PAGERANK PARAMETERS

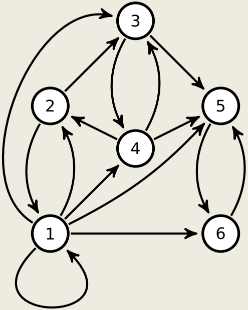
David F. Gleich
Amy N. Langville

American Institute of Mathematics
Workshop on Ranking
Palo Alto, CA
August 17th, 2010

The most important page on the web

The most important page on the web

PageRank details



$$\rightarrow \underbrace{\begin{bmatrix} 1/6 & 1/2 & 0 & 0 & 0 & 0 \\ 1/6 & 0 & 0 & 1/3 & 0 & 0 \\ 1/6 & 1/2 & 0 & 1/3 & 0 & 0 \\ 1/6 & 0 & 1/2 & 0 & 0 & 0 \\ 1/6 & 0 & 1/2 & 1/3 & 0 & 1 \\ 1/6 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{P}}$$

$P_{ij} \geq 0$
$\mathbf{e}^T \mathbf{P} = \mathbf{e}^T$

"jump" $\rightarrow$ $\mathbf{v} = [\, \frac{1}{n} \; \dots \; \frac{1}{n} \,]^T$

$v_i \geq 0$
$\mathbf{e}^T \mathbf{v} = 1$

Markov chain $\quad \left[\alpha \mathbf{P} + (1-\alpha)\mathbf{v}\mathbf{e}^T\right] \mathbf{x} = \mathbf{x}$

unique $\mathbf{x} \Rightarrow x_j \geq 0$, $\mathbf{e}^T \mathbf{x} = 1$.

Linear system $\quad (\mathbf{I} - \alpha \mathbf{P})\mathbf{x} = (1-\alpha)\mathbf{v}$

Ignored $\quad$ *dangling nodes patched back to* $\mathbf{v}$

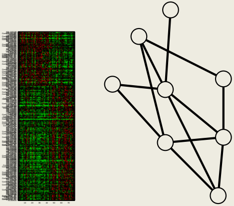
algorithms later

Other uses for PageRank

What else people use PageRank to do

GeneRank

Morrison et al. GeneRank, 2005



Use $(\mathbf{I} - \alpha \mathbf{G}\mathbf{D}^{-1})\mathbf{x} = \mathbf{w}$ to find "nearby" important genes.

Note New paper LabRank with a random scientist?

ProteinRank
ObjectRank
EventRank
IsoRank
Clustering
Sports ranking
Food webs
Centrality
Reverse PageRank
FutureRank
SocialPageRank
BookRank
ArticleRank
ItemRank
SimRank
DiffusionRank
TrustRank
TweetRank

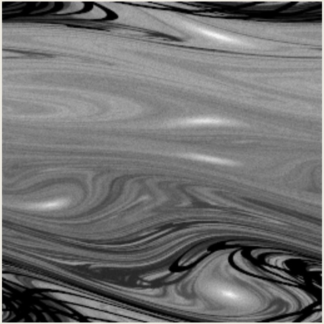
Ulam Networks

Chirikov map

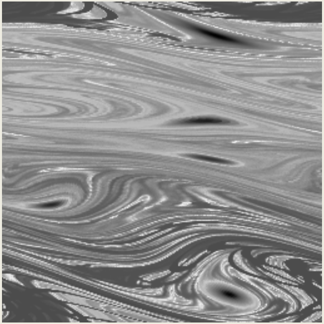
$y_{t+1} = \eta y_t + k \sin(x_t + \theta_t)$
$x_{t+1} = x_t + y_{t+1}$

Ulam network

1. divide phase space into uniform cells
2. form $\mathbf{P}$ based on trajectories.



$\log(\mathrm{E}[\mathbf{x}(A)])$



$\log(\mathrm{Std}[\mathbf{x}(A)])/\log(\mathrm{E}[\mathbf{x}(A)])$

$A \sim \mathrm{Beta}(2, 16)$

Note *White is larger, black is smaller*

Google matrix, dynamical attractors, and Ulam networks, Shepelyansky and Zhirov, arXiv

Choosing alpha

Slide 6 of 21

What is alpha? There's no single answer.

Ask yourself, why am I computing PageRank? Then use the best value for your application.

web-search	$\rightarrow$	tune α for the best feature vector
node centrality	$\rightarrow$	understand what random jumps mean in your graph
find important nodes in a web-graph	$\rightarrow$	use the random surfer interpretation

Author	$\boldsymbol{\alpha}$
Brin and Page (1998)	0.85
Najork et al. (2007)	0.85
Litvak et al. (2006)	0.5
Pan el al. (2004)	0.15
Algorithms (...)	≥ 0.85
Experiment	???

The PageRank limit value

Singular? $(\mathbf{I} - \alpha \mathbf{P})\mathbf{x} = (1-\alpha)\mathbf{v}$

$$\mathbf{P} = \mathbf{X} \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{J}_1 \end{bmatrix} \mathbf{X}^{-1}$$

$$\left(\mathbf{I} - \alpha \mathbf{X} \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{J}_1 \end{bmatrix} \mathbf{X}^{-1} \right) \mathbf{x} = (1-\alpha)\mathbf{v}$$

$$\mathbf{X} \left(\mathbf{I} - \alpha \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{J}_1 \end{bmatrix} \right) \mathbf{X}^{-1}\mathbf{x} = (1-\alpha)\mathbf{v}$$

$$\left(\mathbf{I} - \alpha \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{J}_1 \end{bmatrix} \right) \mathbf{y} = (1-\alpha)\mathbf{z}$$

$$(1-\alpha)\mathbf{y}_1 = (1-\alpha)\mathbf{z}_1$$

$$(\mathbf{I} - \alpha \mathbf{J}_2)\mathbf{y}_2 = (1-\alpha)\mathbf{z}_2$$

Boldi et al. 2003: PageRank as a function of the damping parameter

TotalRank

$$\mathbf{t} = \int_0^1 \mathbf{x}(\alpha)\, d\alpha$$

Proposed by Boldi et al. (2005) as a parameter free PageRank.

Generalized PageRank

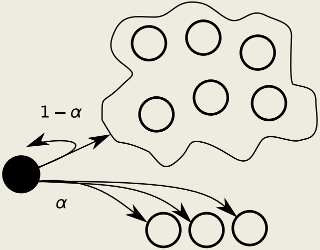
PageRank	$(\mathbf{I} - \alpha \mathbf{P})\mathbf{x} = (1-\alpha)\mathbf{v}$ $\mathbf{x} = \sum_{i=0}^{\infty} (1-\alpha)(\alpha^i)\mathbf{P}^i \mathbf{v}$
Generalized PageRank	$\mathbf{y} = \sum_{i=0}^{\infty} f(i)\mathbf{P}^i \mathbf{v}$ $\sum_i f(i) < \infty$
TotalRank	$f(i) = \frac{1}{i+1} - \frac{1}{i+2}$
LinearRank	$\ldots$
HyperRank	$\ldots$

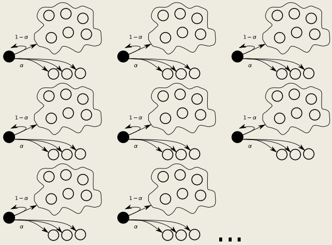
Baeza-Yates et al. 2006

Pick a distribution

Multiple surfers should have an impact!

Each person picks α_i from distribution A





$\downarrow$ $\qquad\qquad\qquad\qquad$ $\downarrow$

$\mathbf{x}(\mathrm{E}[A])$ $\qquad\qquad\qquad\qquad$ $\mathrm{E}[\mathbf{x}(A)]$

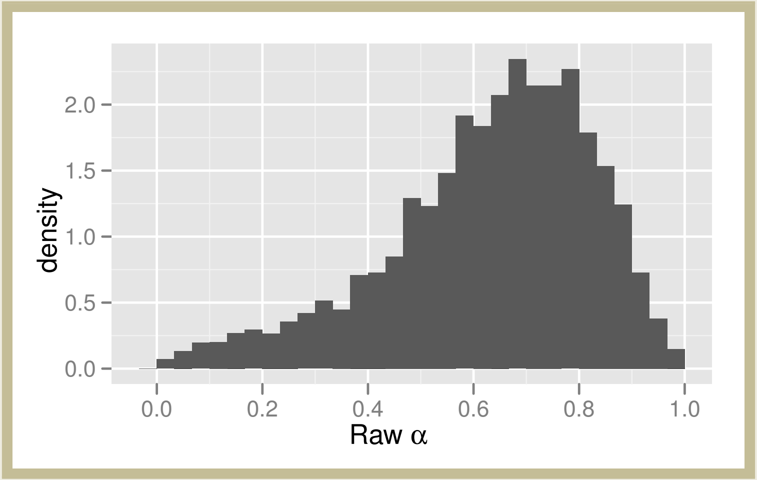
$\searrow$ $\qquad\qquad\qquad\qquad$ $\swarrow$

$$\mathbf{x}(\mathrm{E}[A]) \neq \mathrm{E}[\mathbf{x}(A)]$$

$\text{TotalRank} : \mathrm{E}[\mathbf{x}(A)] : A \sim U[0, 1]$

Constantine & Gleich, Internet Mathematics, in press.

From users



Sample mean $\bar{\mu} = 0.631$.

Gleich et al., WWW2010

Note *257,664 users from Microsoft toolbar data*

Choosing personalization

Slide 13 of 21

Personalization choices

Application specific

- GeneRank : $\mathbf{v}$ = normalized microarray weights
- TopicRank: $\mathbf{v}$ = pages on the same topic
- TrustRank: $\mathbf{v}$ = only pages known to be good
- BadRank: $\mathbf{v}$ = only pages known to be bad (an reverse the graph)

Super-personalized

- Set $\mathbf{v}$ to have only a single non-zero : $\mathbf{v} = \mathbf{e}_i$.

Personalized PageRank

$$\mathbf{B} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}$$

B_{ij} = "personalized score of page i when jumping to page i"

Related methods

PageRank history

See Vigna 2010: Spectral Ranking and
Franceschet 2010: PageRank: Standing on the shoulder of giants.

Let $\mathbf{A}$ be the adjacency matrix of a graph.

PageRank	$(\mathbf{I} - \alpha\mathbf{P})\mathbf{x} = (1-\alpha)\mathbf{v}$	$(\alpha\mathbf{P} + (1-\alpha)\mathbf{v}\mathbf{e}^T)\mathbf{x} = \mathbf{x}$
Seeley 1949		$\mathbf{P}\mathbf{x} = \mathbf{x}$
Wei 1952		$\mathbf{A}^T\mathbf{x} = \mathbf{x}$
Katz 1953	$(\mathbf{I} - \alpha\mathbf{A})\mathbf{x} = \mathbf{e}$	
Hubbell 1965		$\mathbf{A}^T\mathbf{x} = \mathbf{x} + \mathbf{v}$

Graph centrality

For a graph G, a score assigned to each vertex $v \in V$ is a *centrality score* if larger scores are "more central" vertices and the score is independent of the labeling on the vertices.

Open issues

Slide 19 of 21

The Problem

- We can derive gazillions of small variants
- Which ones are meaningful?
- Justify your existence!
- But nobody does :(
- Note: the same happens for the web

From Vigna, A history of spectral ranking, MMDS2010

Other issues

QUESTIONS?