



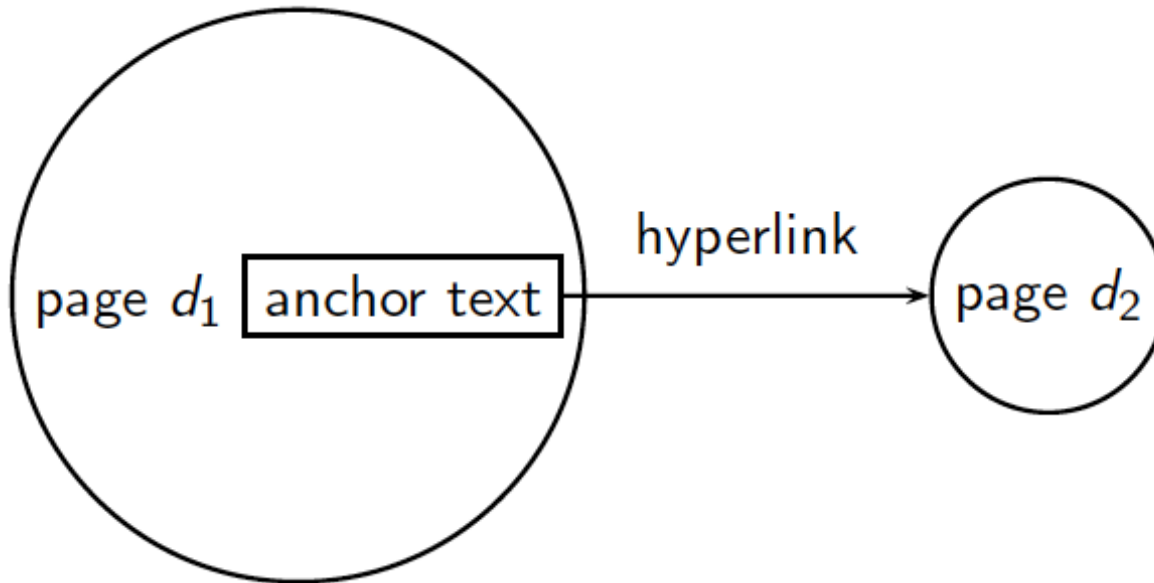
# Análise de Links

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)



# Web como Grafo Direcionado

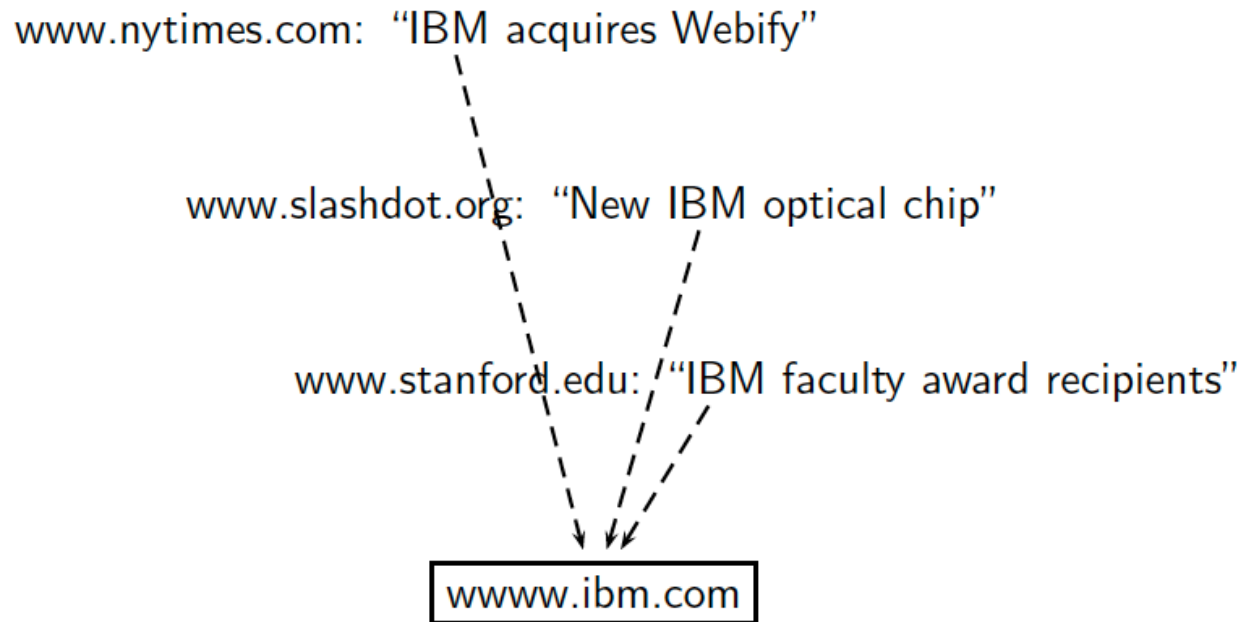


- Suposições:
  - Um hiperlink é um sinal de qualidade
  - Âncora: descreve resumidamente o conteúdo da página que aponta
    - Exemplo: “<a href=http://...> you can find cheap cars here</a>.”
    - Anchor text: “you can find cheap cars here”



# Âncora

- Útil para consultas navegacionais
  - Ex: consulta “IBM”



- Usualmente descrevem bem a página
- Pode ter peso maior no ranking do que o texto da página



# Google Bombs

- Busca com resultados “ruins” por causa de âncoras manipuladas
- Ex: “dangerous cult”
  - Primeiro resultado no Google: Scientology

The screenshot shows a Google search interface with the query "dangerous cult" entered in the search bar. Below the search bar, the "All" tab is selected. The search results indicate "About 14,200,000 results (0.39 seconds)". The first result is titled "1. **Scientology**. The infamous **Church of Scientology** was formed in 1953 by the writer L. Ron Hubbard. Arguably one of the richest cults, with high-profile members like Tom Cruise, it has remained controversial since its origin. Aug 22, 2015". Below the title is a link to "11 Religious Cults That Are Extremely Dangerous For Humanity" from "www.storypick.com/dangerous-cults/". At the bottom of the screenshot, a second result is partially visible, titled "10 of the Most Dangerous Religious Cults - TheRichest" from "www.therichest.com/rich-list/most.../10-of-the-most-dangerous-religious-cults/".

Google dangerous cult

All Videos Images News Shopping More Search tools

About 14,200,000 results (0.39 seconds)

1. **Scientology**. The infamous **Church of Scientology** was formed in 1953 by the writer L. Ron Hubbard. Arguably one of the richest cults, with high-profile members like Tom Cruise, it has remained controversial since its origin. Aug 22, 2015

11 Religious Cults That Are Extremely Dangerous For Humanity  
[www.storypick.com/dangerous-cults/](http://www.storypick.com/dangerous-cults/)

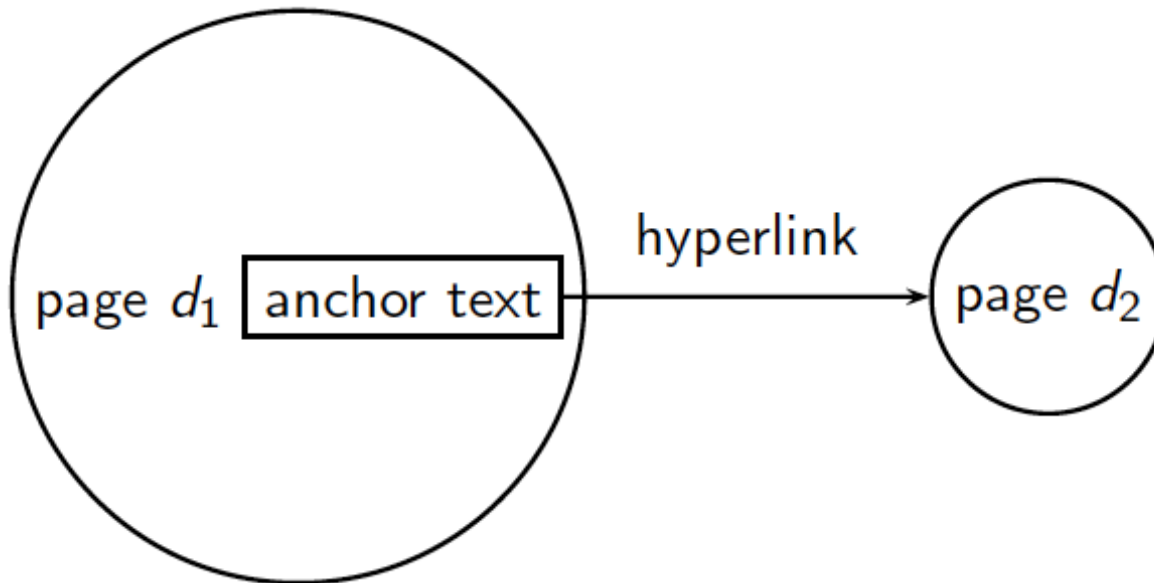
About this result • Feedback

10 of the Most Dangerous Religious Cults - TheRichest  
[www.therichest.com/rich-list/most.../10-of-the-most-dangerous-religious-cults/](http://www.therichest.com/rich-list/most.../10-of-the-most-dangerous-religious-cults/)  
Jun 3, 2014 - They talk of open threats and other dangerous methods which cult leaders use to ensure loyalty. The basis of the cult is a confusing mess of alien influence and the human psyche. But at the core, **Scientology** seems to be about a lot of money.



# Análise de Links

- Objetivo: computar a importância de páginas baseado na topologia





# Análise de Citações

- Oriundo da literatura científica
- Exemplo: “Miller (2001) has shown that physical”
- Citação: referência entre artigos científicos
- Pode medir
  - Similaridade de co-citação: dois artigos são similares se são citados pelos mesmos artigos
    - Na Web: operador “related:” do Google



# “Related:” do Google



[All](#) [Images](#) [Maps](#) [Shopping](#) [More ▾](#) [Search tools](#)

---

About 50 results (0.08 seconds)

**globo.com - Absolutamente tudo sobre notícias, esportes e ...**  
[www.globo.com/](http://www.globo.com/) ▾ [Translate this page](#)  
Só na globo.com você encontra tudo sobre o conteúdo e marcas do Grupo Globo. O melhor acervo de vídeos online sobre entretenimento, esportes e ...

**Google**  
<https://www.google.com.br/> ▾ [Translate this page](#)  
Versão brasileira do popular buscador e diretório. Utiliza também os dados do ODP.

**POP**  
[www.pop.com.br/](http://www.pop.com.br/) ▾ [Translate this page](#)  
COMUNICADO IMPORTANTE. O portal POP foi DESATIVADO no dia 29 de Abril de 2016. Agradecemos sua audiência durante todo o período em que o site ...

**R7 – Últimas notícias, vídeos, esportes, entretenimento e mais**  
[www.r7.com/](http://www.r7.com/) ▾ [Translate this page](#)  
Acompanhe as últimas notícias e vídeos, além de tudo sobre esportes e entretenimento. Conheça o conteúdo e os serviços do R7, o portal da Rede Record.

**Oi | Combo, TV, Celular, Internet, Fixo, Recarga**  
[www.oi.com.br/](http://www.oi.com.br/) ▾ [Translate this page](#)  
Descubra na Oi os melhores combos e planos para TV, Banda Larga, Internet Móvel, Celular e Fixo. Acesse a 2ª via da sua conta e muitos outros serviços.



# Análise de Citações

- Oriundo da literatura científica
- Exemplo: “Miller (2001) has shown that physical”
- Citação: referência entre artigos científicos
- Pode medir
  - Similaridade de co-citação: dois artigos são similares se são citados pelos mesmos artigos
    - Na Web: operador “related:” do Google
  - Impacto de um artigo científico: número de citações a ele
    - Na Web: inlinks ou backlinks
    - Um alto número de inlinks não necessariamente indica alta qualidade (link spam)





# PageRank

- Criado nos anos de 1960 por Pinsker e Narin
- Inspirado em citações científicas
- Citação = hiperlink
- Frequência ponderada de citações

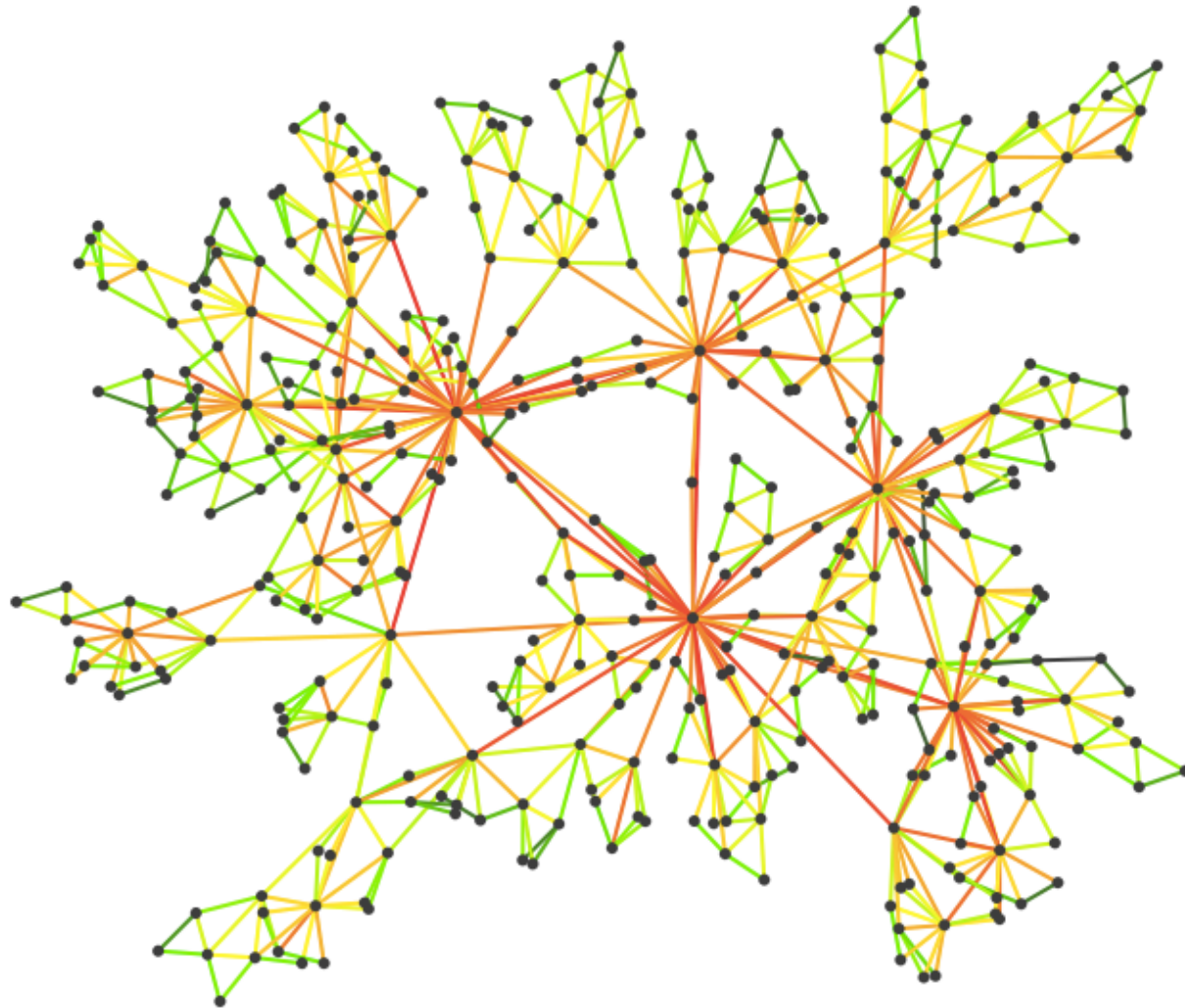


# Modelo por trás: Random Walk

- Inicia de uma página aleatória
- Em cada passo, segue os outlinks da página com igual probabilidade
- Após muitas visitas, cada página tem uma taxa de visitação estável
- Taxa de visitação = PageRank
- Probabilidade de estado estável



# Exemplo





# Cadeias de Markov

- Grafo com  $N$  estados (páginas)
- Começa em um estado e move para outro com uma certa probabilidade
- Matrix de transição  $N \times N$ 
  - Probabilidade de transição entre estados
  - Soma das probabilidades é 1



# Cadeias de Markov

- Exemplo:

**Example 11.1** According to Kemeny, Snell, and Thompson,<sup>2</sup> the Land of Oz is blessed by many things, but not by good weather. They never have two nice days in a row. If they have a nice day, they are just as likely to have snow as rain the next day. If they have snow or rain, they have an even chance of having the same the next day. If there is change from snow or rain, only half of the time is this a change to a nice day. With this information we form a Markov chain as follows. We take as states the kinds of weather R, N, and S. From the above information we determine the transition probabilities. These are most conveniently represented in a square array as

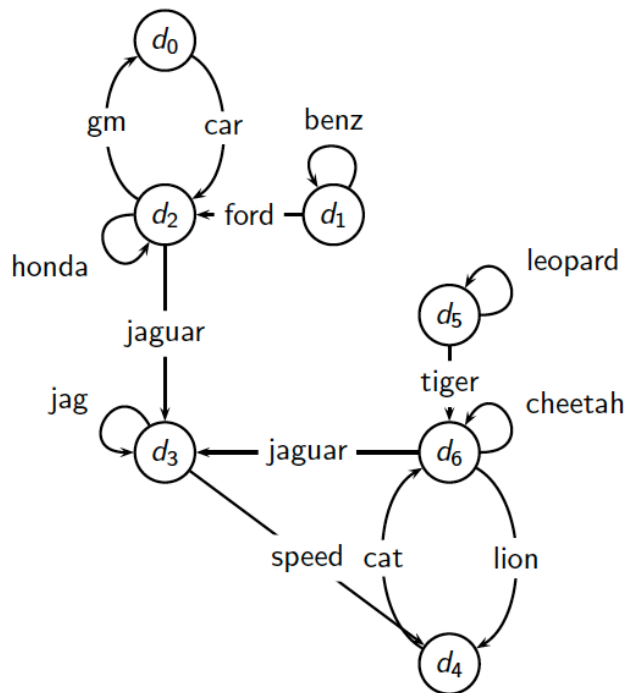
$$\mathbf{P} = \begin{matrix} & \begin{matrix} R & N & S \end{matrix} \\ \begin{matrix} R \\ N \\ S \end{matrix} & \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{matrix} .$$

Fonte :[https://www.dartmouth.edu/~chance/teaching\\_aids/books\\_articles/probability\\_book/Chapter11.pdf](https://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/Chapter11.pdf)



# Cadeias de Markov

- Grafo com N estados (páginas)
- Matrix de transição N x N
  - Probabilidade de transição entre páginas
  - Soma das probabilidades dos outlinks de uma página é 1

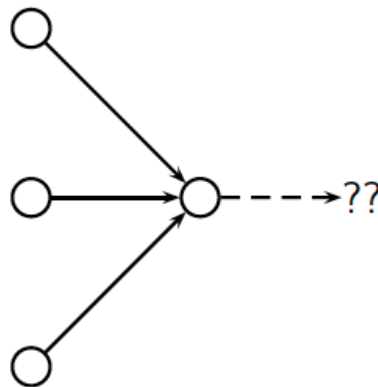


|       | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $d_0$ | 0.00  | 0.00  | 1.00  | 0.00  | 0.00  | 0.00  | 0.00  |
| $d_1$ | 0.00  | 0.50  | 0.50  | 0.00  | 0.00  | 0.00  | 0.00  |
| $d_2$ | 0.33  | 0.00  | 0.33  | 0.33  | 0.00  | 0.00  | 0.00  |
| $d_3$ | 0.00  | 0.00  | 0.00  | 0.50  | 0.50  | 0.00  | 0.00  |
| $d_4$ | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 1.00  |
| $d_5$ | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 0.50  | 0.50  |
| $d_6$ | 0.00  | 0.00  | 0.00  | 0.33  | 0.33  | 0.00  | 0.33  |



# Cadeias de Markov

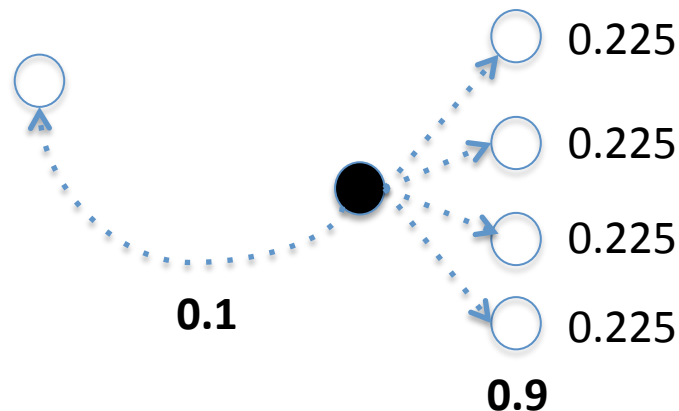
- PageRank = taxa de visitação após muitas visitas
- Grafo tem que ser ergótico
  1. Existe um caminho de qualquer página para outra (irredutível): sem dead-ends
    - A Web possui dead ends
    - Random walk pode ficar presa em dead ends





# Solução para Dead Ends: Teletransporte

- Não fica preso em dead ends
- Em dead-end: Pular para uma página aleatória com  $\text{prob} = 1/N$  ( $N$ =número de nós)
- Qdo não:
  - Pular com uma probabilidade  $p$  (taxa de teletransporte) para uma página aleatória
  - Ir com probabilidade  $1-p$  para um dos outlinks do nó
    - Por exemplo: se  $p = 0.1$  e um nó com 4 outlinks:  $(1-0.1)/4=0.225$

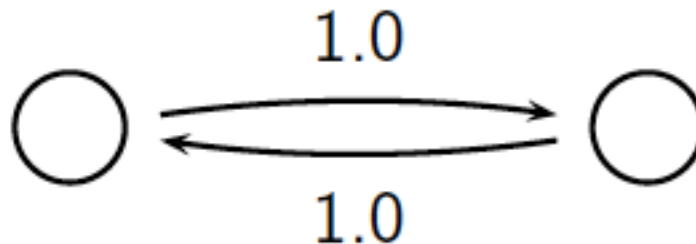






# Cadeias de Markov

- PageRank = taxa de visitação após muitas visitas
- Grafo tem que ser ergótico
  1. Existe um caminho de qualquer página para outra (irredutível): sem dead-ends
  2. Aperiódico: o grafo não pode ser particionado tal que o random walker visite as partições sequencialmente



Grafo periódico



# Cadeias de Markov Ergóticas

- Após um longo período, cada estado é visitado na proporção da taxa de visitação
- Não importa onde começa
- Teletransporte faz o grafo ergótico
- Probabilidade “estável” de visitação: PageRank



# Cálculo do PageRank

- Vetor de probabilidade diz onde o random walker está:

$$\vec{x} = (x_1, \dots, x_N)$$

- Exemplo:

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & 0 \\ 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{pmatrix}$$

$$\begin{pmatrix} 0.05 & 0.01 & 0.0 & \dots & 0.2 & \dots & 0.01 & 0.05 & 0.03 \\ 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{pmatrix}$$

$$\sum x_i = 1$$

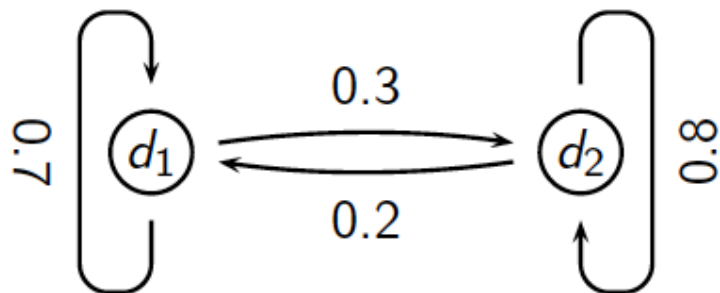


# Passos do PageRank

- Matrix de transição  $P$
- Próximo passo:  $\vec{x}P$
- Depois de muitos passos (multiplicação por  $P$  até convergir):
  - Steady state:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$
  - Taxa de visitação no longo prazo (PageRank)
  - Uma entrada por página



# Computando PageRank



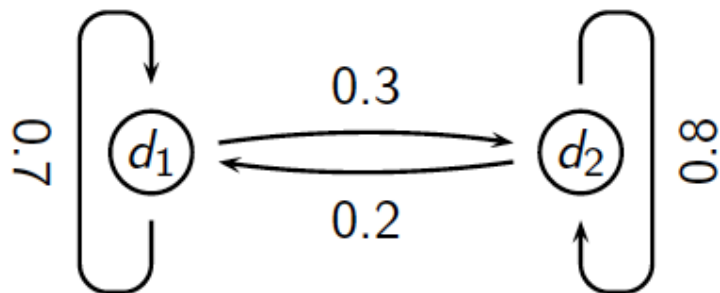
|            | $x_1$<br>$P_t(d_1)$ | $x_2$<br>$P_t(d_2)$ |                                  |                                  |
|------------|---------------------|---------------------|----------------------------------|----------------------------------|
|            |                     |                     | $P_{11} = 0.7$<br>$P_{21} = 0.2$ | $P_{12} = 0.3$<br>$P_{22} = 0.8$ |
| $t_0$      | 0                   | 1                   | 0.2                              | 0.8                              |
| $t_1$      | 0.2                 | 0.8                 | 0.3                              | 0.7                              |
| $t_2$      | 0.3                 | 0.7                 | 0.35                             | 0.65                             |
| $t_3$      | 0.35                | 0.65                | 0.375                            | 0.625                            |
|            |                     |                     |                                  | ...                              |
| $t_\infty$ | 0.4                 | 0.6                 | 0.4                              | 0.6                              |

← Matriz de Transição

$$\text{vector} = \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$$



# Passo 1



|            | $x_1$<br>$P_t(d_1)$ | $x_2$<br>$P_t(d_2)$ |                                  |                                  |
|------------|---------------------|---------------------|----------------------------------|----------------------------------|
|            |                     |                     | $P_{11} = 0.7$<br>$P_{21} = 0.2$ | $P_{12} = 0.3$<br>$P_{22} = 0.8$ |
| $t_0$      | 0                   | 1                   | 0.2                              | 0.8                              |
| $t_1$      | 0.2                 | 0.8                 | 0.3                              | 0.7                              |
| $t_2$      | 0.3                 | 0.7                 | 0.35                             | 0.65                             |
| $t_3$      | 0.35                | 0.65                | 0.375                            | 0.625                            |
|            |                     |                     | ...                              |                                  |
| $t_\infty$ | 0.4                 | 0.6                 | 0.4                              | 0.6                              |

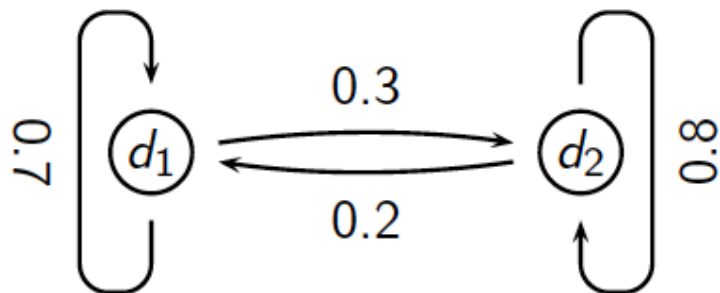
Matriz de Transição

$\begin{bmatrix} 0 & 1 \end{bmatrix} * \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}$   
 $\vec{x}P$

vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$



## Passo 2



|            | $x_1$<br>$P_t(d_1)$ | $x_2$<br>$P_t(d_2)$ |                                  |                                  |
|------------|---------------------|---------------------|----------------------------------|----------------------------------|
|            |                     |                     | $P_{11} = 0.7$<br>$P_{21} = 0.2$ | $P_{12} = 0.3$<br>$P_{22} = 0.8$ |
| $t_0$      | 0                   | 1                   | 0.2                              | 0.8                              |
| $t_1$      | 0.2                 | 0.8                 | 0.3                              | 0.7                              |
| $t_2$      | 0.3                 | 0.7                 | 0.35                             | 0.65                             |
| $t_3$      | 0.35                | 0.65                | 0.375                            | 0.625                            |
|            |                     |                     | ...                              | ...                              |
| $t_\infty$ | 0.4                 | 0.6                 | 0.4                              | 0.6                              |

Matriz de Transição

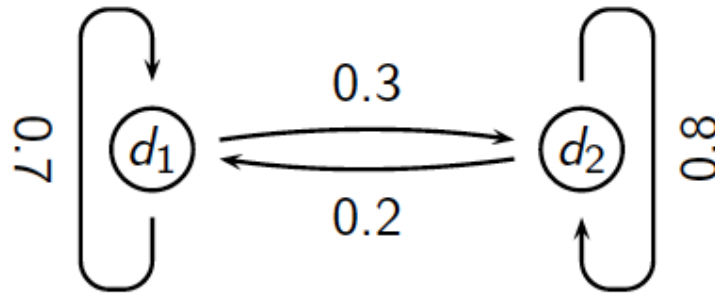
$\begin{bmatrix} 0.2 & 0.8 \end{bmatrix} * \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}$

$\vec{x}P$

vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$



## Passo 3



|            | $x_1$<br>$P_t(d_1)$ | $x_2$<br>$P_t(d_2)$ |                                  |                                  |
|------------|---------------------|---------------------|----------------------------------|----------------------------------|
|            |                     |                     | $P_{11} = 0.7$<br>$P_{21} = 0.2$ | $P_{12} = 0.3$<br>$P_{22} = 0.8$ |
| $t_0$      | 0                   | 1                   | 0.2                              | 0.8                              |
| $t_1$      | 0.2                 | 0.8                 | 0.3                              | 0.7                              |
| $t_2$      | 0.3                 | 0.7                 | 0.35                             | 0.65                             |
| $t_3$      | 0.35                | 0.65                | 0.375                            | 0.625                            |
|            |                     |                     | ...                              |                                  |
| $t_\infty$ | 0.4                 | 0.6                 | 0.4                              | 0.6                              |

Matriz de Transição

$\begin{bmatrix} 0.3 & 0.7 \end{bmatrix} * \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}$

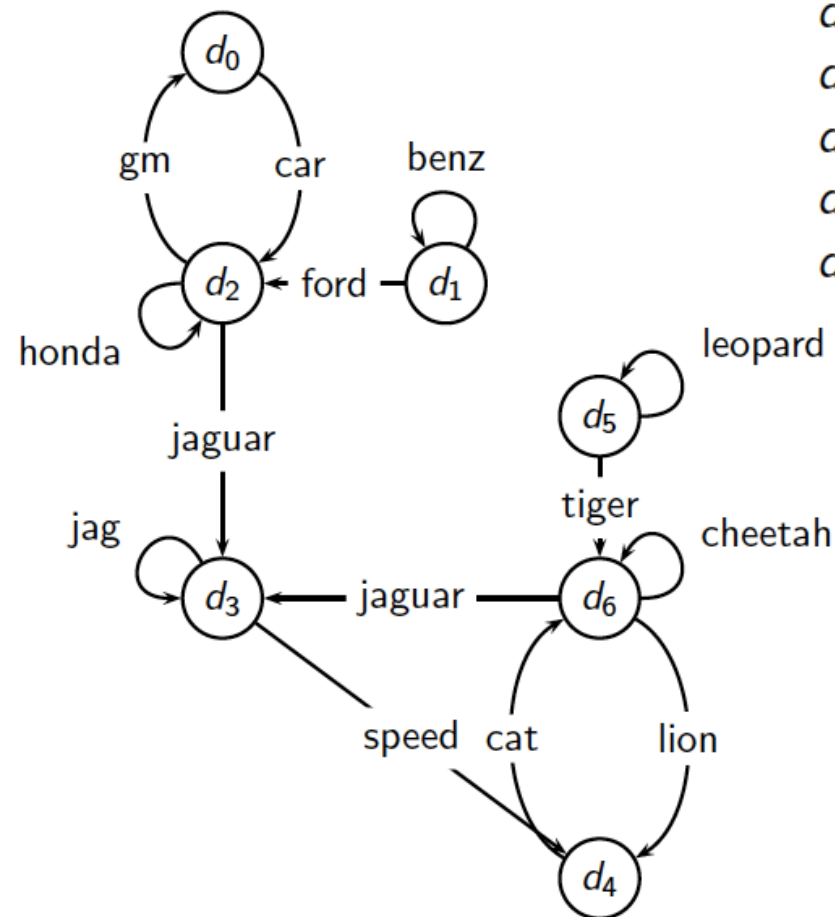
$\vec{x}P$

vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$





# Outro Exemplo

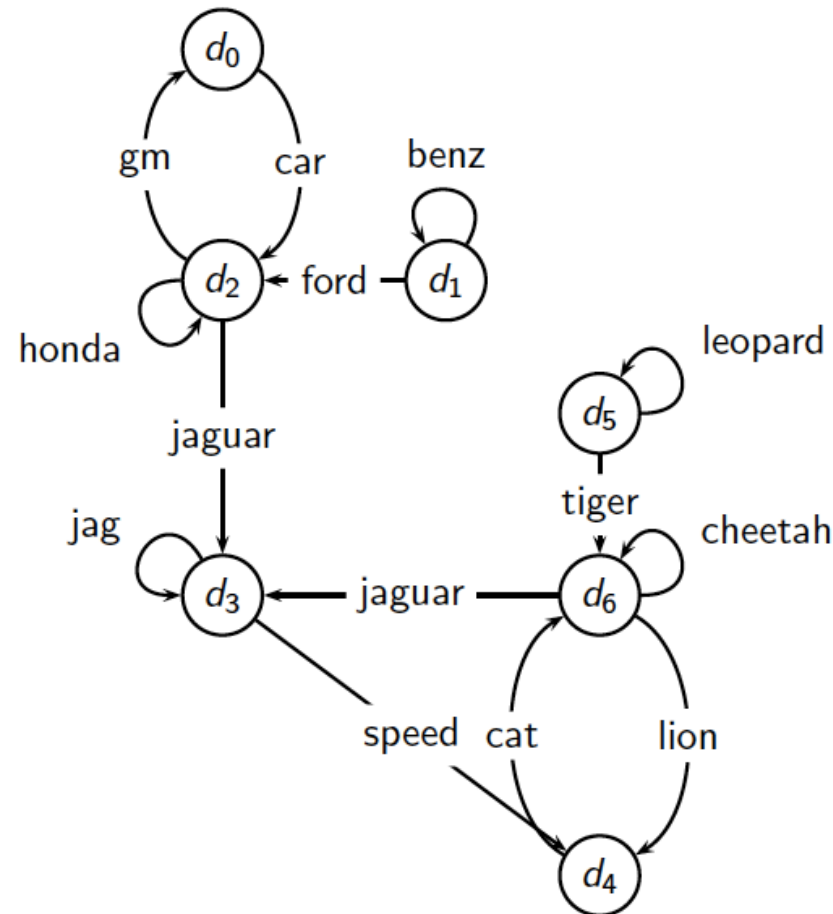


|       | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $d_0$ | 0.00  | 0.00  | 1.00  | 0.00  | 0.00  | 0.00  | 0.00  |
| $d_1$ | 0.00  | 0.50  | 0.50  | 0.00  | 0.00  | 0.00  | 0.00  |
| $d_2$ | 0.33  | 0.00  | 0.33  | 0.33  | 0.00  | 0.00  | 0.00  |
| $d_3$ | 0.00  | 0.00  | 0.00  | 0.50  | 0.50  | 0.00  | 0.00  |
| $d_4$ | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 1.00  |
| $d_5$ | 0.00  | 0.00  | 0.00  | 0.00  | 0.00  | 0.50  | 0.50  |
| $d_6$ | 0.00  | 0.00  | 0.00  | 0.33  | 0.33  | 0.00  | 0.33  |

**Matriz de transição**



# Exemplo com Teletransporte



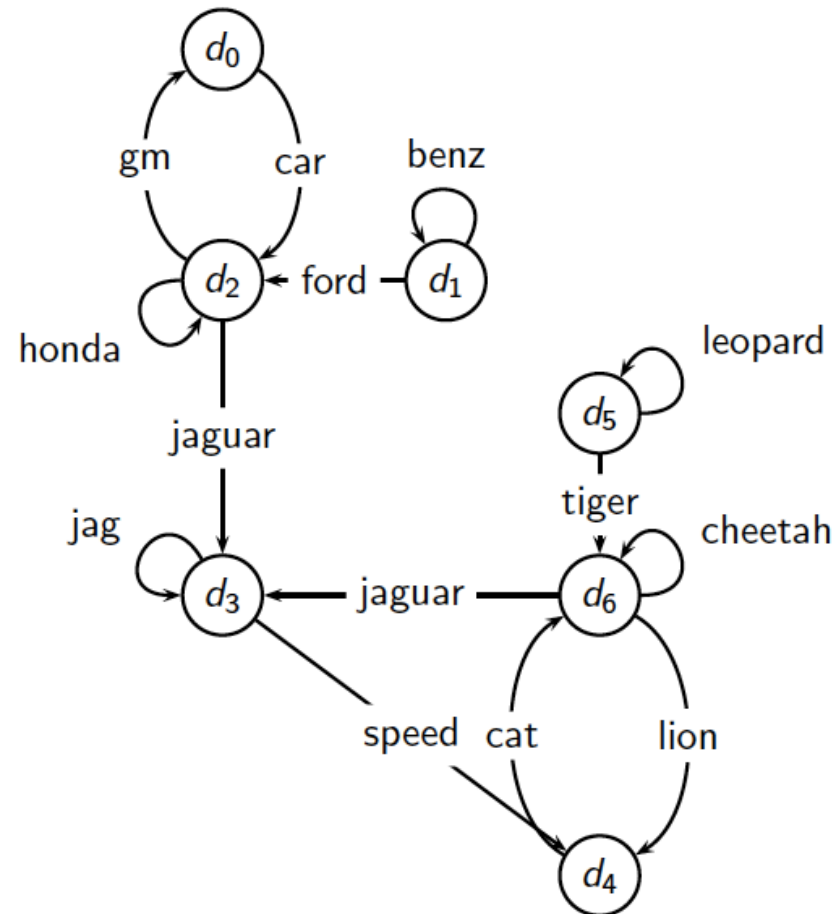
|       | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $d_0$ | 0.02  | 0.02  | 0.88  | 0.02  | 0.02  | 0.02  | 0.02  |
| $d_1$ | 0.02  | 0.45  | 0.45  | 0.02  | 0.02  | 0.02  | 0.02  |
| $d_2$ | 0.31  | 0.02  | 0.31  | 0.31  | 0.02  | 0.02  | 0.02  |
| $d_3$ | 0.02  | 0.02  | 0.02  | 0.45  | 0.45  | 0.02  | 0.02  |
| $d_4$ | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.88  |
| $d_5$ | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.45  | 0.45  |
| $d_6$ | 0.02  | 0.02  | 0.02  | 0.31  | 0.31  | 0.02  | 0.31  |

**Matriz de transição com teletransporte (P)**

Taxa de teletransporte = 0.12



# Exemplo com Teletransporte



|       | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $d_0$ | 0.02  | 0.02  | 0.88  | 0.02  | 0.02  | 0.02  | 0.02  |
| $d_1$ | 0.02  | 0.45  | 0.45  | 0.02  | 0.02  | 0.02  | 0.02  |
| $d_2$ | 0.31  | 0.02  | 0.31  | 0.31  | 0.02  | 0.02  | 0.02  |
| $d_3$ | 0.02  | 0.02  | 0.02  | 0.45  | 0.45  | 0.02  | 0.02  |
| $d_4$ | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.88  |
| $d_5$ | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.45  | 0.45  |
| $d_6$ | 0.02  | 0.02  | 0.02  | 0.31  | 0.31  | 0.02  | 0.31  |

**Matriz de transição com teletransporte (P)**

Taxa de teletransporte = 0.12



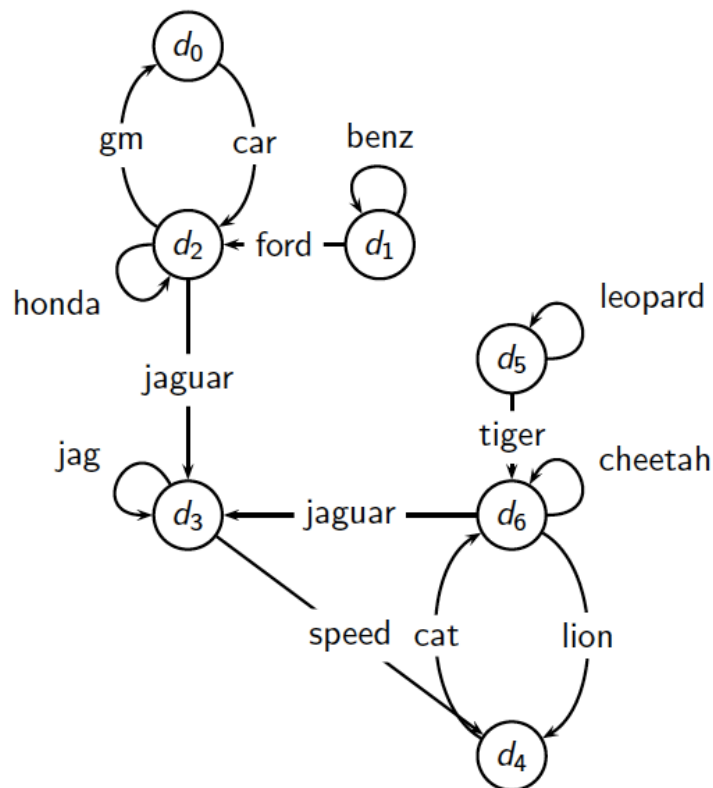
# Iterações

|       | $\bar{x}$ | $\bar{x}P^1$ | $\bar{x}P^2$ | $\bar{x}P^3$ | $\bar{x}P^4$ | $\bar{x}P^5$ | $\bar{x}P^6$ | $\bar{x}P^7$ | $\bar{x}P^8$ | $\bar{x}P^9$ | $\bar{x}P^{10}$ | $\bar{x}P^{11}$ | $\bar{x}P^{12}$ | $\bar{x}P^{13}$ |
|-------|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-----------------|-----------------|-----------------|-----------------|
| $d_0$ | 0.14      | 0.06         | 0.09         | 0.07         | 0.07         | 0.06         | 0.06         | 0.06         | 0.06         | 0.05         | 0.05            | 0.05            | 0.05            | 0.05            |
| $d_1$ | 0.14      | 0.08         | 0.06         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04            | 0.04            | 0.04            | 0.04            |
| $d_2$ | 0.14      | 0.25         | 0.18         | 0.17         | 0.15         | 0.14         | 0.13         | 0.12         | 0.12         | 0.12         | 0.12            | 0.11            | 0.11            | 0.11            |
| $d_3$ | 0.14      | 0.16         | 0.23         | 0.24         | 0.24         | 0.24         | 0.24         | 0.25         | 0.25         | 0.25         | 0.25            | 0.25            | 0.25            | 0.25            |
| $d_4$ | 0.14      | 0.12         | 0.16         | 0.19         | 0.19         | 0.20         | 0.21         | 0.21         | 0.21         | 0.21         | 0.21            | 0.21            | 0.21            | 0.21            |
| $d_5$ | 0.14      | 0.08         | 0.06         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04            | 0.04            | 0.04            | 0.04            |
| $d_6$ | 0.14      | 0.25         | 0.23         | 0.25         | 0.27         | 0.28         | 0.29         | 0.29         | 0.30         | 0.30         | 0.30            | 0.30            | 0.31            | 0.31            |

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|       |       | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
| $P =$ | $d_0$ | 0.02  | 0.02  | 0.88  | 0.02  | 0.02  | 0.02  | 0.02  |
|       | $d_1$ | 0.02  | 0.45  | 0.45  | 0.02  | 0.02  | 0.02  | 0.02  |
|       | $d_2$ | 0.31  | 0.02  | 0.31  | 0.31  | 0.02  | 0.02  | 0.02  |
|       | $d_3$ | 0.02  | 0.02  | 0.02  | 0.45  | 0.45  | 0.02  | 0.02  |
|       | $d_4$ | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.88  |
|       | $d_5$ | 0.02  | 0.02  | 0.02  | 0.02  | 0.02  | 0.45  | 0.45  |
|       | $d_6$ | 0.02  | 0.02  | 0.02  | 0.31  | 0.31  | 0.02  | 0.31  |



# Resultado



|       | $\vec{x}$ | $\vec{x}P^1$ | $\vec{x}P^2$ | $\vec{x}P^3$ | $\vec{x}P^4$ | $\vec{x}P^5$ | $\vec{x}P^6$ | $\vec{x}P^7$ | $\vec{x}P^8$ | $\vec{x}P^9$ | $\vec{x}P^{10}$ | $\vec{x}P^{11}$ | $\vec{x}P^{12}$ | $\vec{x}P^{13}$ |
|-------|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-----------------|-----------------|-----------------|-----------------|
| $d_0$ | 0.14      | 0.06         | 0.09         | 0.07         | 0.07         | 0.06         | 0.06         | 0.06         | 0.06         | 0.05         | 0.05            | 0.05            | 0.05            | 0.05            |
| $d_1$ | 0.14      | 0.08         | 0.06         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04            | 0.04            | 0.04            | 0.04            |
| $d_2$ | 0.14      | 0.25         | 0.18         | 0.17         | 0.15         | 0.14         | 0.13         | 0.12         | 0.12         | 0.12         | 0.12            | 0.11            | 0.11            | 0.11            |
| $d_3$ | 0.14      | 0.16         | 0.23         | 0.24         | 0.24         | 0.24         | 0.24         | 0.25         | 0.25         | 0.25         | 0.25            | 0.25            | 0.25            | 0.25            |
| $d_4$ | 0.14      | 0.12         | 0.16         | 0.19         | 0.19         | 0.20         | 0.21         | 0.21         | 0.21         | 0.21         | 0.21            | 0.21            | 0.21            | 0.21            |
| $d_5$ | 0.14      | 0.08         | 0.06         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04         | 0.04            | 0.04            | 0.04            | 0.04            |
| $d_6$ | 0.14      | 0.25         | 0.23         | 0.25         | 0.27         | 0.28         | 0.29         | 0.29         | 0.30         | 0.30         | 0.30            | 0.30            | 0.31            | 0.31            |



# Sumário

- Pré-processamento
  - Dado um grafo, construir matriz de transição
  - Aplicar teletransporte
  - Computar pagerank
- Processamento de consultas
  - Recuperar páginas que satisfazem a consulta
  - Ranquear baseado em pagerank



# Problemas

- Usuários da Web não são random walkers
  - Botão de back, bookmarks e buscas
- PageRank isolado pode gerar resultados ruins
  - Considere a consulta: serviço de vídeo
  - Página do Yahoo contém as duas palavras e tem alto pagerank
  - Seria bem ranqueada
- Na prática: ranqueamento de acordo com a combinação de vários fatores

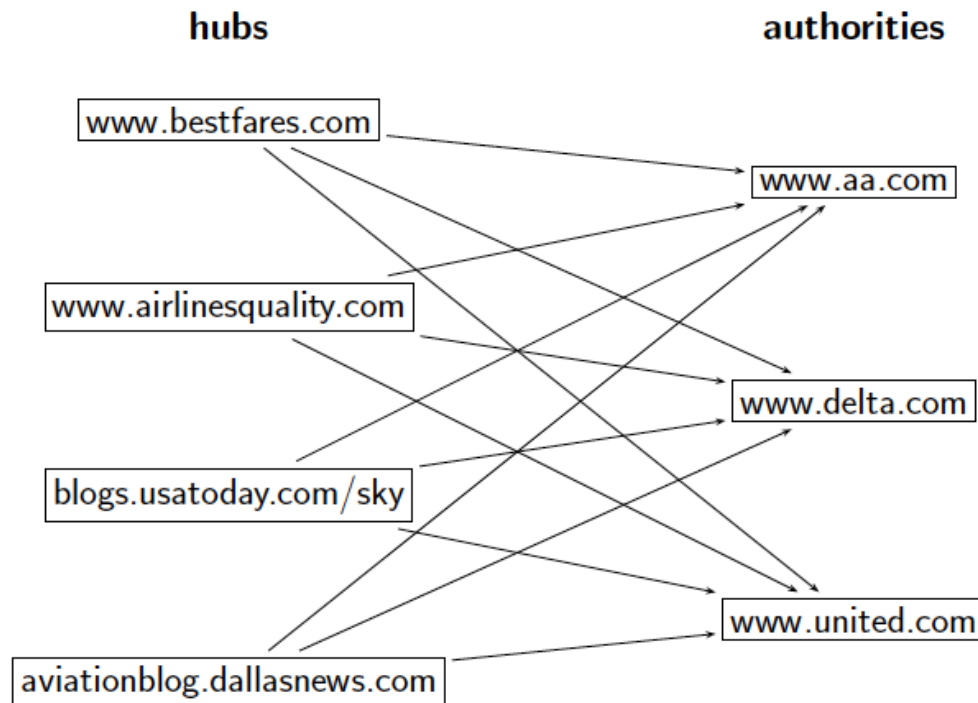


# Importância do PageRank

- Conhecido como o componente mais importante do ranking
- Na realidade:
  - Outros componentes tão importantes quanto: âncora, frases, proximidade, tiered indexes
  - Problema com link spam
- Também usado em crawling



- Premissa: dois tipos diferentes de relevância na Web
  - Hub: página que aponta para várias páginas importantes (autoridades)
  - Autoridade: páginas importantes apontadas por hubs





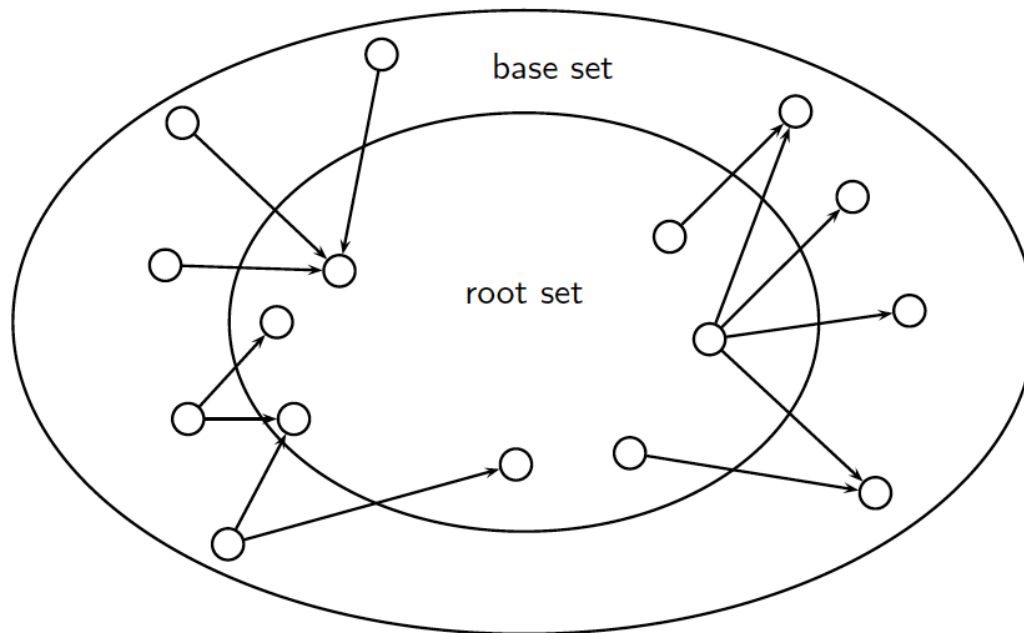
# HITS

- Premissa: dois tipos diferentes de relevância na Web
  - Hub: página que aponta para várias páginas importantes (autoridades)
  - Autoridade: páginas importantes apontadas por hubs
- PageRank não faz essa suposição
- Definição circular:
  - Um bom hub aponta para várias autoridades
  - Uma boa autoridade é apontada por vários hubs



# Computação do HITS

1. Usuário realiza uma busca
2. O resultado da busca é o conjunto raiz
3. Adicionam-se os inlinks e outlinks do conjunto raiz (conjunto base)
4. Computam-se hubs e autoridades para esse grafo

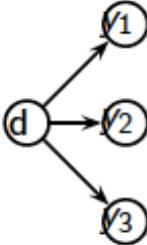




# Computação de HITS

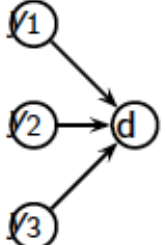
- Processo iterativo

For all  $d$ :  $h(d) = \sum_{d \mapsto y} a(y)$



```
graph LR; d((d)) --> y1((y1)); d --> y2((y2)); d --> y3((y3));
```

For all  $d$ :  $a(d) = \sum_{y \mapsto d} h(y)$



```
graph LR; y1((y1)) --> d((d)); y2((y2)) --> d; y3((y3)) --> d;
```

- Iterar nesses dois passos até convergir
- Em notação de matrix:  $\vec{h} = A\vec{a}$   $\vec{a} = A^T\vec{h}$ 
  - $A$  : matriz de adjacência

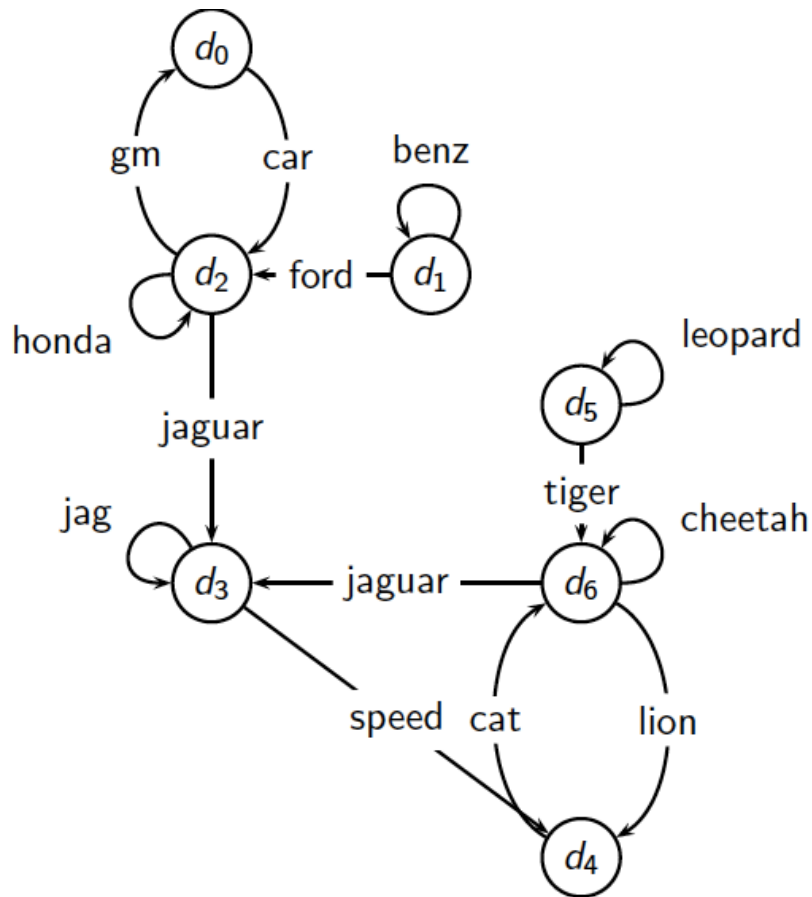


# Computação de HITS

- Depois de convergir (dois rankings)
  - Páginas com  $h$  mais altos são os hubs
  - Páginas com  $a$  mais altos são os autoridades
- Valores relativos
- Em geral, converge em poucas iterações



# Computando HITS: Exemplo

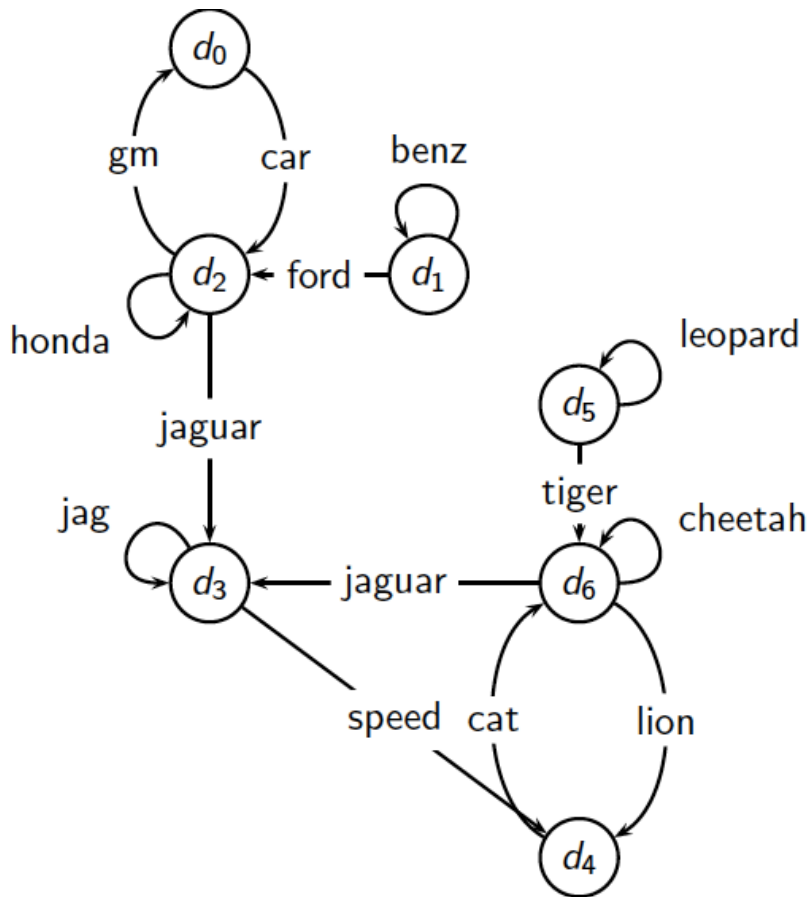


|       | $d_0$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|-------|
| $d_0$ | 0     | 0     | 1     | 0     | 0     | 0     | 0     |
| $d_1$ | 0     | 1     | 1     | 0     | 0     | 0     | 0     |
| $d_2$ | 1     | 0     | 1     | 2     | 0     | 0     | 0     |
| $d_3$ | 0     | 0     | 0     | 1     | 1     | 0     | 0     |
| $d_4$ | 0     | 0     | 0     | 0     | 0     | 0     | 1     |
| $d_5$ | 0     | 0     | 0     | 0     | 0     | 1     | 1     |
| $d_6$ | 0     | 0     | 0     | 2     | 1     | 0     | 1     |

**Matriz de adjacência: A**



# Computando HITS: Exemplo



|       | $\vec{h}_0$ | $\vec{h}_1$ | $\vec{h}_2$ | $\vec{h}_3$ | $\vec{h}_4$ | $\vec{h}_5$ |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|
| $d_0$ | 0.14        | 0.06        | 0.04        | 0.04        | 0.03        | 0.03        |
| $d_1$ | 0.14        | 0.08        | 0.05        | 0.04        | 0.04        | 0.04        |
| $d_2$ | 0.14        | 0.28        | 0.32        | 0.33        | 0.33        | 0.33        |
| $d_3$ | 0.14        | 0.14        | 0.17        | 0.18        | 0.18        | 0.18        |
| $d_4$ | 0.14        | 0.06        | 0.04        | 0.04        | 0.04        | 0.04        |
| $d_5$ | 0.14        | 0.08        | 0.05        | 0.04        | 0.04        | 0.04        |
| $d_6$ | 0.14        | 0.30        | 0.33        | 0.34        | 0.35        | 0.35        |

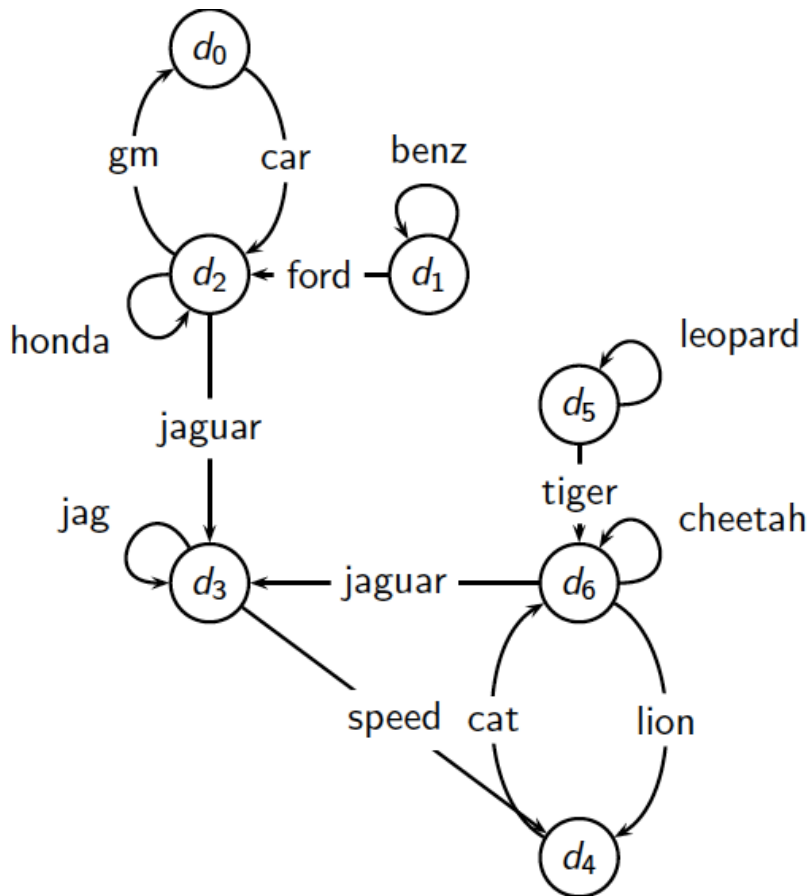
**Vetores de hubs**

|       | $\vec{a}_1$ | $\vec{a}_2$ | $\vec{a}_3$ | $\vec{a}_4$ | $\vec{a}_5$ | $\vec{a}_6$ | $\vec{a}_7$ |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| $d_0$ | 0.06        | 0.09        | 0.10        | 0.10        | 0.10        | 0.10        | 0.10        |
| $d_1$ | 0.06        | 0.03        | 0.01        | 0.01        | 0.01        | 0.01        | 0.01        |
| $d_2$ | 0.19        | 0.14        | 0.13        | 0.12        | 0.12        | 0.12        | 0.12        |
| $d_3$ | 0.31        | 0.43        | 0.46        | 0.46        | 0.46        | 0.47        | 0.47        |
| $d_4$ | 0.13        | 0.14        | 0.16        | 0.16        | 0.16        | 0.16        | 0.16        |
| $d_5$ | 0.06        | 0.03        | 0.02        | 0.01        | 0.01        | 0.01        | 0.01        |
| $d_6$ | 0.19        | 0.14        | 0.13        | 0.13        | 0.13        | 0.13        | 0.13        |

**Vetores de autoridades**



# Computando HITS: Exemplo



|       | $\vec{h}_0$ | $\vec{h}_1$ | $\vec{h}_2$ | $\vec{h}_3$ | $\vec{h}_4$ | $\vec{h}_5$ |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|
| $d_0$ | 0.14        | 0.06        | 0.04        | 0.04        | 0.03        | 0.03        |
| $d_1$ | 0.14        | 0.08        | 0.05        | 0.04        | 0.04        | 0.04        |
| $d_2$ | 0.14        | 0.28        | 0.32        | 0.33        | 0.33        | 0.33        |
| $d_3$ | 0.14        | 0.14        | 0.17        | 0.18        | 0.18        | 0.18        |
| $d_4$ | 0.14        | 0.06        | 0.04        | 0.04        | 0.04        | 0.04        |
| $d_5$ | 0.14        | 0.08        | 0.05        | 0.04        | 0.04        | 0.04        |
| $d_6$ | 0.14        | 0.30        | 0.33        | 0.34        | 0.35        | 0.35        |

**Vetores de hubs**

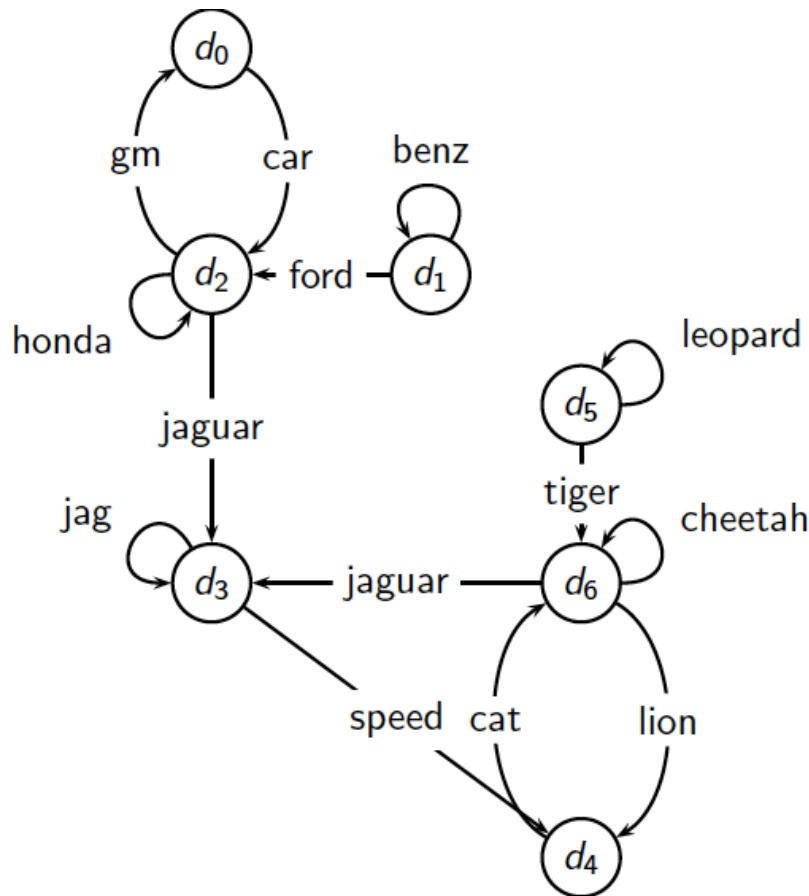
|       | $\vec{a}_1$ | $\vec{a}_2$ | $\vec{a}_3$ | $\vec{a}_4$ | $\vec{a}_5$ | $\vec{a}_6$ | $\vec{a}_7$ |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| $d_0$ | 0.06        | 0.09        | 0.10        | 0.10        | 0.10        | 0.10        | 0.10        |
| $d_1$ | 0.06        | 0.03        | 0.01        | 0.01        | 0.01        | 0.01        | 0.01        |
| $d_2$ | 0.19        | 0.14        | 0.13        | 0.12        | 0.12        | 0.12        | 0.12        |
| $d_3$ | 0.31        | 0.43        | 0.46        | 0.46        | 0.46        | 0.47        | 0.47        |
| $d_4$ | 0.13        | 0.14        | 0.16        | 0.16        | 0.16        | 0.16        | 0.16        |
| $d_5$ | 0.06        | 0.03        | 0.02        | 0.01        | 0.01        | 0.01        | 0.01        |
| $d_6$ | 0.19        | 0.14        | 0.13        | 0.13        | 0.13        | 0.13        | 0.13        |

**Vetores de autoridades**





# Computando HITS: Exemplo



|       | <i>a</i> | <i>h</i> |
|-------|----------|----------|
| $d_0$ | 0.10     | 0.03     |
| $d_1$ | 0.01     | 0.04     |
| $d_2$ | 0.12     | 0.33     |
| $d_3$ | 0.47     | 0.18     |
| $d_4$ | 0.16     | 0.04     |
| $d_5$ | 0.01     | 0.04     |
| $d_6$ | 0.13     | 0.35     |

- Páginas com mais in-degree:  $d_2$ ,  $d_3$ ,  $d_6$
- Páginas com mais out-degree:  $d_2$ ,  $d_6$
- Página com maior PageRank:  $d_6$
- Página com maior hub:  $d_6$
- Páginas com maior autoridade:  $d_3$



# Autoridades para a Consulta “Chicago Bulls”

- 0.85 [www.nba.com/bulls](http://www.nba.com/bulls)
- 0.25 [www.essex1.com/people/jmiller/bulls.htm](http://www.essex1.com/people/jmiller/bulls.htm)  
“da Bulls”
- 0.20 [www.nando.net/SportServer/basketball/nba/chi.html](http://www.nando.net/SportServer/basketball/nba/chi.html)  
“The Chicago Bulls”
- 0.15 [users.aol.com/rynocub/bulls.htm](http://users.aol.com/rynocub/bulls.htm)  
“The Chicago Bulls Home Page”
- 0.13 [www.geocities.com/Colosseum/6095](http://www.geocities.com/Colosseum/6095)  
“Chicago Bulls”

(Ben-Shaul et al, WWW8)



# Exemplo de Autoridade

The screenshot displays the Bulls.com website interface. At the top, a navigation bar includes links for NBA, D-LEAGUE, WNBA, GLOBAL, TEAMS, MOBILE, NBA TICKETS, FANTASY, NBATV, STORE, and VIDEO. Below this is a secondary bar with NEWSLETTER and CONTACT US links. The main banner features the Bulls.com logo and the text 'THE OFFICIAL SITE OF THE CHICAGO BULLS' and 'Delivered by at&t'. A red navigation bar contains links for TICKETS, TEAM, NEWS, SCHEDULE, FEATURES, GAME NIGHT, INSIDE THE BULLS, HISTORY, and STORE, followed by a search bar and a SEARCH button. The main content area is divided into three columns. The left column has a section titled 'Fore!!! Golf with the Bulls!' with text about tickets for a charity golf outing and a list of links including '2009-10: Season & Group Tickets', 'Mobile Alerts', 'Facebook', 'Twitter', 'RSS', 'News Caps', 'mybulls', and 'Sam Smith'. Below this is a list of links for Bulls to compete in NBA Summer League, Chicago Bulls Draft Central 2009, Pre-draft Ask Sam mailbag special, and several pre-draft interviews. The middle column features a large photo of a man in a Bulls polo shirt speaking at a microphone. The right column has a 'BULLSEYE' section powered by KIA, with links for CALENDAR, TICKETS, SEASON TICKETS, TICKETEXCHANGE, GROUP TICKETS, and E-NEWSLETTER. Below this is a 'SEASON TICKETS' section with a photo of a player and the text 'CHICAGO BULLS PRESENTED BY HARRIS'. At the bottom, there is a 'Draft Workouts' section and a 'verizon wireless FAN POLL' section.



# Hubs para a Consulta


## “Chicago Bulls”

- 1.62 [www.geocities.com/Colosseum/1778](http://www.geocities.com/Colosseum/1778)  
“Unbelieveabulls!!!!”
- 1.24 [www.webring.org/cgi-bin/webring?ring=chbulls](http://www.webring.org/cgi-bin/webring?ring=chbulls)  
“Erin’s Chicago Bulls Page”
- 0.74 [www.geocities.com/Hollywood/Lot/3330/Bulls.html](http://www.geocities.com/Hollywood/Lot/3330/Bulls.html)  
“Chicago Bulls”
- 0.52 [www.nobull.net/web\\_position/kw-search-15-M2.htm](http://www.nobull.net/web_position/kw-search-15-M2.htm)  
“Excite Search Results: bulls”
- 0.52 [www.halcyon.com/wordsltd/bball/bulls.htm](http://www.halcyon.com/wordsltd/bball/bulls.htm)  
“Chicago Bulls Links”

(Ben-Shaul et al, WWW8)



# Exemplo de Hub

**COAST TO COAST TICKETS**  
great tickets from nice people

Returning Customer

City Guide | View

[Minnesota Timberwolves Tickets](#)  
[New Jersey Nets Tickets](#)  
[New Orleans Hornets Tickets](#)  
[New York Knicks Tickets](#)  
[Oklahoma City Thunder Tickets](#)  
[Orlando Magic Tickets](#)  
[Philadelphia 76ers Tickets](#)  
[Phoenix Suns Tickets](#)  
[Portland Trail Blazers Tickets](#)  
[Sacramento Kings Tickets](#)  
[San Antonio Spurs Tickets](#)  
[Toronto Raptors Tickets](#)  
[Utah Jazz Tickets](#)  
[Washington Wizards Tickets](#)  
[NBA All-Star Weekend](#)  
[NBA Finals Tickets](#)  
[NBA Playoffs Tickets](#)  
[All NBA Tickets](#)

**Official Website Links:**  
[Chicago Bulls \(official site\)](#)  
<http://www.nba.com/bulls/>

**Fan Club - Fan Site Links:**  
[Chicago Bulls](#)  
Chicago Bulls Fan Site with Bulls Blog, News, Bulls Forum, Wallpapers and all your basic Chicago Bulls essentials!!  
<http://www.bullscentral.com>  
[Chicago Bulls Blog](#)  
The place to be for news and views on the Chicago Bulls and NBA Basketball!  
<http://chi-bulls.blogspot.com>

**News and Information Links:**  
[Chicago Sun-Times \(local newspaper\)](#)  
<http://www.suntimes.com/sports/basketball/bulls/index.html>  
[Chicago Tribune \(local newspaper\)](#)  
<http://www.chicagotribune.com/sports/basketball/bulls/>  
[Wikipedia - Chicago Bulls](#)  
All about the Chicago Bulls from Wikipedia, the free online encyclopedia.  
[http://en.wikipedia.org/wiki/Chicago\\_Bulls](http://en.wikipedia.org/wiki/Chicago_Bulls)

**Merchandise Links:**  
[Chicago Bulls watches](#)  
[http://www.sportimewatches.com/NBA\\_watches/Chicago-Bulls-watches.html](http://www.sportimewatches.com/NBA_watches/Chicago-Bulls-watches.html)

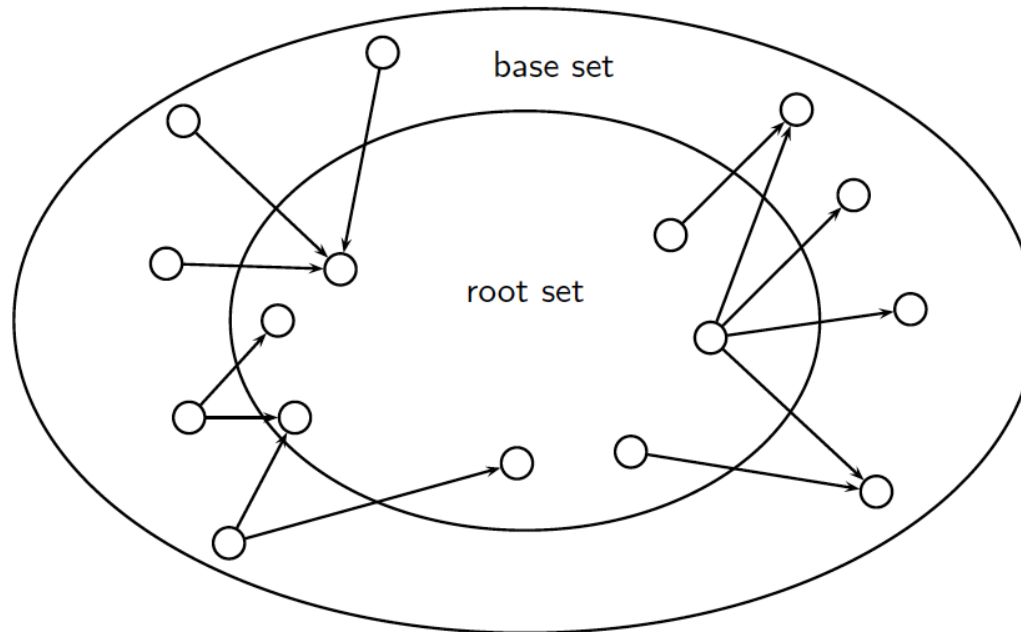
**Event Selections**  
**Sporting Events**  
[MLB Baseball Tickets](#)  
[NFL Football Tickets](#)  
[NBA Basketball Tickets](#)  
[NHL Hockey Tickets](#)  
[NASCAR Racing Tickets](#)  
[PGA Golf Tickets](#)  
[Tennis Tickets](#)  
[NCAA Football Tickets](#)





# HITS: Considerações

- Conteúdo só usado para construir o conjunto raiz
- Páginas no conjunto base podem não ter termos da consulta
- Problema: páginas na base podem não ser no tópico da consulta





# Diferenças HITS e PageRank

- Em termos de computação
  - PageRank: pré-computado
  - HITS: tempo de consulta
- HITS produz 2 rankings, PageRank apenas um
- HITS pode ser aplicado para toda a Web
- PageRank a um pequeno conjunto
- Na Web um bom hub é usualmente uma boa autoridade
- Problema com Web spam