Models and Algorithms for PageRank Sensitivity

David F. Gleich

Stanford University

Ph.D. Oral Defense

Institute for Computational and Mathematical Engineering

May 26, 2009

Outline

PageRank intro

Sensitivity

Random sensitivity

Inner-Outer

Summary

Five years!

2004	2009
Firefox 1.0	Firefox 3.5
Wikipedia?	Wikipedia! YouTube! Hulu!
Facebook?	Facebook! flickr! Twitter!
Gmail?	Gmail! Google Maps!
Yahoo!	Yahoo?
3.0 GHz	3.0 GHz × 4
Google	Google

PageRank intro

Slide 4 of 41

A cartoon websearch primer

1. Crawl webpages
2. Analyze webpage text (information retrieval)
3. **Analyze webpage links**
4. Fit measures to human evaluations
5. Produce rankings
6. Continually update

Gleich's syndrome

From Wikipedia, the free encyclopedia

Gleich's syndrome or **episodic angioedema with eosinophilia** is a rare disease in which the body swells up episodically (angioedema), associated with raised antibodies of the IgM type and increased numbers of eosinophil granulocytes, a type of white blood cells, in the blood (eosinophilia). It was first described in 1984.[1]

Its cause is unknown, but it is unrelated to capillary leak syndrome (which may cause similar swelling episodes) and eosinophilia-myalgia syndrome (which features eosinophilia but alternative symptoms). Moreover, it is not a form of hypereosinophilic syndrome as there is no evidence that it leads to organ damage. Some studies have shown that edema attacks are associated with degranulation (release of enzymes and mediators from eosinophils), and others have demonstrated antibodies against endothelium (cells lining blood vessels) in the condition.[2]

Gleich syndrome has a good prognosis. Attack severity may improve with steroid treatment.[1][2]

Eosinophilia

From Wikipedia, the free encyclopedia

Eosinophilia is the state of having a high concentration of **eosinophils** (eosinophil granulocytes) in the blood. The normal concentration is between 0 and 0.5 x 10^9 eosinophils per litre of blood. Eosinophilia can be *reactive* (in response to other stimuli such as allergy or infection) or *non reactive*.

The release of interleukin 5 by T cells, mast cells and macrophages stimulates the production of eosinophils.

Eosinophilia	
Classification and external resources	
ICD-10	D72.1
ICD-9	288.3
DiseasesDB	4328
eMedicine	med/685
MeSH	D004802

Causes [edit]

Diseases that feature eosinophilia:

Hypereosinophilic syndrome

From Wikipedia, the free encyclopedia

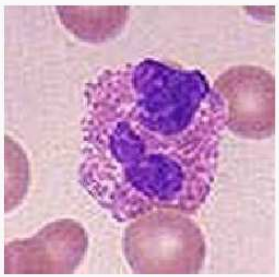
The **hypereosinophilic syndrome** (HS) is a disease characterized by a persistently elevated eosinophil count (≥ 1500 eosinophils/mm³) in the blood for at least six months without any recognizable cause, with involvement of either the heart, nervous system, or bone marrow.[1]

HS is a diagnosis of exclusion, after clonal eosinophilia (such as leukemia) and reactive eosionophilia (in response to infection, autoimmune disease, atopy, hypoadrenalism or cancer) have been ruled out. [2]

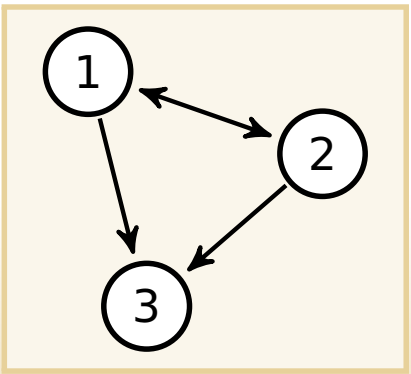
There are some associations with chronic

Hypereosinophilic syndrome

Classification and external resources

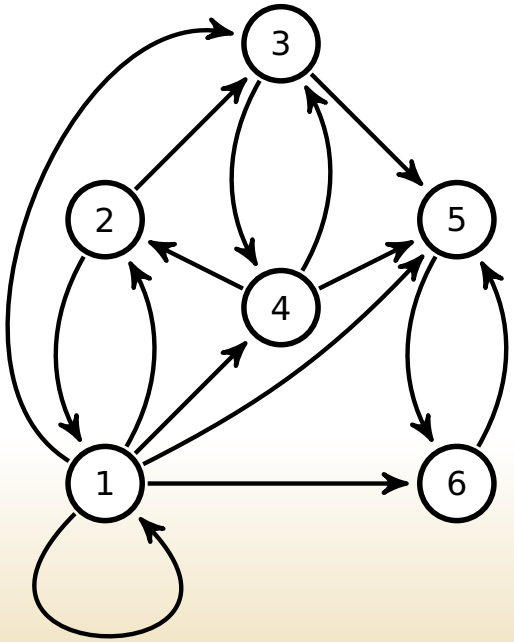


to



PageRank by Google

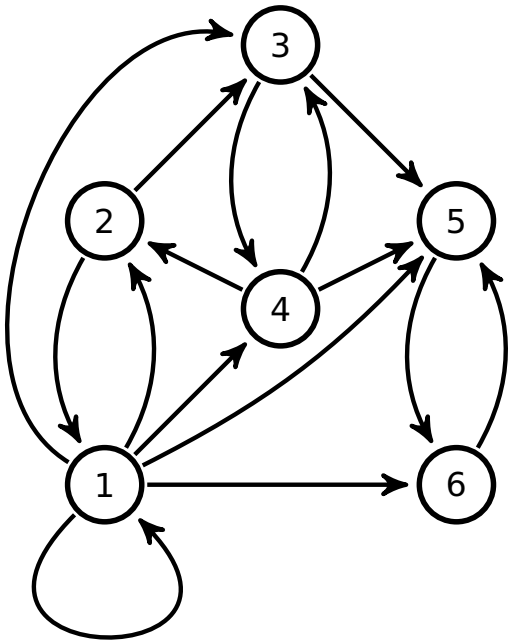
The places we find the surfer most often are important pages.



The Model

1. follow edges uniformly with probability α, and
2. randomly jump with probability $1-\alpha$, we'll assume everywhere is equally likely

Some PageRank details



$$\rightarrow \underbrace{\begin{bmatrix} 1/6 & 1/2 & 0 & 0 & 0 & 0 \\ 1/6 & 0 & 0 & 1/3 & 0 & 0 \\ 1/6 & 1/2 & 0 & 1/3 & 0 & 0 \\ 1/6 & 0 & 1/2 & 0 & 0 & 0 \\ 1/6 & 0 & 1/2 & 1/3 & 0 & 1 \\ 1/6 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{P}} \qquad P_{ij} \geq 0 \quad \mathbf{e}^T\mathbf{P} = \mathbf{e}^T$$

"jump" $\rightarrow$ $\mathbf{v} = \left[\frac{1}{n} \ \cdots \ \frac{1}{n}\right]^T$ $\qquad v_i \geq 0 \quad \mathbf{e}^T\mathbf{v} = 1$

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

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

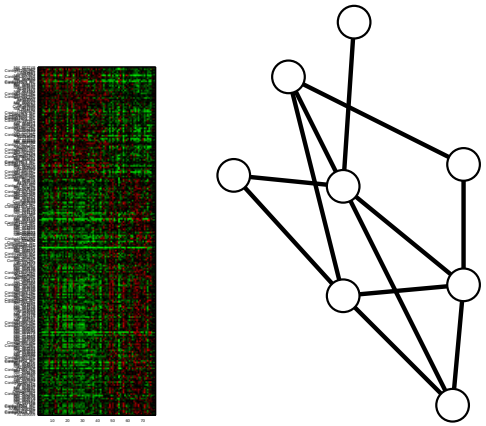
Linear system $(\mathbf{I} - \alpha\mathbf{P})\mathbf{x} = (1-\alpha)\mathbf{v}$

Small detail *dangling nodes patched back to* $\mathbf{v}$

Other uses for PageRank

What else people use PageRank to do

GeneRank



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

ProteinRank

IsoRank

Clustering
(graph partitioning)

Sports ranking

Teaching

Morrison et al. GeneRank, 2005.

My "other projects"

Prior PageRank

Parallel Krylov Methods

Gleich, Zhukov, and Berkhin , Yahoo! Research Labs Technical Report, YRL-2004-038; Gleich and Zhukov, SuperComputing poster, 2005.

"Does existing software work for computing PageRank on a cluster?"

Approximate Personal PageRank

Gleich and Polito, Internet Math. 3(3):257–294, 2007.

"Can you build a web search engine on your PC?"

Ongoing

Parameterized Matrix Problems

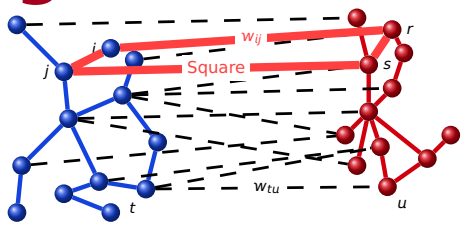
(with Paul Constantine)

$$\mathbf{A}(s)\mathbf{x}(s) = \mathbf{b}(s)$$

Come back here for his defense on Monday, June 1st at 1:30pm!

Network Alignment

(with Mohsen Bayati, Margot Gerritsen, Amin Saberi, and Ying Wang)



My Software

Packages

MatlabBGL

libbvg

gaimc

vismatrix

parameterized matrix package (with Paul)

Publications

Random α PageRank

Inner-Outer PageRank

Sensitivity

Which sensitivity?

Sensitivity to the links : examined and understood

Sensitivity to the jump : examined, understood, and useful

Sensitivity to α : less well understood

PageRank on Wikipedia

$\alpha = 0.50$	$\alpha = 0.85$	$\alpha = 0.99$
United States	United States	C:Contents
C:Living people	C:Main topic classif.	C:Main topic classif.
France	C:Contents	C:Fundamental
Germany	C:Living people	United States
England	C:Ctgs. by country	C:Wikipedia admin.
United Kingdom	United Kingdom	P:List of portals
Canada	C:Fundamental	P:Contents/Portals
Japan	C:Ctgs. by topic	C:Portals
Poland	C:Wikipedia admin.	C:Society
Australia	France	C:Ctgs. by topic

Note *Top 10 articles on Wikipedia with highest PageRank*

The PageRank function

Look at the PageRank vector as a function of α

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

and examine its derivative.

My Contributions

Gleich, Glynn, Golub, Greif, Dagstuhl proceedings, 2007.

Compute the derivative with **just simple PageRank solves**.

Empirically evaluated the derivative as a rank change predictor.

Others

PageRank becomes more sensitive as $\alpha \to 1$.

PageRank vector at $\alpha = 1$ well defined.

α matters!

Golub and Greif, 2004; Boldi et al., 2005; Berkhin, 2005; Langville and Meyer, 2006.

Random sensitivity

What is alpha?

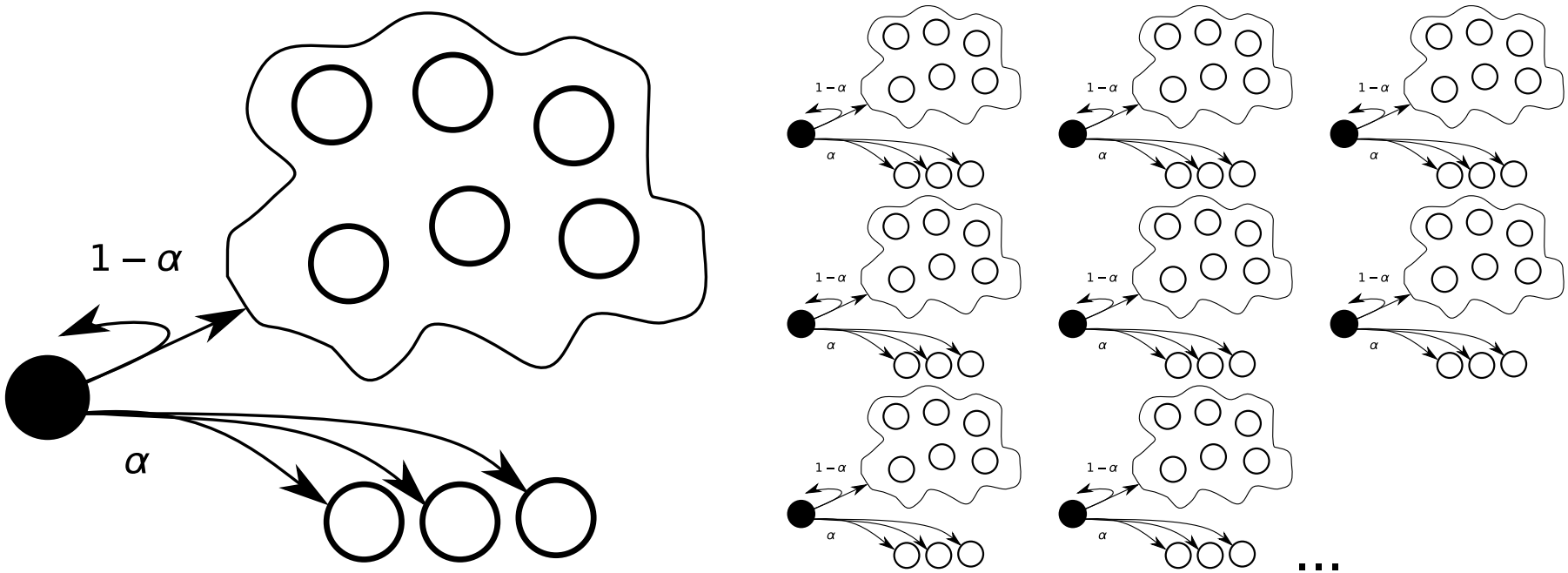
Author	α
Brin and Page (1998)	0.85
Najork et al. (2007)	0.85
Litvak et al. (2006)	0.5
Experiment (slide 20)	0.375
Algorithms (...)	$\geq$ 0.85

For **you**, α is clear

Google wants PageRank for **everyone**

Multiple surfers

Each person picks α_i from distribution A



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

Random alpha PageRank

RAPr

Model PageRank as the random variables

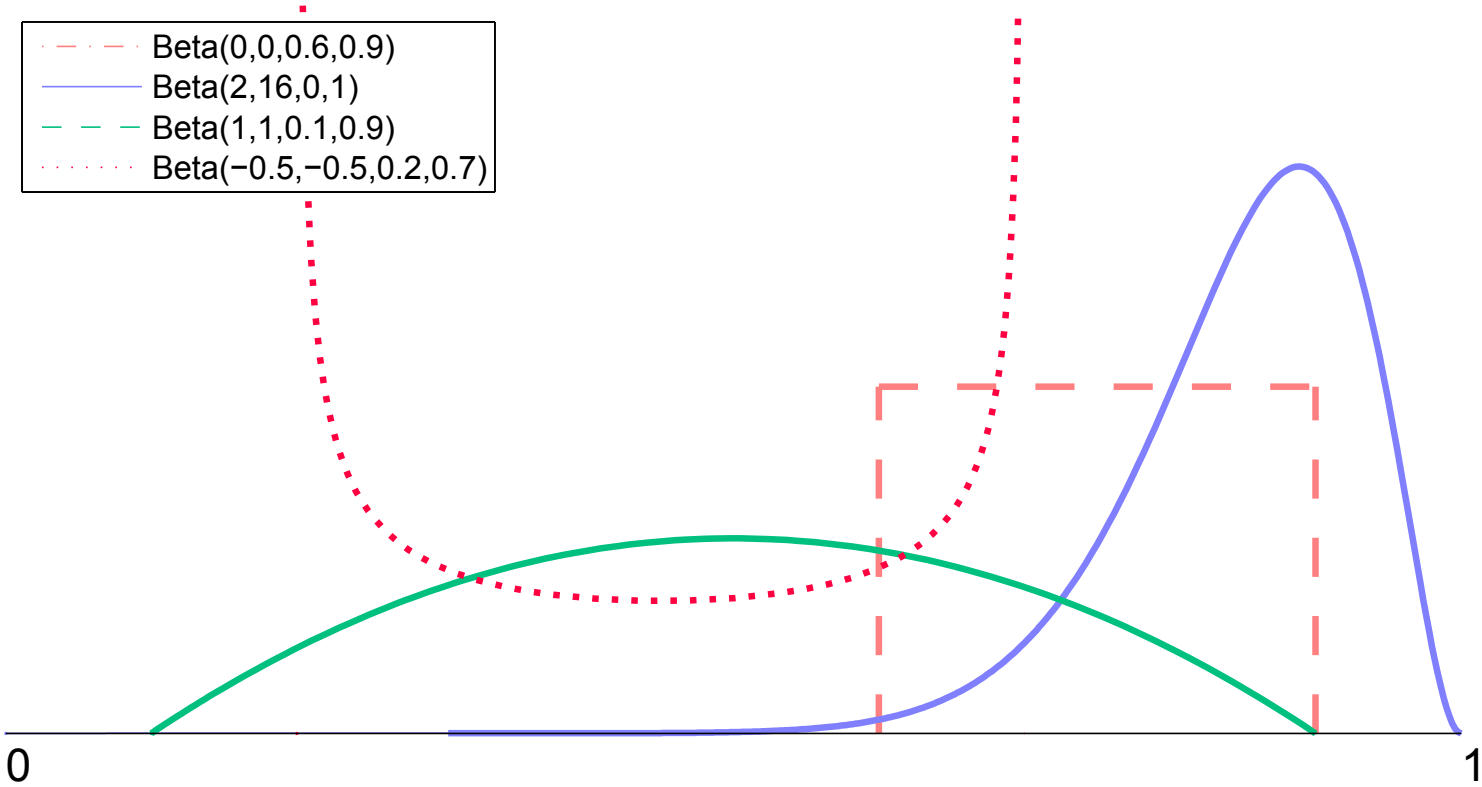
$$\mathbf{x}(A)$$

and look at

$$\mathrm{E}\left[\mathbf{x}(A)\right] \text{ and } \mathrm{Std}\left[\mathbf{x}(A)\right].$$

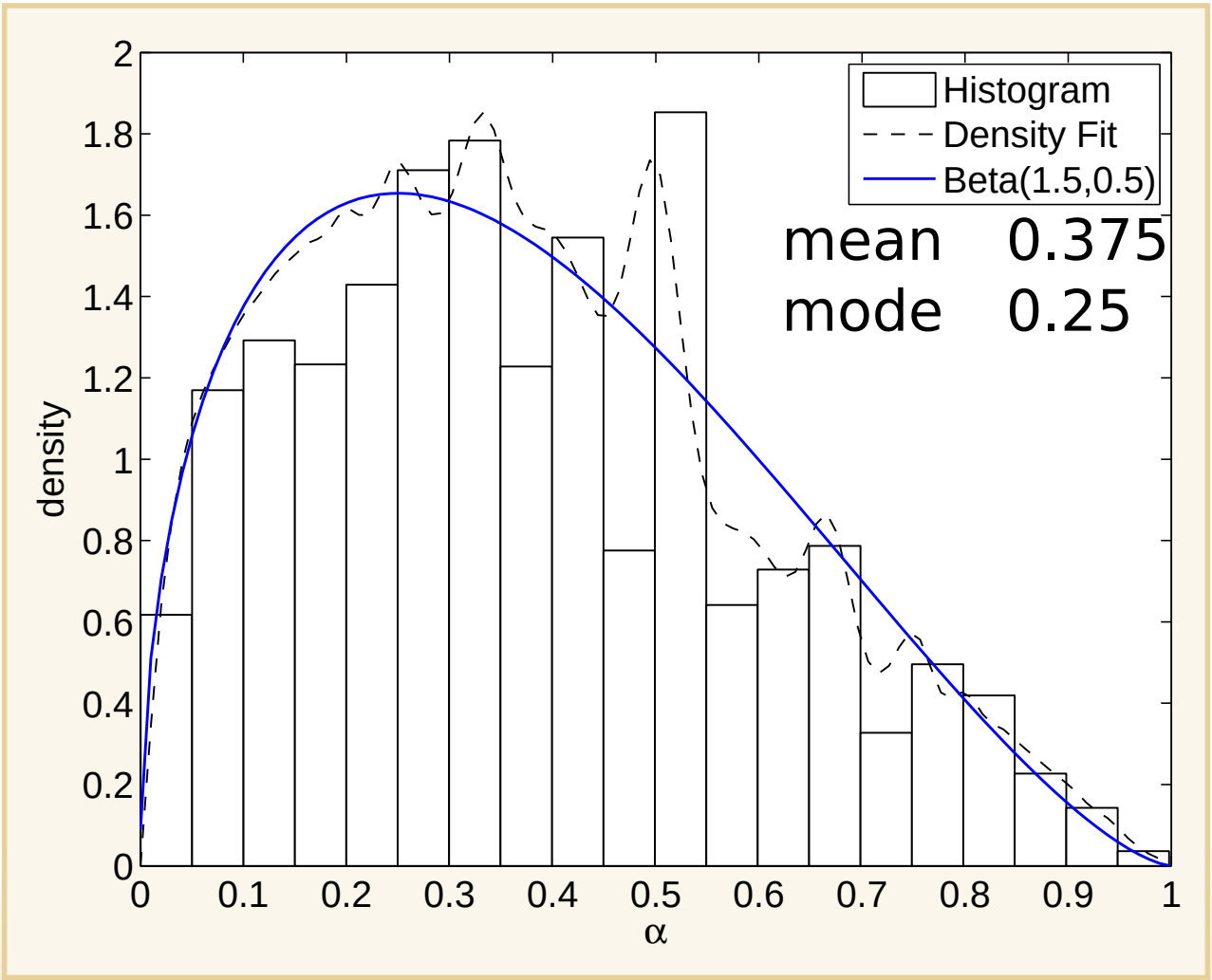
Gleich and Constantine, Workshop on Algorithms on the Web Graph, 2007

What is A?



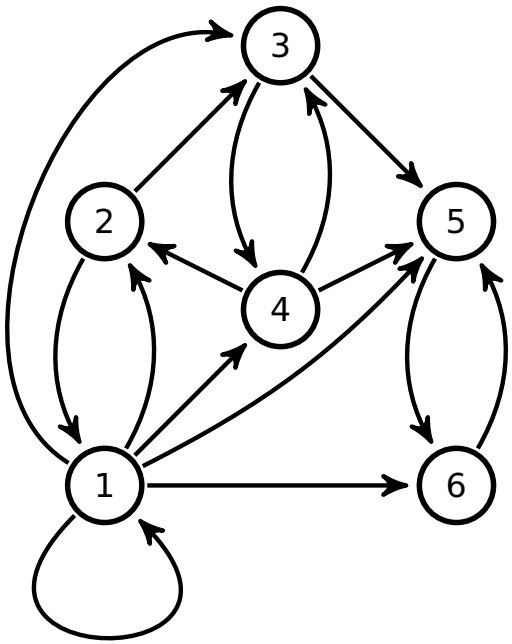
Beta(a, b, l, r)

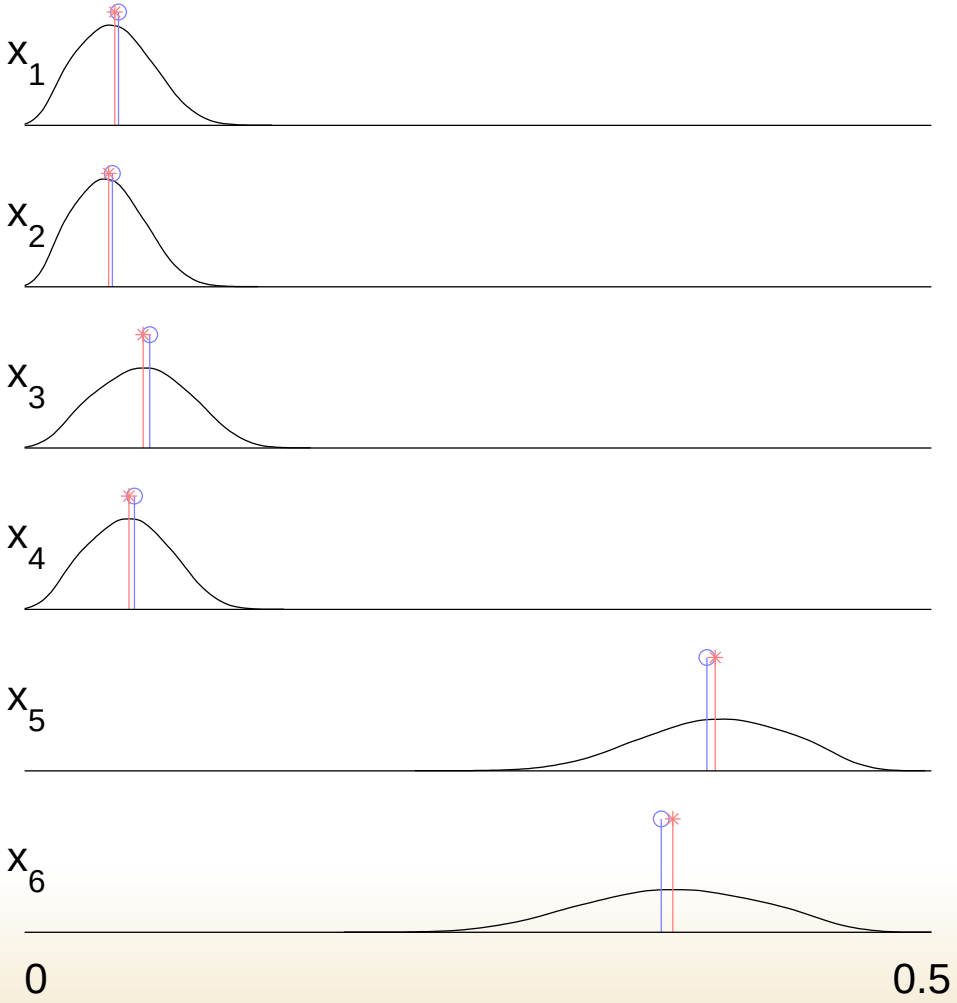
Alpha is



Data provided by Abraham Flaxman and Asela Gunawardana at Microsoft.

Example





What changes?

$\mathbf{x}(A)$ $\quad A \sim \text{Beta}(a, b, l, r)$ with $0 \leq l < r \leq 1$

1. $\mathrm{E}[x_i(A)] \geq 0$ and $\|\mathrm{E}[\mathbf{x}(A)]\| = 1$;

   thus $\mathrm{E}[\mathbf{x}(A)]$ is a probability distribution.

2. $\mathrm{E}[\mathbf{x}(A)] = \sum_{\ell=0}^{\infty} \mathrm{E}\left[A^\ell - A^{\ell+1}\right] \mathbf{P}^\ell \mathbf{v}$;

   thus we can interpret $\mathrm{E}[\mathbf{x}(A)]$ in length-ℓ paths.

3. for page i with no in-links, $x_i(A) = (1 - A)v_i$;

   thus $\mathrm{E}[x_i(A)] = x_i(\mathrm{E}[A])$ and $\mathrm{Std}[x_i(A)] = v_i \mathrm{Std}[A]$

   But is this one useful?

RAPr on Wikipedia

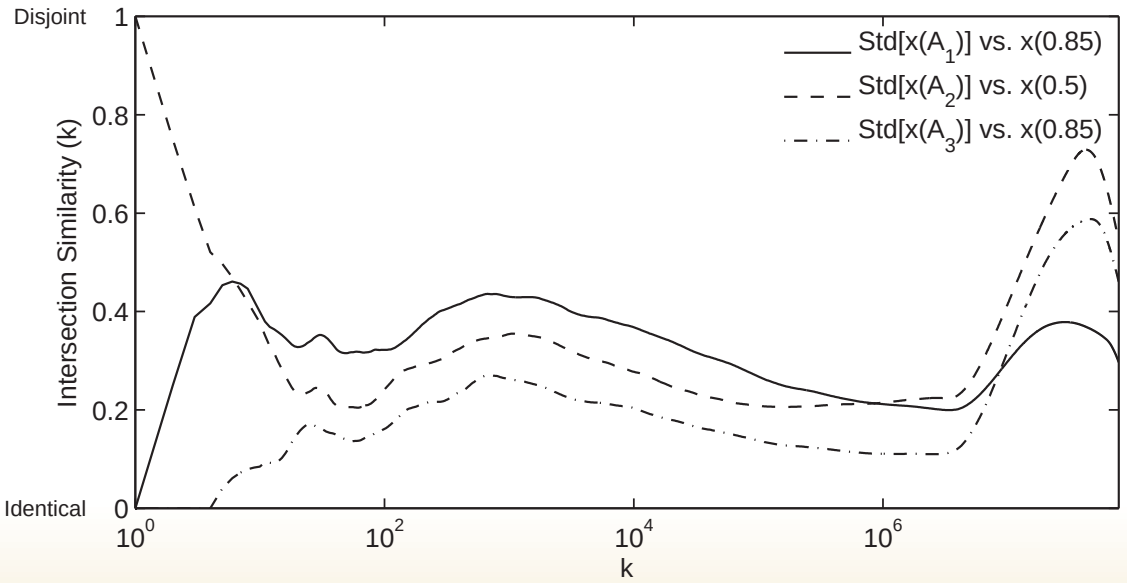
E [**x**(*A*)]	Std [**x**(*A*)]
United States	United States
C:Living people	C:Living people
France	C:Main topic classif.
United Kingdom	C:Contents
Germany	C:Ctgs. by country
England	United Kingdom
Canada	France
Japan	C:Fundamental
Poland	England
Australia	C:Ctgs. by topic

Std vs. PageRank

Does it tell us more than just PageRank?

uk2006 — 77M nodes and 2B edges

$$\text{isim}(k) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{2i}\,|\text{Diff}[Y(1{:}i), Z(1{:}i)]|$$



Kendall's τ

$\tau(\mathbf{x}(E_1), S_1) = +0.3$

$\tau(\mathbf{x}(E_2), S_2) = -0.5$

$\tau(\mathbf{x}(0.85), S_3) = -0.2$

$A_1 \sim \text{Beta}(2, 16, [0, 1])$ $\quad$ $A_2 \sim \text{Beta}(1, 1, [0, 1])$

$A_3 \sim \text{Beta}(0.5, 1.5, [0, 1])$

Computation

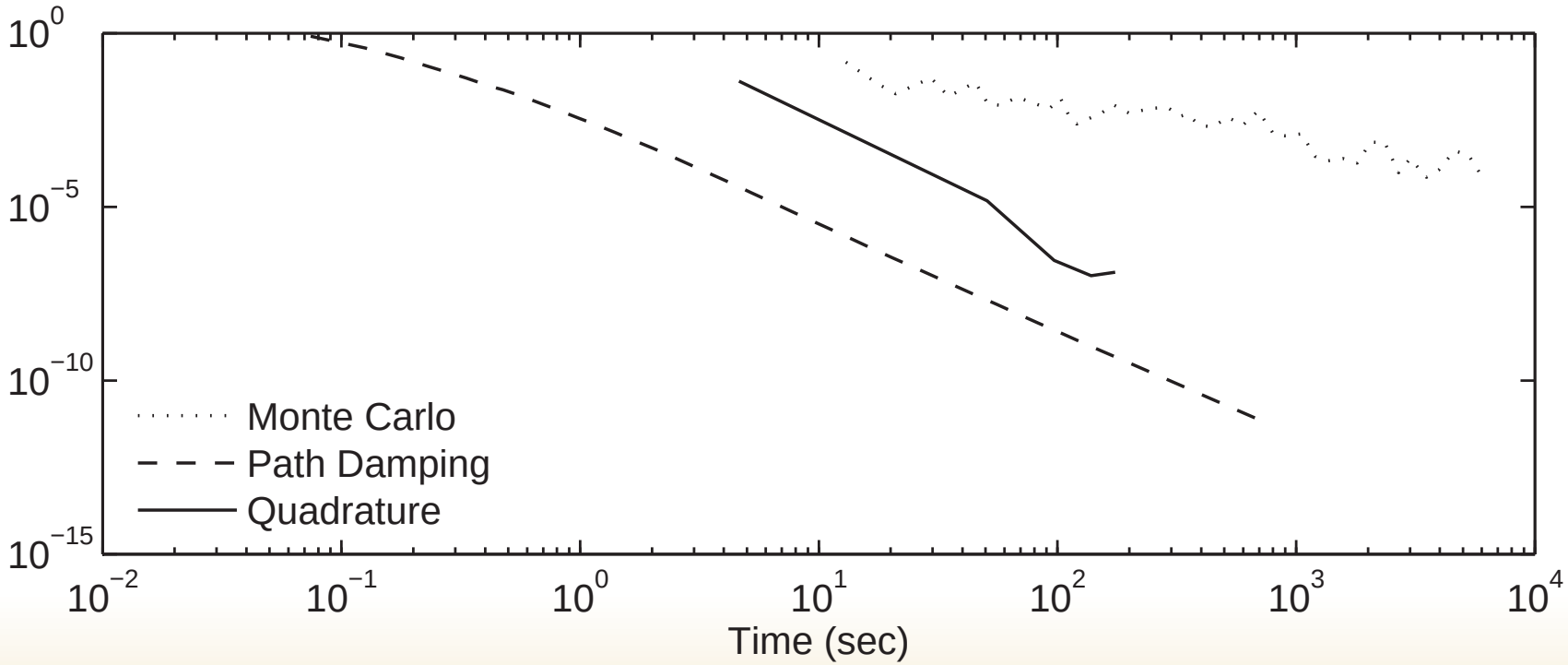
1. monte carlo
   $\mathrm{E}\left[\mathbf{x}(A)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}(\alpha_i) \qquad \alpha_i \sim A$

2. path damping
   $\mathrm{E}\left[\mathbf{x}(A)\right] \approx \sum_{i=0}^{N}\mathrm{E}\left[A^i - A^{i+1}\right]\mathbf{P}^i\mathbf{v}$

3. quadrature
   $\mathrm{E}\left[\mathbf{x}(A)\right] = \int_l^r \mathbf{x}(\alpha)\,d\rho(\alpha) \approx \sum_{i=1}^{N}\mathbf{x}(\zeta_i)\omega_i$

Time

cnr2000 — 325k nodes and 3M edges



Convergence theory

Method	Conv.	Work Required	What is N?
Monte Carlo	$\frac{1}{\sqrt{N}}$	N PageRank systems	number of samples from A
Path Damping (without Std $[\mathbf{x}(A)]$)	$\frac{r^{N+2}}{N^{1+\alpha}}$	$N+1$ matrix vector products	terms of Neumann series
Gaussian Quadrature	r^{2N}	N PageRank systems	number of quadrature points

α and r are parameters from Beta(α, b, l, r)

Webspam application

Hosts of uk-2006 are labeled as **spam**, **not-spam**, **other**

	P	R	f	FP	FN
Baseline	0.694	0.558	0.618	0.034	0.442
Beta(0.5,1.5)	0.695	0.561	0.621	0.034	0.439
Beta(1,1)	0.698	0.562	0.622	0.033	0.438
Beta(2,16)	0.699	0.562	0.623	0.033	0.438

Note Bagged (10) J48 decision tree classifier in Weka, mean of 50 repetitions from 10-fold cross-validation of 4948 non-spam and 674 spam hosts (5622 total).

Becchetti et al. Link analysis for Web spam detection, 2008.

Inner-Outer

Motivation

Why another PageRank algorithm?

For the RAPr codes, we need

1. reliable code
2. fast code over a range of α's
   → Use Matlab's "\"
3. code for big problems
   → **Use a Gauss-Seidel or custom Richardson method**
4. code with only matvec products
   → **Use the inner-outer iteration**
5. code with only 2 vectors of memory
   → **Use the power method**

Inner-Outer

Note PageRank is easier when α is smaller

Thus Solve PageRank with itself using $\beta < \alpha$!

Outer $$(\mathbf{I} - \beta\mathbf{P})\mathbf{x}^{(k+1)} = (\alpha - \beta)\mathbf{P}\mathbf{x}^{(k)} + (1-\alpha)\mathbf{v} \equiv \mathbf{f}^{(k)}$$

Inner $$\mathbf{y}^{(j+1)} = \beta\mathbf{P}\mathbf{y}^{(j)} + (\alpha - \beta)\mathbf{P}\mathbf{x}^{(k)} + (1-\alpha)\mathbf{v}$$

A new parameter? What is β? 0.5

How many inner iterations? Until a residual of 10^{-2}

Gray, Greif, Lau, 2007.

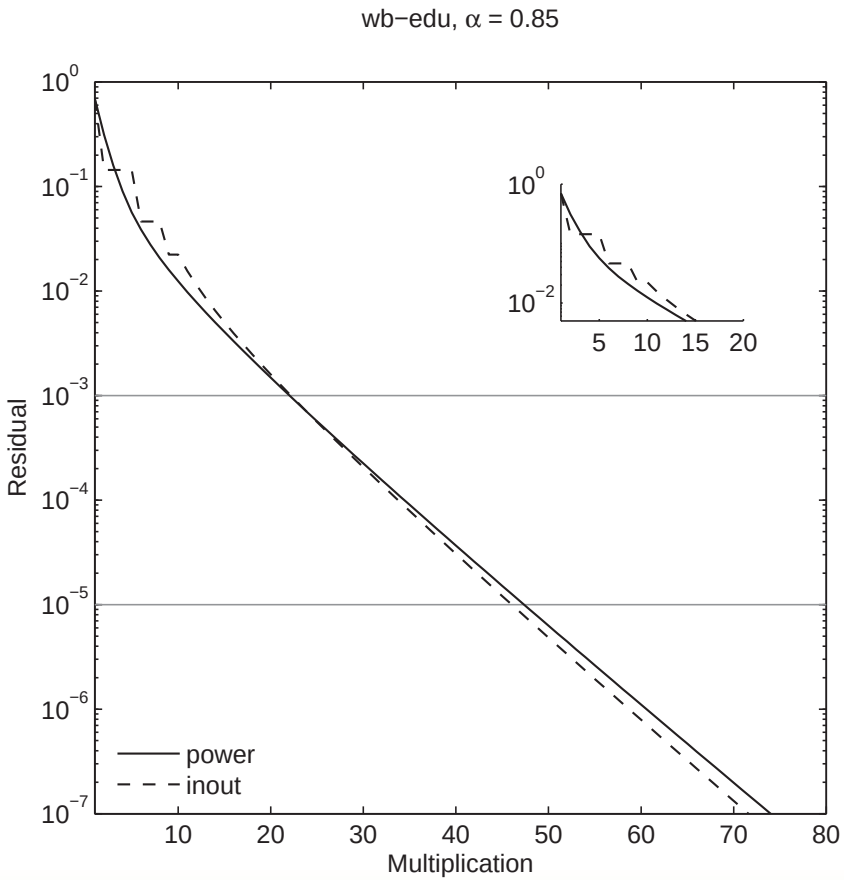
Inner-Outer algorithm

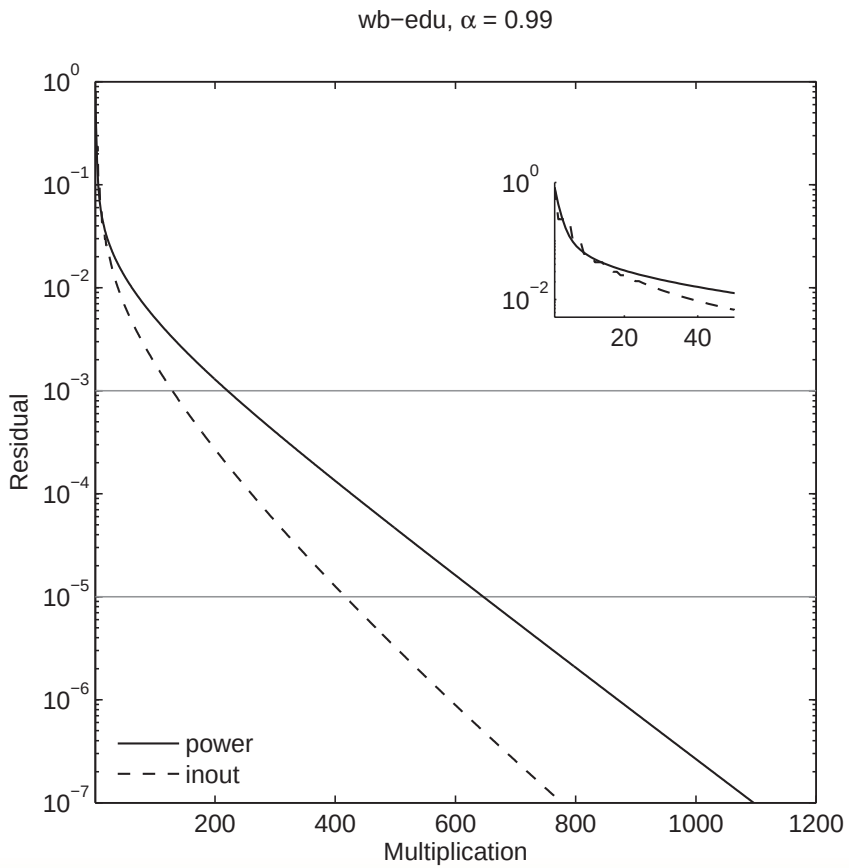
Input: $\mathbf{P}, \mathbf{v}, \alpha, \tau, (\beta = 0.5, \eta = 10^{-2})$
Output: $\mathbf{x}$
1: $\mathbf{x} \leftarrow \mathbf{v}$
2: $\mathbf{y} \leftarrow \mathbf{P}\mathbf{x}$
3: **while** $\|\alpha\mathbf{y} + (1-\alpha)\mathbf{v} - \mathbf{x}\|_1 \geq \tau$
4: $\quad \mathbf{f} \leftarrow (\alpha - \beta)\mathbf{y} + (1-\alpha)\mathbf{v}$
5: $\quad$ **repeat**
6: $\quad\quad \mathbf{x} \leftarrow \mathbf{f} + \beta\mathbf{y}$
7: $\quad\quad \mathbf{y} \leftarrow \mathbf{P}\mathbf{x}$
8: $\quad$ **until** $\|\mathbf{f} + \beta\mathbf{y} - \mathbf{x}\|_1 < \eta$
9: **end while**
10: $\mathbf{x} \leftarrow \alpha\mathbf{y} + (1-\alpha)\mathbf{v}$

- if $0 \leq \beta \leq \alpha$, convergence with any η
- uses only three vectors of memory
- $\beta = 0.5, \eta = 10^{-2}$ often faster than the power method
  (or just a **titch** slower)

Note *Note that the inner-loop checks its condition after doing one iteration.*

Performance





$\tau = 10^{-7}, \beta = 0.5, \eta = 10^{-2}$;
wb-edu graph (9.8M nodes, 57.M edges)

Extensions

1. A large scale shared-memory parallel version on compressed web graphs
2. A Gauss-Seidel variant
3. A BiCG-STAB preconditioner
4. A conjecture about the performance of the iteration
5. Showed the algorithm converges for "any" β, η

Gleich, Gray, Greif, Lau, submitted.

Convergence Result

Sketch of convergence result

1. error after j steps of the inner iteration

$$\mathbf{f}^{(j)} = \left(\alpha\beta^{j-1}\mathbf{P}^j + \left(\frac{\alpha-\beta}{\beta}\right) \sum_{\ell=1}^{j-1} \beta^\ell \mathbf{P}^\ell \right) \mathbf{f}^{(0)}$$

2. upper bound error by

$$\left\|\mathbf{f}^{(j)}\right\| \leq \frac{(\alpha-\beta)+(1-\alpha)\beta^j}{1-\beta} \left\|\mathbf{f}^{(0)}\right\|.$$

3. notice

$$\left\|\mathbf{f}^{(j)}\right\| \leq \alpha \left\|\mathbf{f}^{(0)}\right\|, j \geq 1$$

4. hence, convergence as long as $\beta \leq \alpha$

Summary

Conclusions

- α matters
- sensitivity is useful
- everything is just PageRank

Contributions

1. Derivative

Gleich, Glynn, Golub, Greif, 2007.

- *New technique to compute the derivative using just PageRank*

2. RAPr

Constantine and Gleich, 2007; Constantine, Gleich, and Iaccarino, submitted.

- New PageRank model and sensitivity measure
- Range of algorithms and algorithmic analysis
- Empirically helpful for spam identification
- *Robust software*

3. Inner-Outer

Gleich, Gray, Greif, Lau, submitted.

- Improved convergence analysis
- *Gauss-Seidel and preconditioning variants*
- *Shared-memory parallel implementation*
- *Robust software*

Thanks!

Michael Saunders (My Advisor)
Hector Garcia-Molina
Chen Greif
Art Owen
Amin Saberi

Thanks Gene!

Margot Gerritsen
Peter Glynn
Walter Murray
Reid Andersen
Pavel Berkhin
Kevin Lang
Amy Langville
Matthew Rasmussen
Sebastiano Vigna
Leonid Zhukov
Indira Choudhury
Seth Tornborg
Brian Tempero
Prisilla Williams
Deb Michael
Mayita Romero
Les Fletcher
Hugh Fletcher
Lindsey Fletcher
Jane Fletcher

Debbie Heimowitz
Jason Azicri
Steven Fan
Paul Constantine
Michael Atkinson
Jeremy Kozdon
Esteban Arcaute
Adam Guetz
Will Fong
Andrew Bradley
Nick Henderson
Chris Maes
Nicole Taheri
Ying Wang
Nick West
Kaustuv's Rum
Saeco Coffee Machine
Napa Valley
Matlab
superlu

THANK YOU