PageRank

Xiaolin Yang

2015/12/09

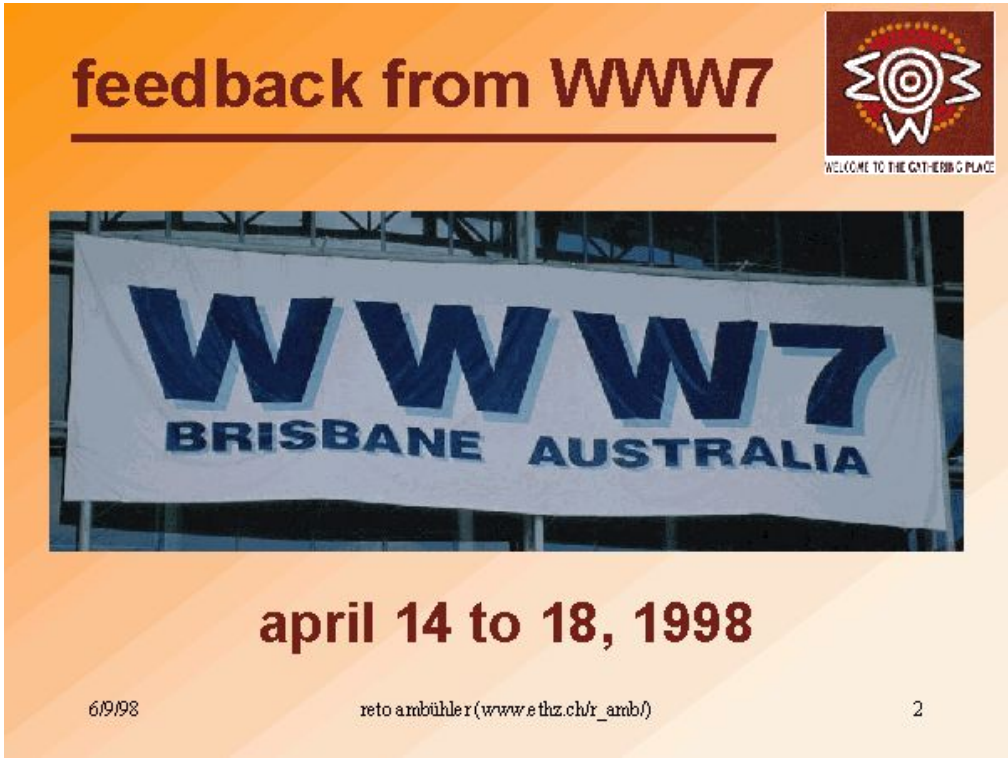
Data Mining Lab, Big Data Research Center, UESTC
School of Computer Science and Engineering
Email: xiaolinyn@gmail.com

CONTENTS

01 Background

Background

At the Seventh International World Wide Web conference(WWW98), Sergey Brin and Larry Page's paper "**The PageRank citation ranking: Bringing order to the Web**" made small ripples in the information science community that quickly turned into waves.



Background

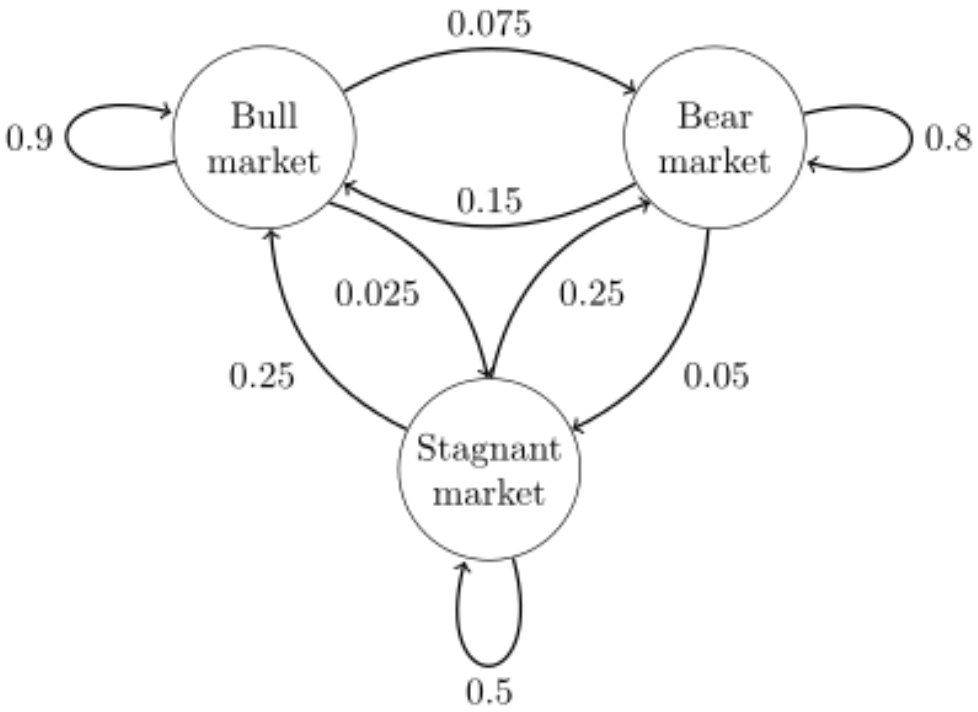
Great Success ⬅ Hyperlink Structure ⬅ PageRank

02 Markov Chain

Markov Chain

Markov Chain：

$$P(X_{n+1} = x | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = P(X_{n+1} = x | X_n = x_n)$$



Stock Market

Irreducible(不可约)：any state can reached from any state.

Absorbing states(吸收态)：the probability of leaving this state is zero

Markov Chain

Markov Chain：

Limit Theorem(极限定理)：

A **homogeneous(齐次), irreducible(不可约), aperiodic(非周期)** and **positive recurrent(正常返)** Markoc Chain has:

a limiting distribution: $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j,\ i,j \in E$

also a stationary distribution: $\pi P = \pi \quad \pi_j \geq 0, \sum_{j\in E} \pi_j = 1$

Markov Chain

Markov Chain：

Limit Theorem(极限定理)：

$$P^{(n)} = (p_{ij}^{(n)}) = \begin{bmatrix} p_{11}^{(n)} & p_{12}^{(n)} & \cdots & p_{1j}^{(n)} & \cdots \\ p_{21}^{(n)} & p_{22}^{(n)} & \cdots & p_{2j}^{(n)} & \cdots \\ p_{31}^{(n)} & p_{32}^{(n)} & \cdots & p_{3j}^{(n)} & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_j & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix} = \begin{bmatrix} \Pi \\ \Pi \\ \Pi \\ \vdots \end{bmatrix}, n \to \infty$$

Markov Chain

Markov Chain(Aperiodic):

Period of a state:

$$k = g.c.d\{n > 0 : \Pr(X_n = i | X_0 = i) > 0\}$$

. . : the greatest common divisor

Aperiodic(非周期): $k = 1$

Markov Chain

Markov Chain(Positive Recurrent):

The first return probability(首返概率):

$$f_{ii}^{(n)} = P\{X_n = i, X_k \neq i, 1 \leq k < n | X_0 = i\}$$

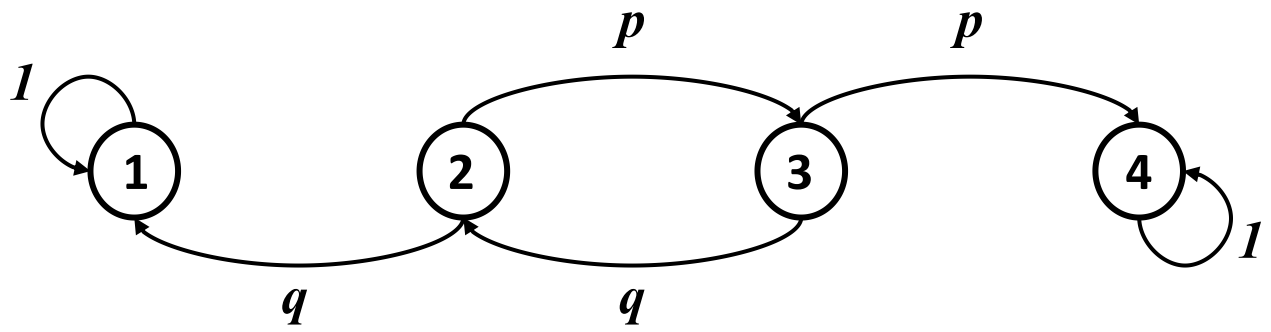
Recurrent (常返): $$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$$

Positive Recurrent (正常返): $$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} < +\infty$$

Markov Chain

Markov Chain：

Positive Recurrent (正常返)：



State 1： $f_{11}^{(1)} = 1, f_{11}^{(n)} = 0(n > 1) \Rightarrow f_{11} = 1, \mu_1 = 1$ ***positive recurrent !***

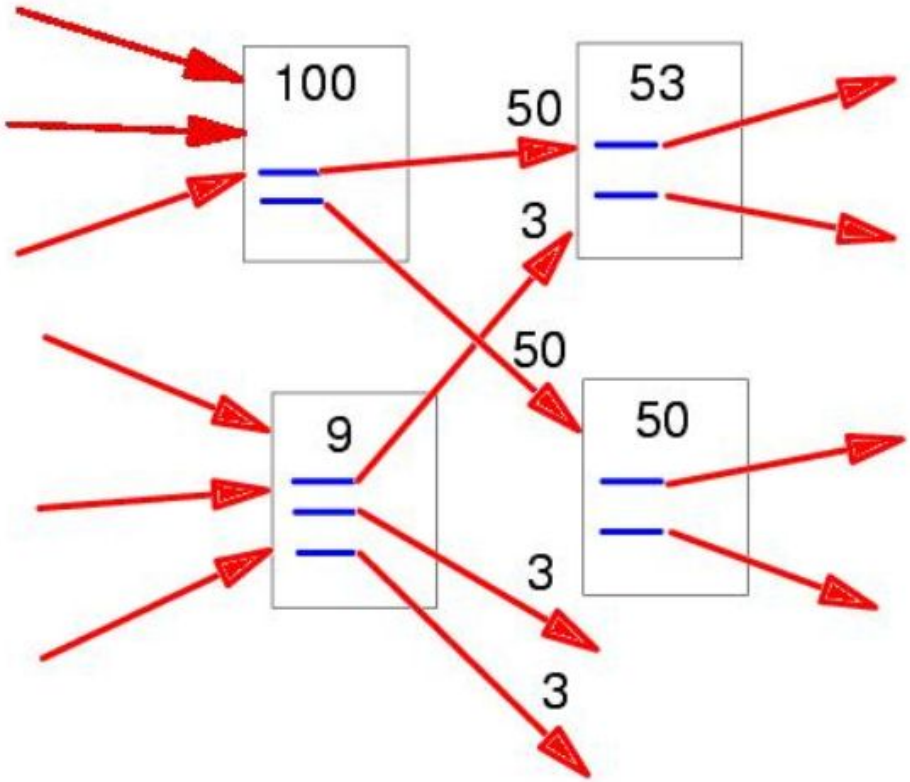
State 2： $f_{22}^{(1)} = 0, f_{22}^{(2)} = pq, f_{22}^{(n)} = 0(n > 2) \Rightarrow f_{22} = \sum_{n=1}^{\infty} n f_{22}^{(n)} = pq < 1$

transient !

03 The Basic PageRank Model

The Basic PageRank Model

Page Links:



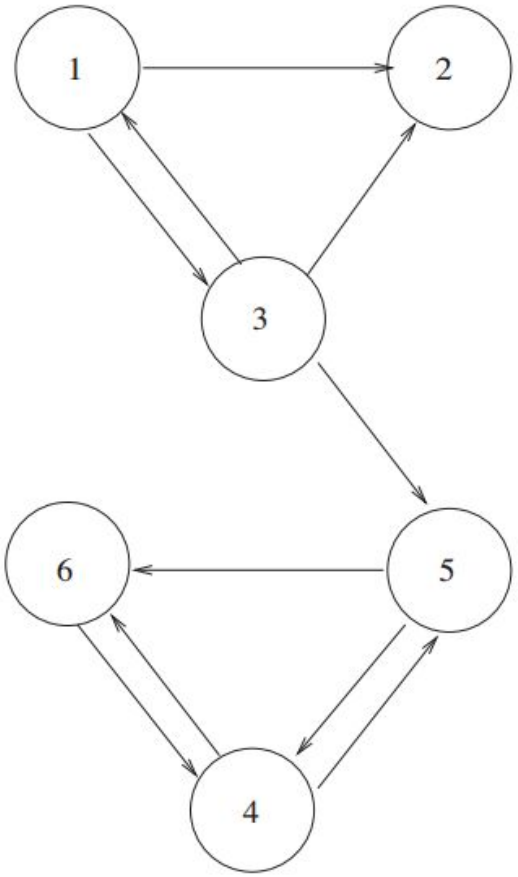
If a page

- is linked by many pages;
- which link to it is authoritative

Then
it will gain high PageRank!

The Basic PageRank Model

Transition Probability Matrix:



node 2(Dangling Node) has no outlinks!

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

$$x^{(1)T} = x^{(0)T} P$$

$$x^{(k+1)T} = x^{(k)T} P$$

$$x^{(k)T} = x^{(0)T} P^k$$

The Basic PageRank Model

Transition Probability Matrix:

Markov matrix/stochastic matrix:

- Every element is Nonnegative
- Row/Column sums are +1

Properties:

- Spectral radius(the supremum among the absolute values of spectrums) is 1;

The Basic PageRank Model

Transition Probability Matrix:

If every element is positive, **Markov matrix** is irreducible (strongly connected):

Every irreducible markov matrix has a stationary vector:

π: does not change under application of the transition matrix:

$$\pi^T P = \pi^T$$

It's independent of the initial state

PageRank vector is π, so this is just an eigenvector problem!

The Basic PageRank Model

Transition Probability Matrix:

The independence of the initial state:

$$u_0 = (0.02, 0.98)$$

$$P = \begin{bmatrix} 0.2 & 0.05 \\ 0.8 & 0.95 \end{bmatrix}$$

$$Px = \lambda x$$

$$P\begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} = 1\begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} \quad P\begin{bmatrix} -1 \\ 1 \end{bmatrix} = 0.75\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$u_0 = \begin{bmatrix} 0.02 \\ 0.98 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} + 0.18\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$u_0 = 1\begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} + 0.75 * 0.18\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$u_k = 1\begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix} + 0.75^k * 0.18\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

The Basic PageRank Model

Transition Probability Matrix:

Practical web graphs are not necessarily strongly connected.

- For Dangling Nodes:

$$\bar{P} = P + av^T$$

a: $a_i = 1$ if row i of **P** corresponds to a dangling node, and 0, otherwise.

v^T: a general probability vector.

- For Irreducible:

$$\bar{\bar{P}} = \alpha\bar{P} + (1-\alpha)ev^T$$

$e = [1,1,1,\ldots,1]^T$

The Basic PageRank Model

Transition Probability Matrix:

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \Rightarrow \bar{P} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\bar{\bar{P}} = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}$$

$$\alpha = 0.9$$

$$v = \left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right]^T$$

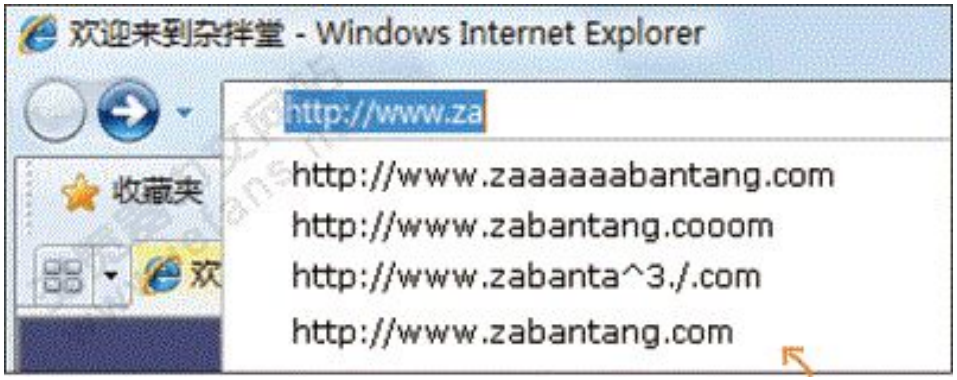
The Basic PageRank Model

Transition Probability Matrix：

$$\bar{\bar{P}} = \alpha \bar{P} + (1 - \alpha) e v^T$$

Another Interpretation for $\bar{\bar{P}}$ is：

Even though the customer always browse webpages by hyperlinks, but he can also use URL to “teleport” to a new page.



04 The Power Method

The Power Method

The Power Method:

$$\begin{aligned}\mathbf{x}^{(k)T} &= \mathbf{x}^{(k-1)T}\bar{\bar{\mathbf{P}}} = \alpha\mathbf{x}^{(k-1)T}\bar{\mathbf{P}} + (1-\alpha)\mathbf{x}^{(k-1)T}\mathbf{e}\mathbf{v}^T \\ &= \alpha\mathbf{x}^{(k-1)T}\bar{\mathbf{P}} + (1-\alpha)\mathbf{v}^T \\ &= \alpha\mathbf{x}^{(k-1)T}\mathbf{P} + (\alpha\mathbf{x}^{(k-1)T}\mathbf{a} + (1-\alpha))\mathbf{v}^T,\end{aligned}$$

$\mathrm{x}^{(k-1)T}e = 1.$

The Power Method

Four advantages:

- $\bar{\bar{P}}$ and $\bar{P}$ are never formed or stored.
- Since P is sparse, each vector-matrix multiplication can be computed in $nnz(P)$ flops.
  $nnz(P)$: the number of nonzeros in P.
- At each iteration, the power method only requires the storage of one vector, the current iterate.
- Converges quickly.

05 Discussion about the Model

Discussion about the Model

α(damping factor):

Google always using = 0.85, so why this choice for ?

1. a trade-off:

- The larger is, the more the true hyperlink structure of the web is used to determine webpage importance.
- The smaller is, the faster the convergence for power method.

Discussion about the Model

α:

Rate of convergence(r) is determined by the subdominant eigenvalue(λ_2) of the transition matrix $\bar{\bar{p}}$, and this eigenvalue is strictly less than α for an irreducible hyperlink matrix.

$$r \propto (\lambda_1 - \lambda_2) = (1 - \lambda_2) \propto (1 - \alpha)$$

for $\tau = x^{(k+1)T} - x^{(k)T} = 10^{-8}$, only need 142 iterations.

Discussion about the Model

α:

2. intuitive reality:

$\alpha = 0.85$ implies that roughly:

- 5/6 of the time a Web surfer randomly clicks on hyperlinks.
- 1/6 of the time this Web surfer will go to the URL line and type the address of a new page.

Discussion about the Model

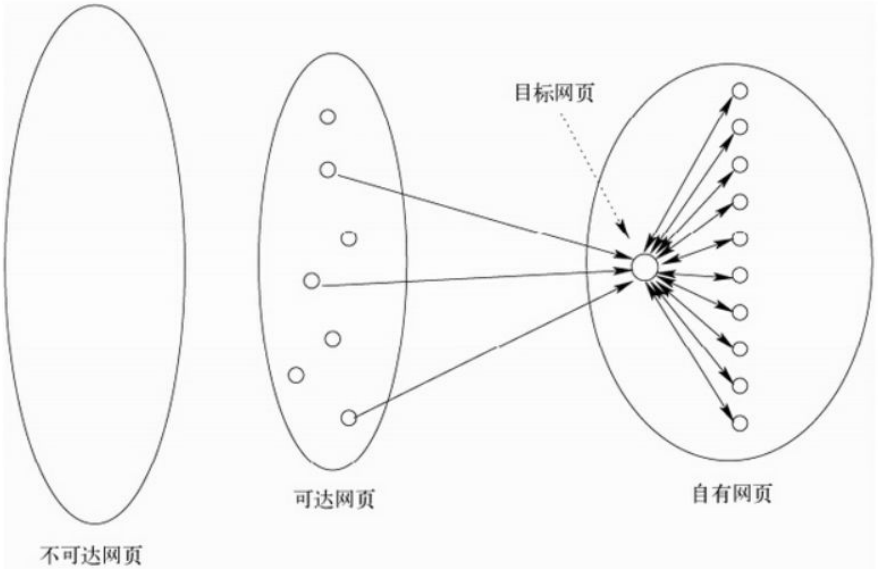
v^T:

Instead of traditional $v^T = \frac{1}{n} e^T$, v^T can be more personalized.

- different classes of surfers have different v^T.

- control spamming done by the so-called link farms.

A link farm



Discussion about the Model

v^T:

	Damping Factor (α)	Personalization Vector (v)	Google Matrix (G)	PageRank Vector ($\approx \pi$)	Ordering of Nodes (1 = Highest)
Model 1	0.85	$\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$	$\begin{pmatrix} \frac{3}{80} & \frac{71}{80} & \frac{3}{80} & \frac{3}{80} \\ \frac{3}{80} & \frac{3}{80} & \frac{71}{80} & \frac{3}{80} \\ \frac{37}{80} & \frac{3}{80} & \frac{3}{80} & \frac{37}{80} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$	$\begin{pmatrix} 0.21 & 0.26 & 0.31 & 0.21 \end{pmatrix}$	$\begin{pmatrix} 3 & 2 & 1 & 3 \end{pmatrix}$
Model 2	0.85	$\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} \frac{3}{20} & \frac{17}{20} & 0 & 0 \\ \frac{3}{20} & 0 & \frac{17}{20} & 0 \\ \frac{23}{40} & 0 & 0 & \frac{17}{40} \\ \frac{29}{80} & \frac{17}{80} & \frac{17}{80} & \frac{17}{80} \end{pmatrix}$	$\begin{pmatrix} 0.30 & 0.28 & 0.27 & 0.15 \end{pmatrix}$	$\begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix}$
Model 3	0.95	$\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$	$\begin{pmatrix} \frac{1}{80} & \frac{77}{80} & \frac{1}{80} & \frac{1}{80} \\ \frac{1}{80} & \frac{1}{80} & \frac{77}{80} & \frac{1}{80} \\ \frac{39}{80} & \frac{1}{80} & \frac{1}{80} & \frac{39}{80} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$	$\begin{pmatrix} 0.21 & 0.26 & 0.31 & 0.21 \end{pmatrix}$	$\begin{pmatrix} 3 & 2 & 1 & 3 \end{pmatrix}$
Model 4	0.95	$\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{20} & \frac{19}{20} & 0 & 0 \\ \frac{1}{20} & 0 & \frac{19}{20} & 0 \\ \frac{21}{40} & 0 & 0 & \frac{19}{40} \\ \frac{23}{80} & \frac{19}{80} & \frac{19}{80} & \frac{19}{80} \end{pmatrix}$	$\begin{pmatrix} 0.24 & 0.27 & 0.30 & 0.19 \end{pmatrix}$	$\begin{pmatrix} 3 & 2 & 1 & 4 \end{pmatrix}$

Discussion about the Model

Forcing Irreducibility:

enforce every node is directly connected to every other node. (alter the true nature of the Web)

add a dummy node to the Web which connects to every other node and to which every other node is connected.

$$\hat{\mathbf{P}} = \left(\begin{array}{c|c} \alpha\bar{\mathbf{P}} & (1-\alpha)\mathbf{e} \\ \hline \mathbf{v}^T & 0 \end{array}\right)$$

Discussion about the Model

Forcing Irreducibility:

LeaderRank:

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{a_{ji}}{k_j^{out}} s_j(t)$$

$$S_i = s_i(t_c) + \frac{s_g(t_c)}{N}$$

- self-adaptive
- parameter-free

06 Other Topics

Other Topics

- **Storage and Speed**

  “The World’s Largest Matrix Computation”

- **Spam**
- **The evolution and dynamics of the Web**
- **Web’s structure：**

  how to use the *scale-free* structure to improve PageRank computations

- **Community：**

  how do changes within the community affect the PageRank of community pages?

Thanks!