College of Information Studies

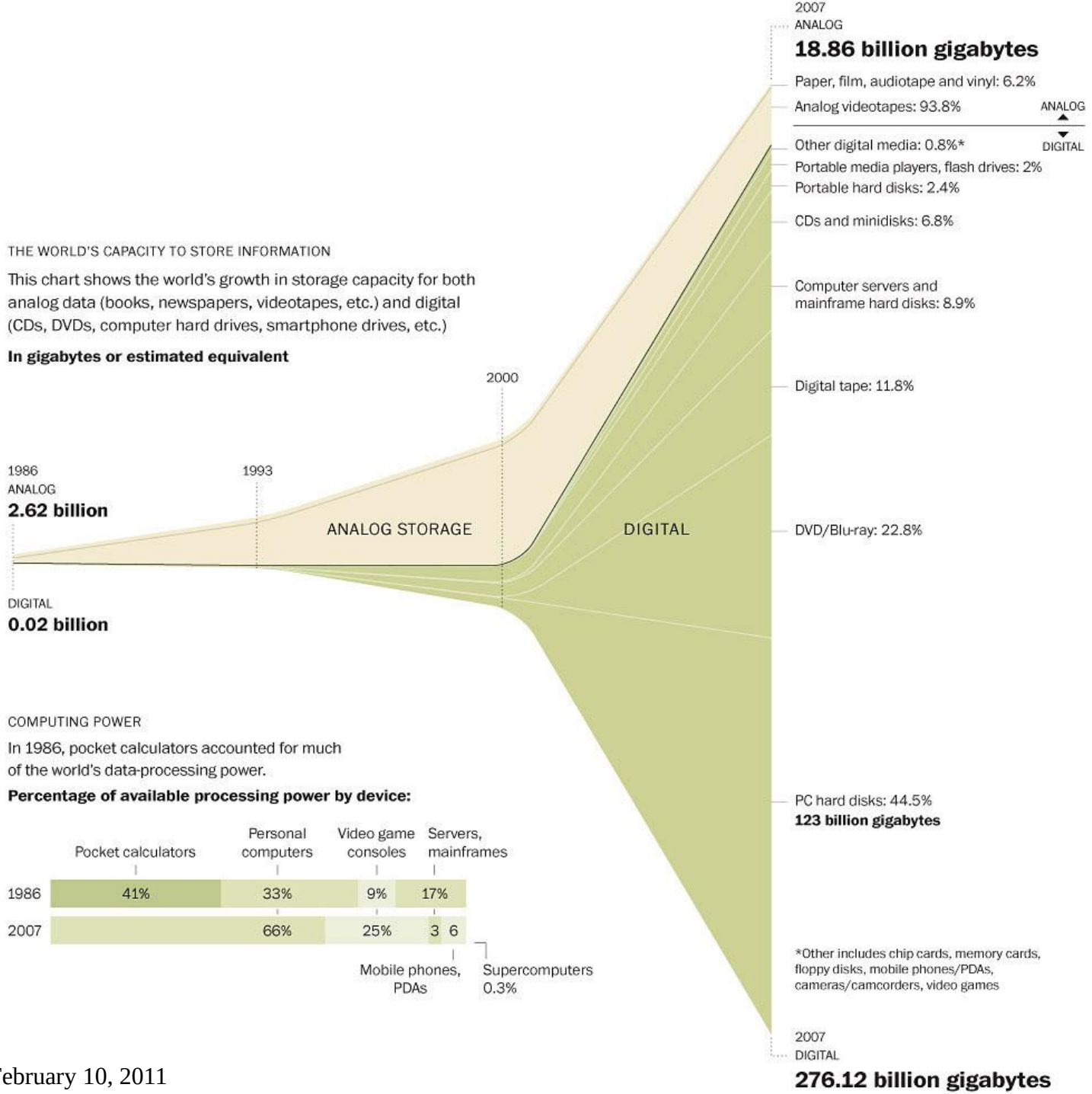
University of Maryland Hornbake Library Building College Park, MD 20742-4345

Web Search

Week 6

LBSC 796/INFM 718R

March 9, 2011



Washington Post, February 10, 2011

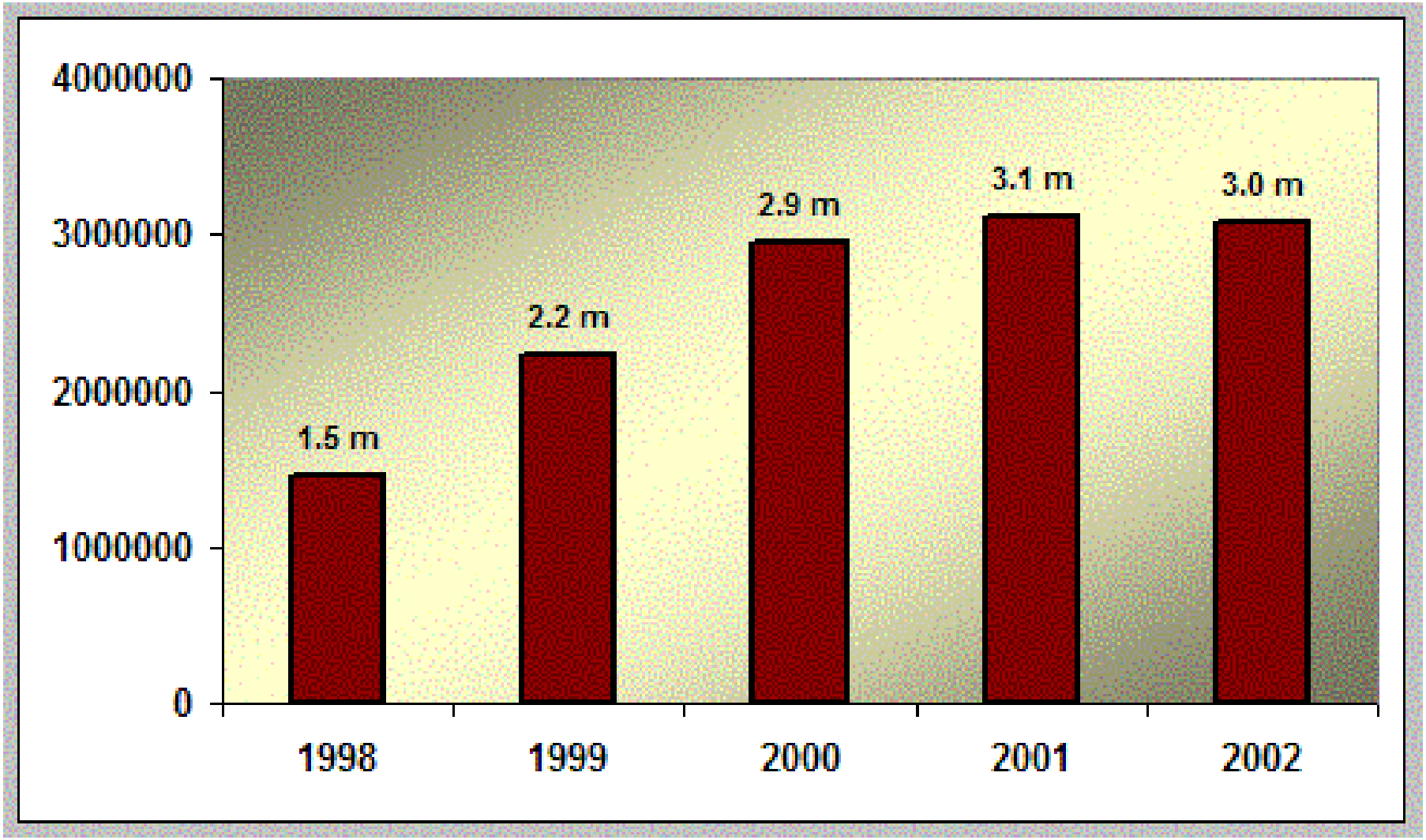
What "Caused" the Web?

- Affordable storage
  - 300,000 words/$ by 1995
- Adequate backbone capacity
  - 25,000 simultaneous transfers by 1995
- Adequate "last mile" bandwidth
  - 1 second/screen (of text) by 1995
- Display capability
  - 10% of US population could see images by 1995
- Effective search capabilities
  - Lycos and Yahoo! achieved useful scale in 1994-1995

Defining the Web

- HTTP, HTML, or URL?
- Static, dynamic or streaming?
- Public, protected, or internal?

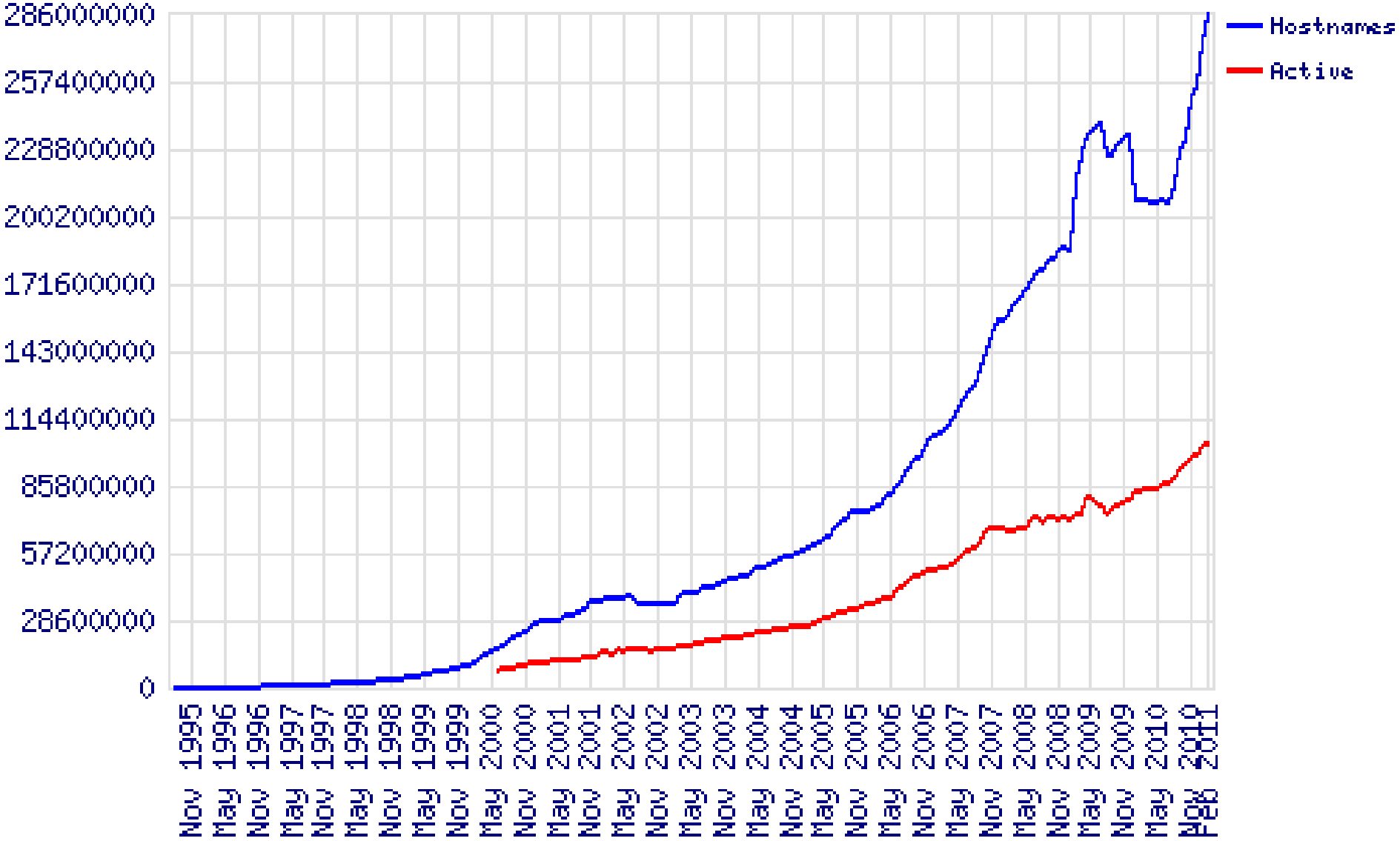
Number of Web Sites



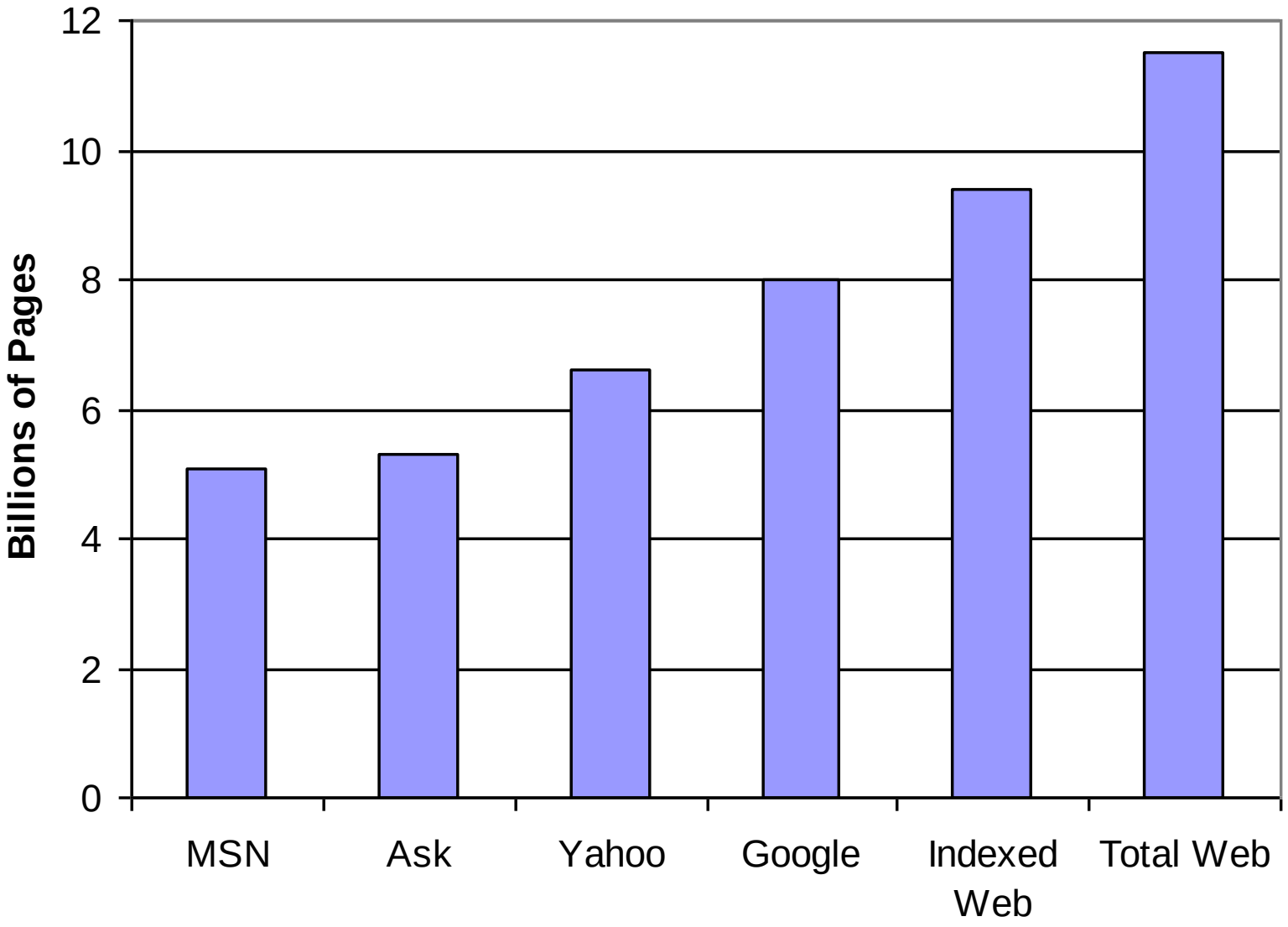
What's a Web "Site"?

- Any server at port 80?
  - Misses many servers at other ports
- Some servers host unrelated content
  - Geocities
- Some content requires specialized servers
  - rtsp

Web Servers

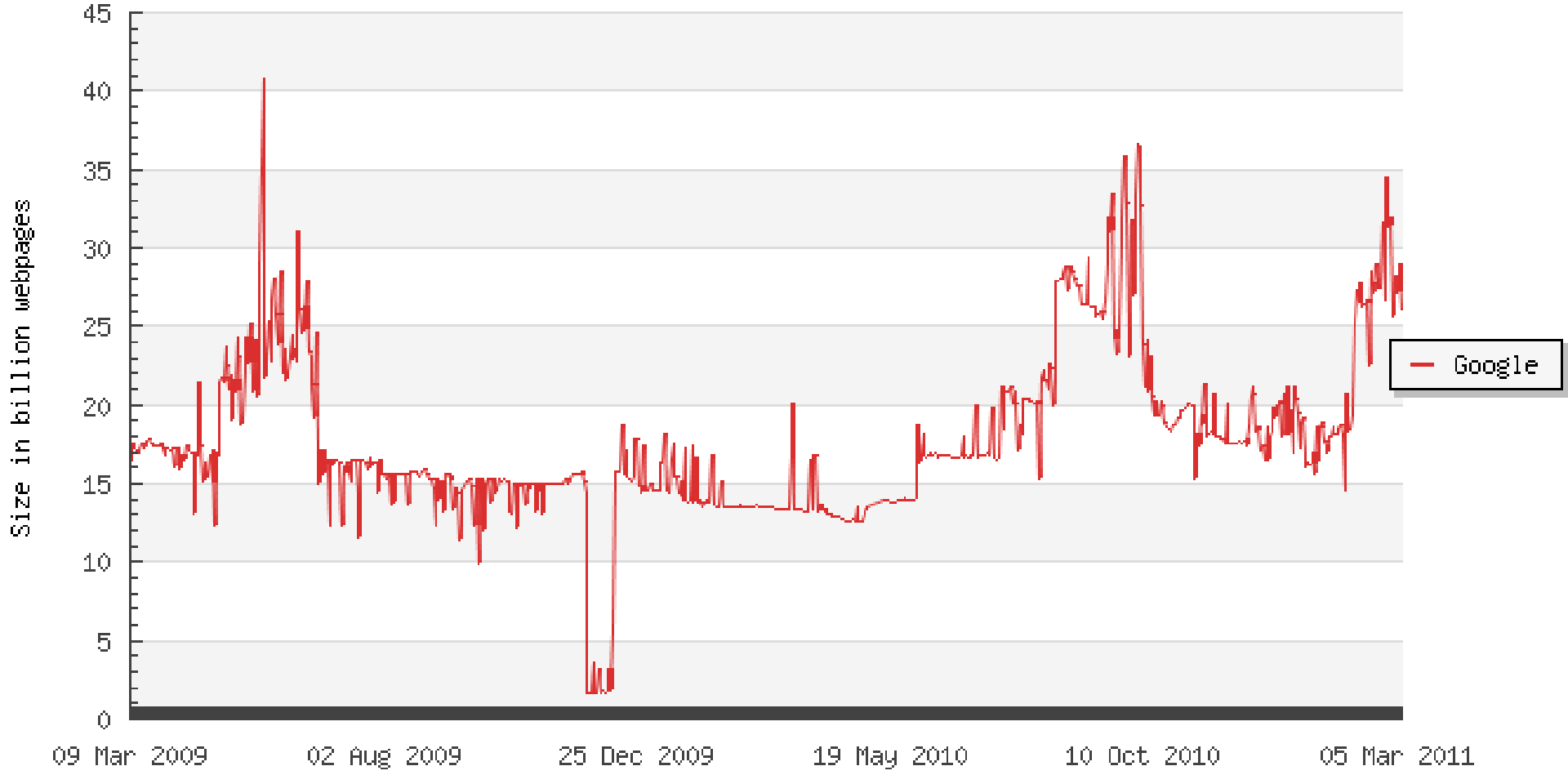

286000000
257400000
228800000
200200000
171600000
143000000
114400000
85800000
57200000
28600000
0
Hostnames
Active
Nov 1995
May 1996
Nov 1996
May 1997
Nov 1997
May 1998
Nov 1998
May 1999
Nov 1999
May 2000
Nov 2000
May 2001
Nov 2001
May 2002
Nov 2002
May 2003
Nov 2003
May 2004
Nov 2004
May 2005
Nov 2005
May 2006
Nov 2006
May 2007
Nov 2007
May 2008
Nov 2008
May 2009
Nov 2009
May 2010
Aug 2011

Web Pages (2005)

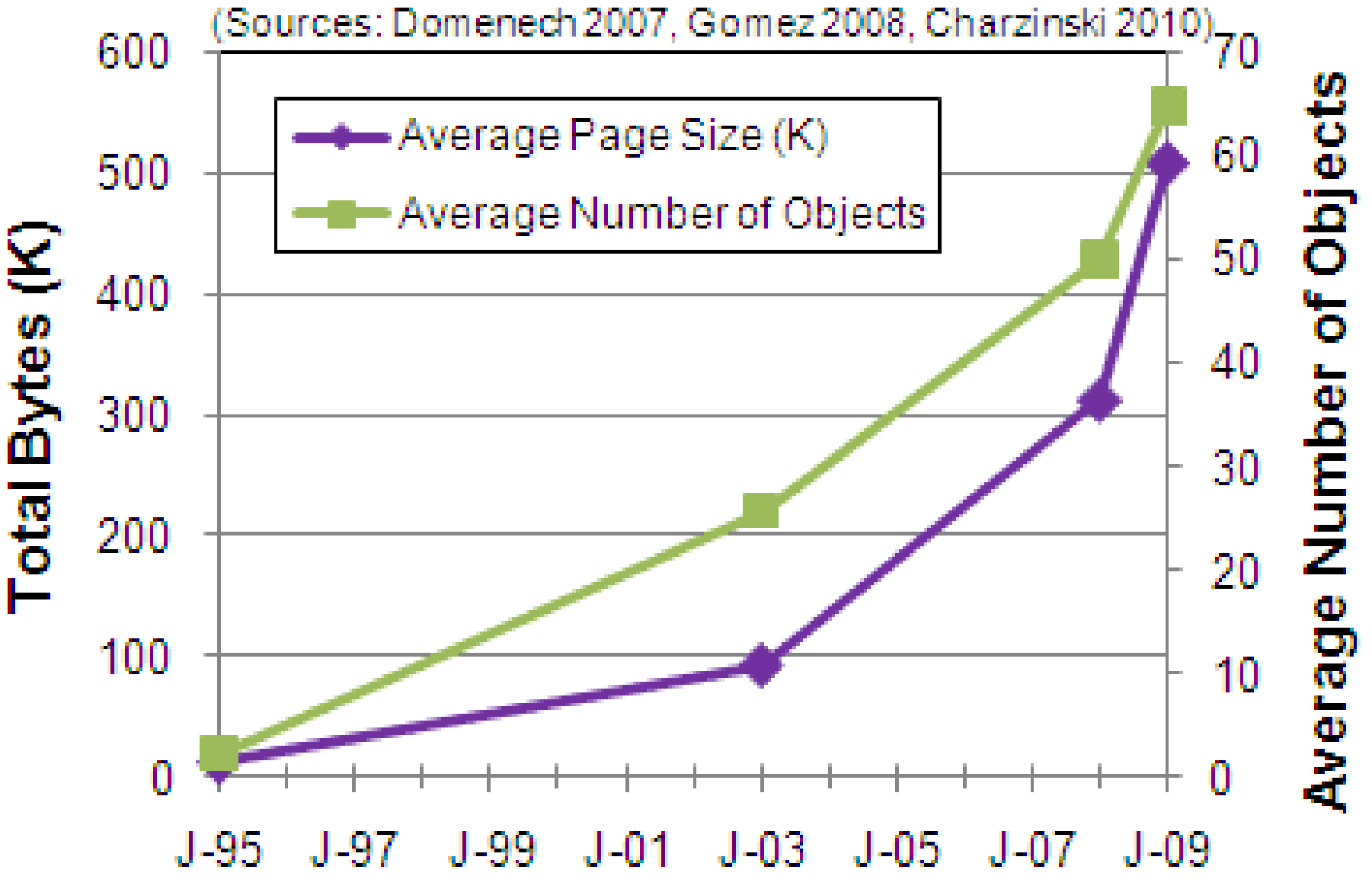


Gulli and Signorini, 2005

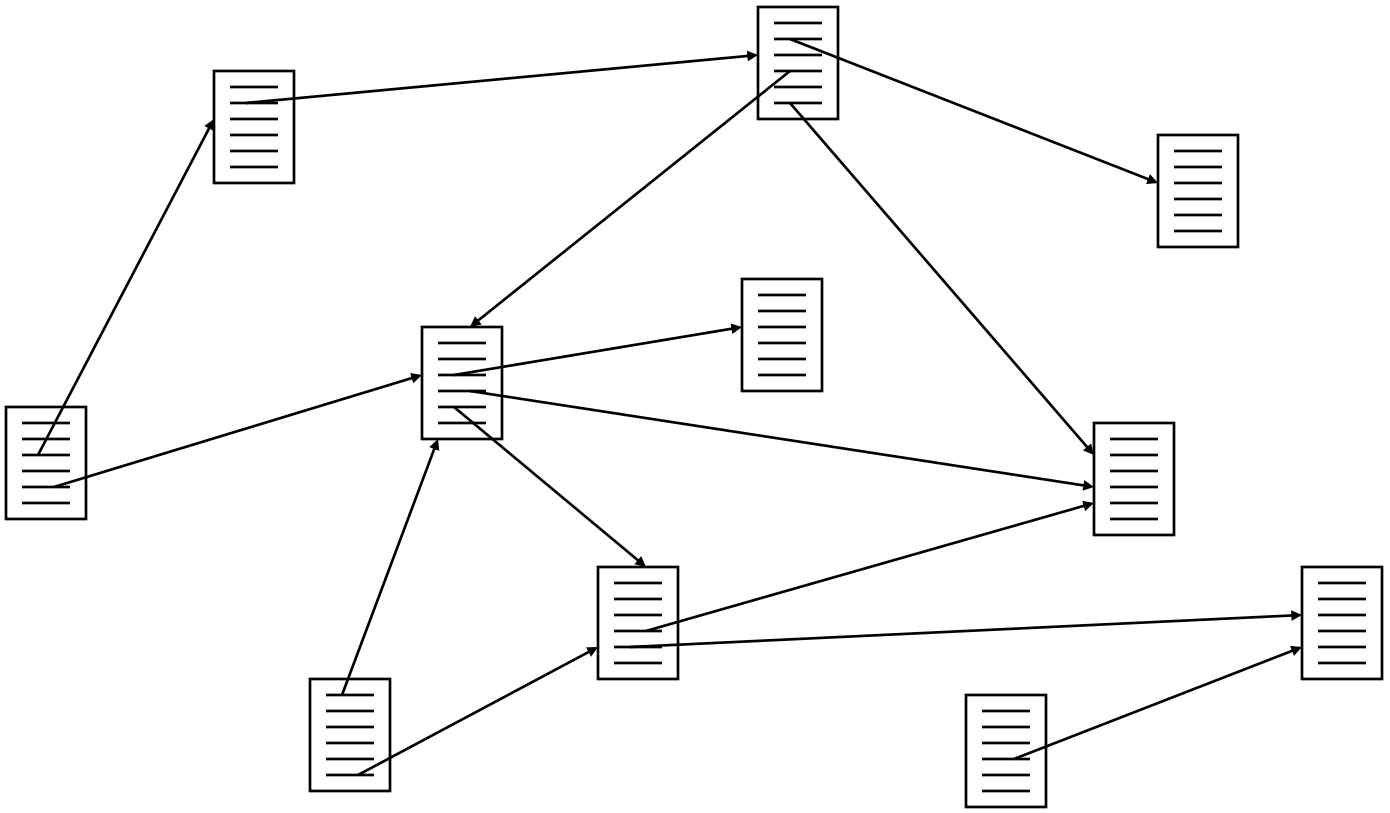
The Indexed Web in 2011



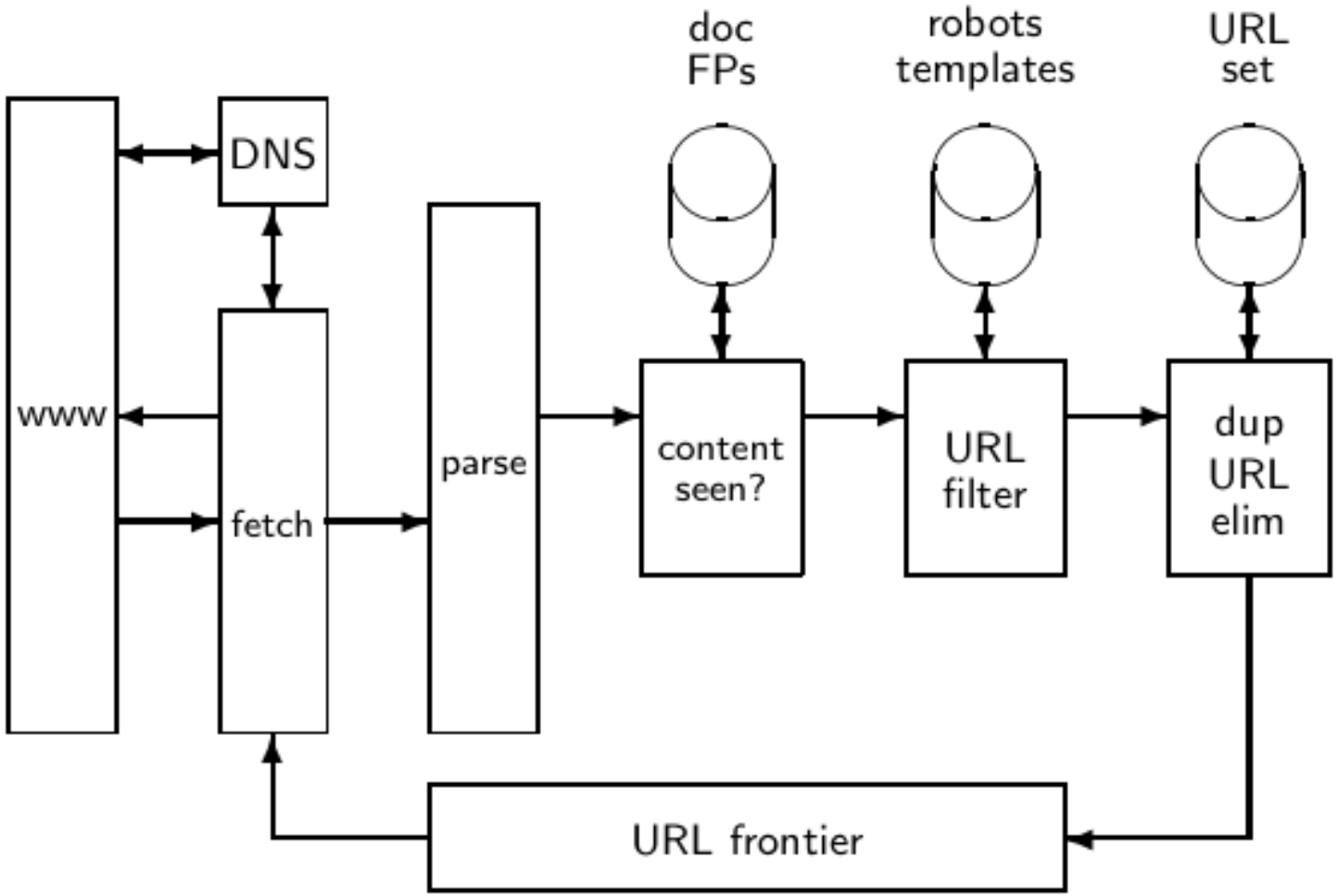
Growth of Average Web Page Size and Number of Objects



Crawling the Web

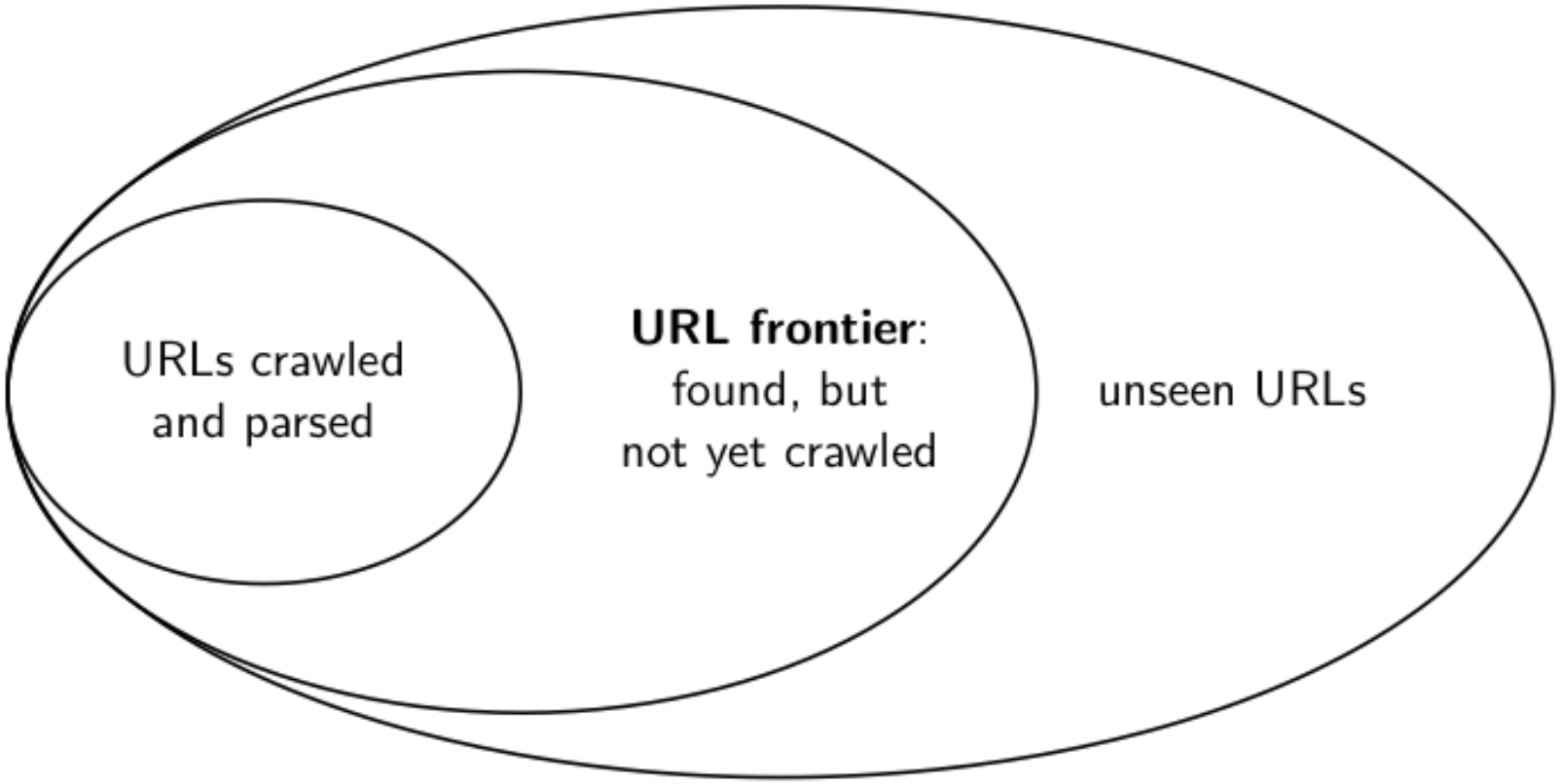


Basic crawl architecture



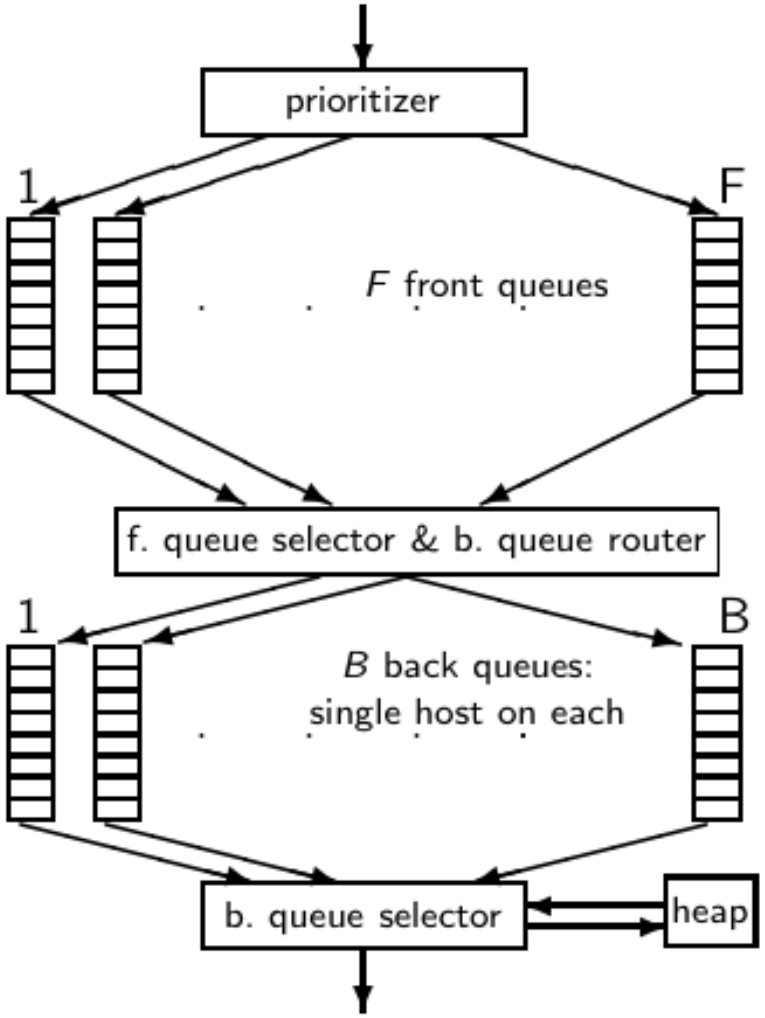
12

URL frontier



13

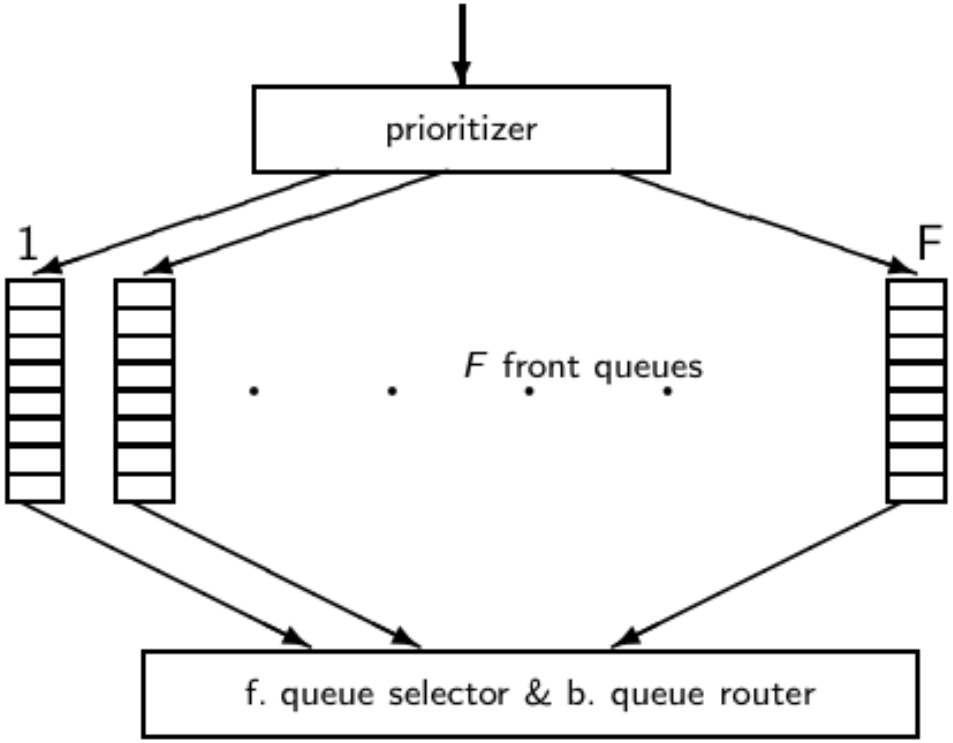
Mercator URL frontier



- URLs flow in from the top into the frontier.
- Front queues manage prioritization.
- Back queues enforce politeness.
- Each queue is FIFO.

14

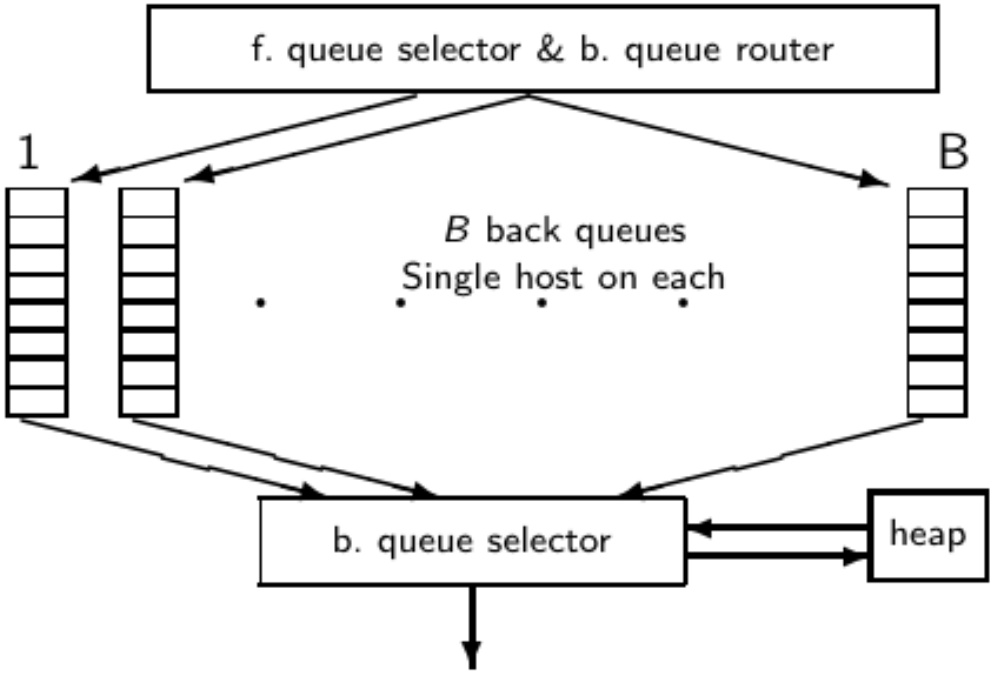
Mercator URL frontier: Front queues



- Selection from front queues is initiated by back queues
- Pick a front queue from which to select next URL: Round robin, randomly, or more sophisticated variant
- But with a bias in favor of high-priority front queues

15

Mercator URL frontier: Back queues



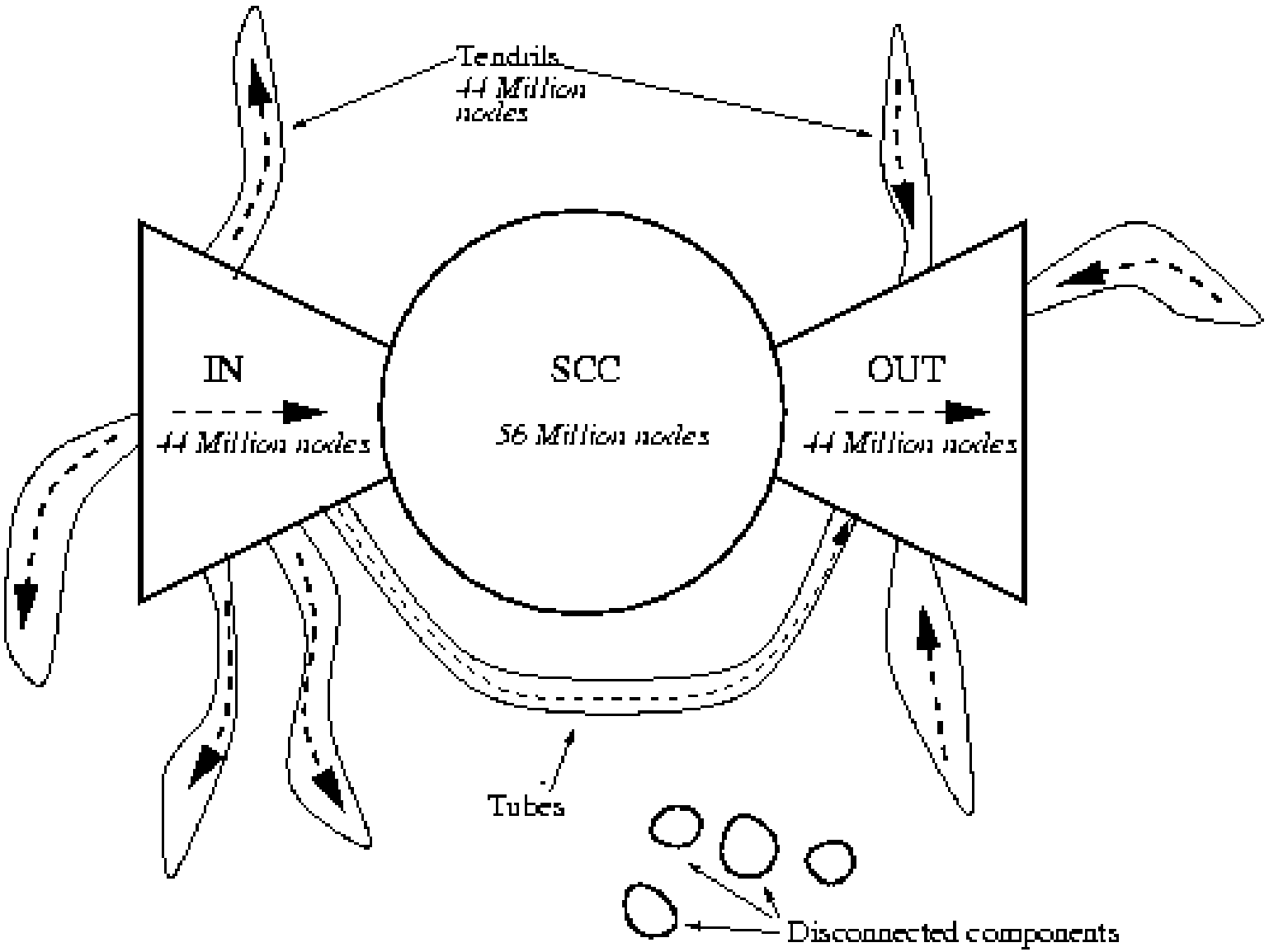
- When we have emptied a back queue *q*:
- Repeat (i) pull URLs *u* from front queues and (ii) add *u* to its corresponding back queue . . .
- . . . until we get a *u* whose host does not have a back queue.
- Then put *u* in *q* and create heap entry for it.

16

Web Crawling Algorithm

- Put a set of known sites on a queue
- Repeat the until the queue is empty:
  - Take the first page off of the queue
  - Check to see if this page has been processed
  - If this page has not yet been processed:
    - Add this page to the index
    - Add each link on the current page to the queue
    - Record that this page has been processed

Link Structure of the Web



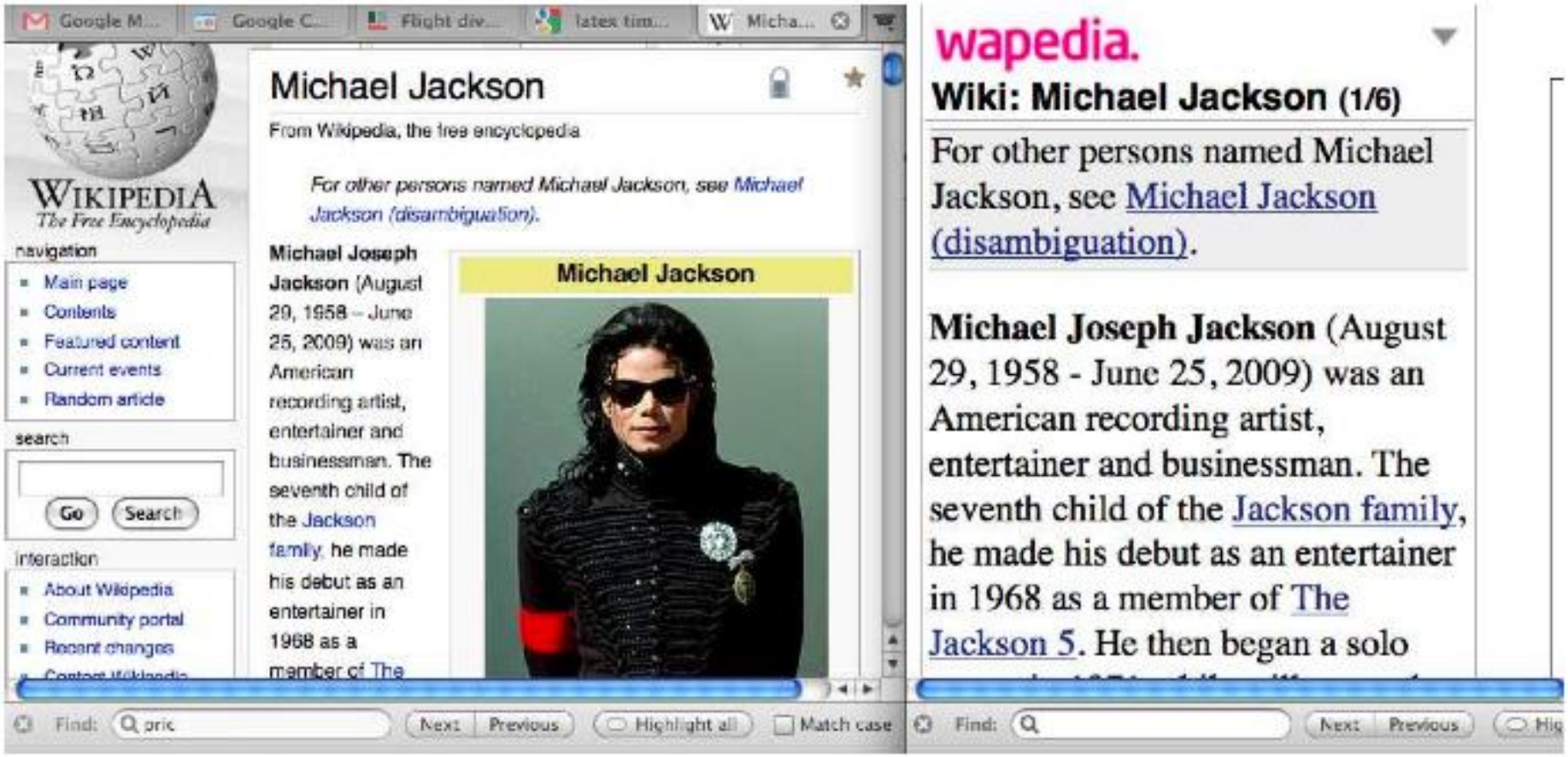
Web Crawl Challenges

- Politeness
- Discovering “islands” and “peninsulas”
- Duplicate and near-duplicate content
  - 30-40% of total content
- Server and network loads
- Dynamic content generation
- Link rot
  - Changes at ~1% per week
- Temporary server interruptions
- Spider traps

Duplicate Detection

- Structural
  - Identical directory structure (e.g., mirrors, aliases)
- Syntactic
  - Identical bytes
  - Identical markup (HTML, XML, …)
- Semantic
  - Identical content
  - Similar content (e.g., with a different banner ad)
  - Related content (e.g., translated)

Near-duplicates: Example



Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want "syntactic" (as opposed to semantic) similarity.
  - True semantic similarity (similarity in content) is too difficult to compute.
- We do not consider documents near-duplicates if they have the same content, but express it with different words.
- Use similarity threshold θ to make the call "is/isn't a near-duplicate".
- E.g., two documents are near-duplicates if similarity $> \theta = 80\%$.

Represent each document as set of **shingles**

- A shingle is simply a word n-gram.
- Shingles are used as features to measure syntactic similarity of documents.
- For example, for $n = 3$, "a rose is a rose is a rose" would be represented as this set of shingles:
  - { a-rose-is, rose-is-a, is-a-rose }
- We can map shingles to $1..2^m$ (e.g., $m = 64$) by fingerprinting.
- From now on: s_k refers to the shingle's fingerprint in $1..2^m$.
- We define the similarity of two documents as the Jaccard coefficient of their shingle sets.

Shingling: Summary

- Input: N documents
- Choose n-gram size for shingling, e.g., $n = 5$
- Pick 200 random permutations, represented as hash functions
- Compute N sketches: $200 \times N$ matrix shown on previous slide, one row per permutation, one column per document
- Compute $\frac{N \cdot (N-1)}{2}$ pairwise similarities
- Transitive closure of documents with similarity $> \theta$
- Index only one document from each equivalence class

Robots Exclusion Protocol

- Depends on voluntary compliance by crawlers
- Exclusion by site
  - Create a robots.txt file at the server's top level
  - Indicate which directories not to crawl
- Exclusion by document (in HTML head)
  - Not implemented by all crawlers

    <meta name="robots“ content="noindex,nofollow">

Hands on:
The Internet Archive

- Web crawls since 1997
  - http://archive.org
- Check out the iSchool's Web site in 1997
- Check out the history of your favorite site

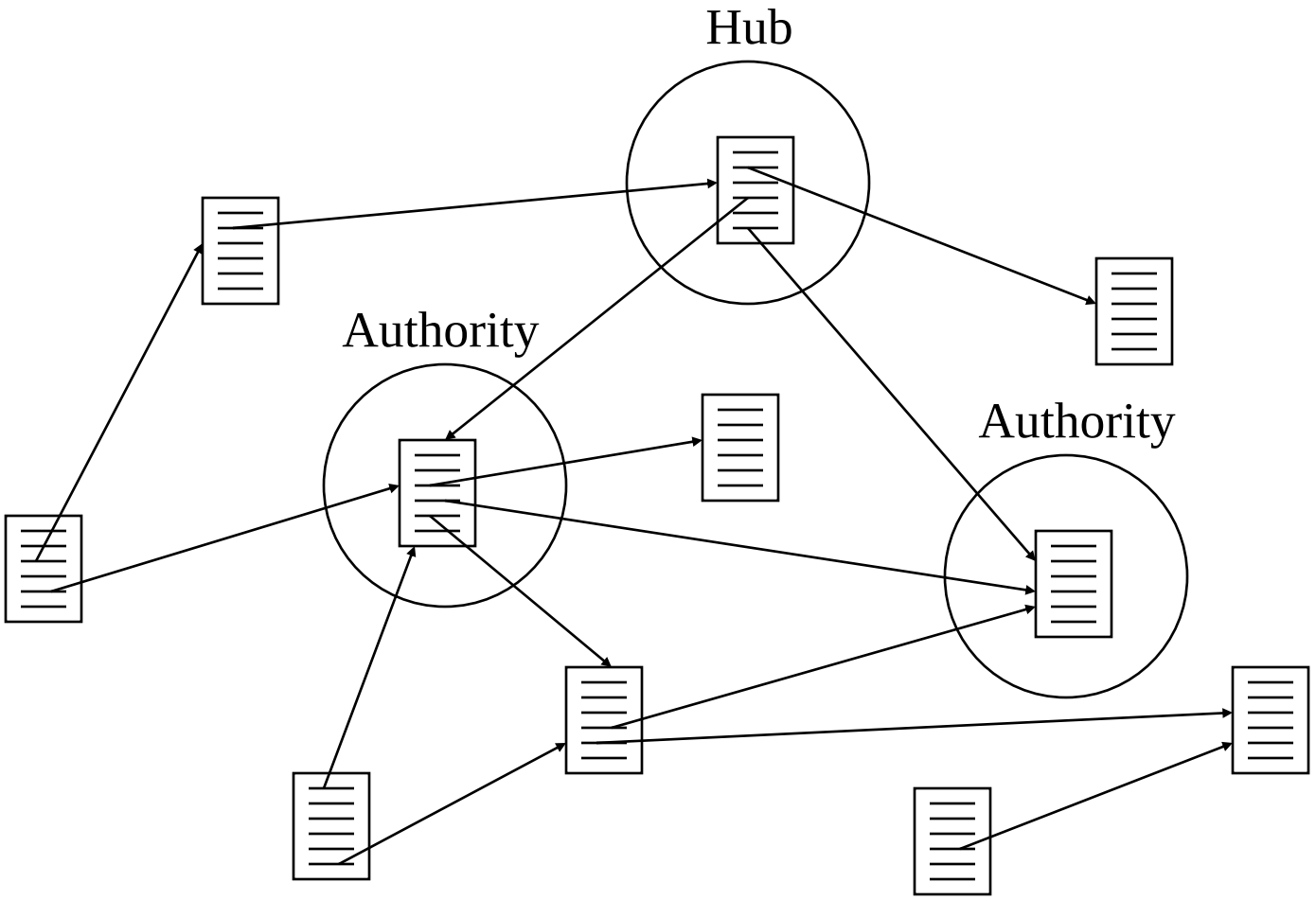
Indexing Anchor Text

- A type of "document expansion"
  - Terms near links describe content of the target
- Works even when you can't index content
  - Image retrieval, uncrawled links, …

[Bean - "And that's the way we tried to do every rock. Because you always had the gnomon. And then we took a photo afterwards."]

[Conrad - "We practiced this...I started out by just laying rocks around on the floor. One of the things was setting the camera deal; we had the three (focus) distances. And what we did was actually take pictures to calibrate ourselves. They developed that film in training to make sure we stood the right distance."]

Estimating Authority from Links

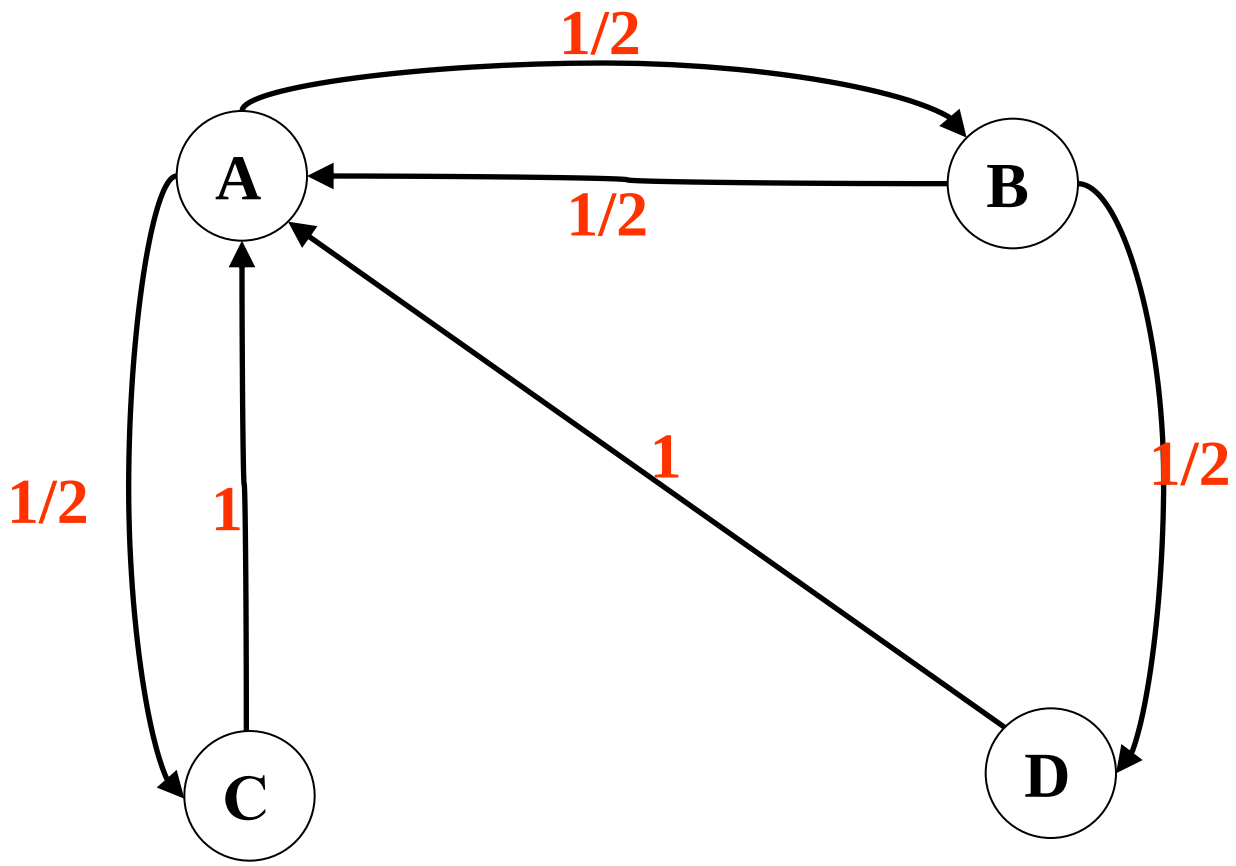


Simplified PageRank Algorithm

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

R(u): PageRank score of page u
B_u: the set of pages that link to u
R(v): PageRank score of page v
N_V: number of links from page v
c: normalization factor

PageRank Algorithm Example

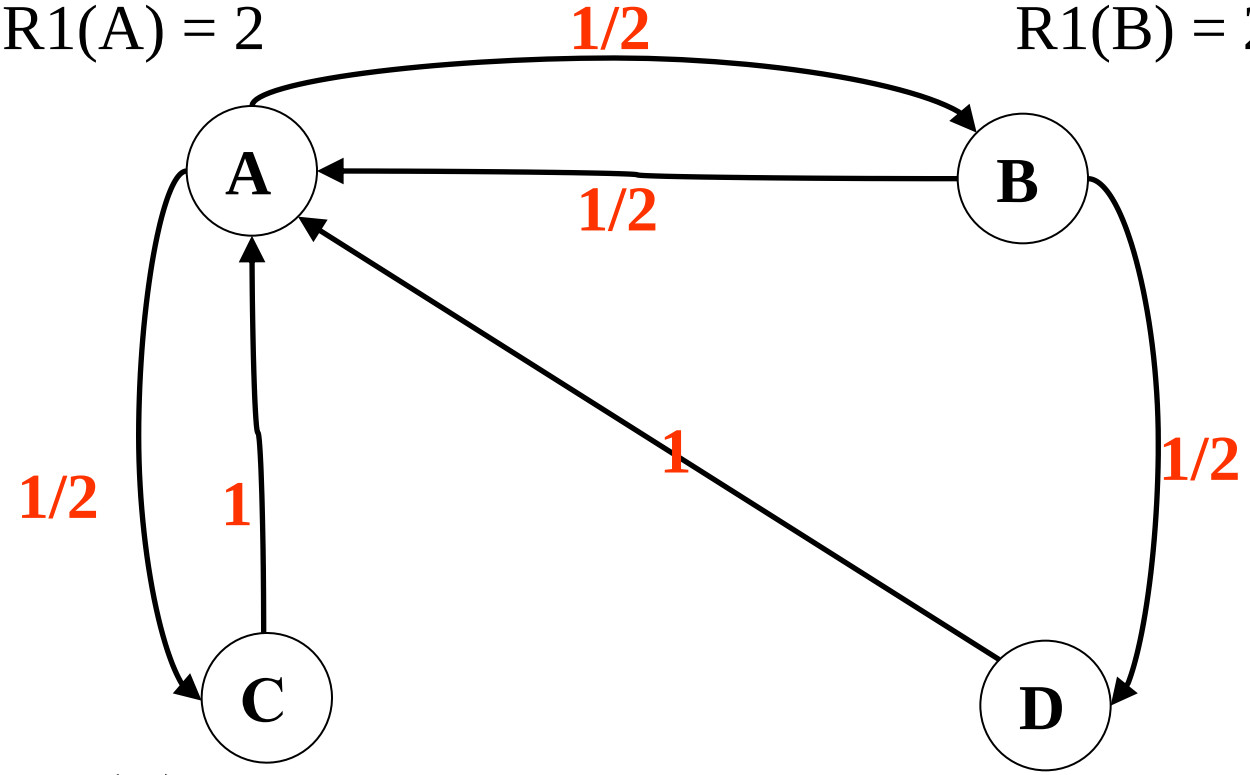


R2(A) = ½ x 2 + 1 x 2 + 1 x 2 = 5

R1(A) = 2

R2(B)’ = ½ x 2 = 1

R1(B) = 2



R2(C) = 2

R1(C)’ = ½ x 2 = 1

R2(D) = 2

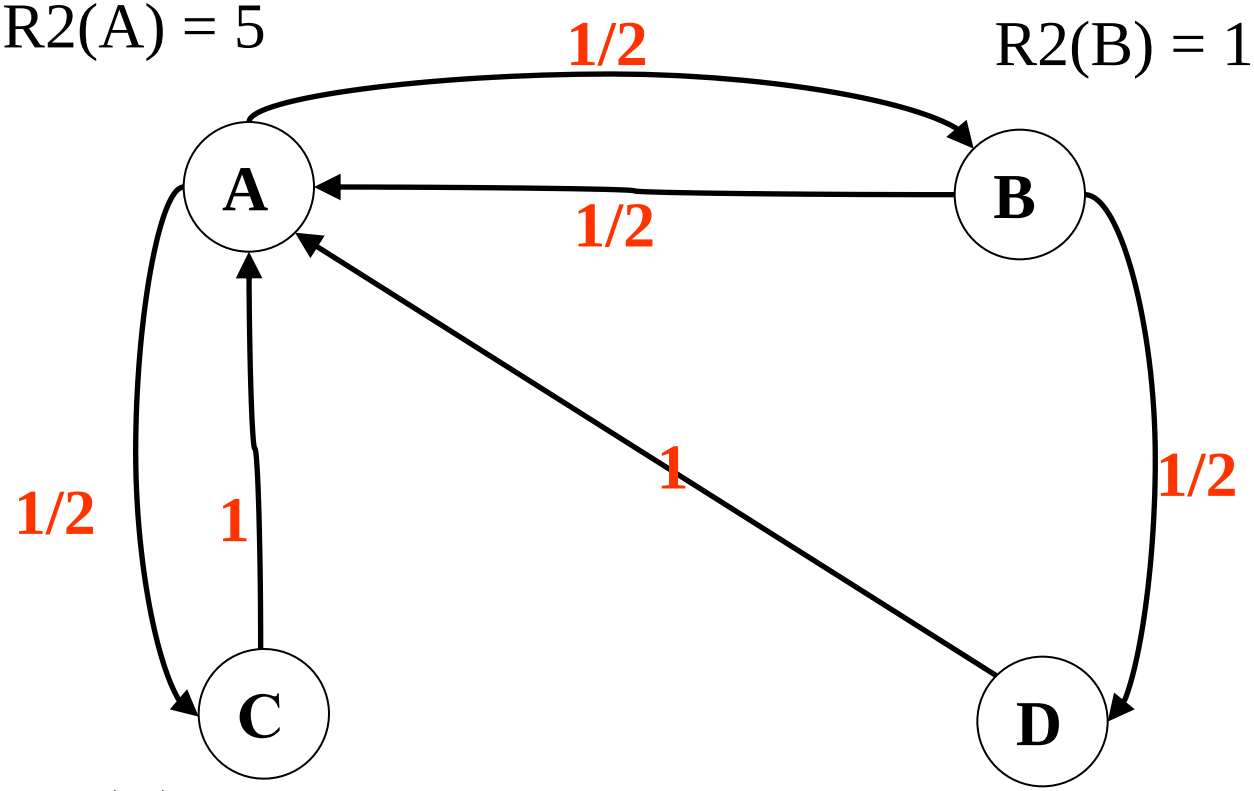
R1(D)’ = ½ x 2 = 1

R3(A) = ½ x 1 + 1 x 1 + 1 x 1 = 2.5

R2(A) = 5

R3(B) = ½ x 5 = 2.5

R2(B) = 1



R2(C) = 1

R3(C) = ½ x 5 = 2.5

R2(D) = 1

R3(D) = ½ x 1 = 0.5

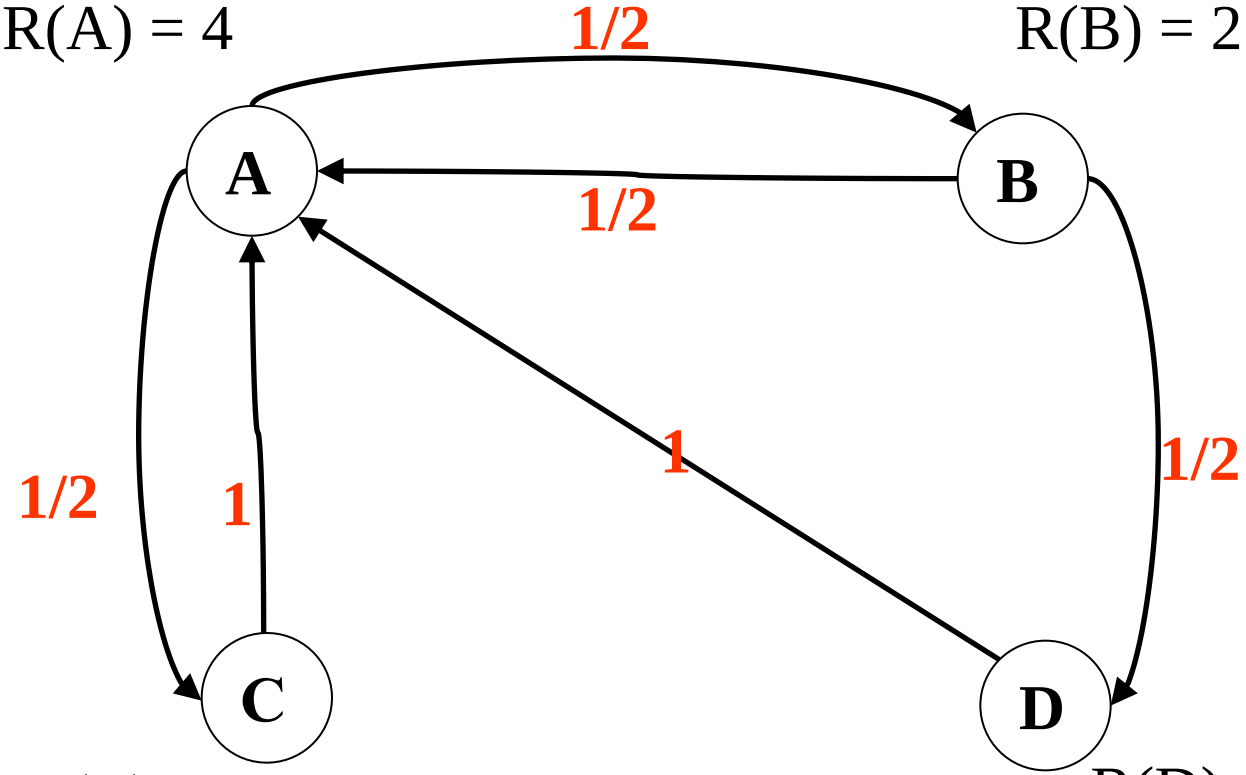
Convergence

R(A)’ = ½ x 2 + 1 x 2 + 1 x 1 = 4

R(A) = 4

R(B)’ = ½ x 4 = 2

R(B) = 2



R(C) = 2

R(C)’ = ½ x 4 = 2

R(D) = 1

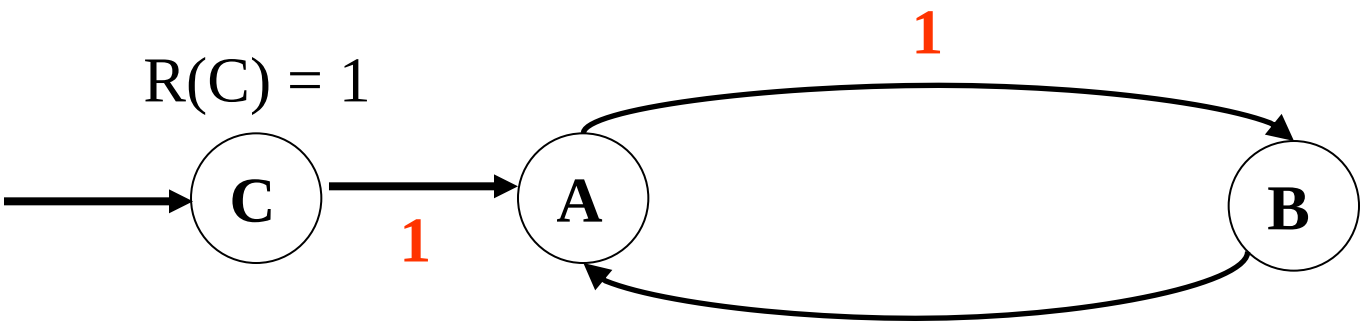
R(D)’ = ½ x 2 = 1

Index Spam

- Goal: Manipulate rankings of an IR system

- Multiple strategies:
  - Create bogus user-assigned metadata
  - Add invisible text (font in background color, …)
  - Alter your text to include desired query terms
  - “Link exchanges” create links to your page
  - Cloaking

Spamming with Link Farms



R(A) = 1
2
2
3
3
…

R(B) = 1
1
2
2
3
…

Adversarial IR

- Search is user-controlled suppression
  - Everything is known to the search system
  - Goal: avoid showing things the user doesn't want
- Other stakeholders have different goals
  - Authors risk little by wasting your time
  - Marketers hope for serendipitous interest

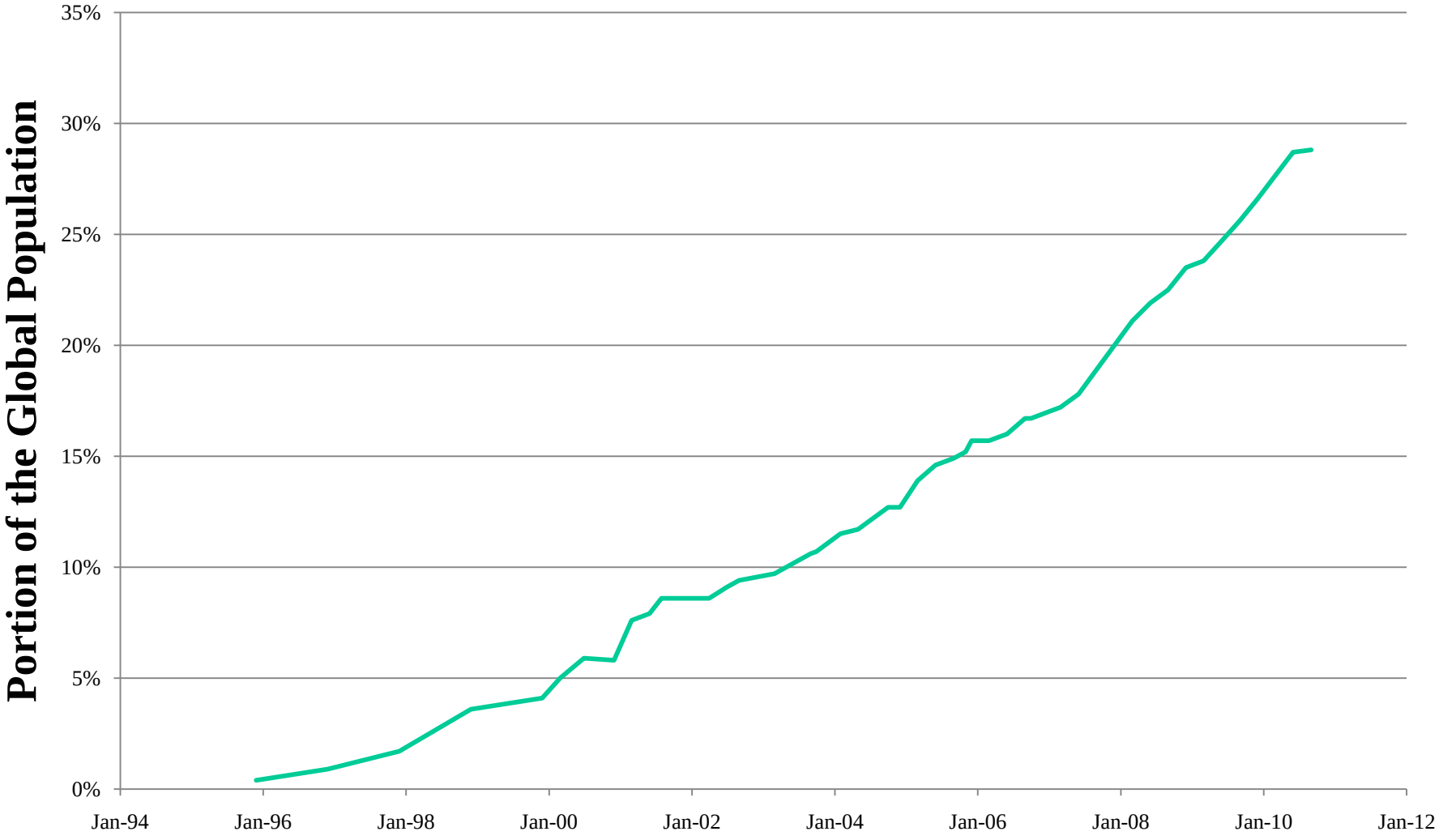
“Safe Search”

- Text
- Whitelists and blacklists
- Link structure
- Image analysis

Computational Advertizing

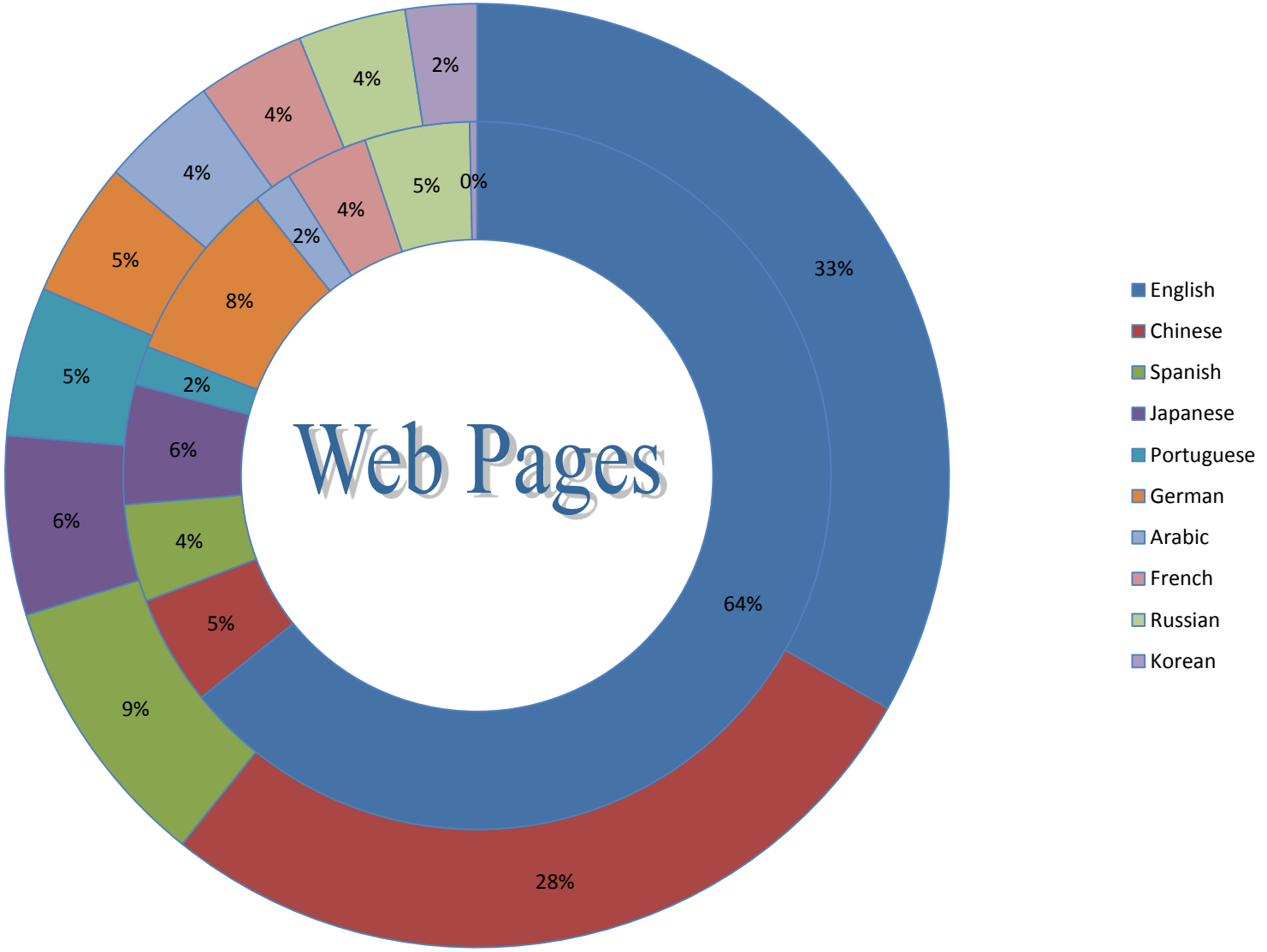
- Variant of a search problem
  - Jointly optimize relevance and revenue
- Triangulating features
  - Queries
  - Clicks
  - Page visits
- Auction markers

Internet Users

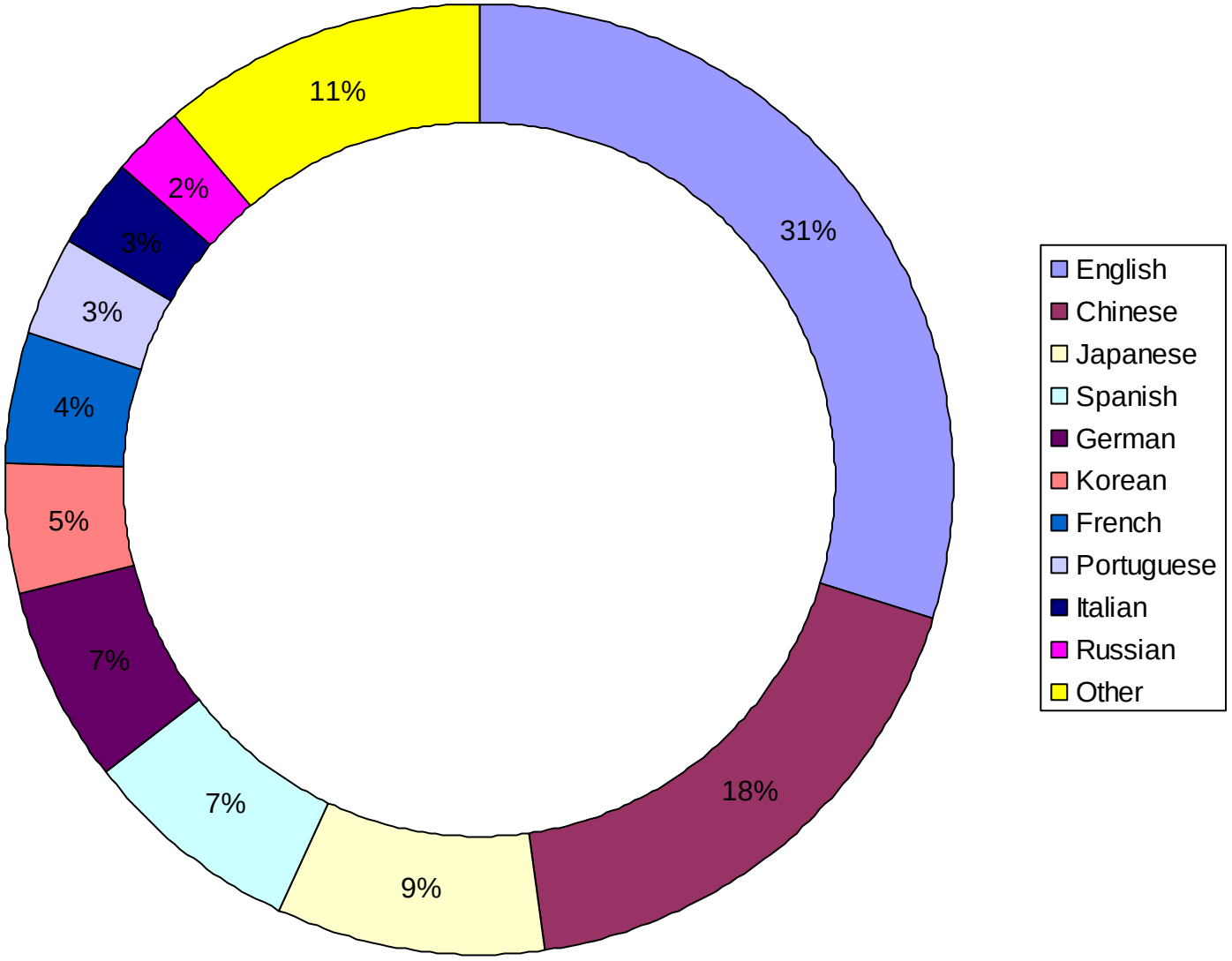


http://www.internetworldstats.com/

Global Internet Users

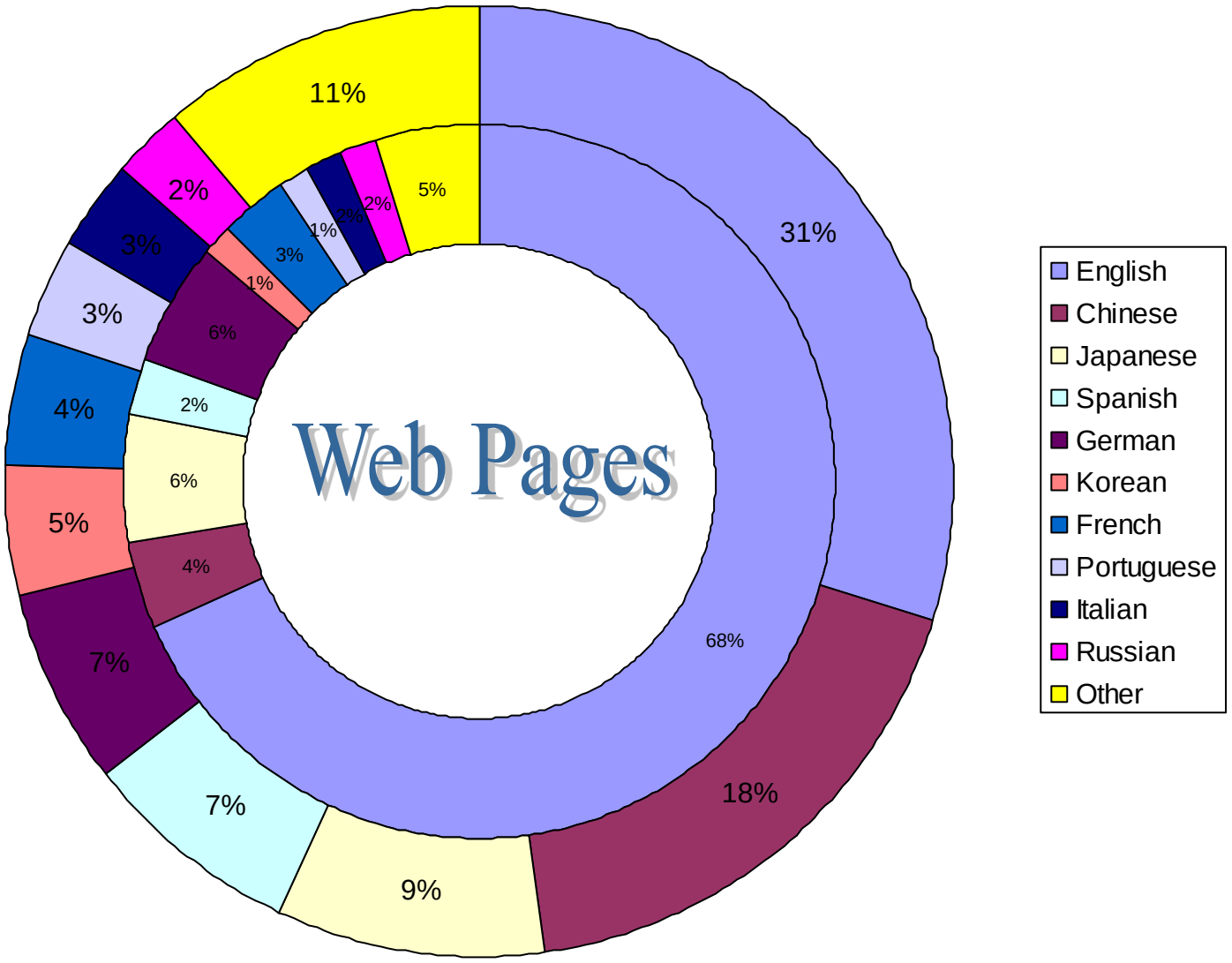


Global Internet Users



Native speakers, Global Reach projection for 2004 (as of Sept, 2003)

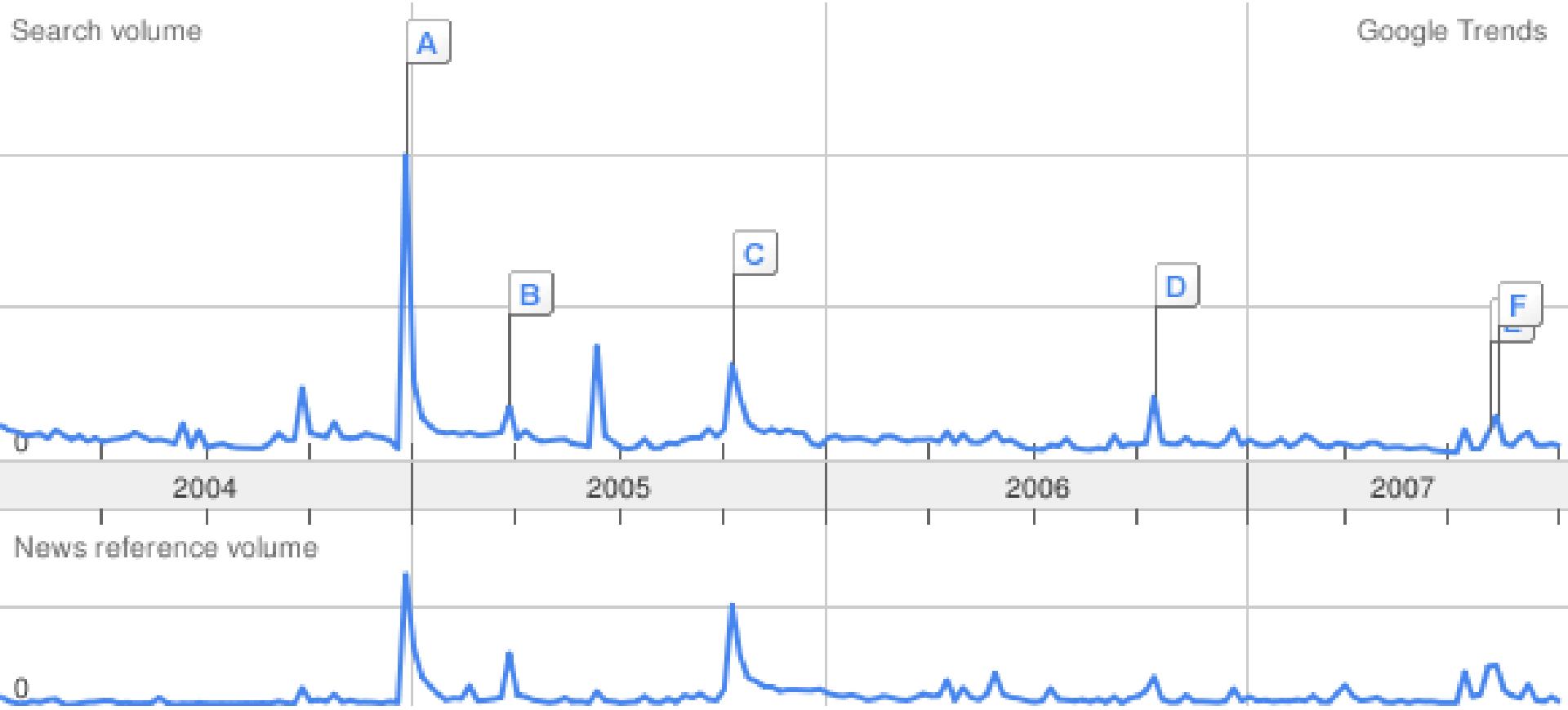
Global Internet Users



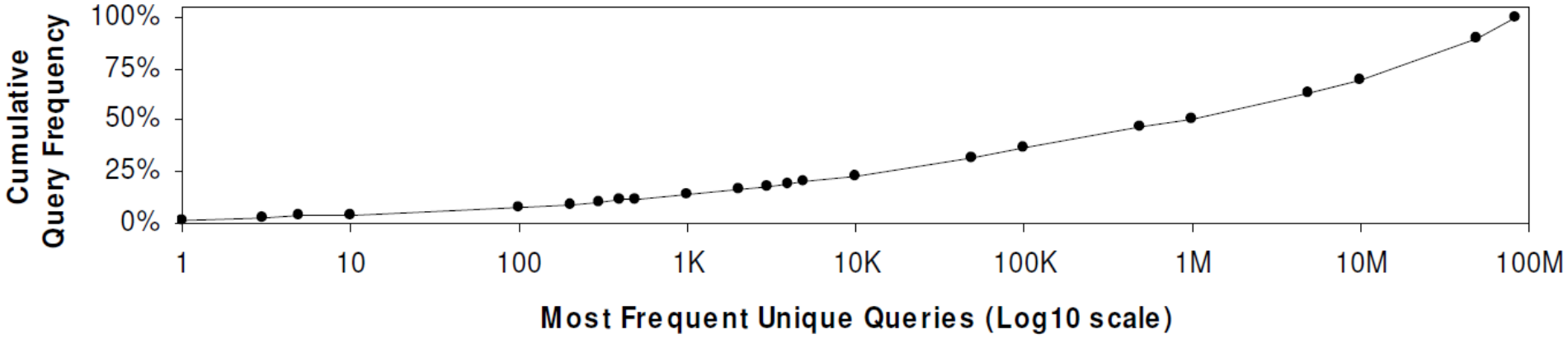
Native speakers, Global Reach projection for 2004 (as of Sept, 2003)

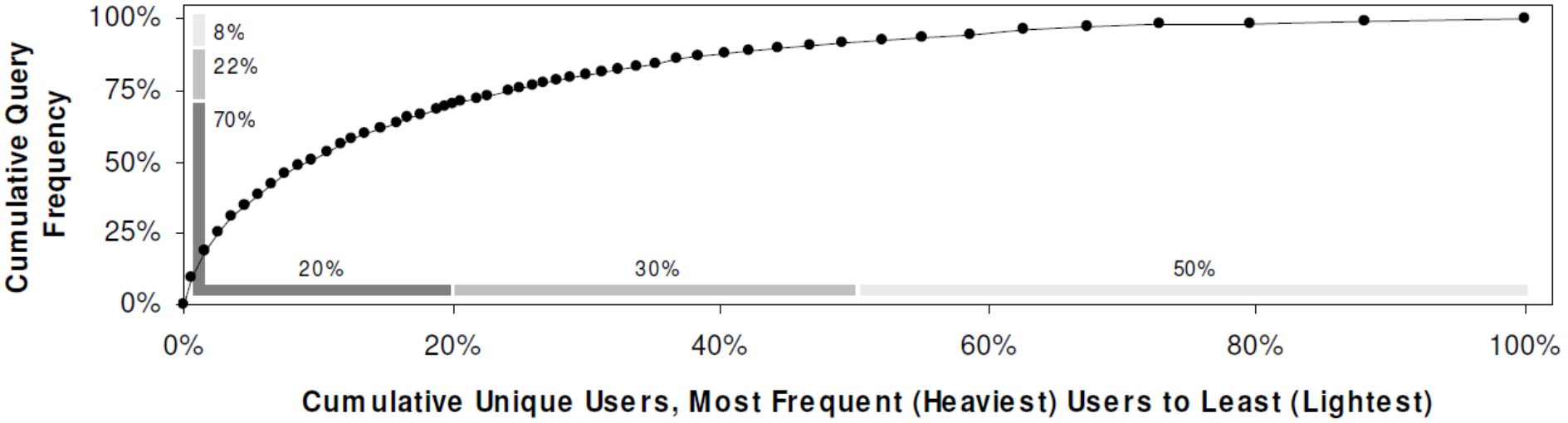
Search Engine Query Logs

A: Southeast Asia (Dec 27, 2004)
B: Indonesia (Mar 29, 2005)
C; Pakistan (Oct 10, 2005)
D; Hawaii (Oct 16, 2006)
E: Indonesia (Aug 8, 2007)
F: Peru (Aug 16, 2007)

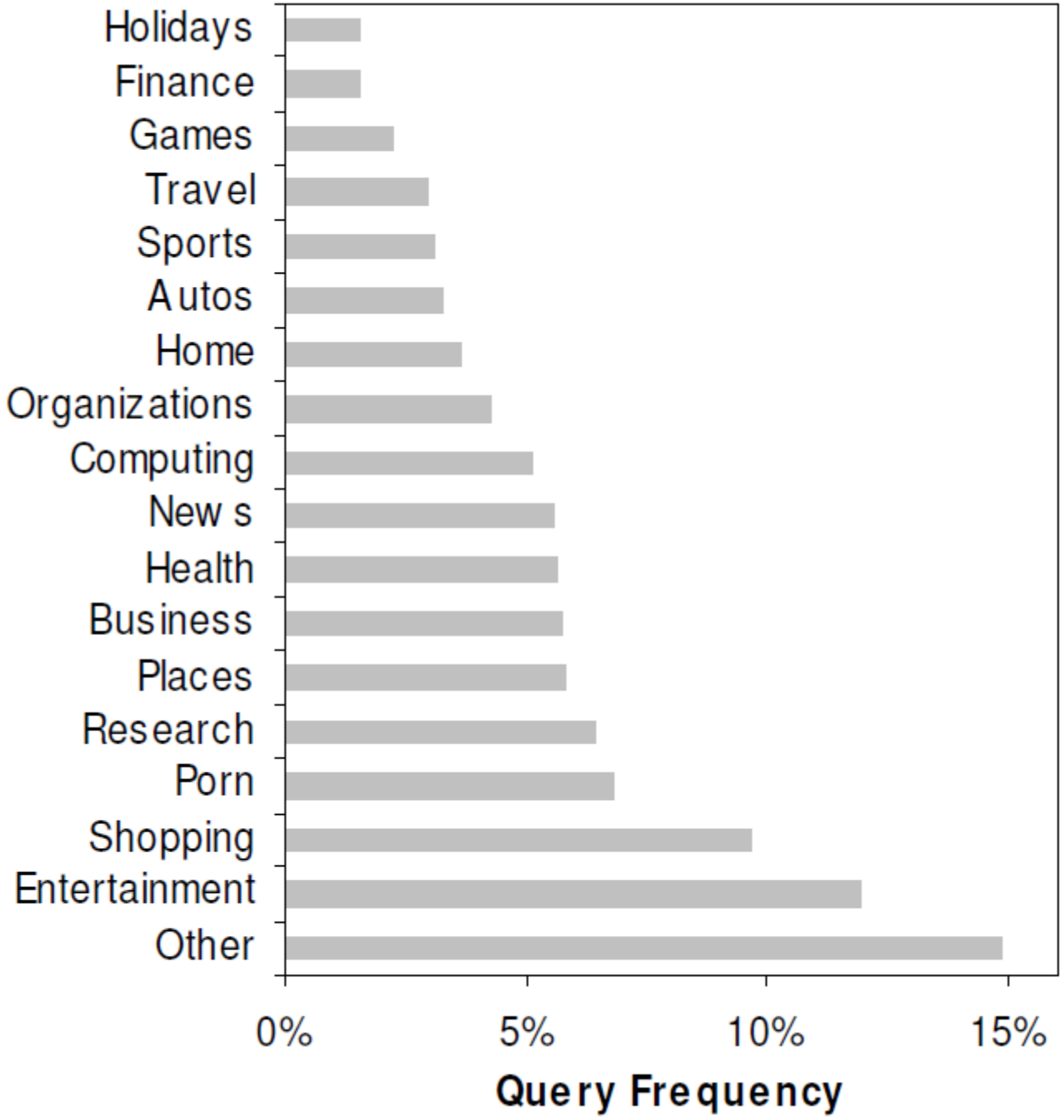


Query Statistics



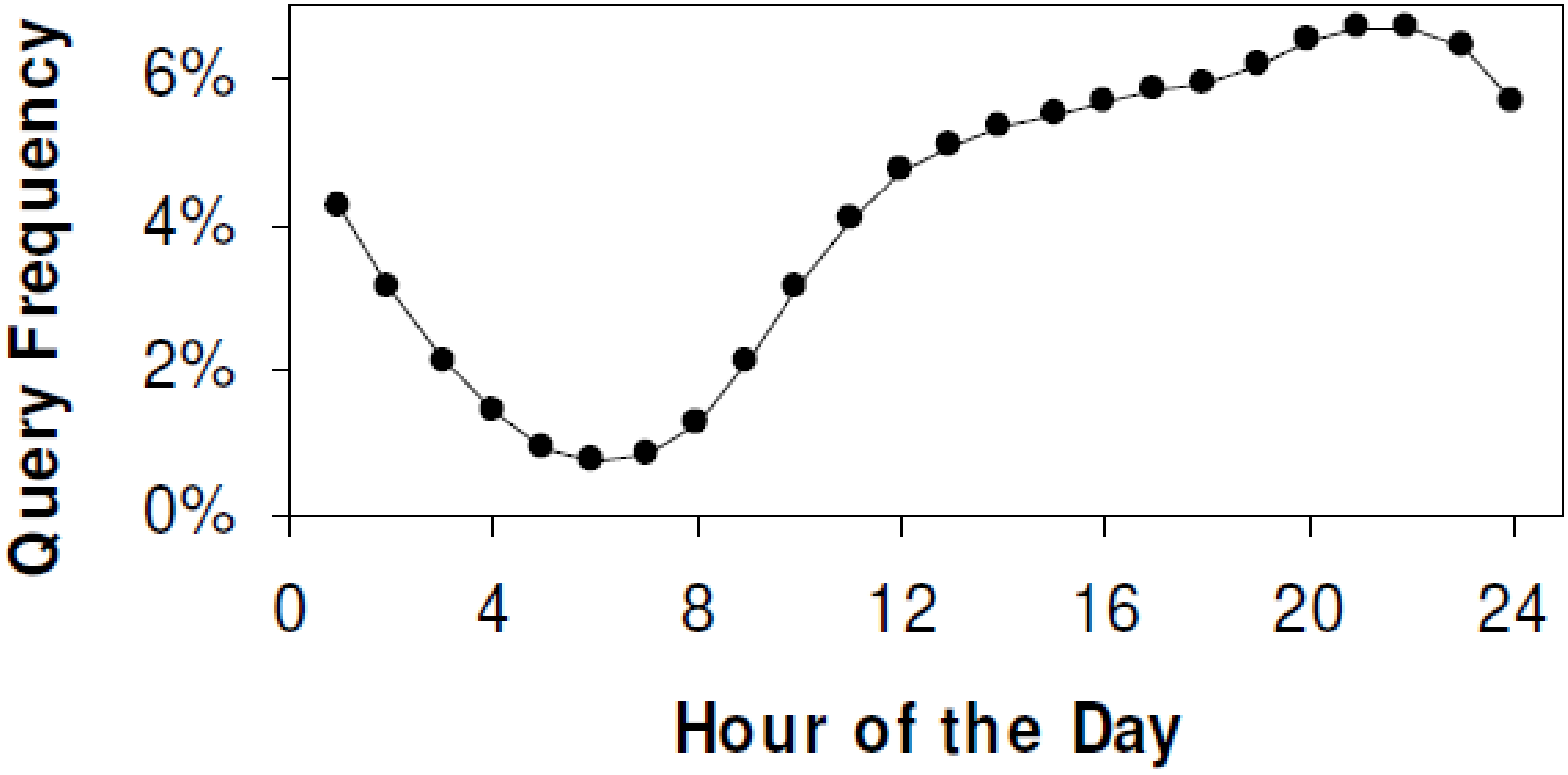


Pass, et al., "A Picture of Search," 2007

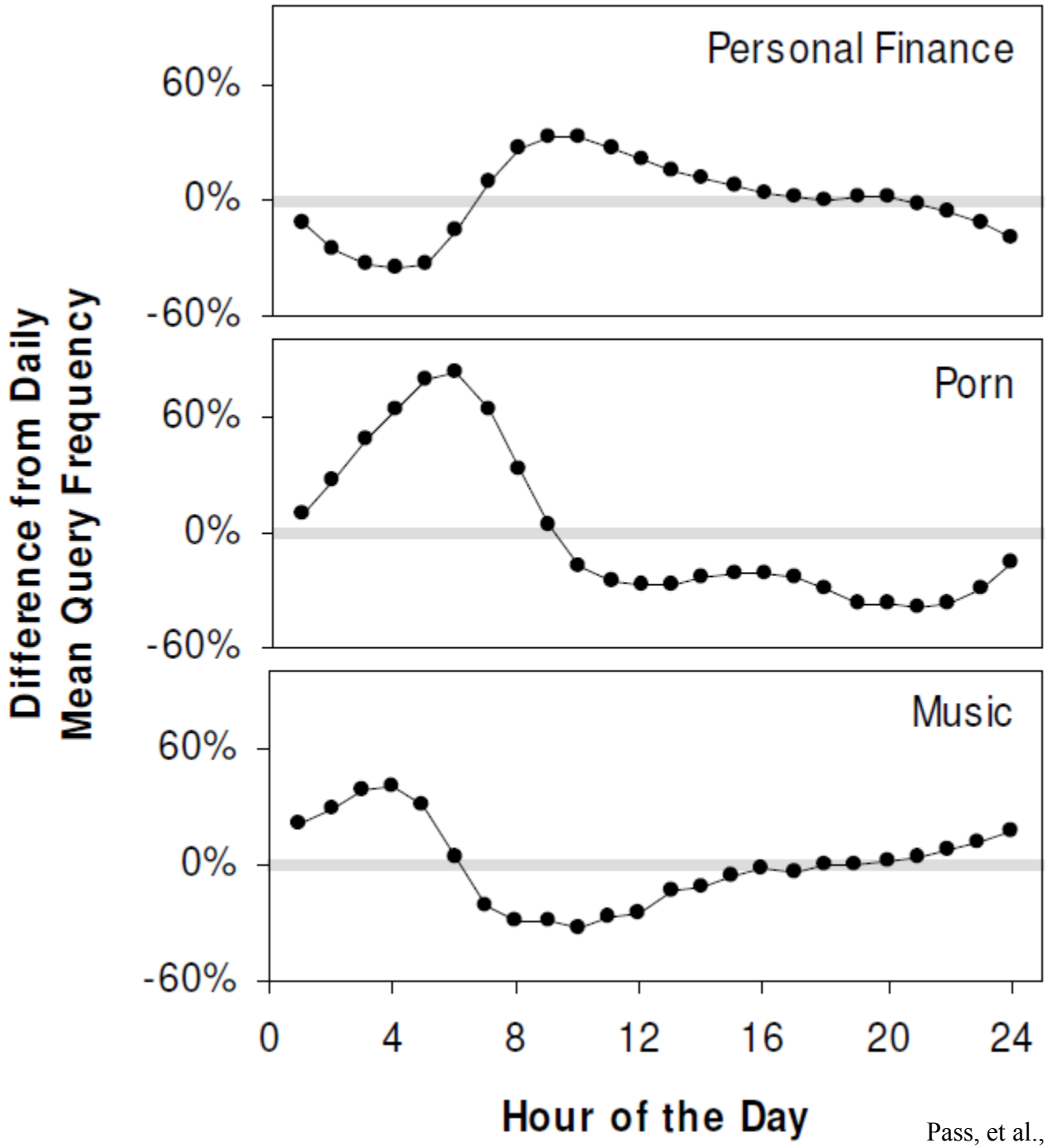


Pass, et al., "A Picture of Search," 2007

Temporal Variation



Pass, et al., "A Picture of Search," 2007



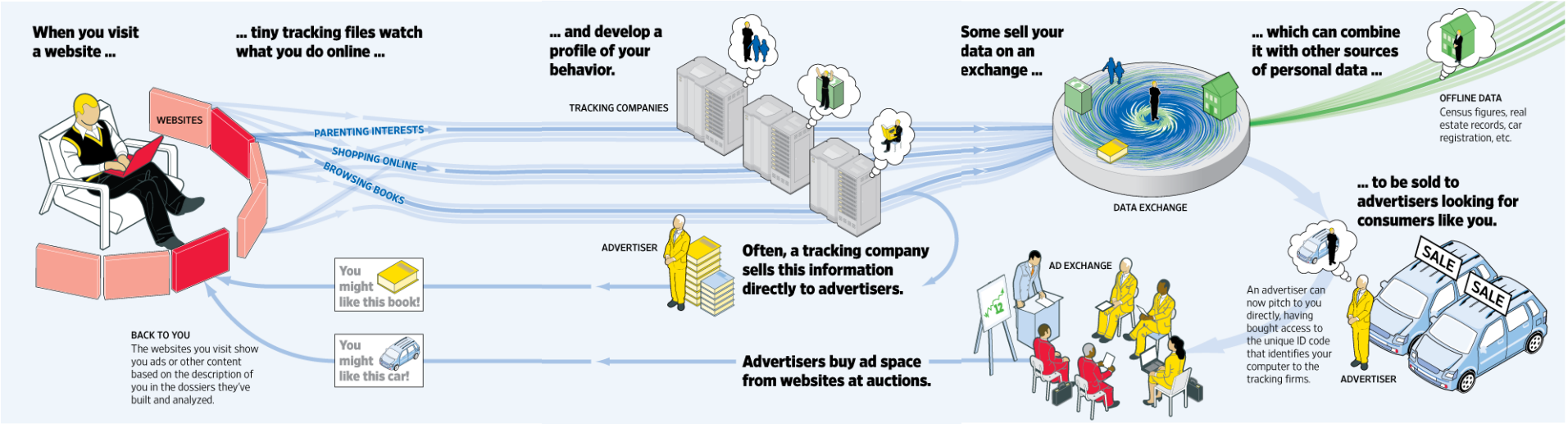
Pass, et al., "A Picture of Search," 2007

28% of Queries are Reformulations

Timeline (mm:ss)		Query
00:00	◯	nursing registry
04:18	c	certified nursing assistant 1
08:48	c	nursing assistant registry
09:48	c	license look up for nursing assistants
10:06	c	nursing assistant 1 certification
11:42	c	nursing assistant 1 license look ups
12:18	c	nursing assistant 1 expiration look up
12:30	c	nursing registry in Raleigh
13:24	c	nursing aide registry of Raleigh
15:00	+	nursing aide registry of Raleigh website
16:06	<	nursing aide registry of Raleigh
19:48	c	north carolina board of nursing information for nursing assistant 1
22:24	c	license look up for nursing assistant 1
24:36	c	license information for nursing assistant 1 expiration
28:30	c	north carolina nursing assistant 1 license information

Pass, et al., “A Picture of Search,” 2007

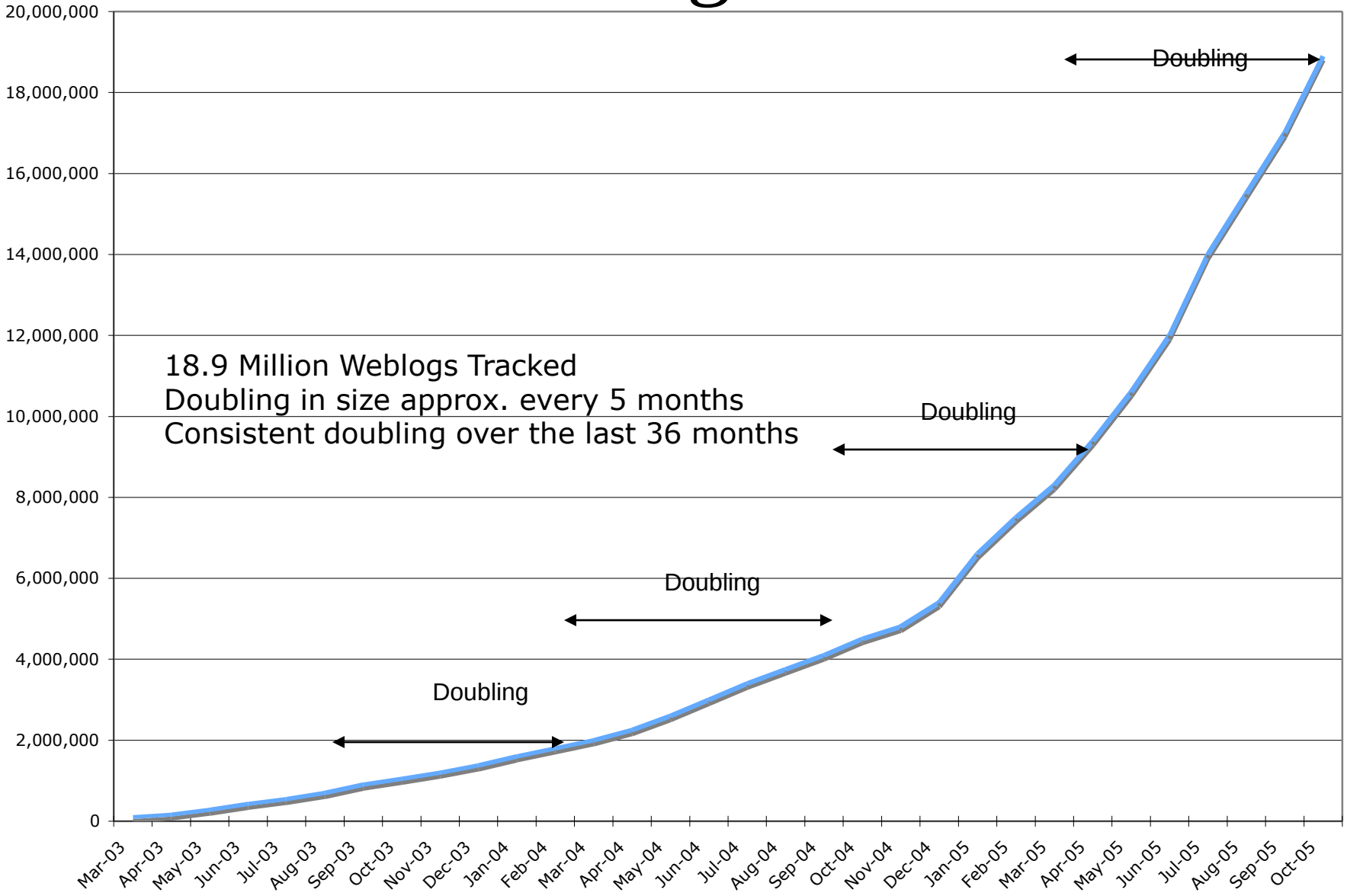
The Tracking Ecosystem

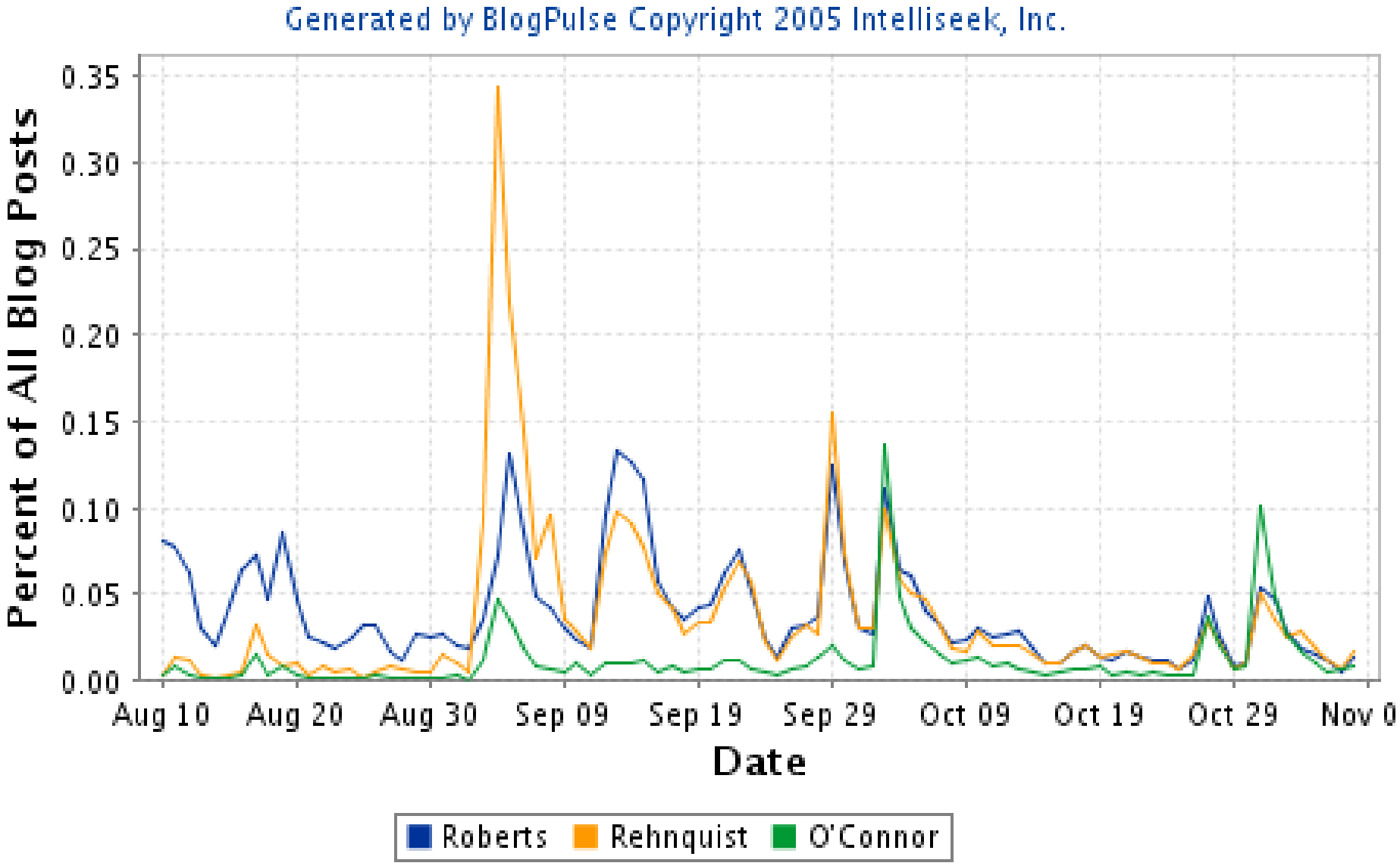


http://wsj.com/wtk

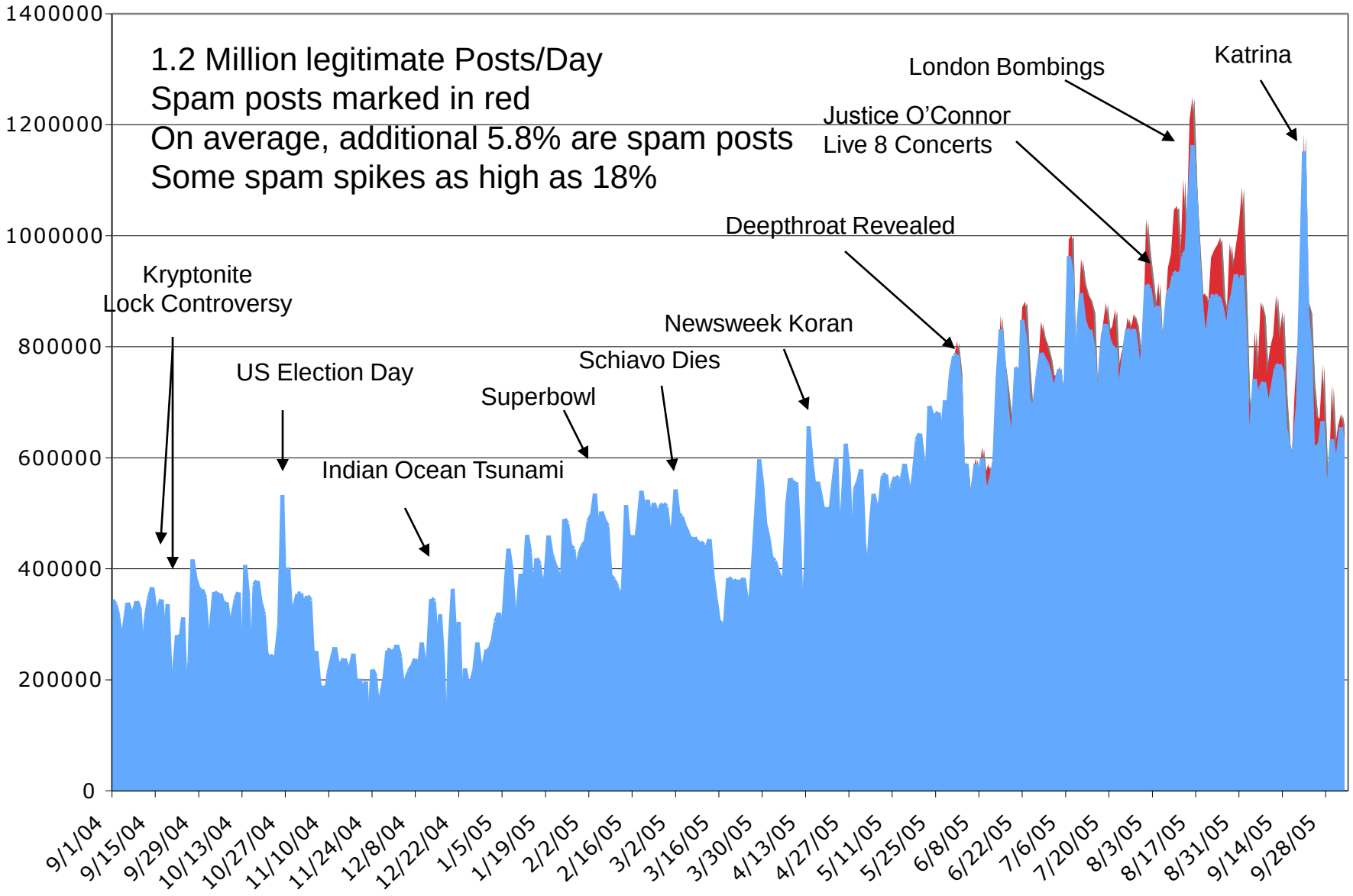
AOL User 4417749

Blogs


20,000,000
18,000,000
16,000,000
14,000,000
12,000,000
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
18.9 Million Weblogs Tracked
Doubling in size approx. every 5 months
Consistent doubling over the last 36 months
Doubling
Doubling
Doubling
Doubling
Mar-03
Apr-03
May-03
Jun-03
Jul-03
Aug-03
Sep-03
Oct-03
Nov-03
Dec-03
Jan-04
Feb-04
Mar-04
Apr-04
May-04
Jun-04
Jul-04
Aug-04
Sep-04
Oct-04
Nov-04
Dec-04
Jan-05
Feb-05
Mar-05
Apr-05
May-05
Jun-05
Jul-05
Aug-05
Sep-05
Oct-05


Generated by BlogPulse Copyright 2005 Intelliseek, Inc.
Percent of All Blog Posts
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Aug 10
Aug 20
Aug 30
Sep 09
Sep 19
Sep 29
Oct 09
Oct 19
Oct 29
Nov 0
Date
Roberts
Rehnquist
O'Connor

Daily Posting Volume



The “Deep Web”

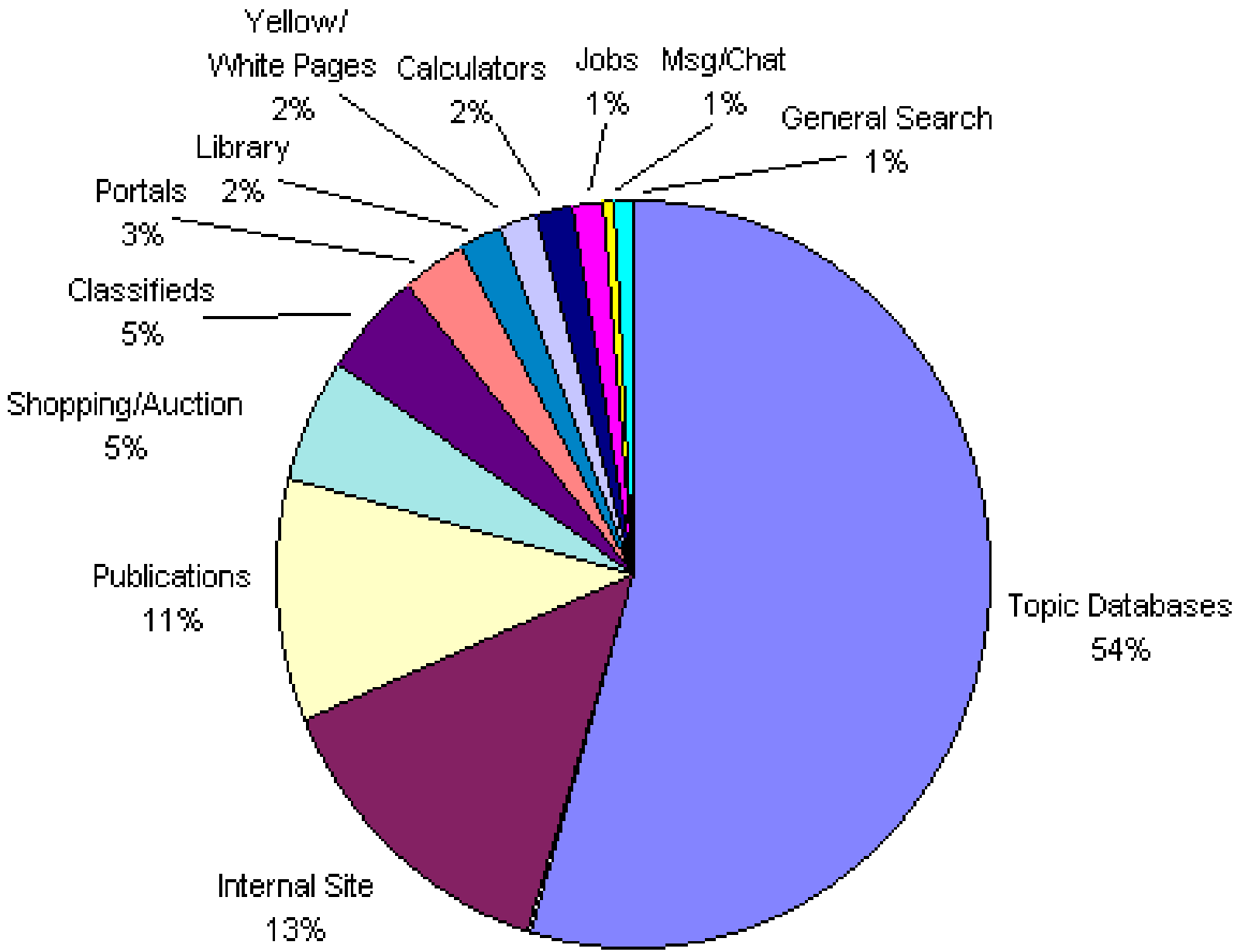
- Dynamic pages, generated from databases
- Not easily discovered using crawling
- Perhaps 400-500 times larger than surface Web
- Fastest growing source of new information

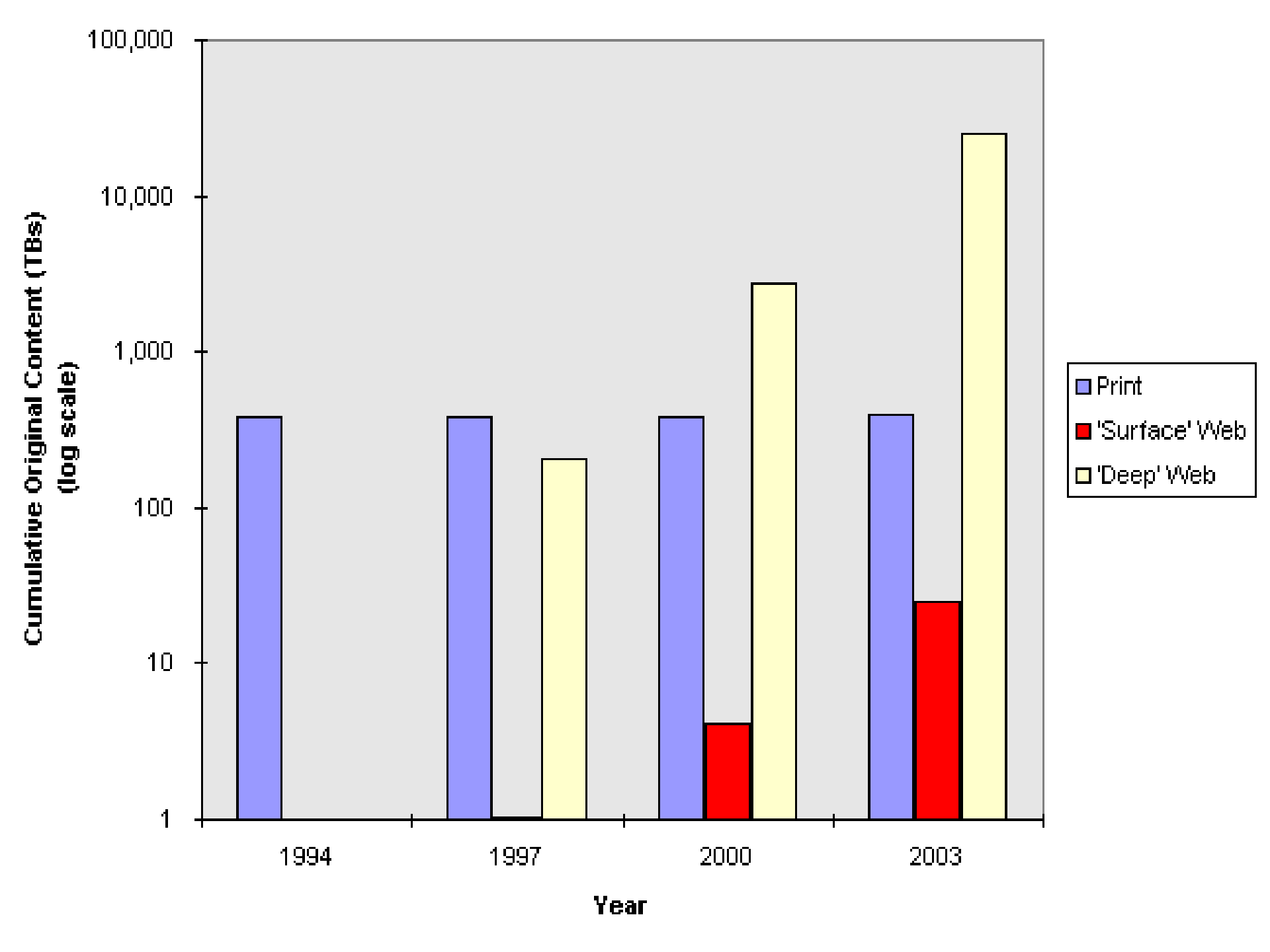
Deep Web

- **60 Deep Sites Exceed Surface Web by 40 Times**

Name	Type	URL	Web Size (GBs)
National Climatic Data Center (NOAA)	Public	http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html	366,000
NASA EOSDIS	Public	http://harp.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html	219,600
National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	http://www.nodc.noaa.gov/, http://www.ngdc.noaa.gov/	32,940
Alexa	Public (partial)	http://www.alexa.com/	15,860
Right-to-Know Network (RTK Net)	Public	http://www.rtk.net/	14,640
MP3.com	Public	http://www.mp3.com/	

Content of the Deep Web




100,000
10,000
1,000
100
10
1
Cumulative Original Content (TBs)
(log scale)
1994
1997
2000
2003
Year
Print
'Surface' Web
'Deep' Web

Semantic Web

- RDF provides the schema for interchange
- Ontologies support automated inference
  - Similar to thesauri supporting human reasoning
- Ontology mapping permits distributed creation
  - This is where the magic happens ☺