Web Science: Measuring the Web

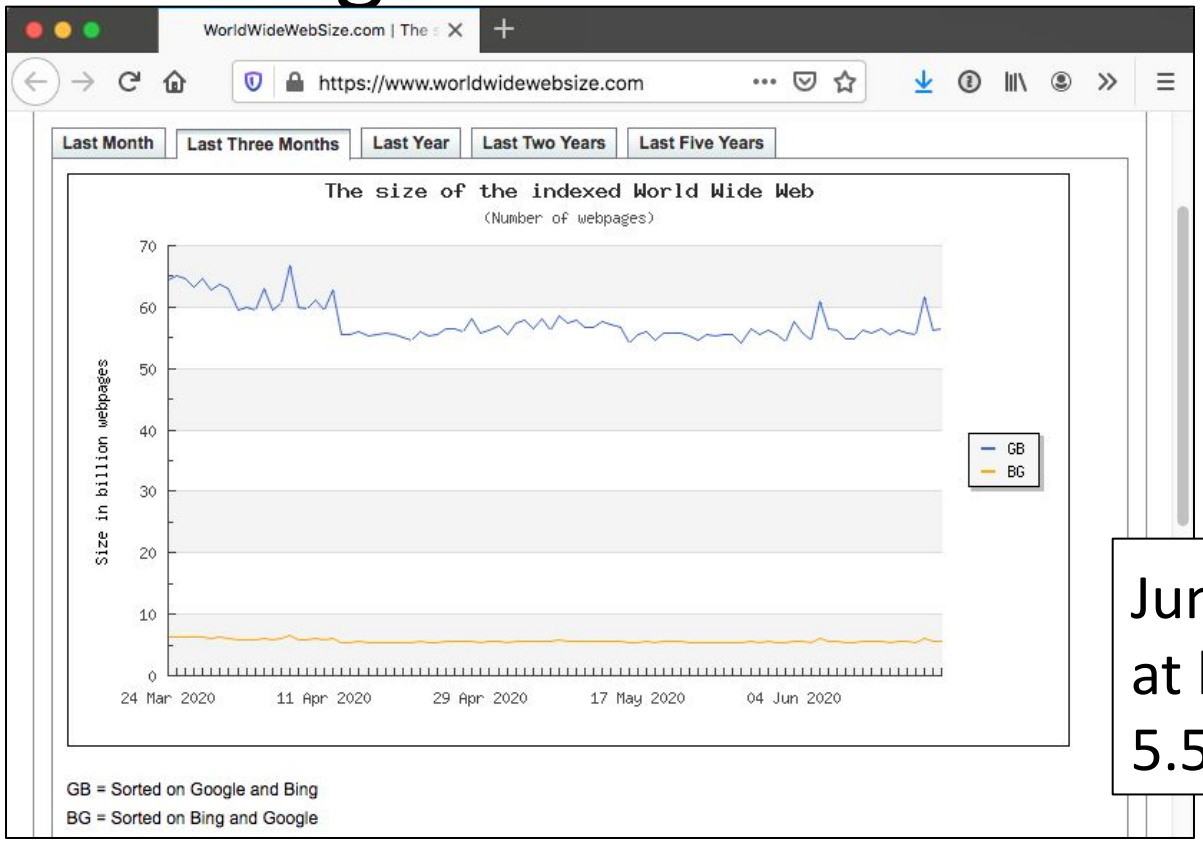
(Part 1 - How Big Is the Web and How Can We Tell?)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle

How Big Is the Web? (2020-06-22)



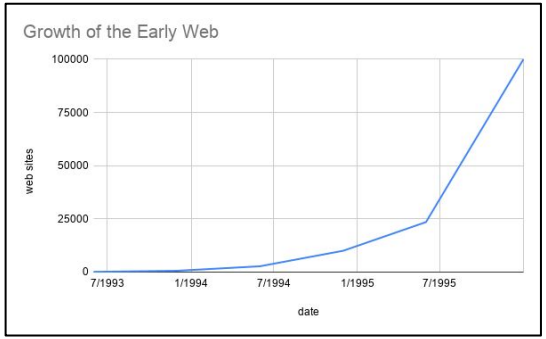
June 2020:
at least
5.56 B pages

source: [World Wide Web Size]

Measuring the Web - Initial Studies

- MIT Study (1993-1995)
  - How fast is the Web growing?
- W3C Characterization (1998-1999)
  - How many web pages are there? How fast is the Web growing?
- OCLC (1998-2002)
  - Analyzed Web samples annually to look for trends
- Baeza-Yates et al. (2000-2005)
  - Examined languages, file sizes, pages per site, link structure, etc. of national domains

MIT Study (1993-1995)



- Crawled the web June 1993 to June 1995
- Used World Wide Web Wanderer, the first automated Web agent or "spider"

Results Summary			
Month	**# of Web sites**	**% .com sites**	**Hosts per Web server**
6/93	130	1.5	13,000
12/93	623	4.6	3,475
6/94	2,738	13.5	1,095
12/94	10,022	18.3	451
6/95	23,500	31.3	270
1/96	100,000	50.0	94

"The growth of the Web has been remarkable even compared to the Internet at large"

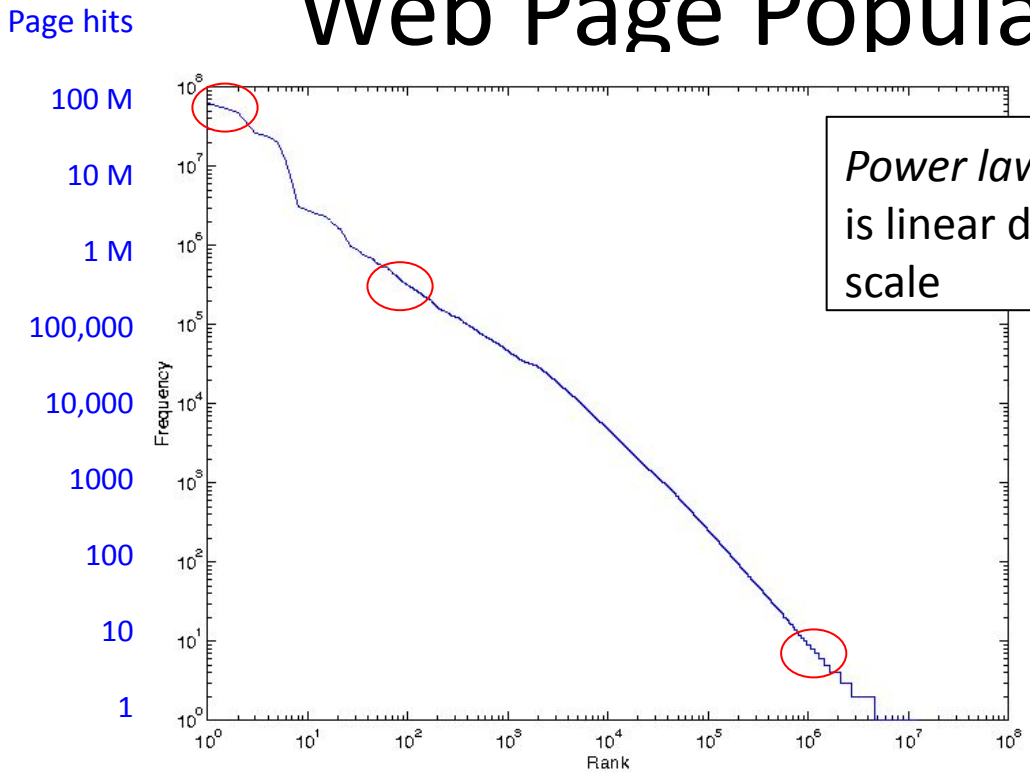
ref: "Measuring the Growth of the Web"

W3C Characterization Activity (1998-1999)

- Provided definitions for common Web terms like resource, link, proxy, server, etc., some of which are now dated (Web Characterization Terminology & Definitions)
- Attempted to answer questions like: How many web pages are there? How fast is the Web growing?

Summary: Pitkow, Summary of WWW Characterizations, *Journal of the World Wide Web*, 1999

Web Page Popularity



Power law distributions - tell-tale is linear distribution on a log-log scale

The Zipf distribution of number of page hits versus rank for five days of AOL December 1997 data

Summary of WWW Characterizations

Zipf's Law

- Originally formulated in study of linguistics
- The frequency of any word is inversely proportional to its rank in the frequency table

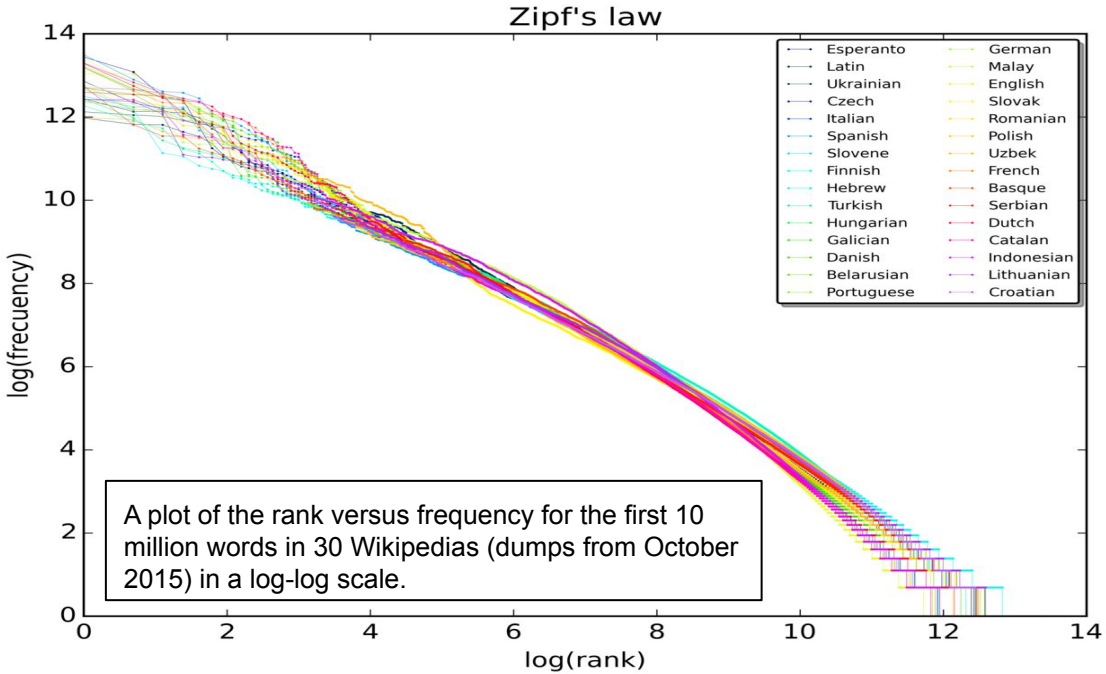
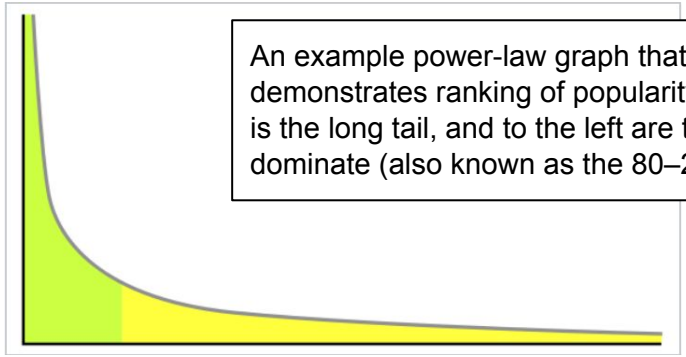


Image credit: By SergioJimenez - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45516736
Zipf's Law (wikipedia)

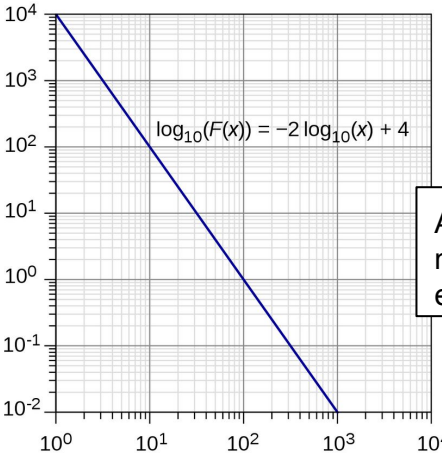
Power Law

- Relative change in one quantity results in a proportional change in the other quantity.
- One quantity varies as a *power*



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

Image credit: By User:Husky - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1449504
Power Law (wikipedia)

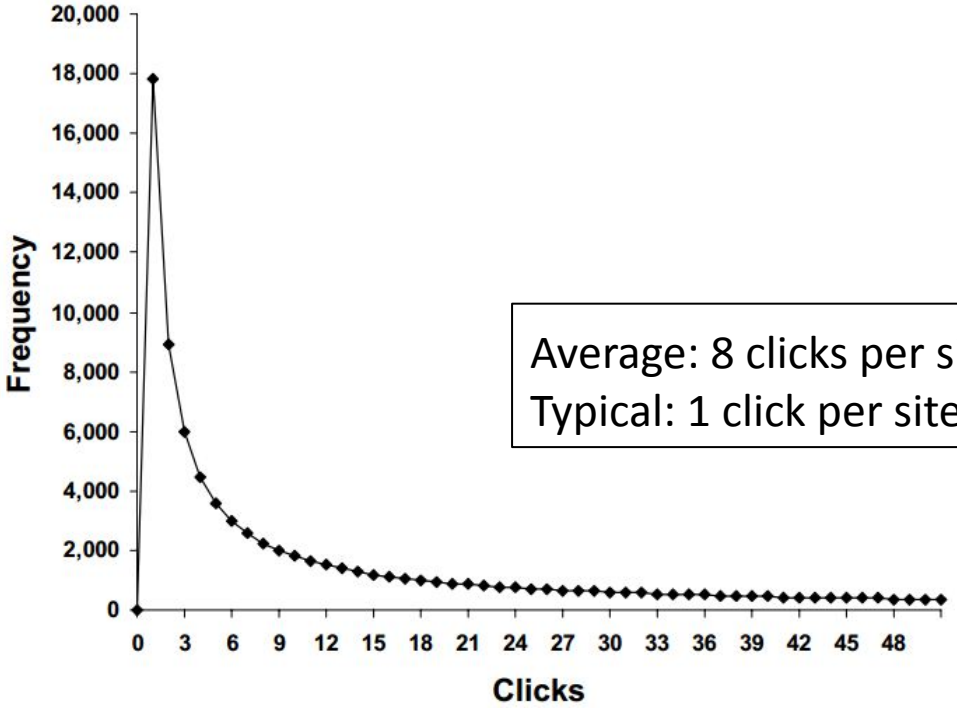


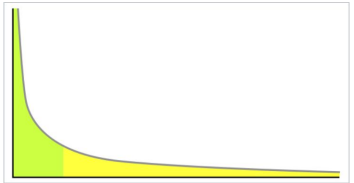
A straight line on a log–log plot is necessary but insufficient evidence for power-laws.

- A small number of items is clustered at the top of a distribution (or at the bottom), taking up 95% of the resources.

Image credit: By M. W. Toews - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=63281920

Web Page Popularity





Average: 8 clicks per site
Typical: 1 click per site

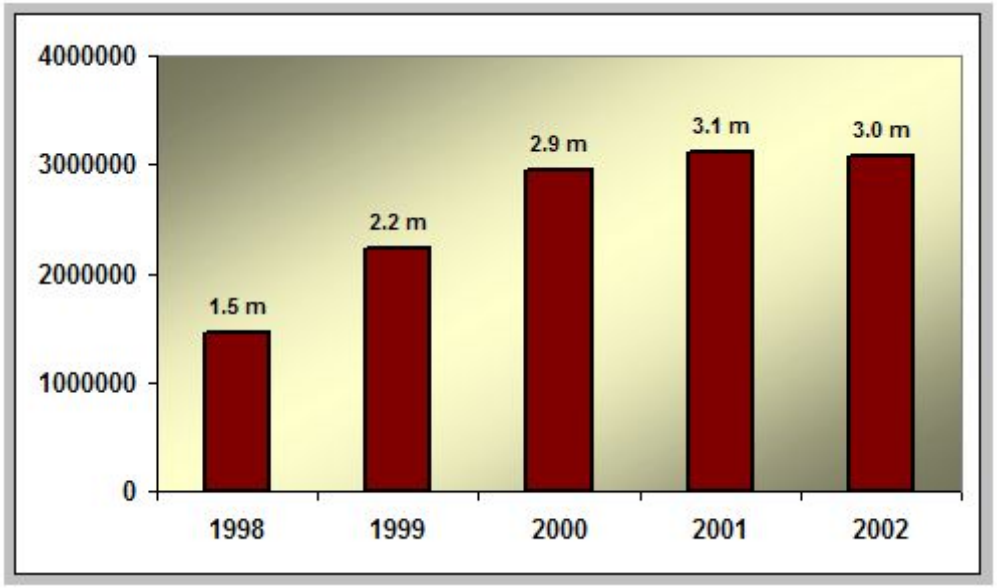
Summary of WWW Characterizations

The number of clicks per user at the Xerox WWW Site during May 1998.
The curve follows an inverse Gaussian Distribution, which has a heavy right tail.

OCLC Characterization Research (1998-2002)

- Work by Online Computer Library Center (OCLC)
- 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 (remember the "Host:" request header?)*
- Findings: O'Neill et al., Trends in the Evolution of the Public Web, *D-Lib Magazine*, Apr 2003

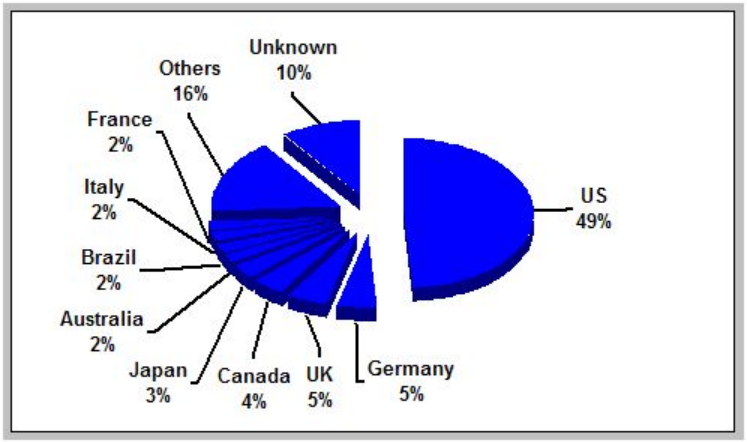
Number of Public Websites Doubled in Five Years



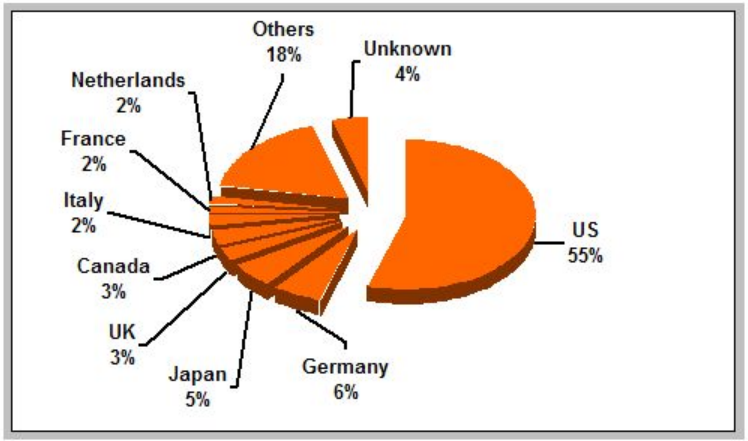
O'Neill et al., Trends in the Evolution of the Public Web, *D-Lib Magazine*, Apr 2003

Distribution of Websites by Country

(this is why you don't use pie charts!)



1999



2002

O'Neill et al., Trends in the Evolution of the Public Web, *D-Lib Magazine*, Apr 2003

Popular Websites by In-Links

OCLC Most Linked-To Websites[1]

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

(Jan 2013)[2]	(Jan 2020)[3]
1. facebook.com	1. google.com
2. twitter.com	2. apple.com
3. google.com	3. youtube.com
4. youtube.com	4. microsoft.com
5. adobe.com	5. play.google.com
6. wordpress.org	6. support.google.com
7. blogspot.com	7. blogger.com
8. wikipedia.org	8. docs.google.com
9. godaddy.com	9. adobe.com
10. wordpress.com	10. plus.google.com

[1] OCLC Top 50 most frequently linked-to sites (archived 2013)

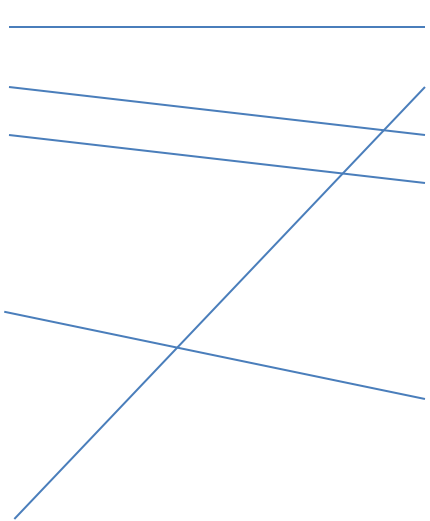
[2] Moz Top 500 Websites (archived Jan 2013)

[3] Moz Top 500 Most Popular Websites

Popular Websites by Visits

Alexa's Top Websites (Jan 2013)[1]

1. facebook.com
2. google.com
3. youtube.com
4. yahoo.com
5. baidu.com
6. wikipedia.org
7. live.com
8. amazon.com
9. qq.com
10. twitter.com



Most Linked-To Websites (Jan 2013)[2]

1. facebook.com
2. twitter.com
3. google.com
4. youtube.com
5. adobe.com
6. wordpress.org
7. blogspot.com
8. wikipedia.org
9. godaddy.com
10. wordpress.com

lots of links, but when is the last time you went here?

[1] Alexa Top 500 sites (archived 2013)

[2] SEOMoz Top 500 domains (archived 2013)

see also: "What is Alexa Traffic Rank?", "How are Alexa traffic rankings determined?"

Characterizing National Web Domains

- A large-scale study by Baeza-Yates et al.[1] analyzed web collections from 10 national domains and multinational Web spaces of African and Indochinese Web sites
  - *Indochina* - Cambodia (KH), Laos (LA), Myanmar (MM), Thailand (TH) and Vietnam (VN)
- Examined languages, file sizes, pages per site, link structure, etc.

[1]Baeza-Yates et al., Characterization of national Web domains, *ACM Trans. Internet Technol.,* May 2007

Distribution of Web Page Languages (English vs. Local Language)

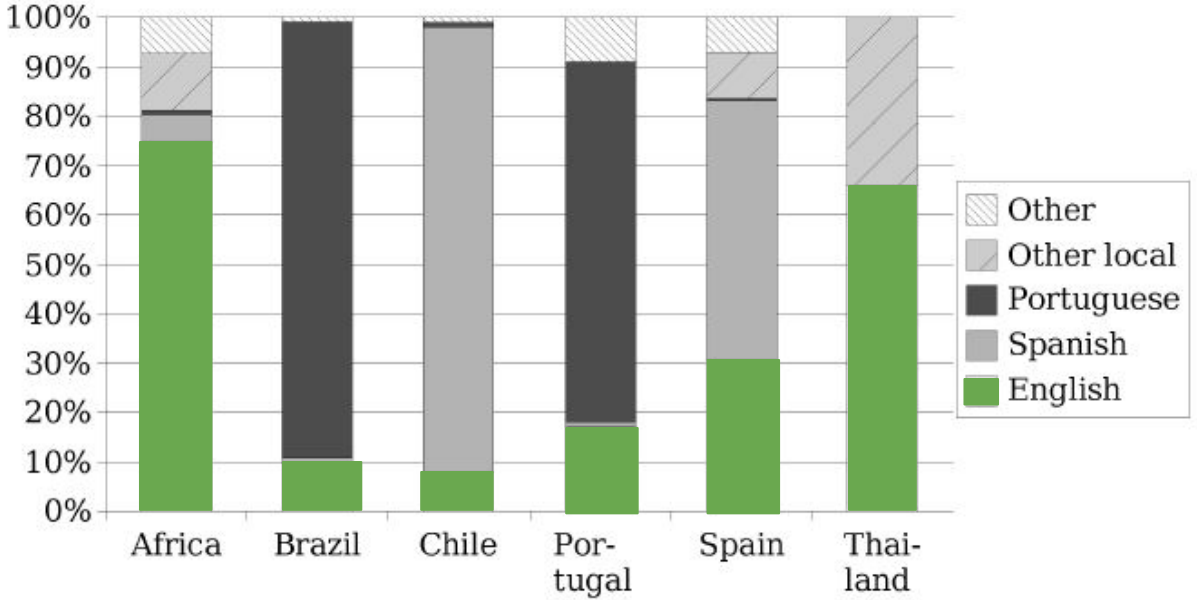
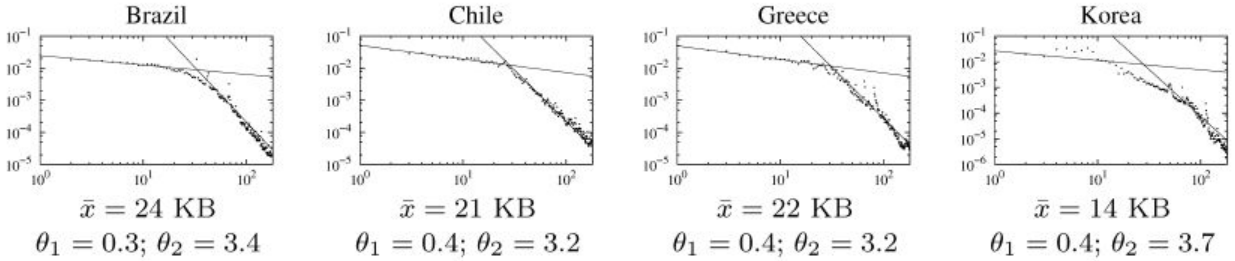


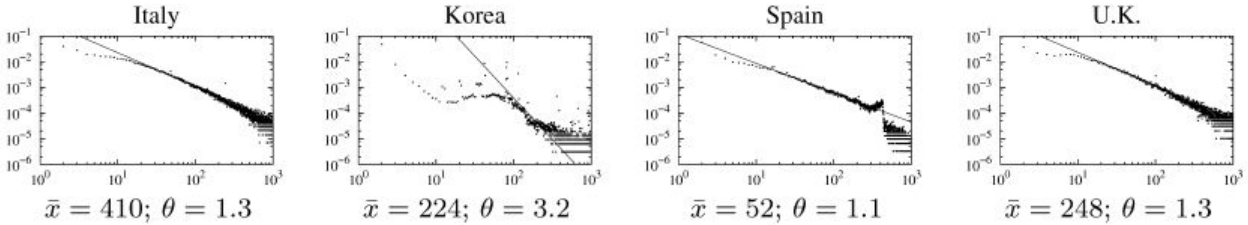
Fig. 2. Distribution of the number of pages in different languages.

Baeza-Yates et al., Characterization of national Web domains, *ACM Trans. Internet Technol.,* May 2007

Some Power-law Distributions



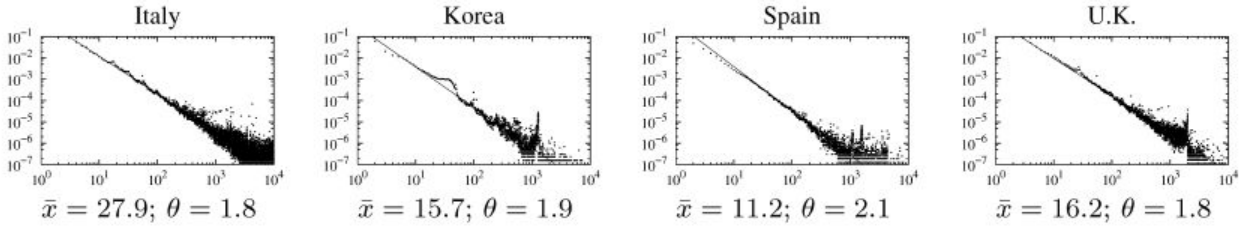
File sizes for small and large files



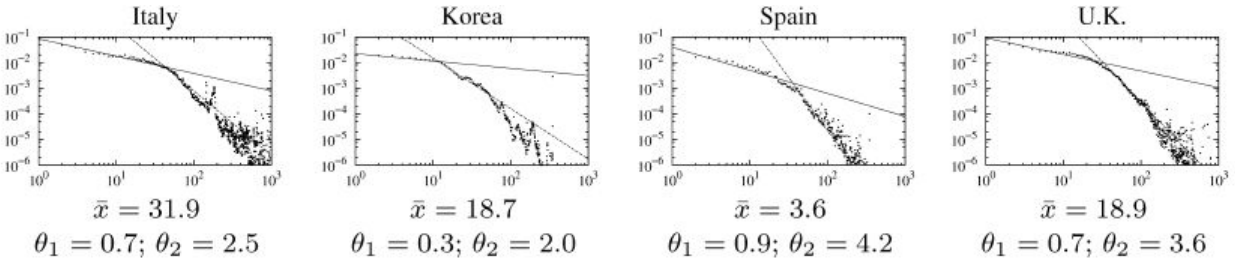
Pages per site

Baeza-Yates et al., Characterization of national Web domains, *ACM Trans. Internet Technol.,* May 2007

In and Out Degree of Web Pages



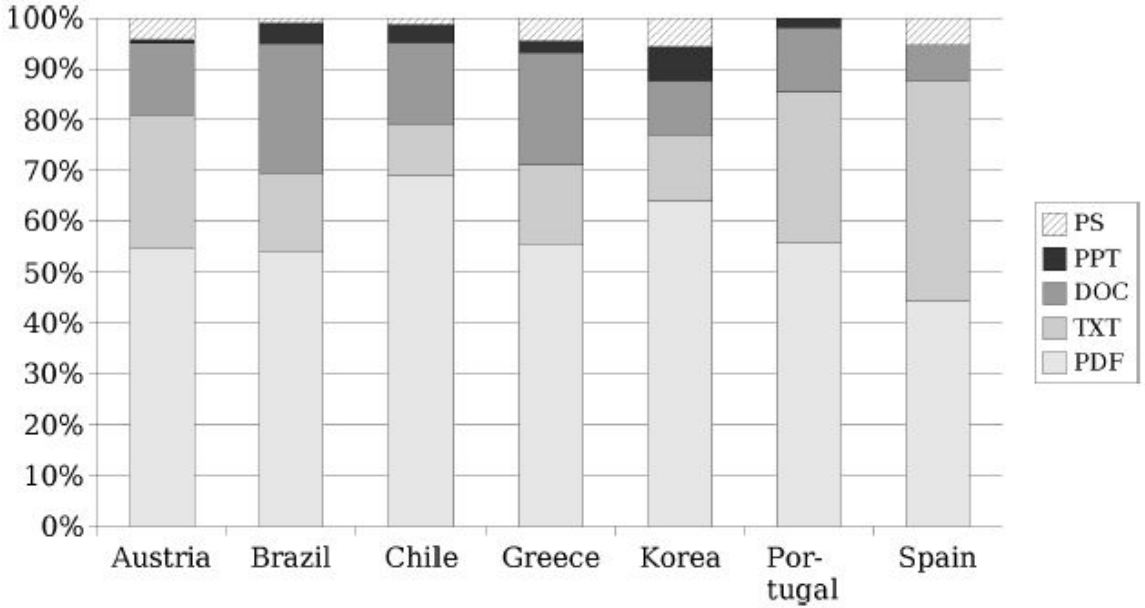
In-degree of web pages



Out-degree of web pages for few and many outlinks

Baeza-Yates et al., Characterization of national Web domains, *ACM Trans. Internet Technol.,* May 2007

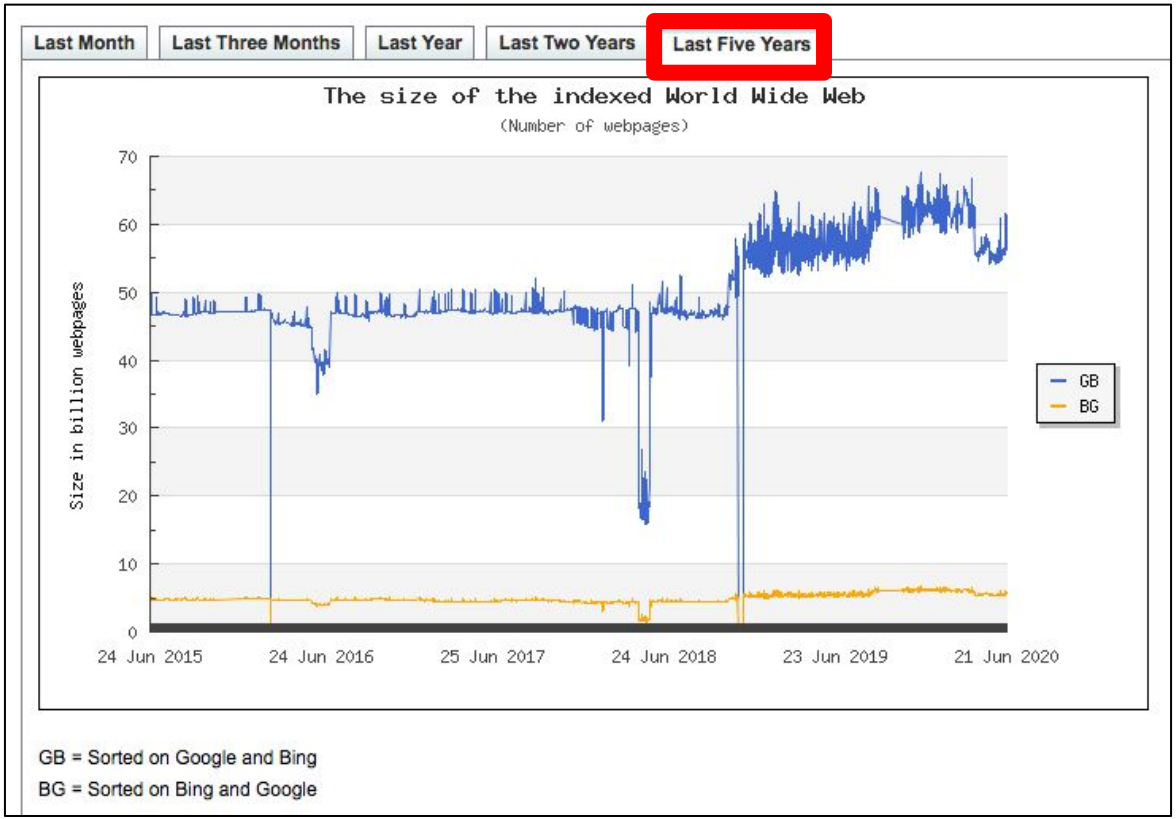
Non-HTML File Content



More than 95% of content was HTML

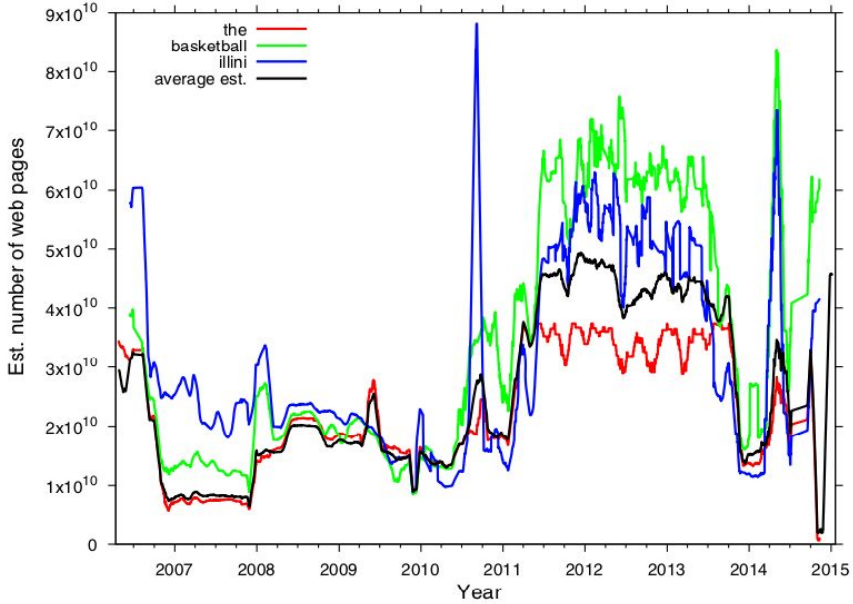
Baeza-Yates et al., Characterization of national Web domains, *ACM Trans. Internet Technol.,* May 2007

How Has the Web Grown?



ref: [World Wide Web Size]

Estimating the Size of the Web



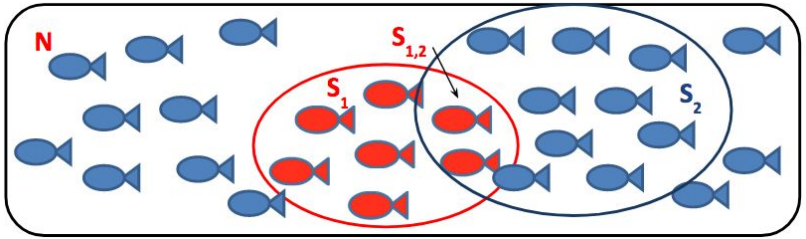
ref: World Wide Web Size

Fig. 3 Estimated size of the Google index from March 2006 to January 2015 for three pivot words, *the*, *basketball*, and *illini*, and the average estimate over all 28 words (*black line*). The *lines* connect the unweighted running daily averages of 31 days

van den Bosch et al., "Estimating search engine index size variability: a 9-year longitudinal study", *Scientometrics*, 2016

Estimate Web Population

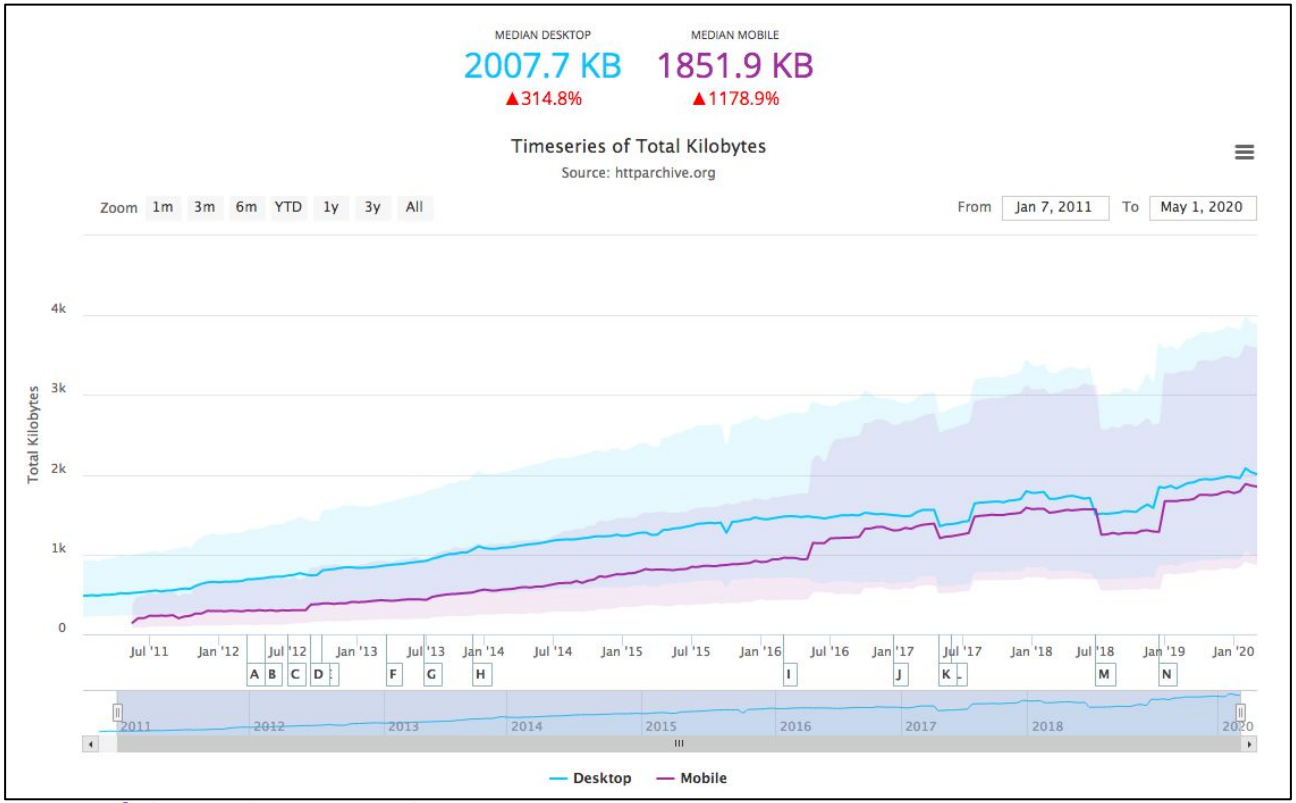
- Lawrence and Giles (1998) used capture-recapture method to estimate web page population
  - Submitted 575 queries to sets of 2 search engines
  - S1 = All pages returned by SE1
  - S2 = All pages returned by SE2
  - S1,2 = All pages returned by both SE1 and SE2
  - Size of indexable Web (N) = $S_1 \times S_2 / S_{1,2}$



- Estimated size of indexable Web in 1998 = 320 million pages

Lawrence & Giles, Searching the World Wide Web, *Science*, 1998

HTTPArchive State of the Web



refs: [HTTPArchive's State of the Web](), [HTTPArchive FAQ]()

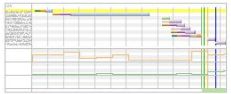
WebPageTest

	First Byte	Start Render	First Contentful Paint	Speed Index	Last Painted Hero	Web Vitals: Largest Contentful Paint	Web Vitals: Cumulative Layout Shift	Document Complete: Time	Document Complete: Requests	Document Complete: Bytes In	Fully Loaded: Time	Fully Loaded: Requests	Fully Loaded: Bytes In	Fully Loaded: Cost
First View (Run 2)	0.742s	1.100s	1.106s	1.100s	1.100s	1.106s	0	1.326s	9	52 KB	1.394s	10	53 KB	$----

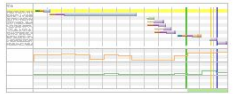
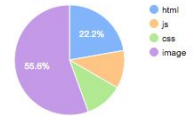
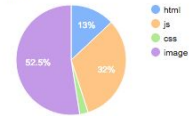
Plot Full Results

Test Results

Run 1:

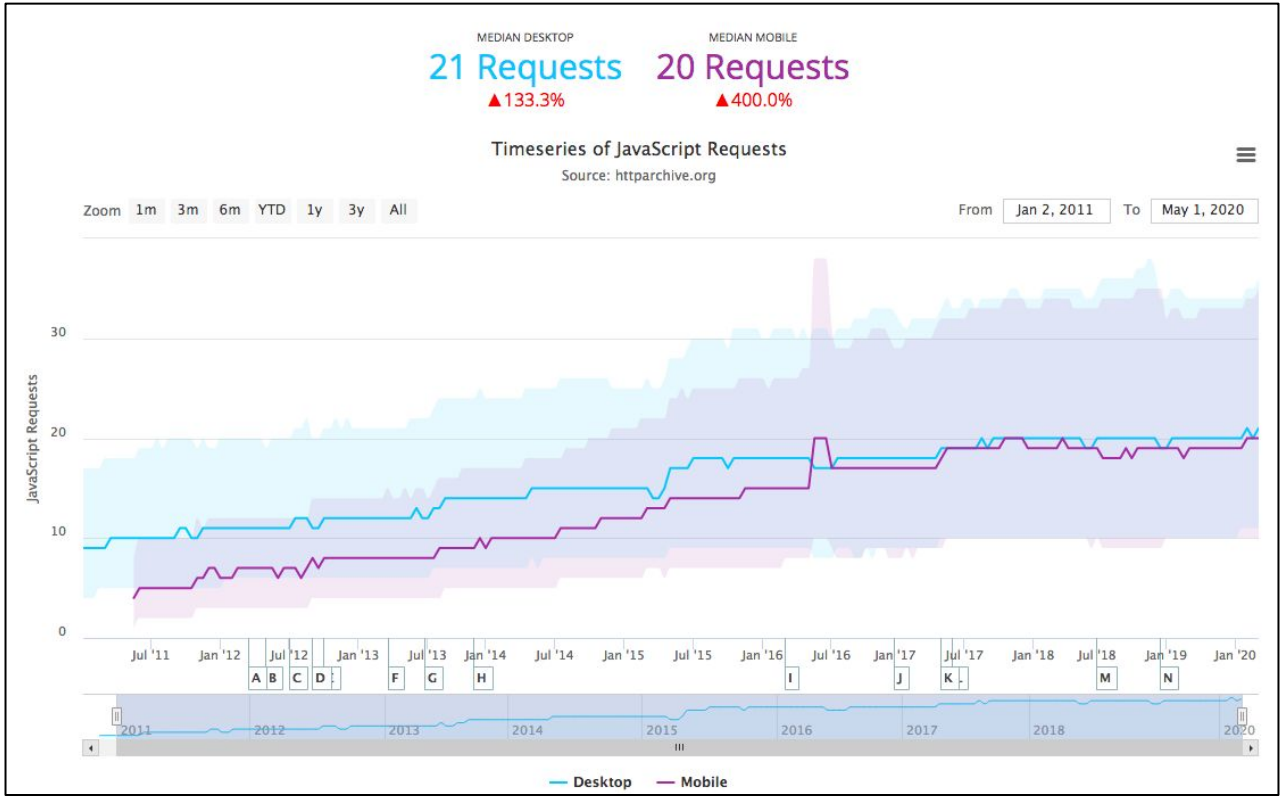
	Waterfall	Screenshot
First View (1.169s)		

Run 2:

	Waterfall	Screenshot
First View (1.326s)		
Content Breakdown		

source: [WebPageTest]

HTTPArchive State of JavaScript



ref: [HTTPArchive's State of JavaScript]

Web Science:
Measuring the Web
(Part 2 - How Dynamic Is the Web?)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle

How dynamic is the Web?

- How often are pages added to the Web?
- How often are pages are 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?

- We focus on two studies that attempt to answer these questions:
  - 2004 study (Fetterly et al.[1]) of 150 million web pages over 11 weeks analyzed weekly snapshots
  - 2004 study (Ntoulas et al.[2]) of 150 websites over one year analyzed weekly snapshots

- What follows are some selected highlights

[1]Fetterly et al., A large-scale study of the evolution of Web pages, *Software Practice & Experience,* 2004
[2]Ntoulas et al., What's new on the web?: the evolution of the web from a search engine perspective, *Proc WWW 2004*

Document Length

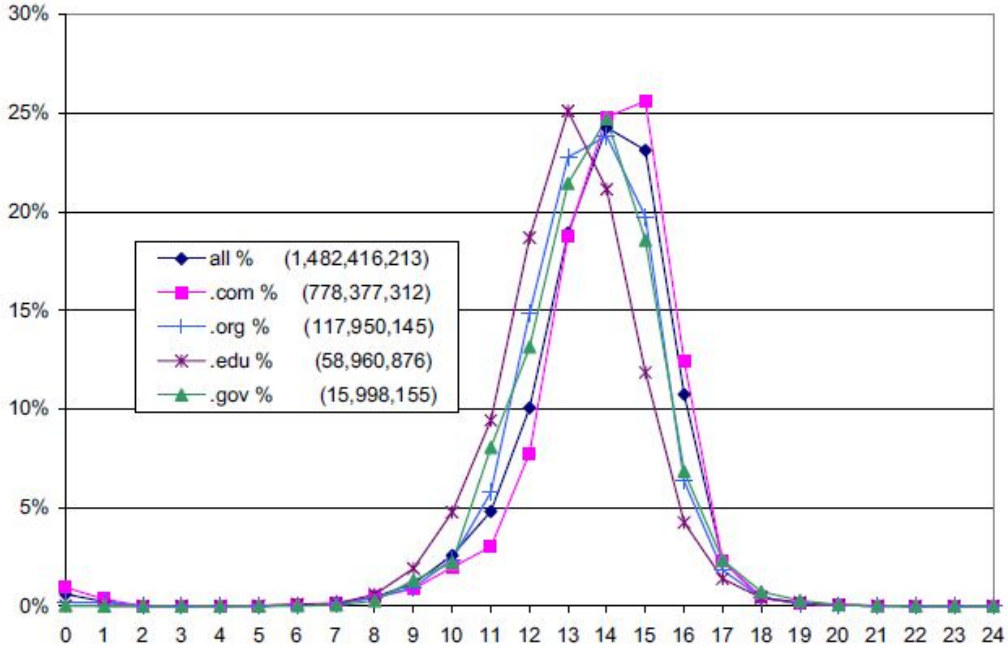


Figure 2: Distribution of documents lengths overall and for selected top-level domains.

Fetterly et al., 2004

Successful Downloads

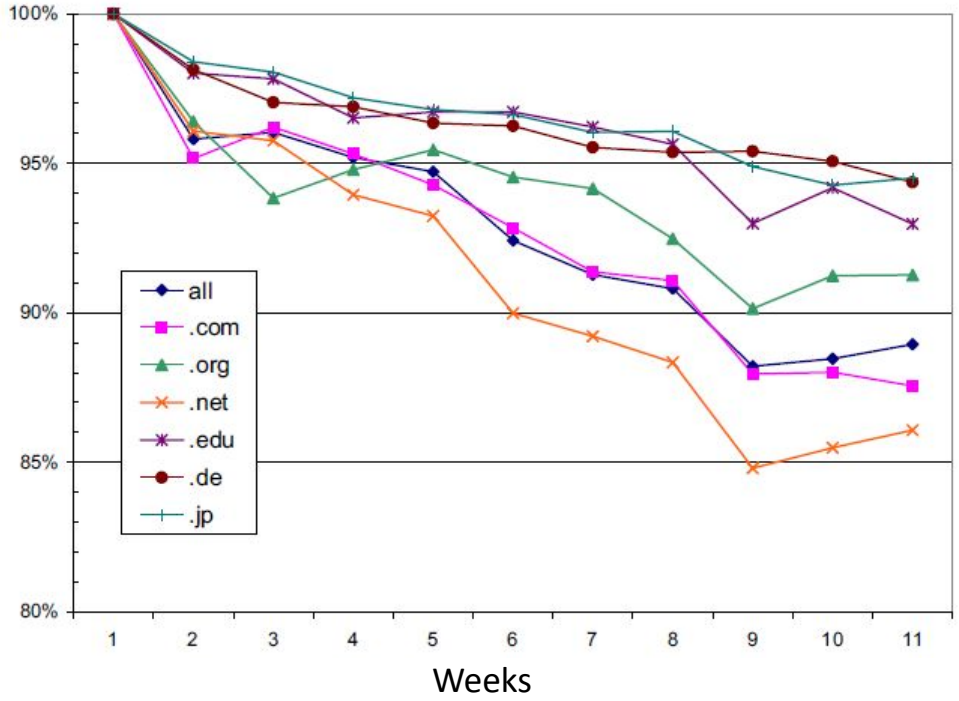


Figure 5: Distribution of successful downloads over crawl generations, broken down by selected top-level domains.

Fetterly et al., 2004

Rates of Change by TLD

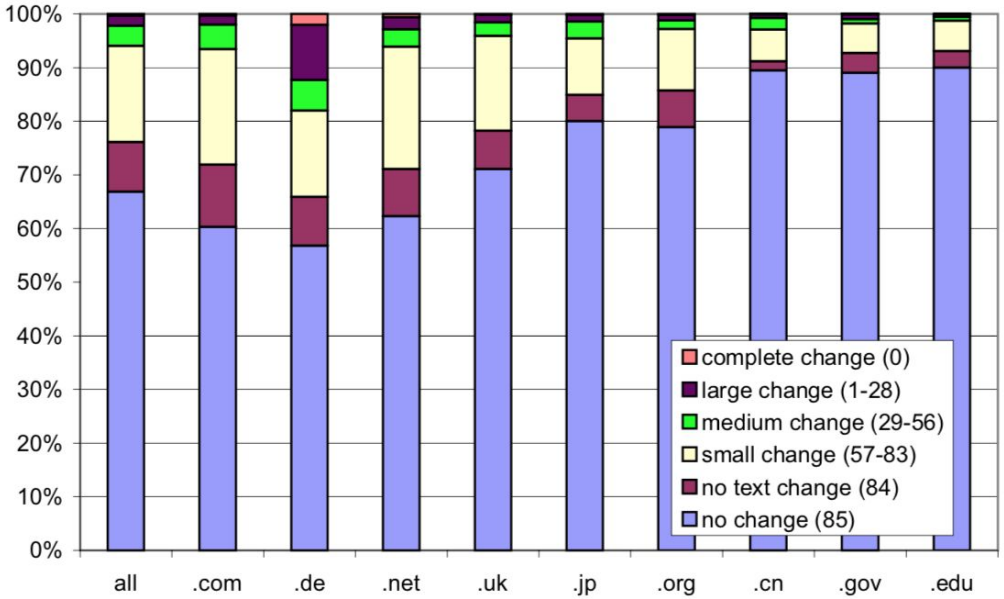


Figure 11: Clustered rates of change, broken down by selected top-level domains, after excluding automatically generated keyword-spam documents.

Fetterly et al., 2004

New Pages

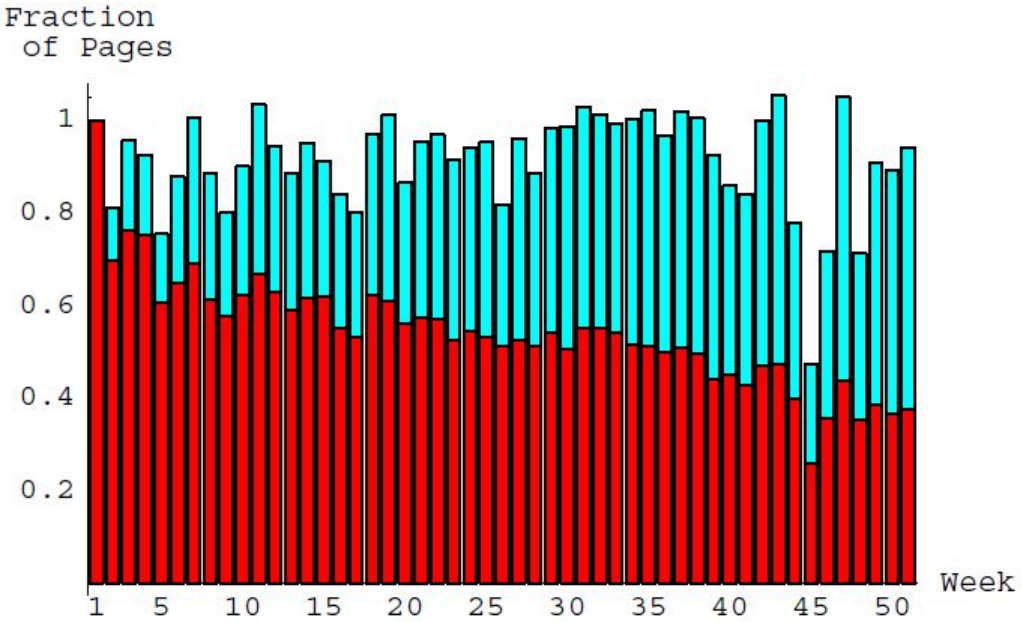


Figure 2: Fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars).

Ntoulas et al. , 2004

Link Evolution

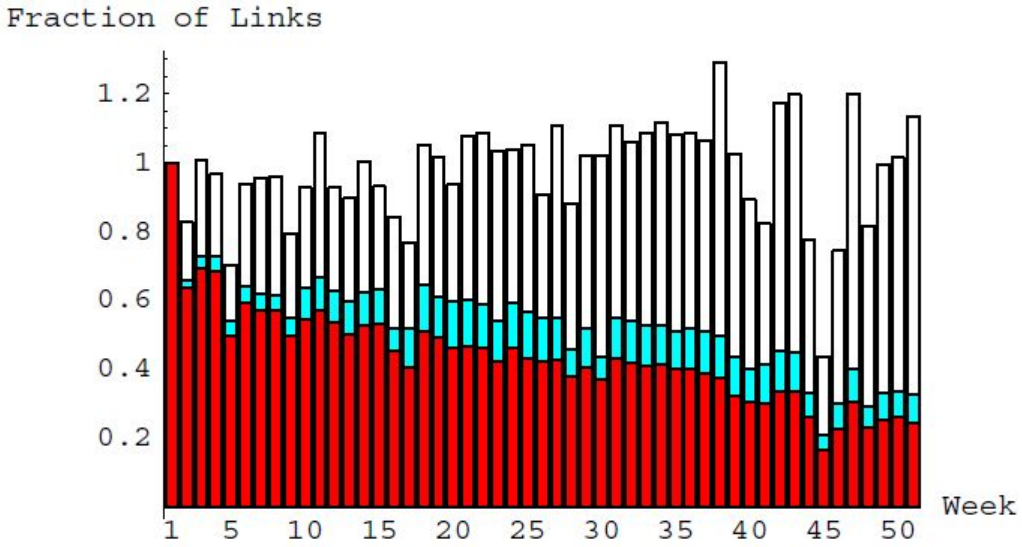


Figure 8: Fraction of links from the first weekly snapshot still existing after n weeks (dark/bottom portion of the bars), new links from existing pages (grey/middle) and new links from new pages (white/top).

Ntoulas et al. , 2004

Summary of Findings

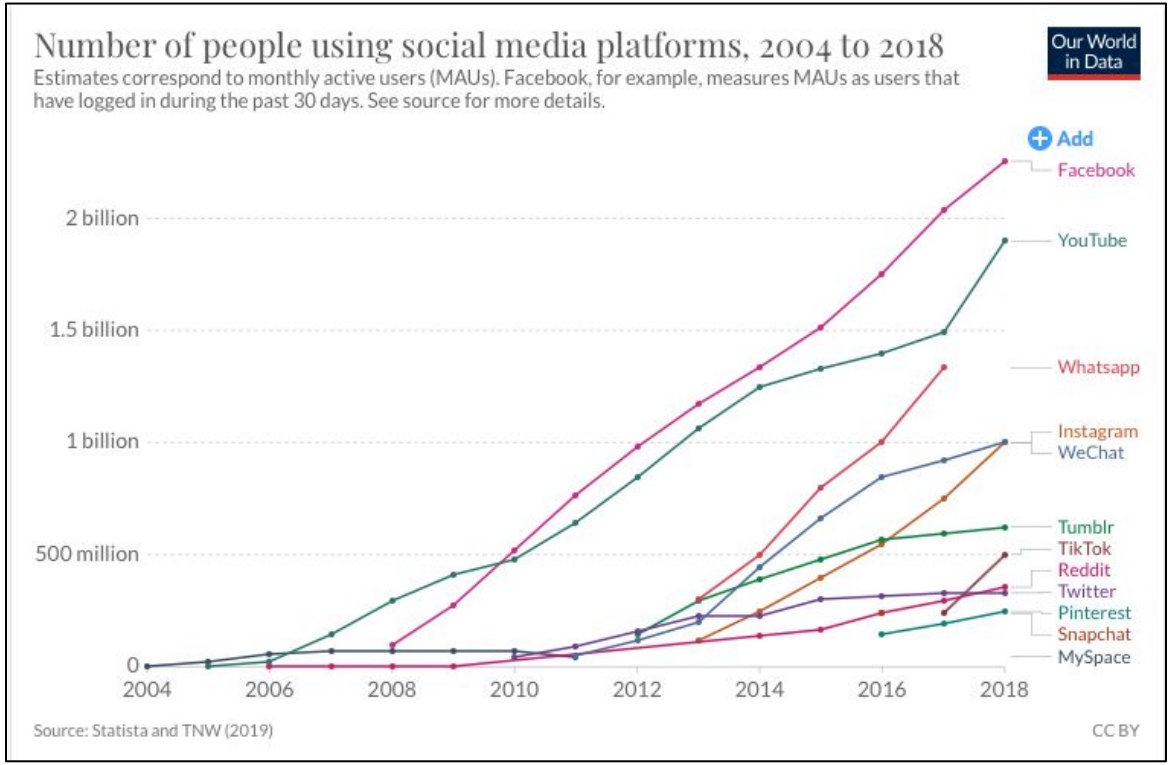
Fetterly et al., 2004

- 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*

Rise of Social Media



ref: [Internet stats from Our World in Data]

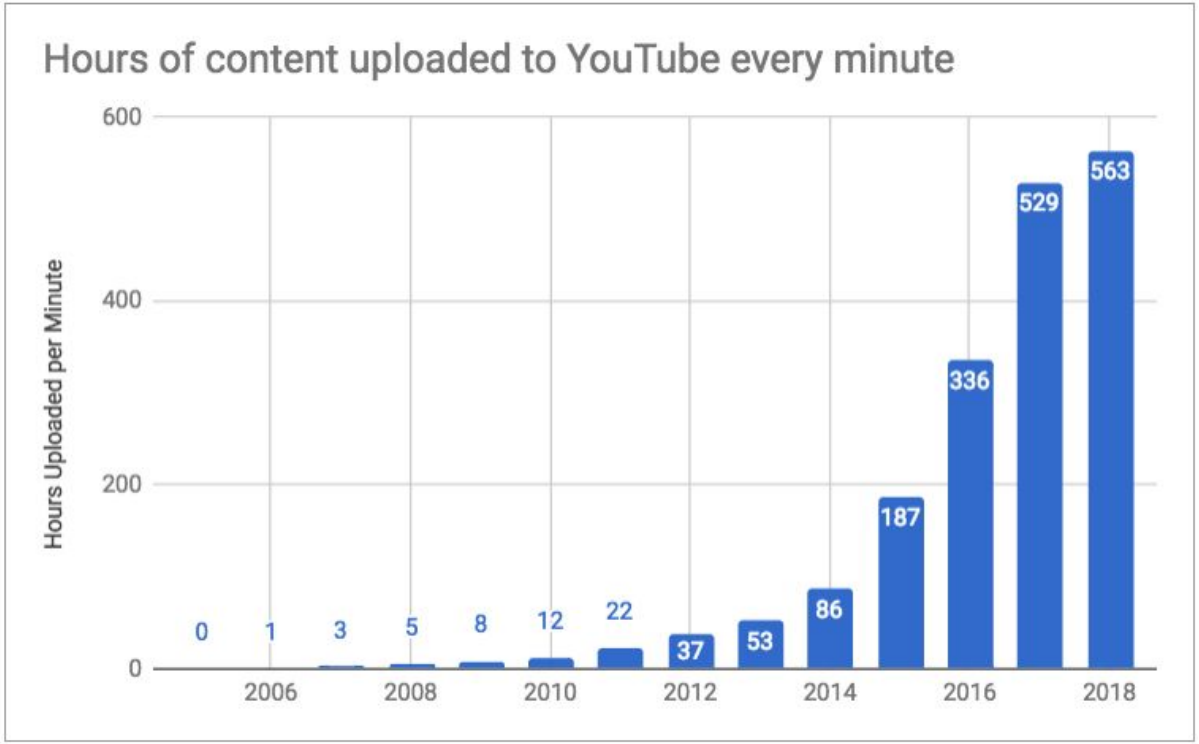
Today, we add lots of content, mainly through social media

- Every day
  - Facebook: more than 300 million photos uploaded
  - Instagram: 95 million photos and videos shared

- Every *minute* of the day
  - Snapchat users: share 527,760 photos
  - Twitter: 456,000 tweets are sent
  - Instagram users: post 46,740 photos
  - Facebook: 510,000 comments posted
  - Wikipedia: 600 new page edits

"How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read" (stats from 2017-2019)

YouTube

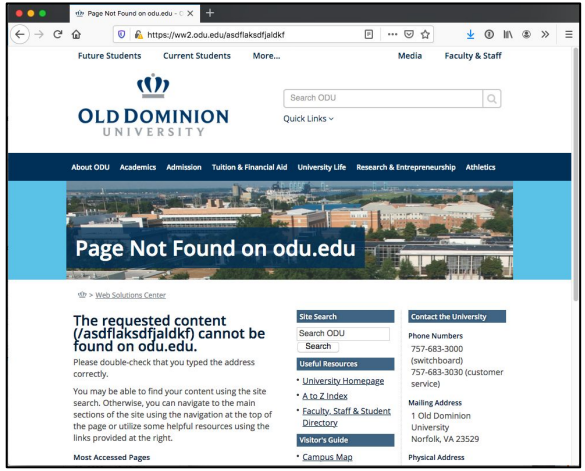


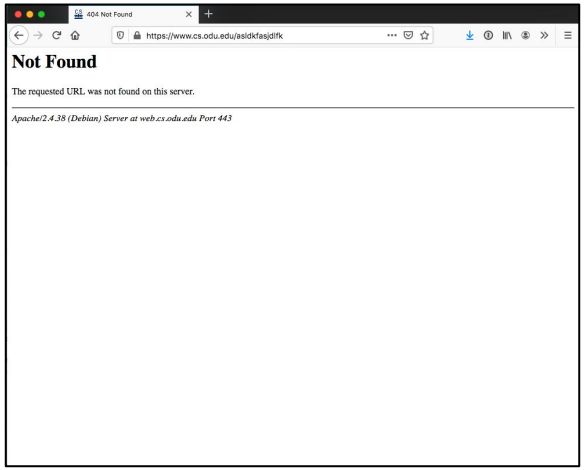
ref: [State of the YouTube Address — an overview of YouTube usage and growth]

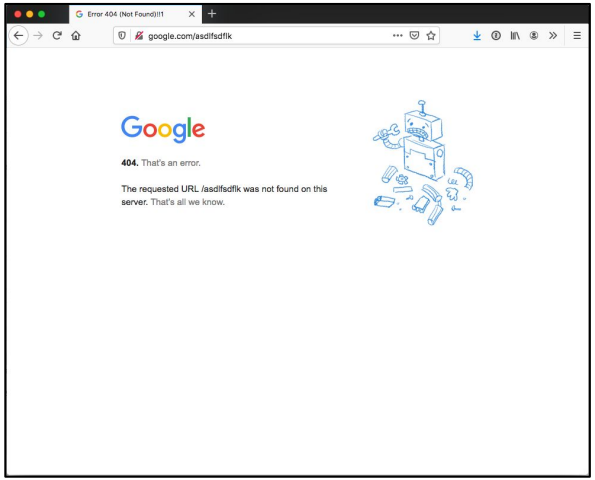
Links and Pages Can Still Disappear

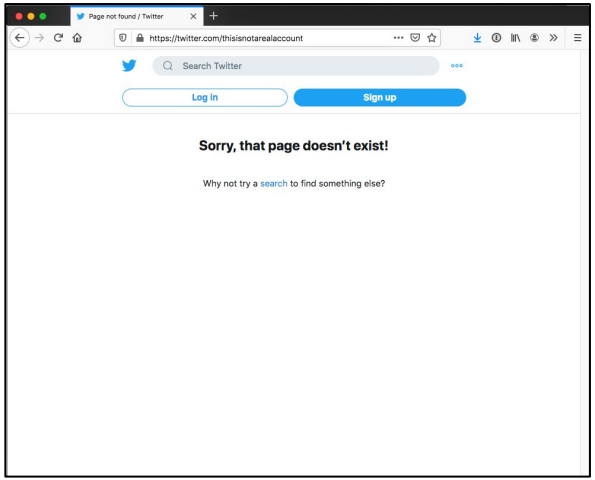
Link Rot

- 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
- SalahEldeen & Nelson ('12) – 11% of URLs from Tweets disappear after 1 year


Page Not Found on odu.edu
https://ww2.odu.edu/asdflaksdfjaldkf
Future Students
Current Students
More...
Media
Faculty & Staff
OLD DOMINION UNIVERSITY
Search ODU
Quick Links
About ODU
Academics
Admission
Tuition & Financial Aid
University Life
Research & Entrepreneurship
Athletics
Page Not Found on odu.edu
Web Solutions Center
The requested content (/asdflaksdfjaldkf) cannot be found on odu.edu.
Please double-check that you typed the address correctly.
You may be able to find your content using the site search. Otherwise, you can navigate to the main sections of the site using the navigation at the top of the page or utilize some helpful resources using the links provided at the right.
Most Accessed Pages
Site Search
Search ODU
Search
Useful Resources
University Homepage
A to Z Index
Faculty, Staff & Student Directory
Visitor's Guide
Campus Map
Contact the University
Phone Numbers
757-683-3000 (switchboard)
757-683-3030 (customer service)
Mailing Address
1 Old Dominion University
Norfolk, VA 23529
Physical Address


404 Not Found
https://www.cs.odu.edu/asldkfasjdlfk
Not Found
The requested URL was not found on this server.
Apache/2.4.38 (Debian) Server at web.cs.odu.edu Port 443


Error 404 (Not Found)!!1
google.com/asdlfsdflk
Google
404. That's an error.
The requested URL /asdlfsdflk was not found on this server. That's all we know.


Page not found / Twitter
https://twitter.com/thisisnotarealaccount
Search Twitter
Log in
Sign up
Sorry, that page doesn't exist!
Why not try a search to find something else?

```
% curl -Ik https://ww2.odu.edu/asdflaksdfjaldkf
HTTP/1.1 404 Not Found
Date: Mon, 29 Jun 2020 17:54:43 GMT
Server: Apache
Vary: Accept-Encoding,User-Agent
X-XSS-Protection: 1; mode=block
Content-Type: text/html
```

```
% curl -Ik https://www.cs.odu.edu/asldkfasjdlfk
HTTP/1.1 404 Not Found
Server: nginx
Date: Mon, 29 Jun 2020 17:54:48 GMT
Content-Type: text/html; charset=iso-8859-1
Connection: keep-alive
```

```
% curl -I http://google.com/asdlfsdflk
HTTP/1.1 404 Not Found
Content-Type: text/html; charset=UTF-8
Referrer-Policy: no-referrer
Content-Length: 1571
Date: Mon, 29 Jun 2020 17:54:04 GMT
```

```
curl -I https://twitter.com/thisisnotarealaccount
HTTP/2 200
content-type: text/html; charset=utf-8
date: Mon, 29 Jun 2020 17:54:13 GMT
expiry: Tue, 31 Mar 1981 05:00:00 GMT
last-modified: Mon, 29 Jun 2020 17:54:13 GMT
pragma: no-cache
server: tsa_b
strict-transport-security: max-age=631138519
vary: Accept-Encoding
x-content-type-options: nosniff
x-frame-options: DENY
x-powered-by: Express
x-response-time: 48
x-xss-protection: 0
```

Content Drift

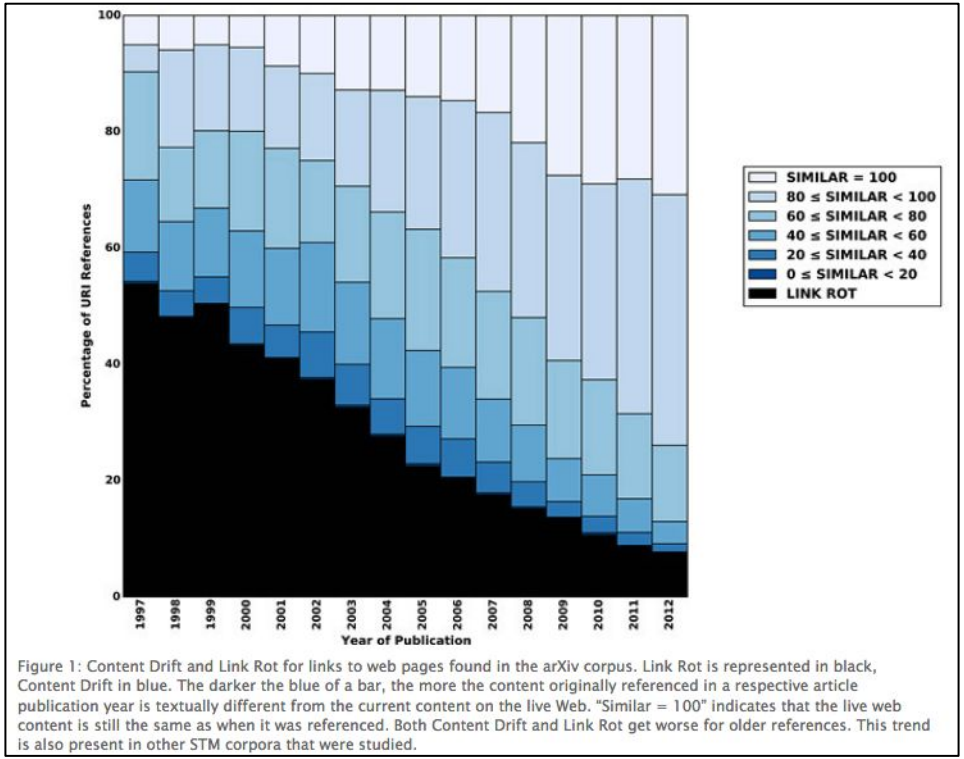


Figure 1: Content Drift and Link Rot for links to web pages found in the arXiv corpus. Link Rot is represented in black, Content Drift in blue. The darker the blue of a bar, the more the content originally referenced in a respective article publication year is textually different from the current content on the live Web. "Similar = 100" indicates that the live web content is still the same as when it was referenced. Both Content Drift and Link Rot get worse for older references. This trend is also present in other STM corpora that were studied.

Jones, Van de Sompel, Shankar, Klein, Tobin, Grover (2016) "Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content". *PLoS ONE* 11(12): e0167475.

Objectives

- Characterize the growth of the Web during the 1990s.
- Explain what it means that some web characteristics exhibit a power law distribution.
- Explain how researchers use search engines to estimate the size of the web.
- Differentiate between the concepts of link rot and content drift.