



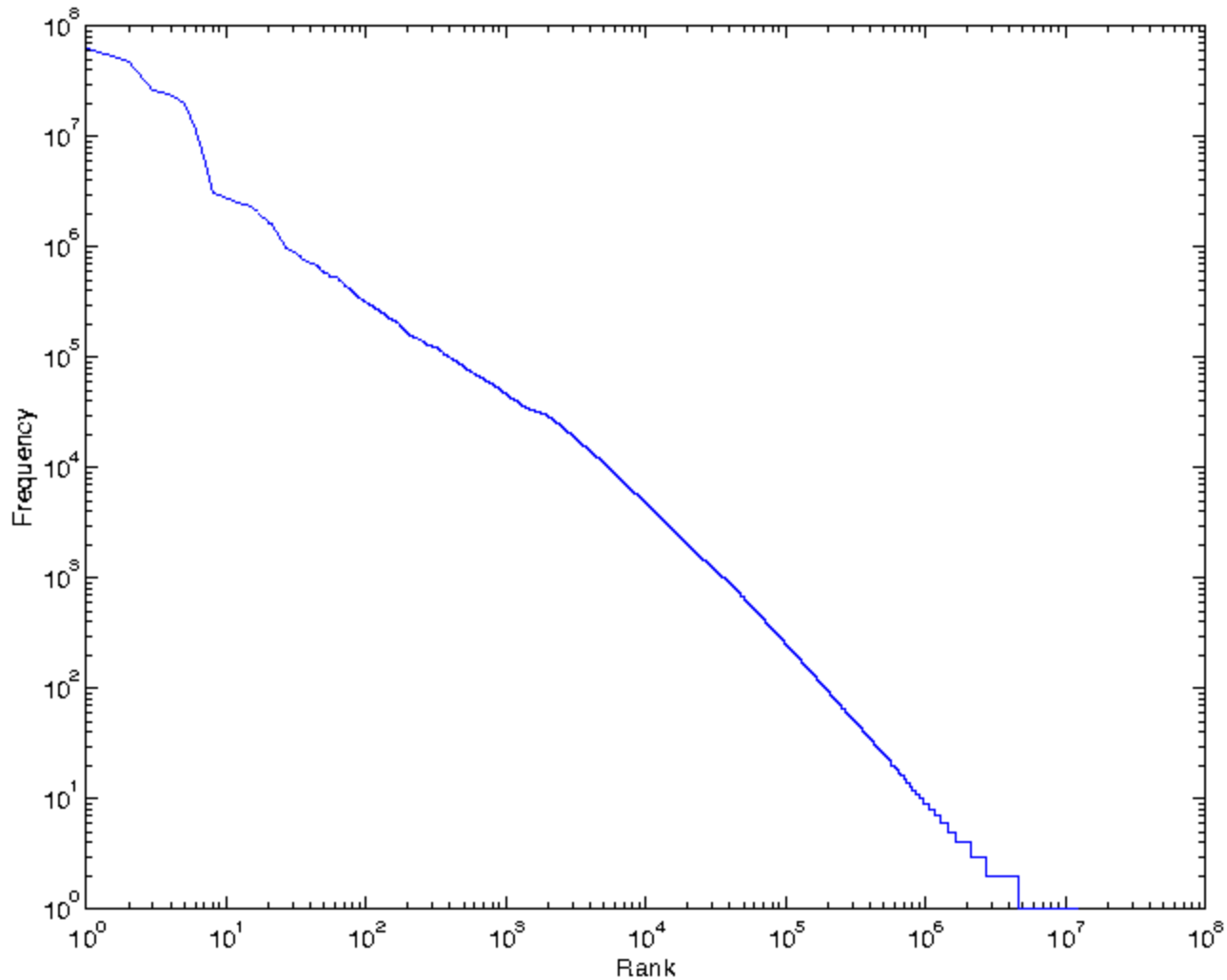
# **Web Characterization: What Does the Web Look Like?**

Frank McCown  
Introduction to Web Science  
Harding University  
Spring 2011

# W3C Characterization Activity

- Work from 1998-1999
- Provided definitions for common Web terms like resource, link, proxy, server, etc., some of which are now dated  
<http://www.w3.org/1999/05/WCA-terms/>
- Attempted to answer questions like: How many web pages are there? and How fast is the Web growing?
- Summary in Pitkow, [Summary of WWW Characterizations](#), Journal of the World Wide Web, 1999

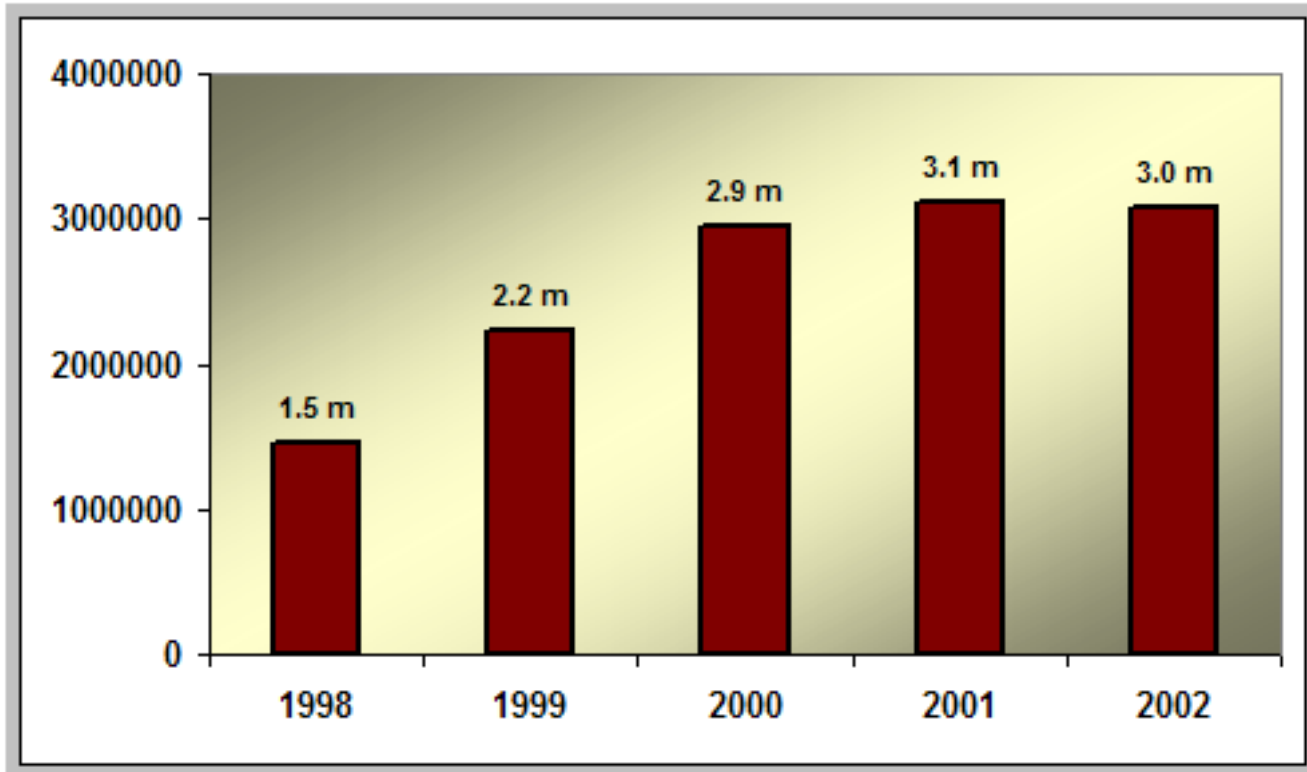
# Web Page Popularity (1994)



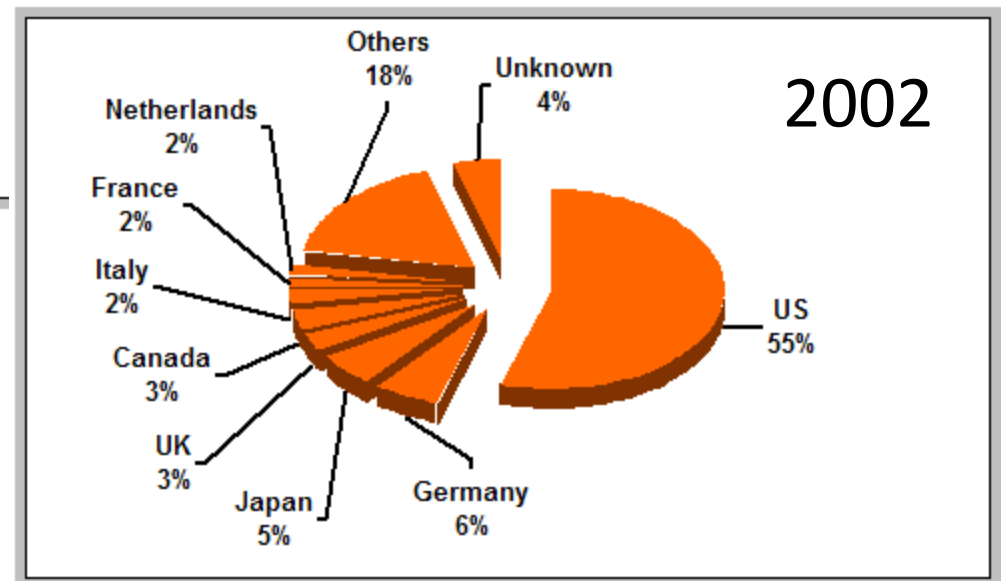
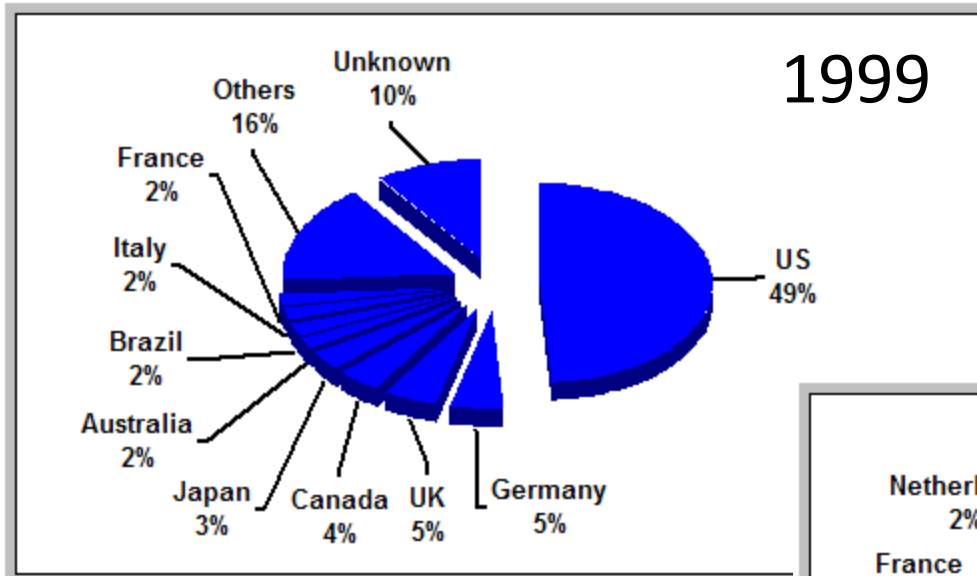
# OC LC Characterization Research

- Work from 1998-2002
- Analyzed Web samples annually to look for trends
- Sample obtained by randomly sampling IP addresses and connecting to port 80
  - Today this method would miss a large number of websites that use *virtual hosting* – multiple domain names hosted on same computer using one IP address
- Findings: O'Neill et al., [Trends in the Evolution of the Public Web](#), *D-Lib Magazine*, Apr 2003

# Number of Public Websites



# Distribution of Websites by Country



# Popular Websites by In-Links

OCLC Most Linked-To Websites<sup>1</sup>

2000		2002	
1	www.microsoft.com	1	www.adobe.com
2	www.netscape.com	2	www.microsoft.com
3	www.geocities.com	3	www.geocities.com
4	members.aol.com	4	www.netscape.com
5	www.yahoo.com	5	members.aol.com
6	www.adobe.com	6	www.yahoo.com
7	www.amazon.com	7	www.amazon.com
8	www.altavista.com	8	www.google.com
9	members.tripod.com	9	www.macromedia.com
10	www.macromedia.com	10	www.cnn.com

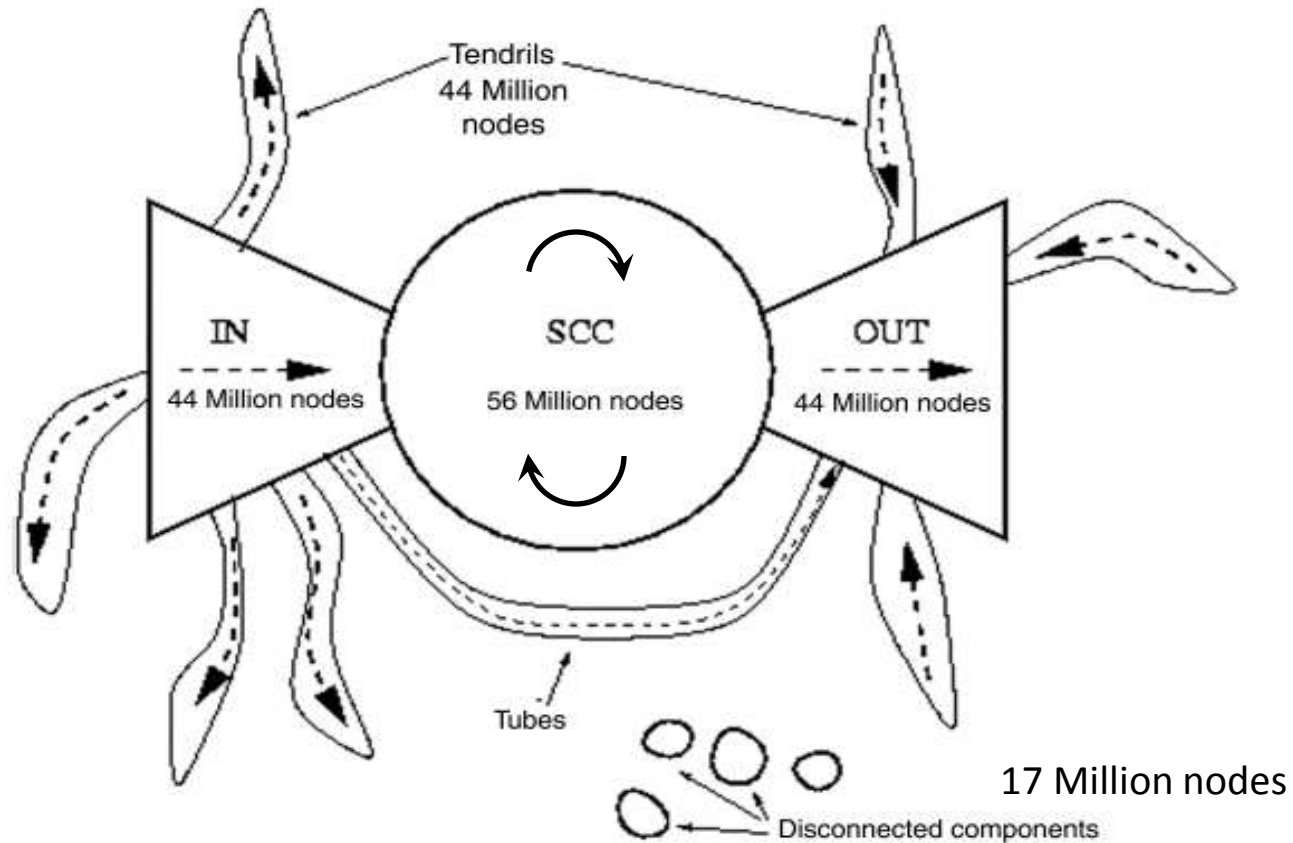
Most Linked-To Websites in 2009<sup>2</sup>

1. en.wikipedia.org
2. www.youtube.com
3. www.dictionary.com
4. www.craigslist.com
5. www.facebook.com
6. www.myspace.com
7. www.twitter.com
8. www.imdb.com
9. www.hulu.com
10. www.perezhilton.com

<sup>1</sup><http://www.oclc.org/research/activities/past/orprojects/wcp/stats/linkage.htm>

<sup>2</sup><http://www.seomoz.org/blog/tangled-web-the-most-linked-to-pages-on-the-internet>

# Bow-Tie Structure of the Web



Broder et. al (Graph Structure of the Web, 2000)  
Examined a large web graph (200M pages, 1.5B links)

# Characterizing National Web Domains

- A large-scale study by Baeza-Yates et al.<sup>1</sup> analyzed web collections from 10 national domains and multinational Web spaces of African and Indochinese Web sites
- Examined languages, file sizes, pages per site, link structure, etc.

<sup>1</sup>Baeza-Yates et al. , Characterization of national Web domains, *ACM Trans. Internet Technol.*, May 2007

# Web Page Languages

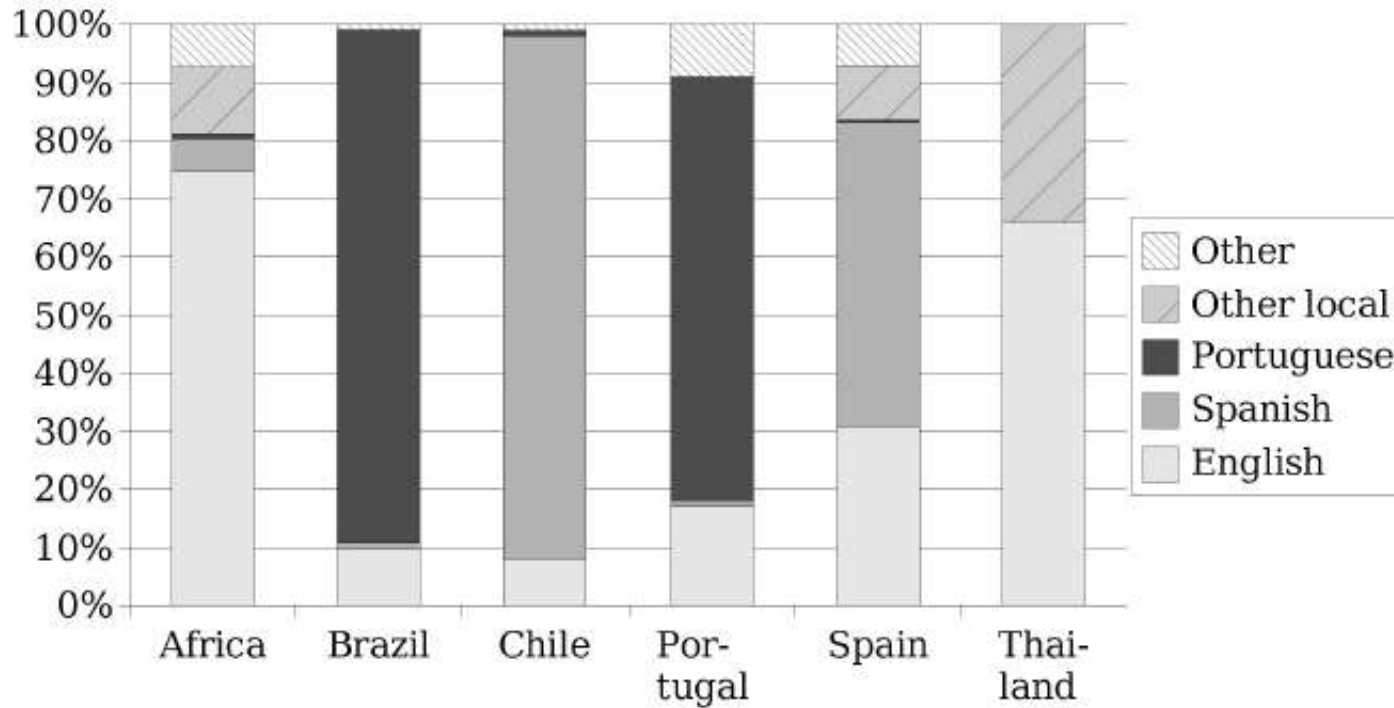
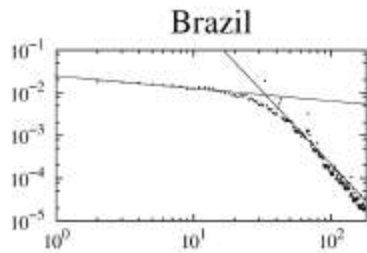


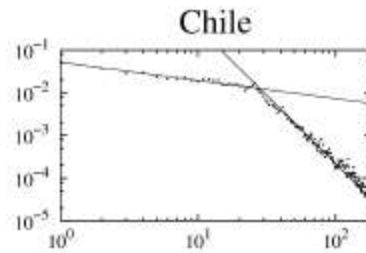
Fig. 2. Distribution of the number of pages in different languages.

# Some Power-law Distributions



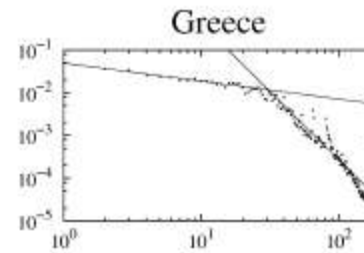
$\bar{x} = 24$  KB

$\theta_1 = 0.3; \theta_2 = 3.4$



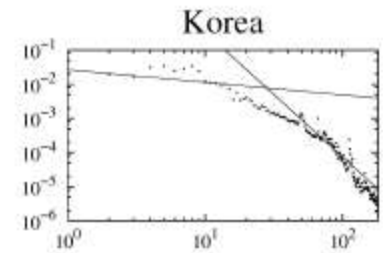
$\bar{x} = 21$  KB

$\theta_1 = 0.4; \theta_2 = 3.2$



$\bar{x} = 22$  KB

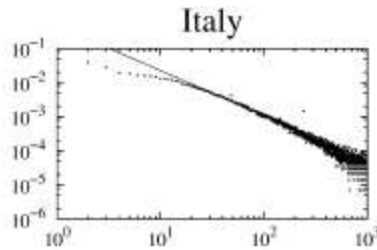
$\theta_1 = 0.4; \theta_2 = 3.2$



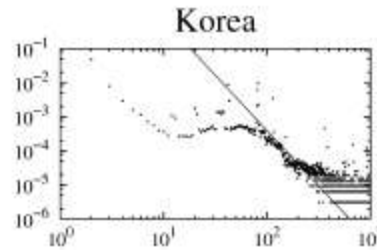
$\bar{x} = 14$  KB

$\theta_1 = 0.4; \theta_2 = 3.7$

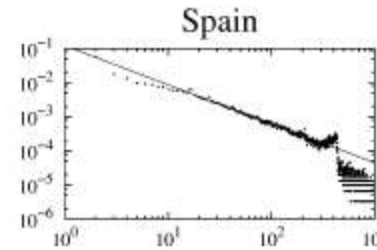
## File sizes for small and large files



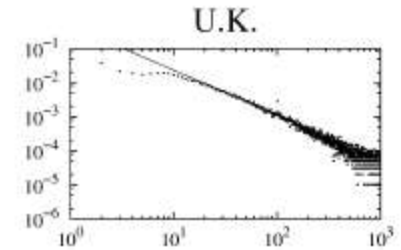
$\bar{x} = 410; \theta = 1.3$



$\bar{x} = 224; \theta = 3.2$



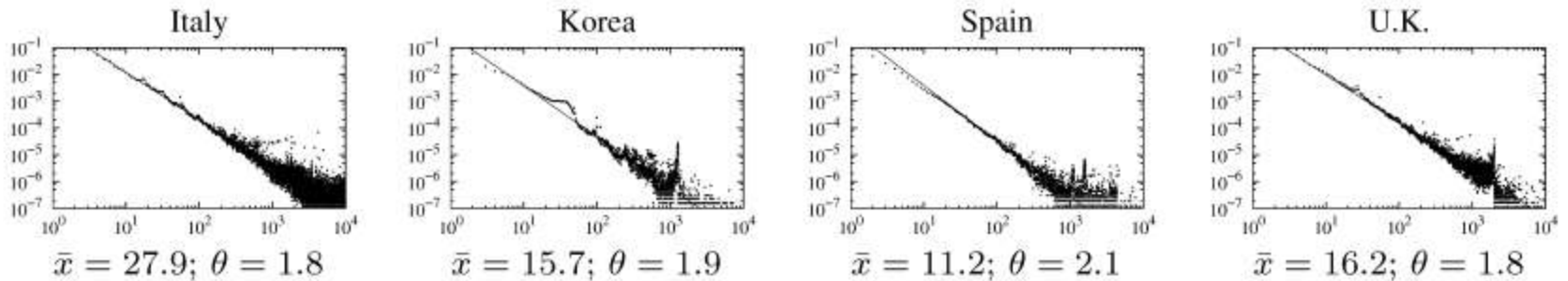
$\bar{x} = 52; \theta = 1.1$



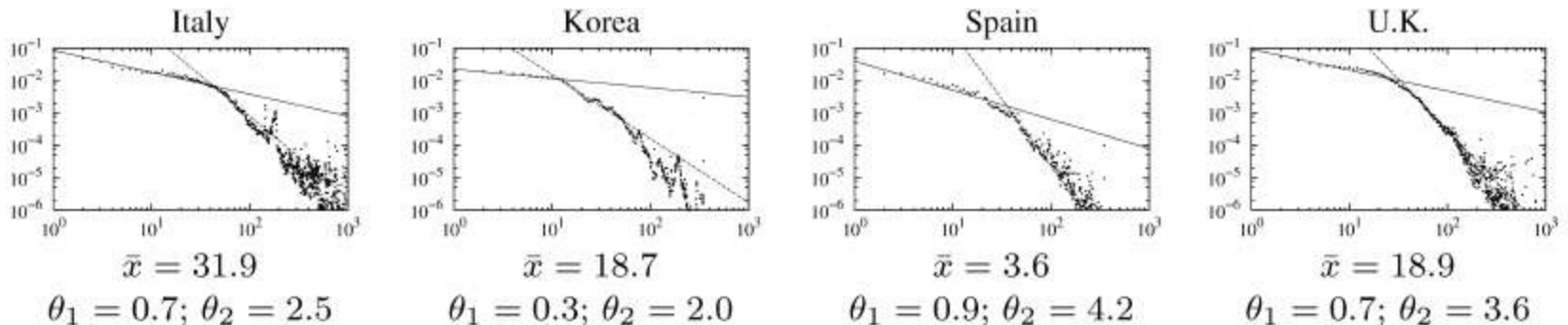
$\bar{x} = 248; \theta = 1.3$

## Pages per site

# In and Out Degree of Web Pages

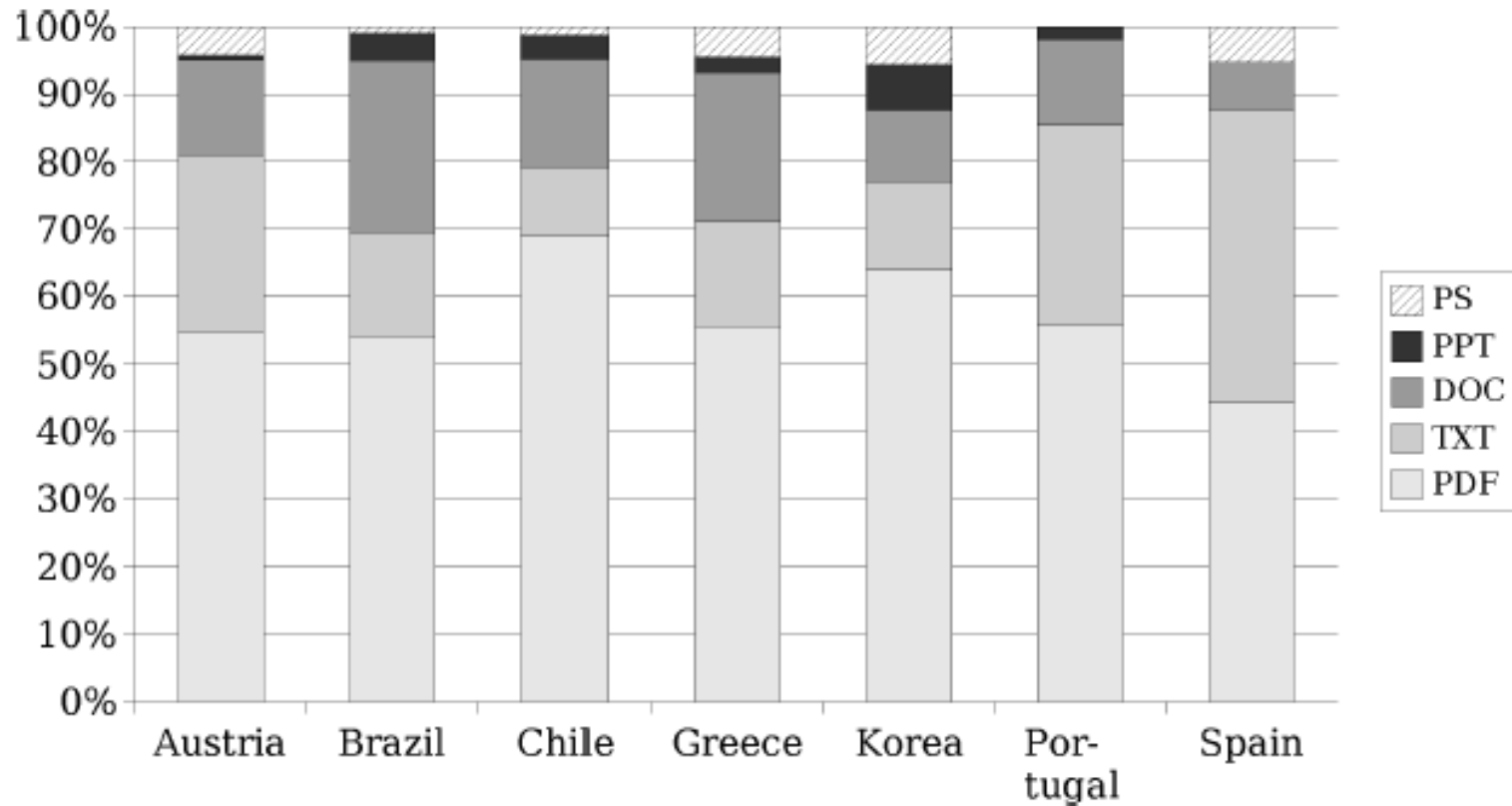


## In-degree of web pages



## Out-degree of web pages for few and many outlinks

# Non-HTML File Content



More than 95% of content was HTML

# How dynamic is the Web?

- How often are pages added to the Web?
- How often are pages removed from the Web?
- How often do pages change?
- What kinds of changes do pages typically exhibit?
- How does the link structure change over time?

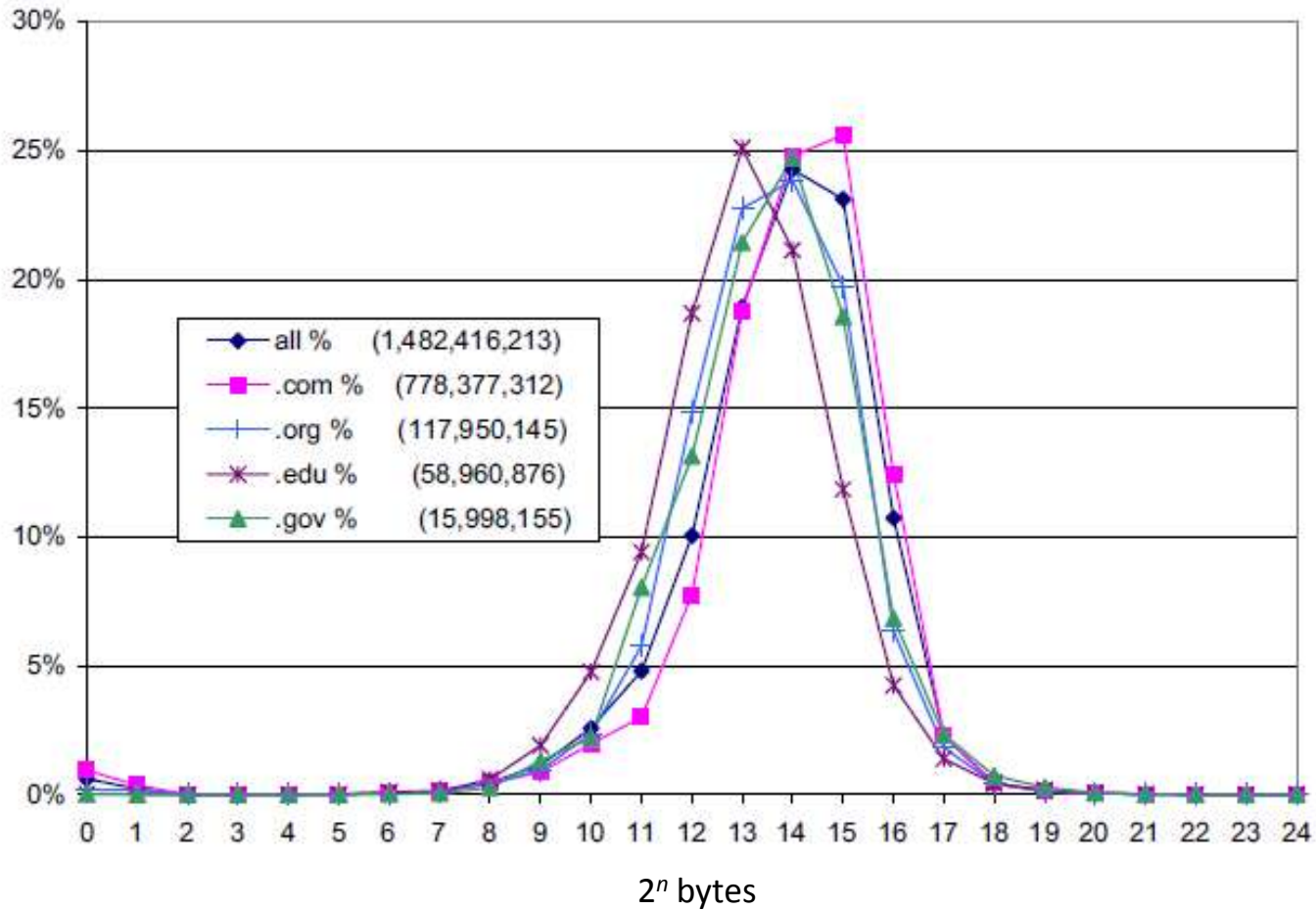
# How dynamic is the Web?

- Two studies attempt to answer these questions:
  - 2004 study (Fetterly et al.<sup>1</sup>) of 150 million web pages over 11 weeks analyzed weekly snapshots
  - 2004 study (Ntoulas et al.<sup>2</sup>) of 150 websites over one year analyzed weekly snapshots
- What follows are some selected highlights

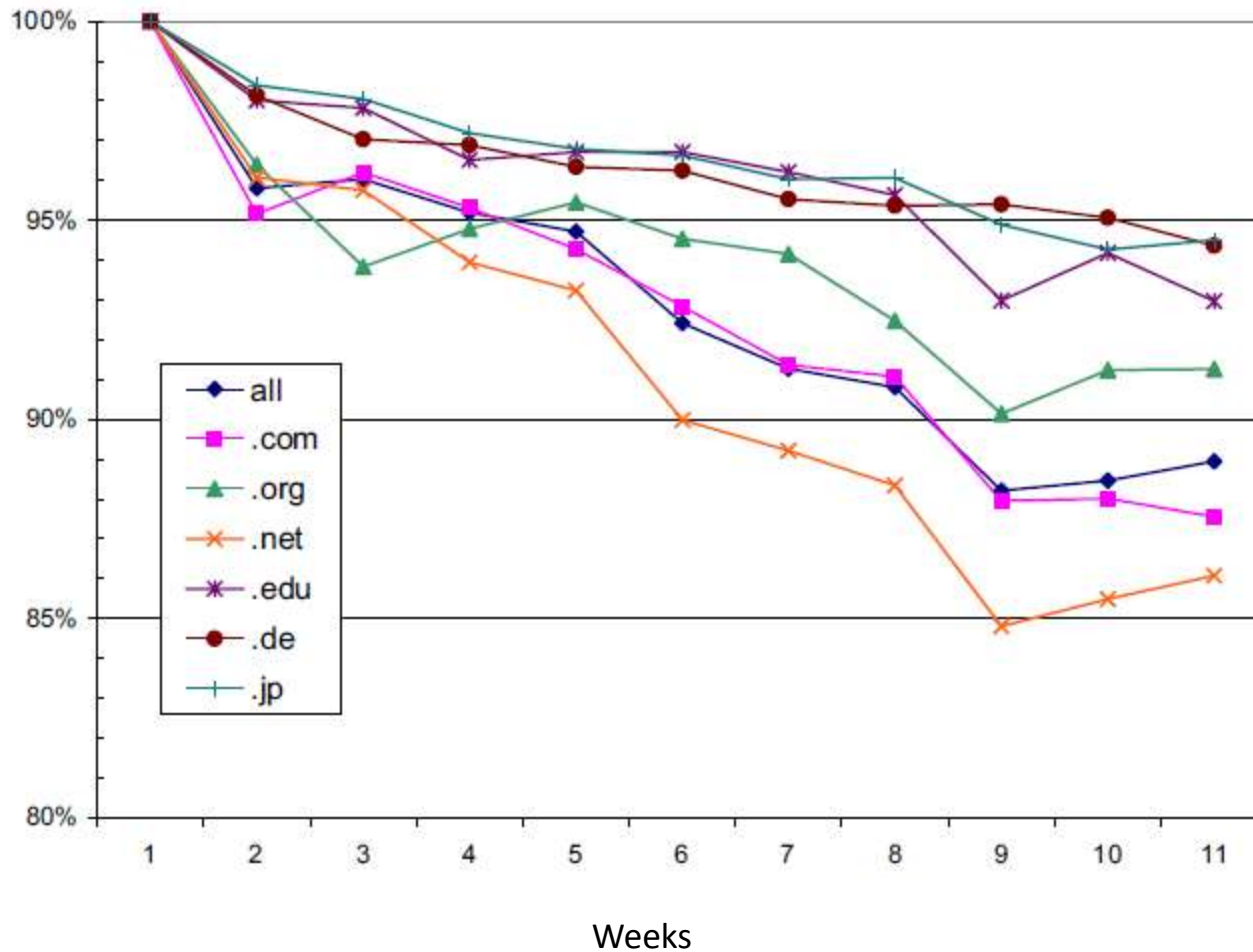
<sup>1</sup>Fetterly et al. , A large-scale study of the evolution of Web pages, *Software Practice & Experience* ,2004

<sup>2</sup>Ntoulas et al. , What's new on the web?: the evolution of the web from a search engine perspective, *Proc WWW 2004*

# Document Length



# Successful Downloads



# Rates of Change by TLD

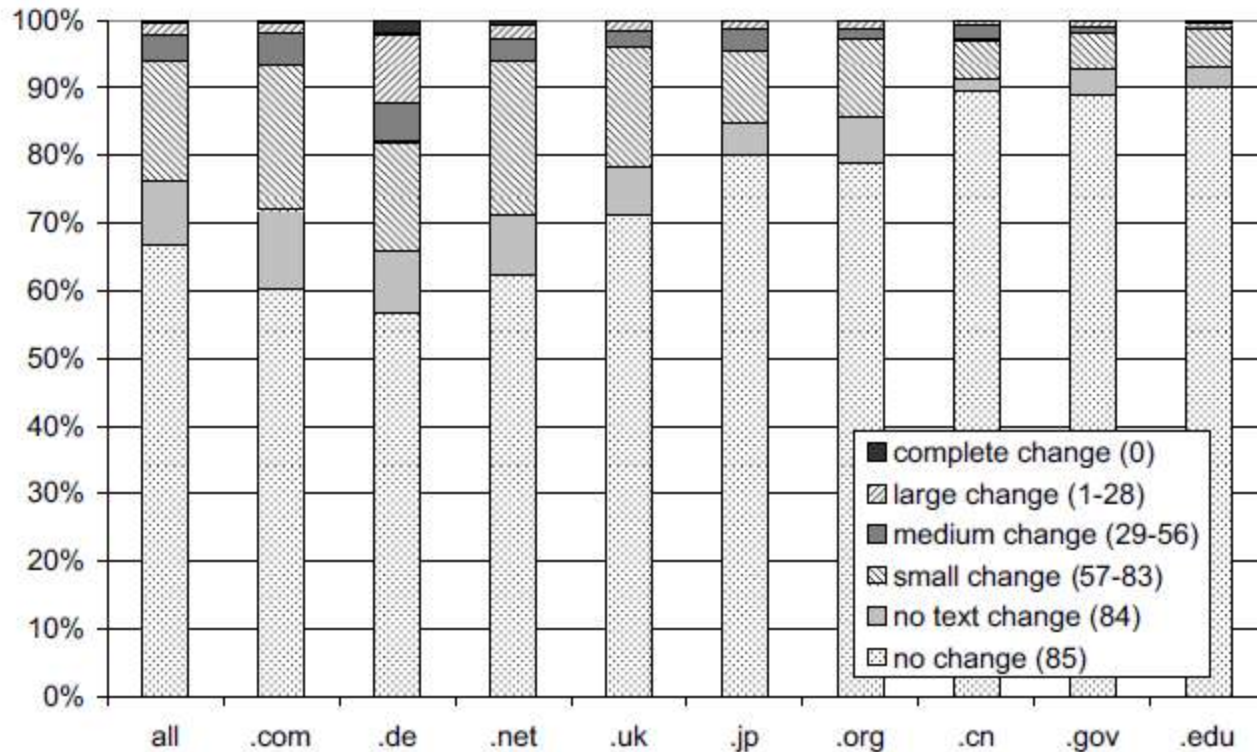
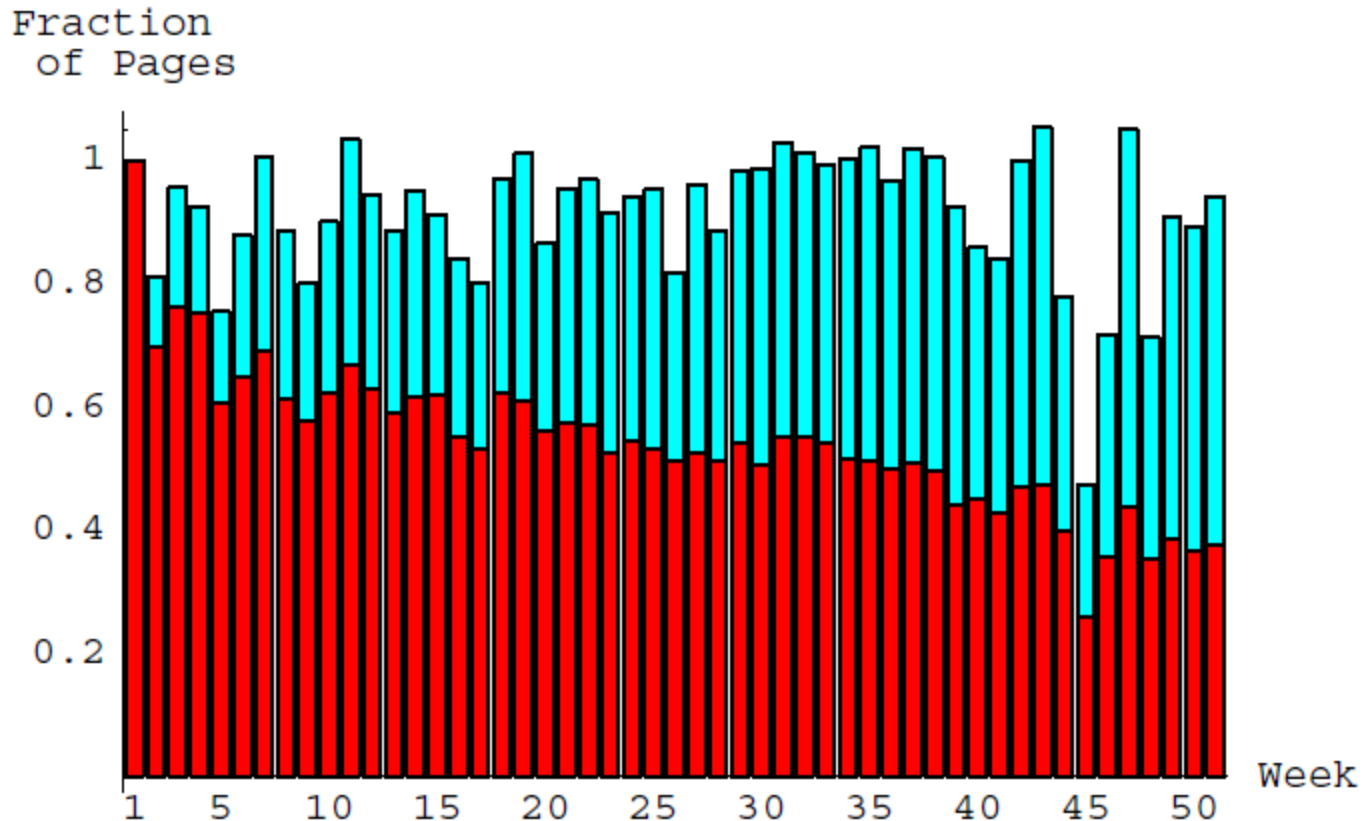


Figure 13. Clustered rates of change, broken down by selected top-level domains, after excluding automatically generated keyword-spam documents.

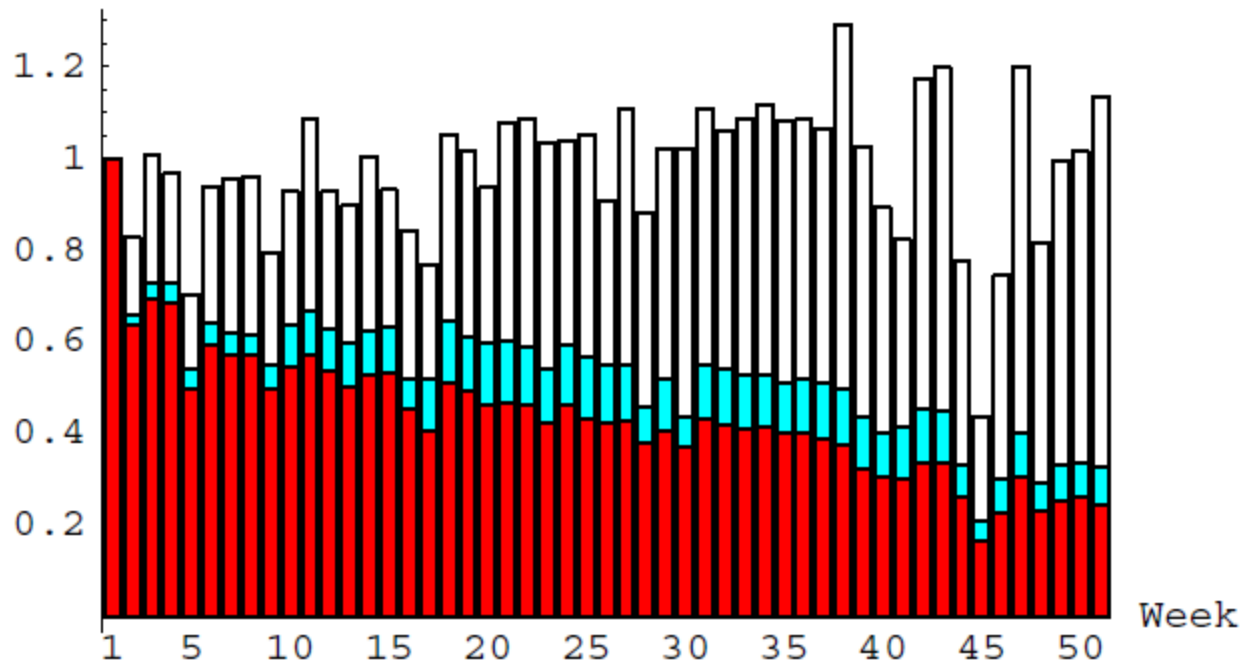
# New Pages



**Figure 2:** Fraction of pages from the first crawl still existing after  $n$  weeks (dark bars) and new pages (light bars).

# Link Evolution

Fraction of Links



**Figure 8:** Fraction of links from the first weekly snapshot still existing after  $n$  weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

# Summary of Findings

## Fetterly et al., 2004

- When pages change, they change in trivial ways or just their markup
- Strong relationship between TLD and *rate* of change but not *degree* of change
- The larger the document, the more likely it is to be changed more frequently and significantly
- Past frequency of changes to a page is good predictor of future page changes

## Ntoulas et al., 2004

- Web page changes are usually minor
- New pages are created at rate of 8% per week
- Only 20% of pages today will be accessible in a year
- Large number of pages borrow content from existing pages
- Every week, 25% new links are created, and after 1 year, 80% of links are replaced with new ones
- Past degree of change to web page is good predictor of future degree of change

# Linkrot: The 404 Problem

- Kahle ('97) - Average page lifetime is 44 days
- Koehler ('99, '04) - 67% URLs lost in 4 years
- Lawrence et al. ('01) - 23%-53% URLs in CiteSeer papers invalid over 5 year span (3% of invalid URLs “unfindable”)
- Spinellis ('03) - 27% URLs in CACM/Computer papers gone in 5 years
- Fetterly et al. ('03) – about 0.5% of web pages disappeared per week
- Ntoulas et al. ('04) – predicted only 20% of pages today will be accessible in a year
- McCown et al. ('05) - 10 year half-life for URLs in D-Lib Magazine articles
- Nelson & Allen ('02) - 3% objects in digital library gone in 1 year

# Blogosphere

**Weblogs Cumulative  
March 2003 - April 2006**

