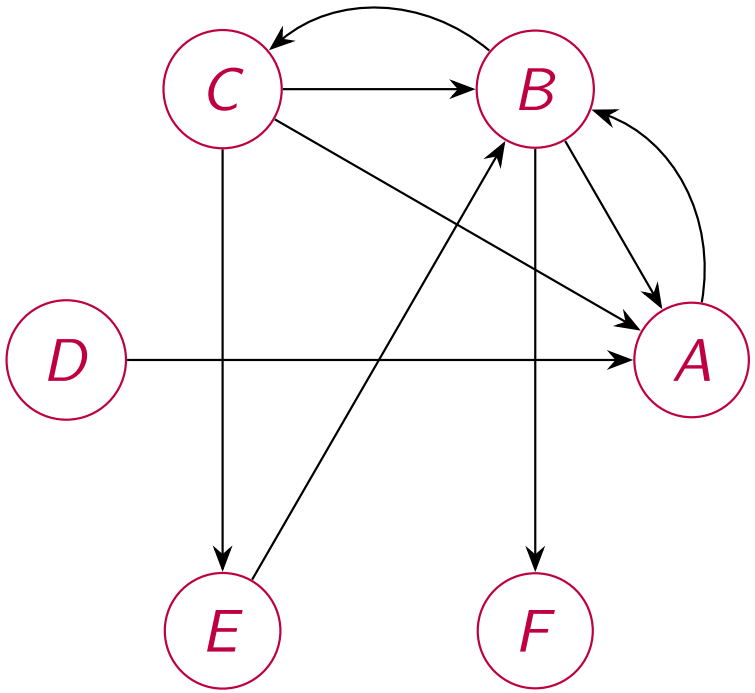
```
# Fazendo o grafo fortemente conexo

p = 0.15;
B = (1-p)*A + p*(1/n)*ones(n,n);

# Calculando o estado estacionário: 1ª coluna de B^50

C = B^50;
pi = C(:,1);
```

Calculando π no Octave: resultado



$$\pi = \begin{pmatrix} 0.225197 \\ 0.351899 \\ 0.145287 \\ 0.045582 \\ 0.086747 \\ 0.145287 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}$$

$$B > A > C = F > E > D$$

Outros exemplos

Ao vivo no Octave

- Times do Campeonato Brasileiro 2017 até a 4ª rodada
- Docentes do SME por colaboração
- Implicações na Lava-Jato