

# **The PageRank Citation Ranking: Bringing Order to the Web**

by

Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd

presented by

Martin Klein, Santosh Vuppala  
{mklein, svuppala}@cs.odu.edu

ODU, Norfolk, 01/31/2007

# Outline

- Background
- PageRank
- Implementation
- PageRank's Convergence
- Searching and other Applications
- Discussion

# Background - Authors



- Larry Page (~Rank)
  - BS in CE from UMich, MS from Stanford

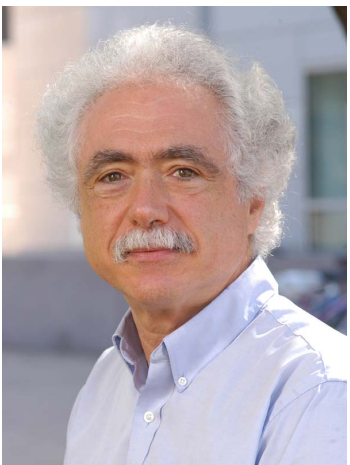


- Sergey Brin
  - BS in Math&CS from UMD, MS from Stanford
- Google Inc. in 09/98 (google.com - 09/97)

# Background - Authors



- Rajeev Motwani
  - Ph.D 1988, CS, UC Berkeley
  - Professor at Stanford U



- Terry Winograd
  - Ph.D. 1970, M.I.T, Applied Mathematics
  - Professor at Stanford U

# Background - Paper

- Stanford WebBase project (1996 - 1999)  
<http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/>  
<http://dbpubs.stanford.edu:8091/diglib/>
- funded by NSF through DLI I  
<http://www.dli2.nsf.gov/dlione/>

“The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.” quote from the DLI I website

# Background - Paper

- it is a technical report! (working paper)  
(Stanford Digital Libraries SIDL-WP-1999-0120)
- from the paper: web size = 150M web pages
- 2005: Google claims to index more than 8B pages  
(<http://blog.searchenginewatch.com/blog/041111-084221>)
- 11.5B overall (<http://www.cs.uiowa.edu/~assignori/web-size/>)

# PageRank - Motivation

“The average web page quality experienced by a user is higher than the quality of the average web page. This is because the simplicity of creating and publishing web pages results in a large fraction of low quality web pages that users are unlikely to read.”

ex #1

- Differentiate Pages

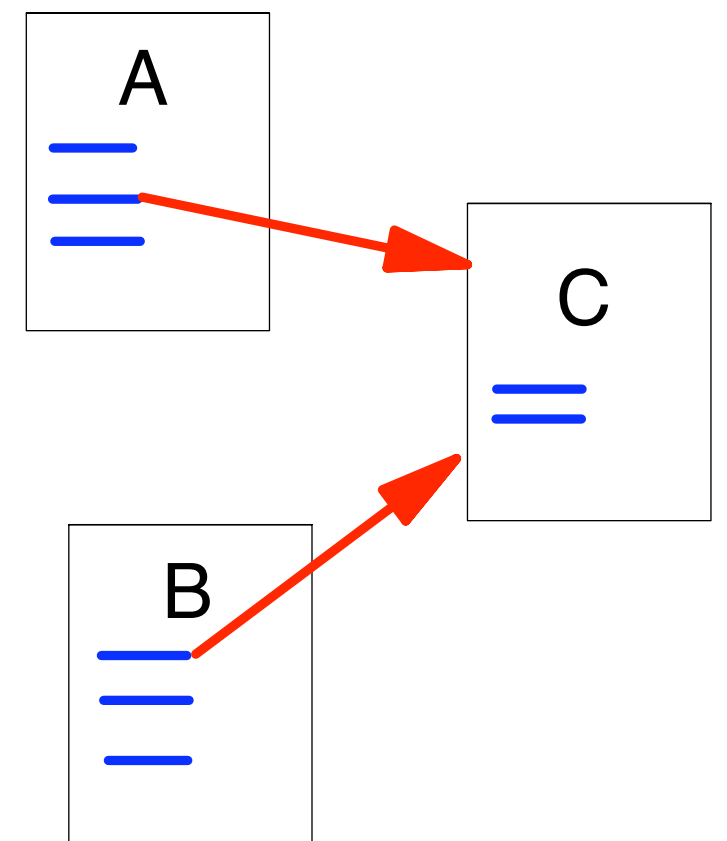
ex #2

- Relative Importance
- Ranking/Search



# PageRank - Basics

- based on link structure of the web
- pages = nodes && links = edges
- forward links = outedges
- backlinks = inedges
- A and B are Backlinks of C



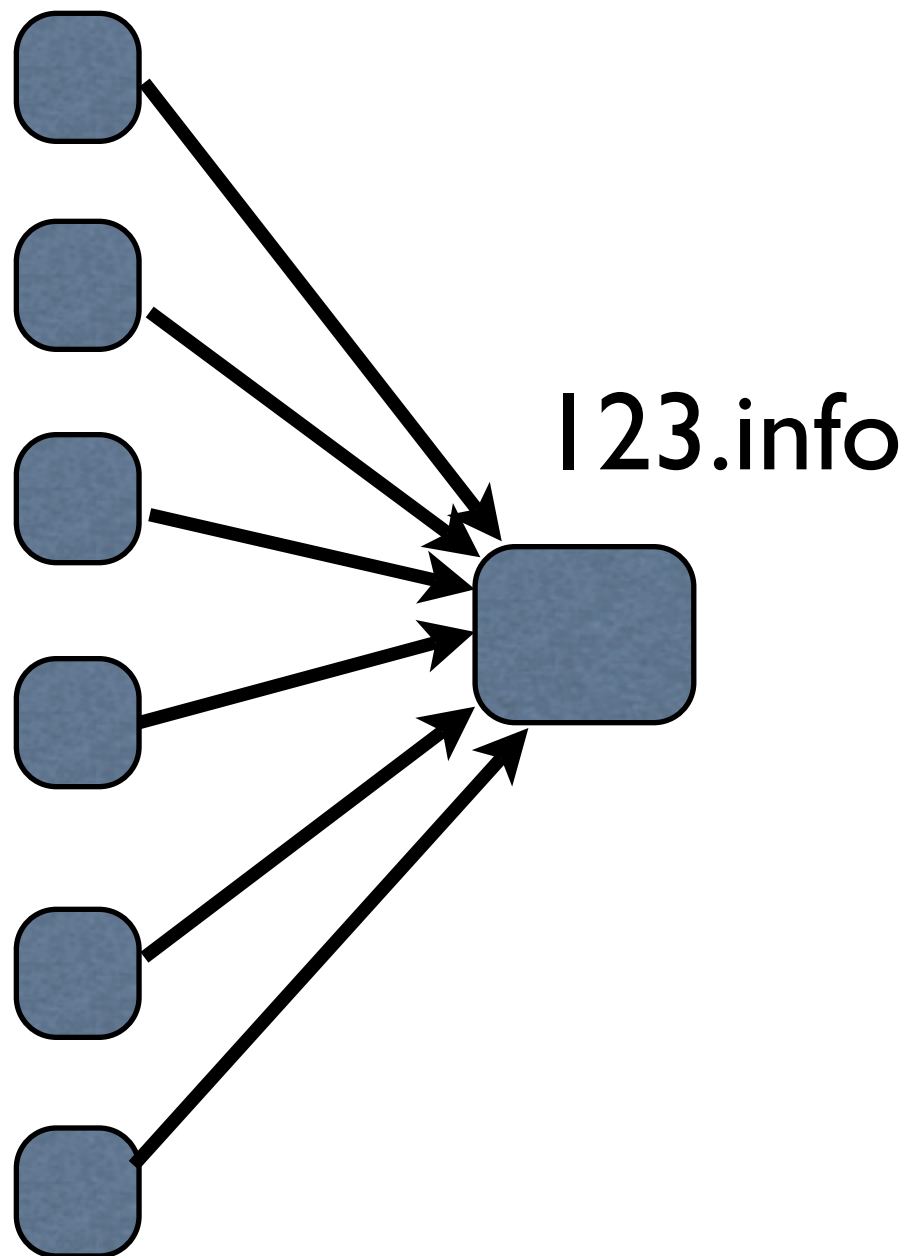


# PageRank - Assumptions

- a link from page A to page B is a vote from A to B
- highly linked pages are more “important” than pages with few links
- backlinks from high PR-pages count more than links from low PR-pages
- combination of PR and text-matching techniques result in highly relevant search results

# PageRank - Assumptions

p1-p6.info



cnn.com



abc.com



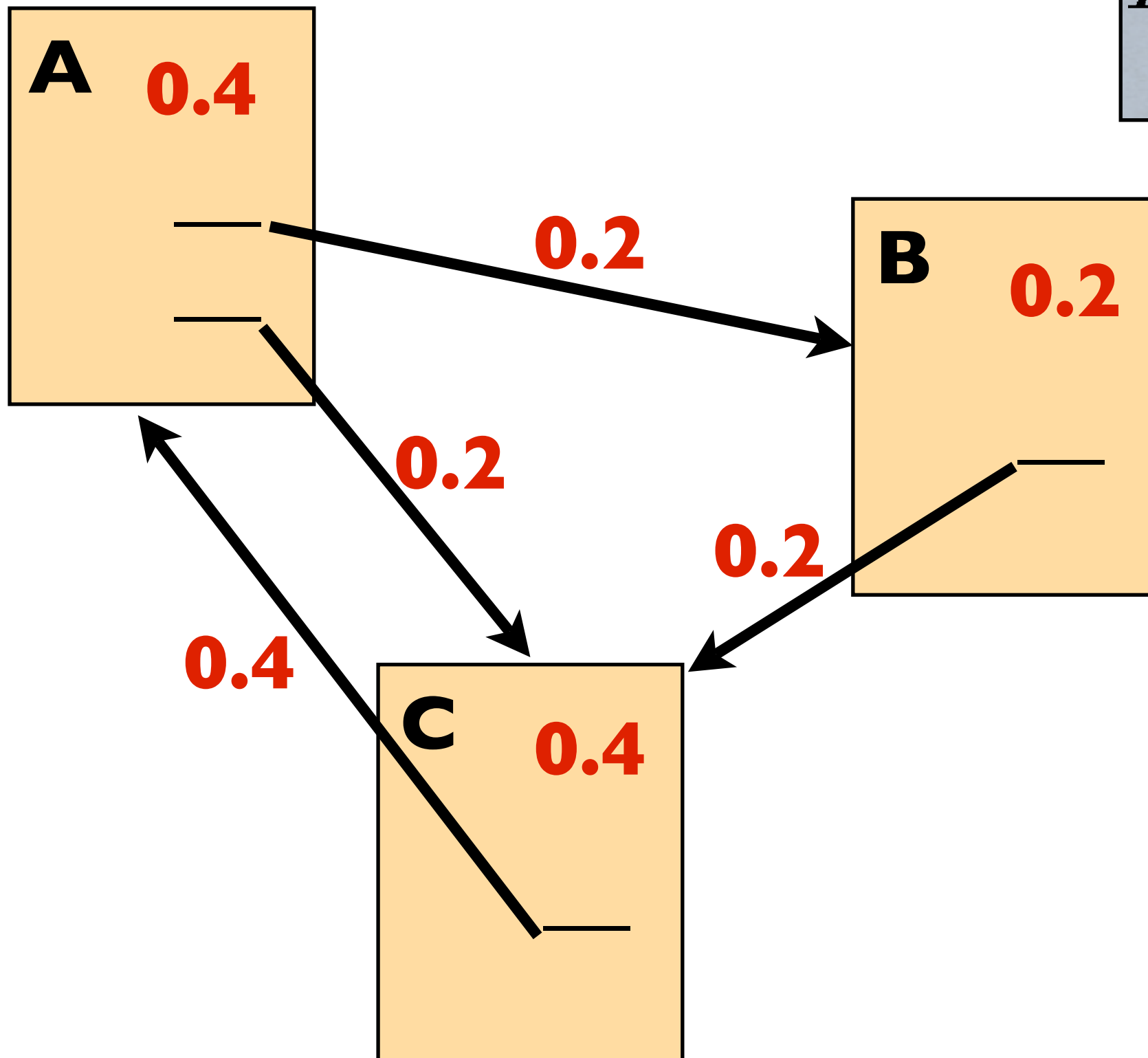
# PageRank - Definition

- $u$  is a web page
- $F_u$  = set of pages  $u$  points to
- $B_u$  = set of pages pointing to  $u$
- $c$  = normalization factor
- $N_u = |F_u|$

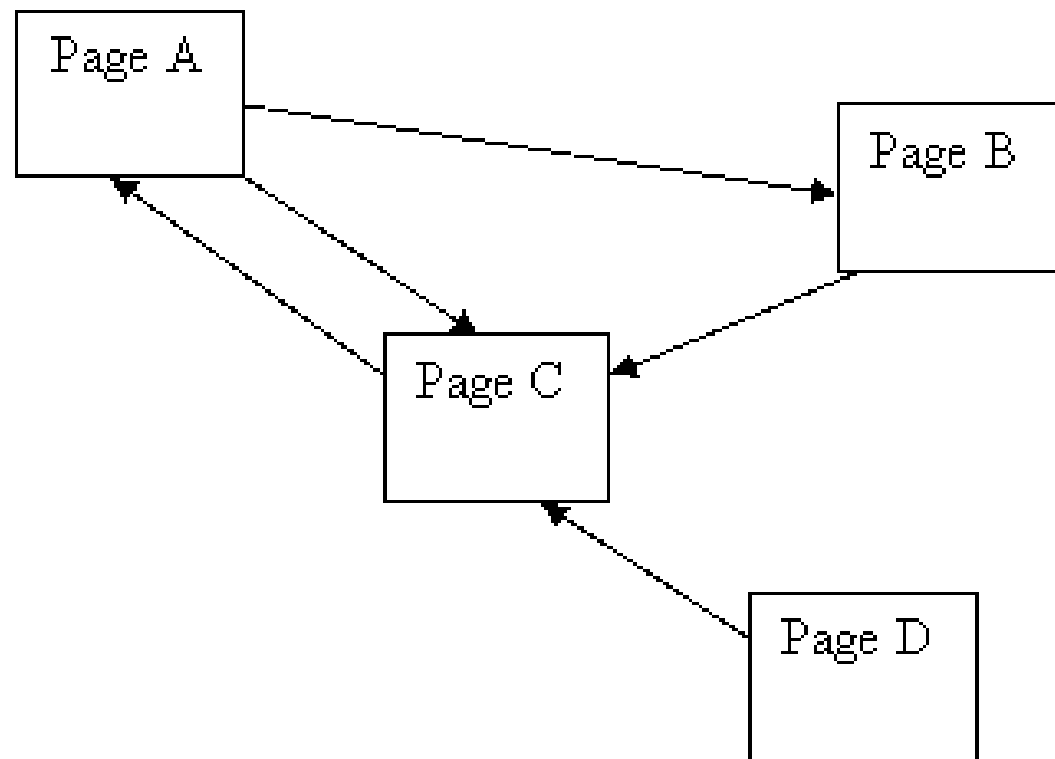
$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

# PageRank - Example

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$



# PageRank - Iteration Example



$d=0.85$

Iteration 1

$PR = 1$  for all nodes

$$PR(A) = 1 - d + d \left( \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \dots \right)$$

Iteration 2	Iteration 3	Iteration 4		Iteration 10
$PR(A)=1.85$	$PR(A)=1.8653$	$PR(A)=1.568$		$PR(A)=1.024$
$PR(B)=1.7225$	$PR(B)=1.735$	$PR(B)=1.4828$	...	$PR(B)=1.0204$
$PR(C)=4.036$	$PR(C)=3.3377$	$PR(C)=2.8706$		$PR(C)=2.057$
$PR(D)=0.15$	$PR(D)=0.15$	$PR(D)=0.15$		$PR(D)=0.15$

# PageRank - Definition

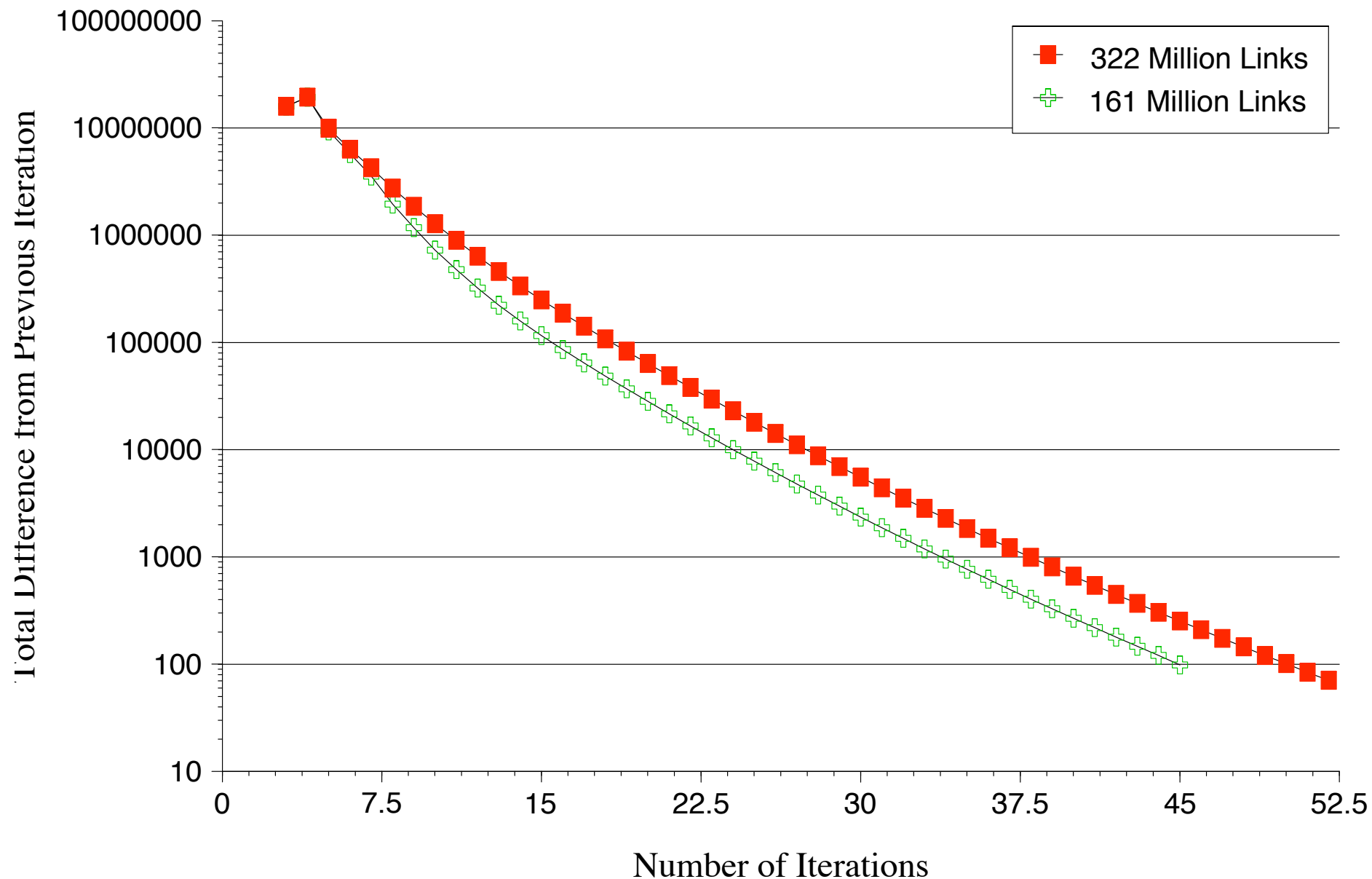
What if two pages only link to each other and some page points to one of them?

- this loop/trap is called rank sink
- based on random surfer model
  - $E$  - probability that a user visits a page

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

# Convergence

## Convergence of PageRank Computation



- PR computation converges very quickly
- scales very well



# Implementation

- built a crawling and indexing system
- repository size: 24M web pages (over 75M unique URLs)
- web crawler keeps index of links
- computing PR of entire repository takes ~5h
- issues: volume(!!!), incorrect HTML, dynamics of the web, page exclusion (robots.txt)

# Search - Background

- title search and full text search (Google)
- ex.: title search
  - 16M pages
  - returns pages where title contains all query words

# Title Search

**Multi Search** university Search Next! [national parks]

10 results clustering on Search

Query: **university**  
11 Results Returned  
Showing Results From 0 to 10

**Stanford University Homepage**  
http://www.stanford.edu/  
74.79% 4k - 2591993 - 010397

**Stanford University: Portfolio Collection**  
http://www.stanford.edu/home/administration/portfolio.html  
65.78% 3k - 2591993 - 010397

**University of Illinois at Urbana-Champaign**  
http://www.uiuc.edu/  
73.26% 13k - 123096 - 010397

**Indiana University**  
http://www.indiana.edu/  
68.38% 1k - 092896 - 010597

**University of California, Irvine**  
http://www.uci.edu/  
68.07% 3k - 123096 - 010397

**University of Minnesota**  
http://www.umn.edu/  
67.05% 0k - 121696 - 010397

**Iowa State University Homepage**  
http://www.iastate.edu/  
66.66% 3k - 121896 - 010397

**The University of Michigan**  
http://www.umich.edu/  
66.35% 1k - 2591993 - 010397

**Mississippi State University**  
http://www.msstate.edu/  
66.35% 3k - 2591993 - 010397

**Northwestern University: NUInfo**  
http://www.nwu.edu/  
66.15% 3k - 121496 - 010597

next 10

**Optical Physics at the University of Oregon**  
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group...  
<http://optics.mcgill.edu/> - size 1K - 16 Dec 96

**Carnegie Mellon University - Campus Networking**  
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...  
<http://www.net.cmu.edu/> - size 4K - 19 Aug 95

**Wesleyan University Computer Science Group Home Page**  
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.  
<http://www.cs.wesleyan.edu/> - size 2K - 15 Apr 96

**Keio University Shonan Fujisawa Campus (SFC)**  
B\$3\$N%ZIEFnF#Bt%-9c%e%Q%9 (B(SFC) \$B\$N (BWWW \$B% \$BCmOU=q\$- (B \$B\$rFI\$s\$G\$/\$@5\$ \$!# (B. Nihongo | English. SFC \$B>pJs (B. [ \$B%a%G%#% "%; %e%?!\*...  
<http://www.sfc.keio.ac.jp/> - size 3K - 5 Feb 97

**School of Chemistry, University of Sydney**  
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.  
<http://www.chem.su.oz.au/> - size 4K - 25 Feb 97

**Mankato State University**  
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...  
<http://www.mankato.msus.edu/> - size 3K - 27 Nov 96

**St. Ambrose University**  
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...  
<http://www.sau.edu/> - size 2K - 4 Feb 97

**University of Washington ECSEL Projects**

# Search - The Common Case

- page with high usage
- PR handles CC queries well
- CC for “wolverine” - U Michigan software system
- else: wiki page, imdb, etc

**“It is important to note that the goal of finding a site that contains a great deal of information about wolverines is a very different task than finding the common case wolverine site.”**

quote taken from the paper

# Personalized PageRank

- E vector - distribution of web pages a random surfer jumps to
- usually E is uniform over all web pages (democratic)
- apply E just for one web page results in high PR value for relevant pages regarding the applied page
- e.g. apply E for web page of faculty from cs@odu results in high PR for CS related pages

# Other Uses of PageRank

- estimating web traffic - compare web page access from proxy vs PR
- PR as backlink predictor
  - efficient web crawling - better docs first
  - PR outperforms citation counts b/c number of citation count is not known in advance
- the PR proxy - annotate links with PR value
- PR is applied to the binary directed network model which is one of the methods used to model the co-authorship networks in relevance to digital libraries



# Unwanted Uses of PageRank

- bmw.de banned from google in early 2006 due to its doorway page  
~ is a page stuffed full of keywords that the site feels a need to be optimized for  
blog: <http://blog.outer-court.com/archive/2006-02-04-n60.html>
- “If an SEO creates deceptive or misleading content on your behalf, such as doorway pages or ‘throwaway’ domains, your site could be removed entirely from Google’s index.” *unknown at Google*
- google's webmaster helpcenter:  
<http://www.google.com/support/webmasters/bin/answer.py?answer=35291>



# Unwanted Uses of PageRank

- “Google Bomb”

<http://searchengineland.com/070125-230048.php>

- create lots of links to one certain destination
- label all of them with the same remarkable terms
- query Google for those terms and you will get the linked page

`<a href="http://www.whitehouse.gov/president/gwbbio.html">Miserable Failure</a>`

# Discussion

## **Question 1:**

PageRank is not optimal! How can it be improved? What can be changed?

## **Question 2:**

Do you think, not publishing the PR value (Google Toolbar) would make it difference in the quest for obtaining a high PR value?

## **Question 3:**

Considering the responsibility Google as a Search Engine has (as a prime source of information), should PageRank plus Google's additional "Ranking-VooDoo" not be more transparent to the public?



Hot News Headlines

Mac OS X Tip of the Week: Search by Colors



Introducing iPhone



Macworld Keynote



Presenting Apple TV



New Mac Ads

Visit an Apple Store

Search Apple.com

[1-800-MY-APPLE](#) | [Find a Reseller](#) | [Site Map](#) | [RSS Feeds](#) | [Job Opportunities](#) | [Media Info](#) | [Terms of Use](#) | [Privacy Policy](#) | [Contact Us](#)

[Redeem iTunes Gift Cards](#) (iTunes 7 required)

[iBook G4 and PowerBook G4 Battery Recall](#)

Copyright © 2007 Apple Inc. All rights reserved.



Apple.com Worldwide



**YAHOO!**

Search:  Web Images Video Local Shopping more

Y! Answers: How do trick candles work? Find out

My Yahoo! My Mail

Page Options

**Featured** Entertainment Sports Life

Jan 31, 2007

**Ban on the bulb?**  
Find out a measure that's being considered in one state to regulate the way homes are lit. » [Read article](#)  
• Comparing old and new light bulbs  
• Eiffel Tower plans to turn off lights

Could one state ban old light bulbs?  
Tyra Banks defends her weight gain  
Sleuth on trail of lost Da Vinci masterpiece  
Barbie does her best to stay in the game

» [More Featured](#)

**In the News** World Local Video

As of 11:01 a.m. EST

- British police arrest eight terror suspects | Attack foiled
- Car bombs, attacks kill 12 in Baghdad | Who is enemy?
- North Korea may plan second nuclear test if talks fail
- Ex-Israeli justice minister convicted of sexual misconduct
- Analysts predict Fed to leave interest rates steady
- Astronauts to try three spacewalks in nine-day period | Photos
- Canada city to declare 'Stephen Colbert Day' after hockey bet

» [More: News | Popular | Business](#)

Markets: Dow: **+0.3%** Nasdaq: **+0.1%** Sponsored by: **Scottrade**

Stock Quotes:

**Marketplace**

GEICO auto insurance

**Small Business**

- Get a Web Site
- Domain Names
- eCommerce
- Search Listings

**Featured Services**

- Downloads

**Autos**

**Finance** EXPANDED!

**Games**

**GeoCities**

**Groups**

**HotJobs**

**Maps**

**Movies**

**Music**

**Personals**

**Photos**

**Real Estate**

**Shopping**

**Sports**

**Tech**

**Travel**

**TV**

**Yellow Pages**

**More Yahoo! Services**

Check your mail status: [Sign In](#) Free mail: [Sign Up](#)

[Mail](#) [Messenger](#) [Radio](#)

[Weather 30°F](#) [Local](#) [Horoscopes](#)

**YOUR PC**  
FIND YOUR MATCH  
[SHOP NOW](#)

© Best Buy

Love your PC - Ad Feedback

**Inside Yahoo! Games**

**Flo returns in Diner Dash 3**

- Diner Dash 3
- Family Feud
- Westward
- Bookworm Adventures
- Slingo Quest
- Myth Match

» [More Games](#)

**Pulse - What Yahoos Are Into**

Yahoo! My Yahoo! Mail Welcome, Guest [\[Sign In\]](#) [Directory Home](#) [Help](#)

**YAHOO! SEARCH** Directory

Search:  the Web | the Directory | this category

**Apple Notebook Computers** [Email this page](#) [Suggest a Site](#) [Advanced Search](#)

Directory > Computers and Internet > Hardware > Notebook Computers > Product Information and Reviews > Apple [Save to My Web](#)

**CATEGORIES** [\(What's This?\)](#)

**SPONSOR RESULTS**

[Auctions@](#)

[Buyer's Guides@](#)

**SITE LISTINGS** [By Popularity](#) | [Alphabetical](#) [\(What's This?\)](#) Sites 1 - 5 of 5

- Apple MacBook**  
Official information for Apple's Macbook notebook computer which features an Intel Core Duo processor an iSight camera, iLife application suite, and a 13-inch widescreen display.  
[www.apple.com/macbook/macbook.html](http://www.apple.com/macbook/macbook.html)
- PowerBook (15)**  
[dir.yahoo.com/.../Apple/PowerBook](http://dir.yahoo.com/.../Apple/PowerBook)
- iBook (3)**  
[dir.yahoo.com/.../Apple/iBook](http://dir.yahoo.com/.../Apple/iBook)
- MacBook Pro**  
Information, pictures, and features of Apple's Intel-based notebook computer.  
[www.apple.com/macbookpro](http://www.apple.com/macbookpro)
- Pimp My Laptop**  
Service offering customized stickers designed to fit your Apple laptop covers. Choose from a pre-made design or upload your own design.  
[www.pimpmylaptop.com](http://www.pimpmylaptop.com)

**SPONSOR RESULTS** [\(What's This?\)](#) [\(Become a Sponsor\)](#)

- Official Apple Mac(TM) Store**  
[store.apple.com](http://store.apple.com) New MacBook with Intel Core 2 Duo.
- MacBook - Cheap Prices**  
[www.nextag.com](http://www.nextag.com) Find notebooks by display, weight, and processor.

**Official Apple Mac(TM) Store**  
New MacBook with Intel Core 2 Duo.  
[store.apple.com](http://store.apple.com)

**MacBook - Cheap Prices**  
Find notebooks by display, weight, and processor.  
[www.nextag.com](http://www.nextag.com)

**MacBooks - Sale Prices**  
Notebook Computers from 104+ Shops.  
[www.calibex.com](http://www.calibex.com)

**Apple MacBook Laptop - Save Online**  
Visit 24HourMall.com affiliate site for special deals, links,....  
[www.24hourmall.com](http://www.24hourmall.com)

**Macbook at MSN Shopping**  
Find Great Deals on a Wide Selection of Computers and Software.  
[shopping.msn.com](http://shopping.msn.com)

[See your message here...](#)

Copyright © 2007 Yahoo! Inc. All rights reserved. [Privacy Policy](#) - [Terms of Service](#) - [Copyright/IP Policy](#)

<http://www.yahoo.com>

[http://dir.yahoo.com/Computers\\_and\\_Internet/Hardware/Notebook\\_Computers/Product\\_Information\\_and\\_Reviews/Apple/](http://dir.yahoo.com/Computers_and_Internet/Hardware/Notebook_Computers/Product_Information_and_Reviews/Apple/)

# References

websites:

<http://www.google.com/corporate/execs.html>

<http://www.google.com/corporate/index.html>

<http://www.iprcom.com/papers/pagerank/>

<http://www.webworkshop.net/pagerank.html>

<http://en.wikipedia.org/wiki/PageRank>

and many more papers....

# PR Computation

$$PR(A) = 1 - d + d \left( \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \dots \right)$$

$$PR(A) = \frac{1 - d}{N} + d \left( \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \dots \right) .$$

where N = number of documents in the collection

# Precision and Recall

