What's New on the Web?

The Evolution of the Web from a Search Engine Perspective

Alexandros Ntoulas[1] Junghoo Cho[1] Christopher Olston[2]

[1]University of California Los Angeles
{ntoulas, cho}@cs.ucla.edu

[2]Carnegie Mellon University
olston@cs.cmu.edu

World Wide Web Conference, New York, 18th May 2004

Motivation

- Search engines crawl the Web to build local indexes
- The Web is constantly evolving: pages appear, disappear, change
- Search engines need to update their index to keep up with the evolving Web
- How does the Web evolution affect search engines?

Outline

- Experimental Setup
- What's new on the Web?
  - Birth, death, replacement of pages
  - Creation of new content
  - Link-structure evolution
- How much do persisting pages change?
  - Frequency and degree of change
- Can we predict the changes?

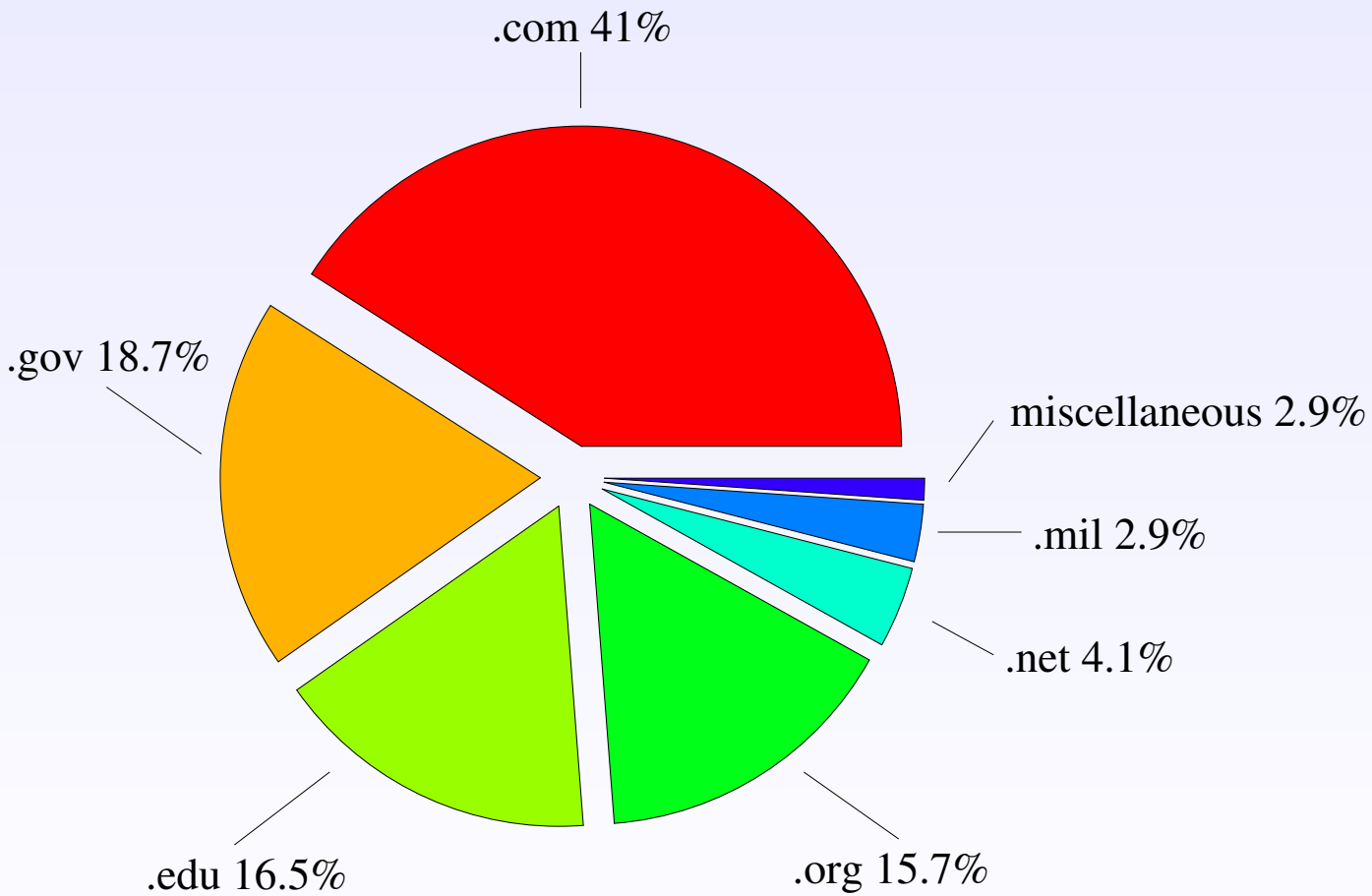
Data selection and crawling

- Picked 5 top-ranked sites from a subset of Google Directory's topical categories
- Crawled pages from 154 sites every week from Oct. 2002 until Oct. 2003
- Crawled in breadth-first manner until we either:
  - downloaded all pages from a site, or
  - reached a limit of 200,000 pages (only 4 such sites)
- Considered a page unavailable after 3 unsuccessful attempts to download it

Data characteristics

- Crawled 4.4 million pages on average every week
- Weekly crawl size: ≈ 65GB
- Total crawl size: ≈ 3.3TB
- Meta-data derived from the crawls (links, shingles etc.): ≈ 4TB

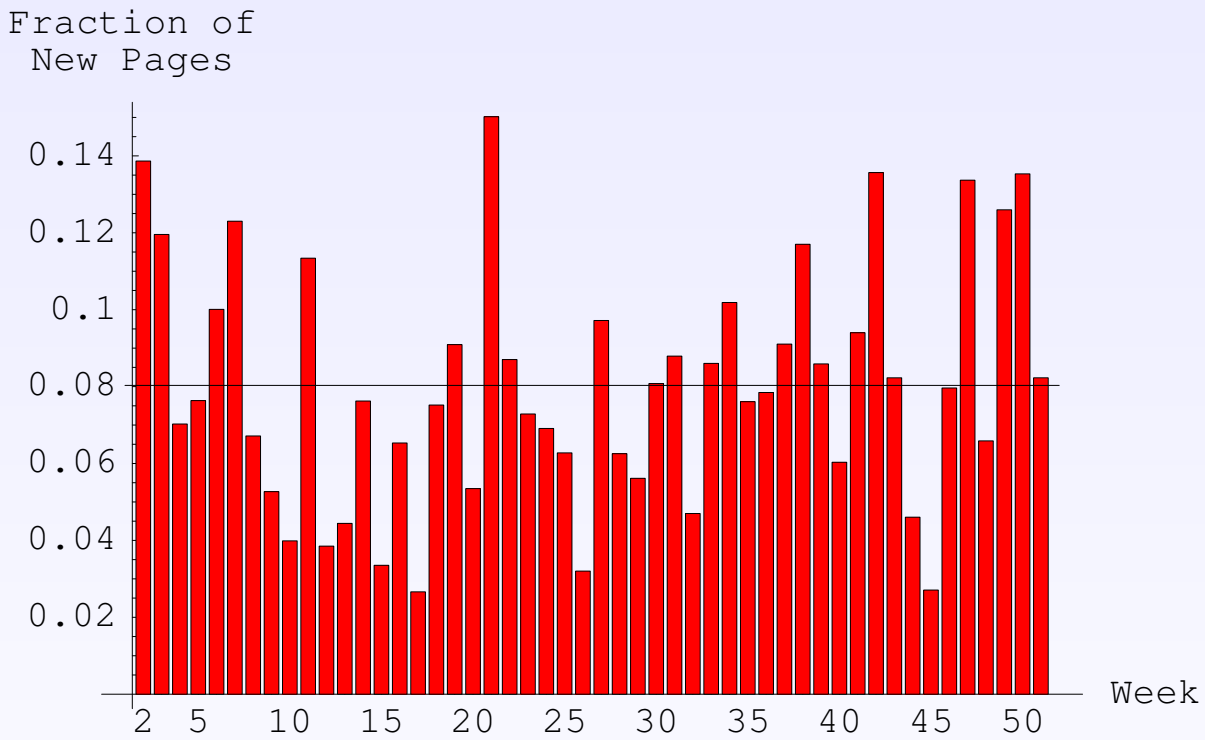
Distribution of pages per domain



Outline

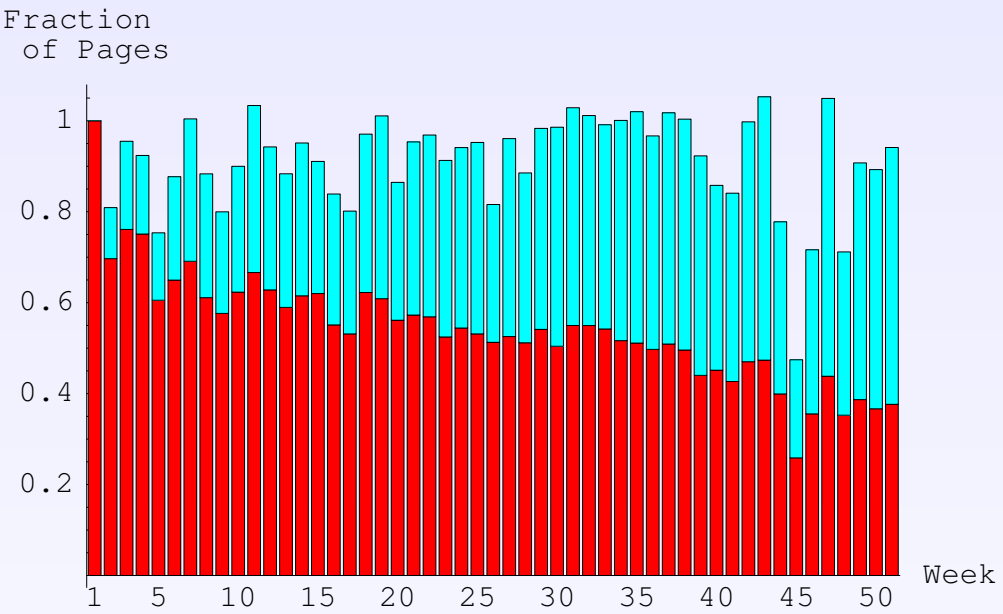
- Experimental Setup
- What's new on the Web?
  - Birth, death, replacement of pages
  - Creation of new content
  - Link-structure evolution
- How much do persisting pages change?
  - Frequency and degree of change
- Can we predict the changes?

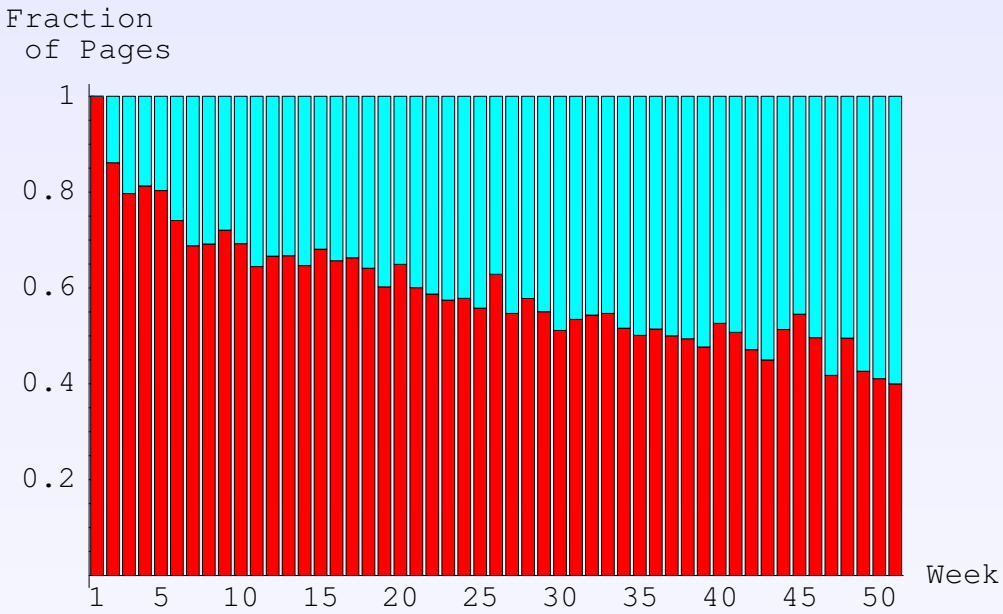
Weekly birth rate of pages



- Average weekly birth rate $\approx 8\%$
- A lot of new pages appear at the end of a calendar month

Birth, death and replacement over time



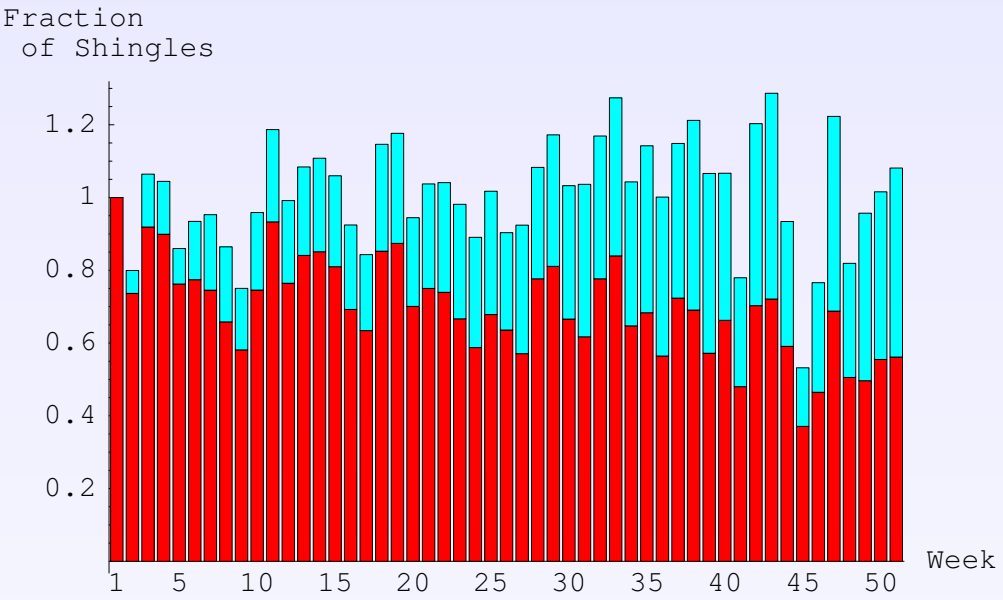


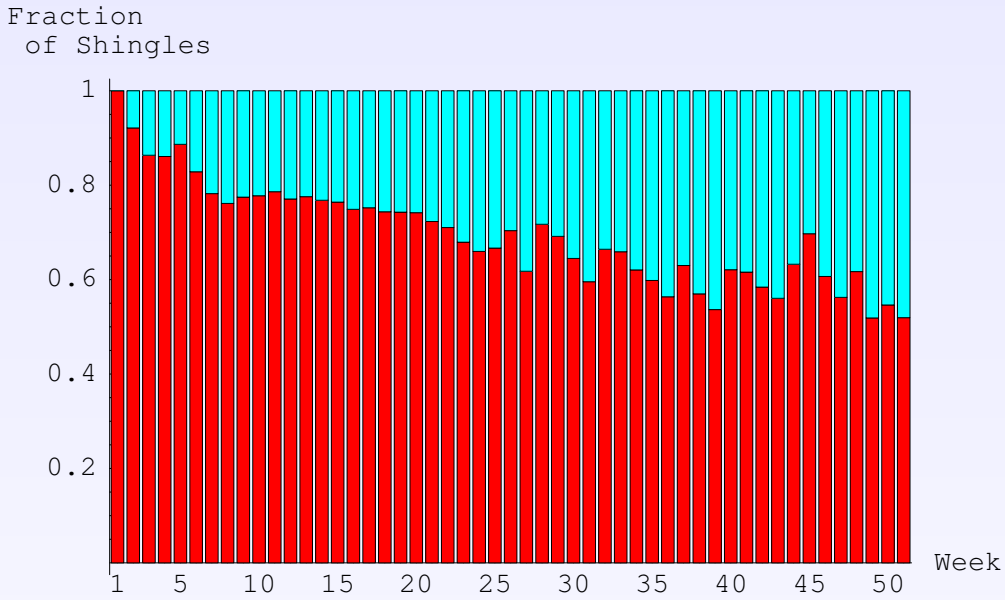
- Total number of pages almost constant
- Half-life of the pages is about 9 months
- Could not find a good fit for the data

Creation of new content

- We discovered the new pages, but how much content is actually new?
- We used the shingling technique [FMNW03]
  - Exclude HTML from the pages
  - Extract the w-shingles from the pages (w=50)
  - Compare how many shingles exist/disappear over time

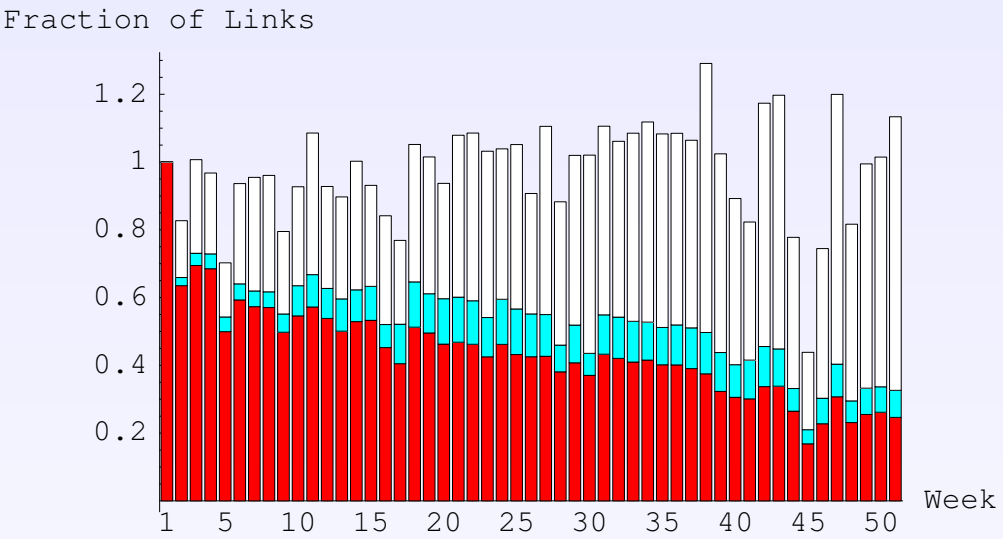
Evolution of content over time

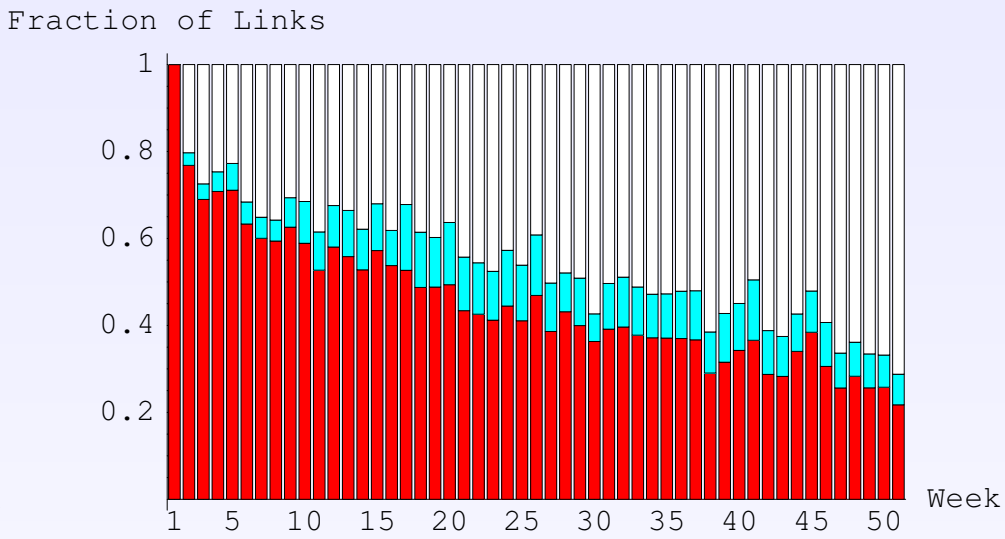




- Shingles are replaced slower than pages
- About 5% of content is new every week
- About 62% of the content in new pages is actually new

Evolution of the link structure





- Link structure is significantly more dynamic than pages
- About 25% of new links every week

Outline

- Experimental Setup
- What's new on the Web?
  - Birth, death, replacement of pages
  - Creation of new content
  - Link-structure evolution
- How much do persisting pages change?
  - Frequency and degree of change
- Can we predict the changes?

Degree of change

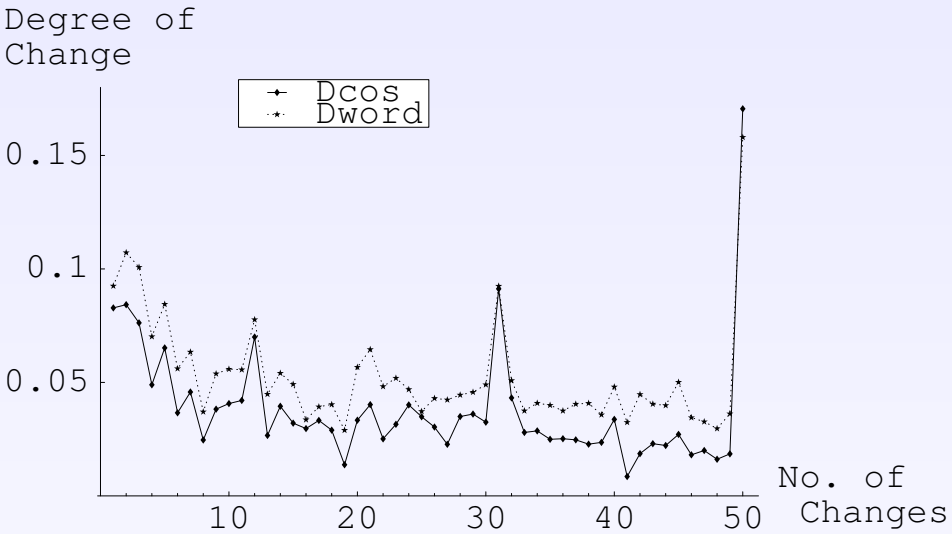
- Search engines care about degree, not presence of change
- We measured degree of change using two metrics
  - **TF.IDF Cosine Distance**

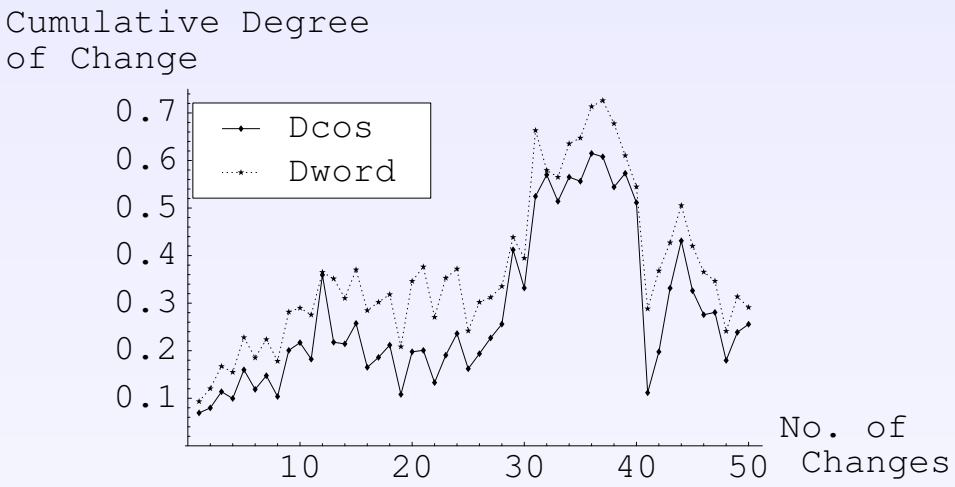
$$D_{cos}(p_1, p_2) = 1 - \frac{v_1 \cdot v_2}{||v_1||_2 ||v_2||_2}$$

  - **Word Distance**

$$D_{word}(p_1, p_2) = 1 - \frac{2 \cdot |\text{common words}|}{|\text{words in } p_1| + |\text{words in } p_2|}$$

Degree of change and frequency of change



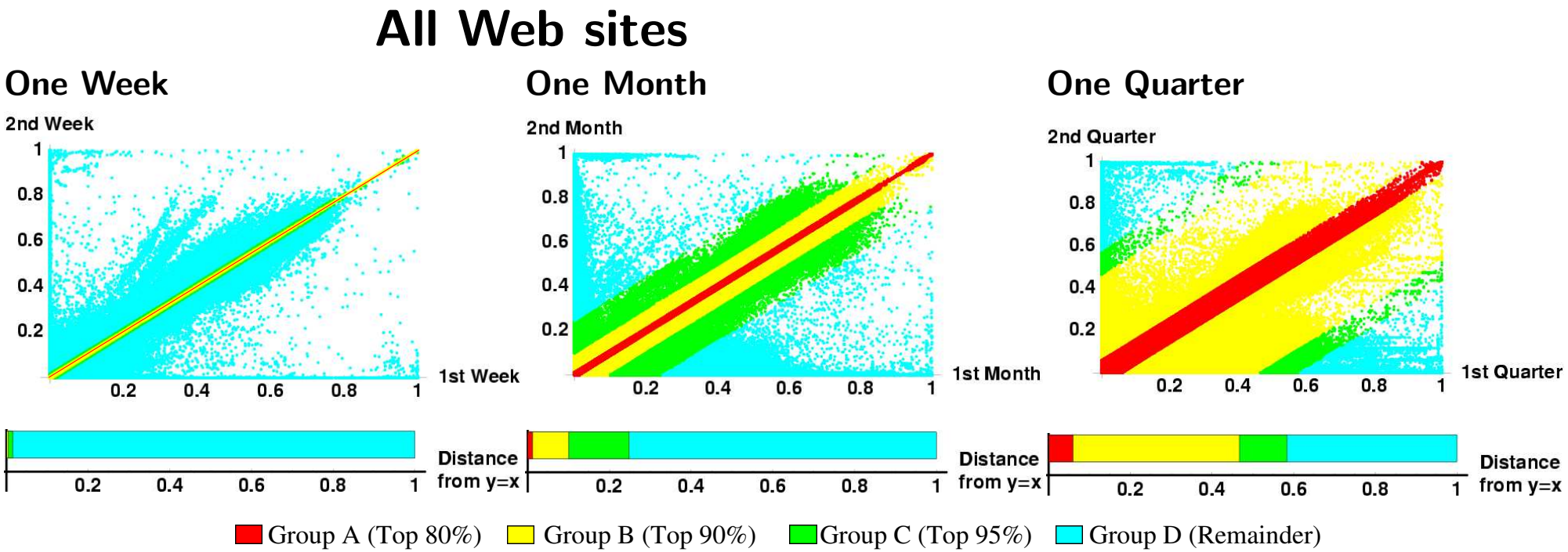


- No correlation between frequency and degree of change
- The same portion of pages changes repeatedly

Outline

- Experimental Setup
- What's new on the Web?
  - Birth, death, replacement of pages
  - Creation of new content
  - Link-structure evolution
- How much do persisting pages change?
  - Frequency and degree of change
- Can we predict the changes?

Can we predict the changes?



- Most pages have highly predictable change patterns
- Predictability decreases with longer intervals
- Predictability can vary per Web site

Conclusions

- Existing pages on the Web are replaced at a high rate
- New pages "borrow" content from existing ones
- Link structure changes faster than pages
- Pages that persist demonstrate minor changes
- The past degree of change is a good predictor for future degree of change

Related work

- Others have studied Web evolution [FMNW03, LWP+01, BC00].
- Theoretical work on the Web graph evolution [BKM+00, CDK+99, KRR+00]
- Our study:
  - spans over a longer period
  - focuses on change metrics which are important for search engines
  - studies link evolution
  - examines predictability of changes

References


B. E. Brewington and G. Cybenko.
How dynamic is the web?
In *Proceedings of the Ninth International World Wide Web Conference*, Amsterdam, The Netherlands, May 2000.


A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener.
Graph structure in the web.
In *Proceedings of the Nineth International World Wide Web Conference*, Amsterdam, Netherlands, May 2000.


S. Chakrabarti, B. E. Dom, S. Ravi Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg.
Mining the Web's link structure.
Computer, 32(8):60–67, 1999.


D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener.
A large-scale study of the evolution of web pages.
In *Proceedings of the Twelfth International World Wide Web Conference*, Budapest, Hungary, May 2003.


R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal.
Stochastic models for the web graph.
In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2000.

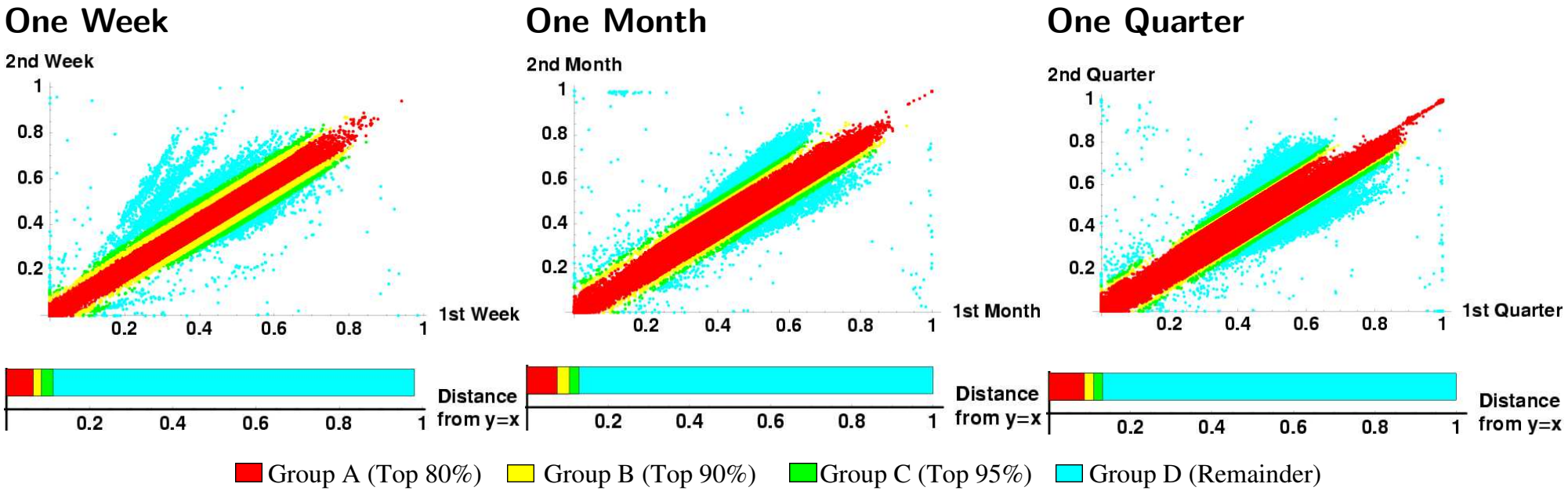

L. Lim, M. Wang, S. Padmanabhan, J. Scott Vitter, and R. C. Agarwal.
Characterizing web document change.
In *Proceedings of the Second International Conference on Advances in Web-Age Information Management*, pages 133–144. Springer-Verlag, 2001.

Thank you

Questions?

Can we predict the changes?

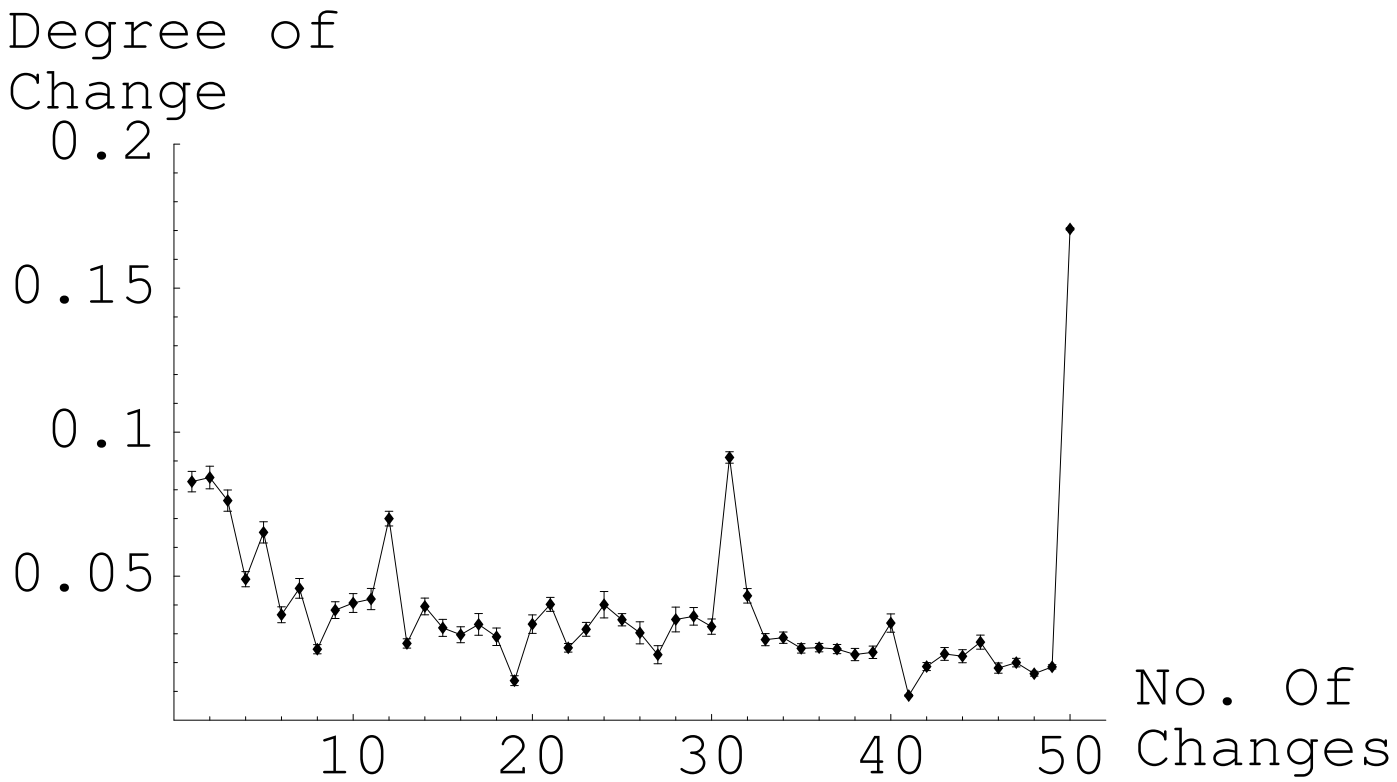
Web site `www.eonline.com`



- Less predictable than overall pages
- The ability to predict can vary per Web site

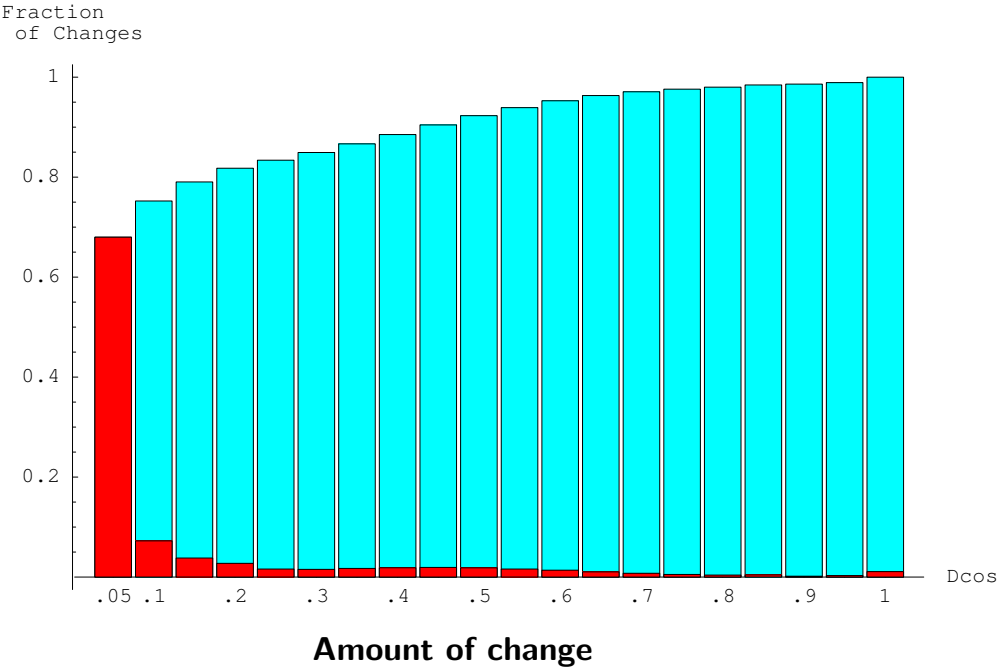
Degree of change and frequency of change

99% confidence interval



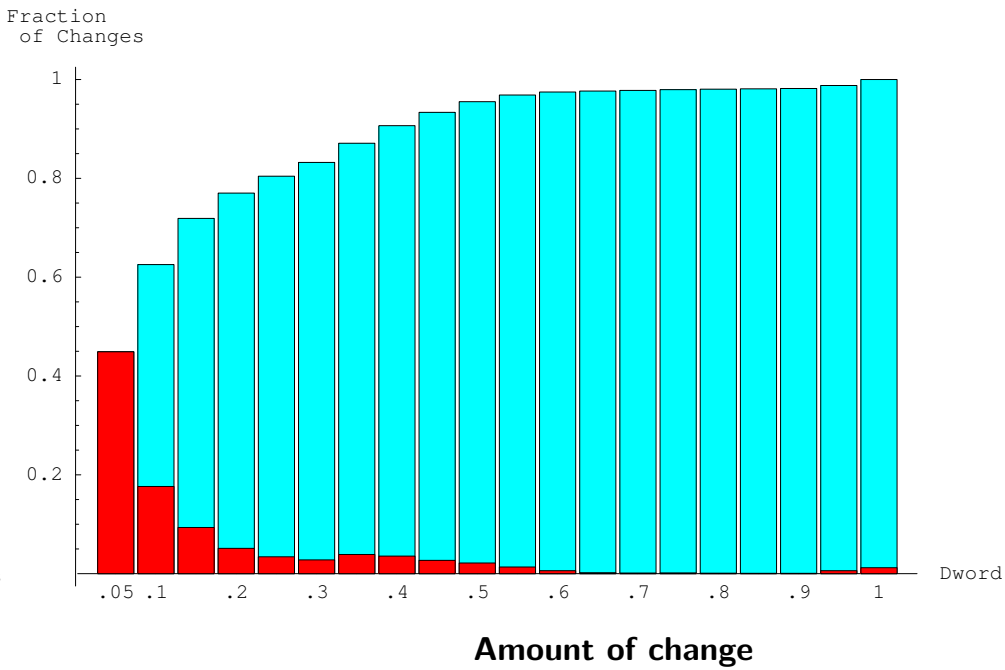
Degree of change

TF.IDF Cosine Distance



Amount of change

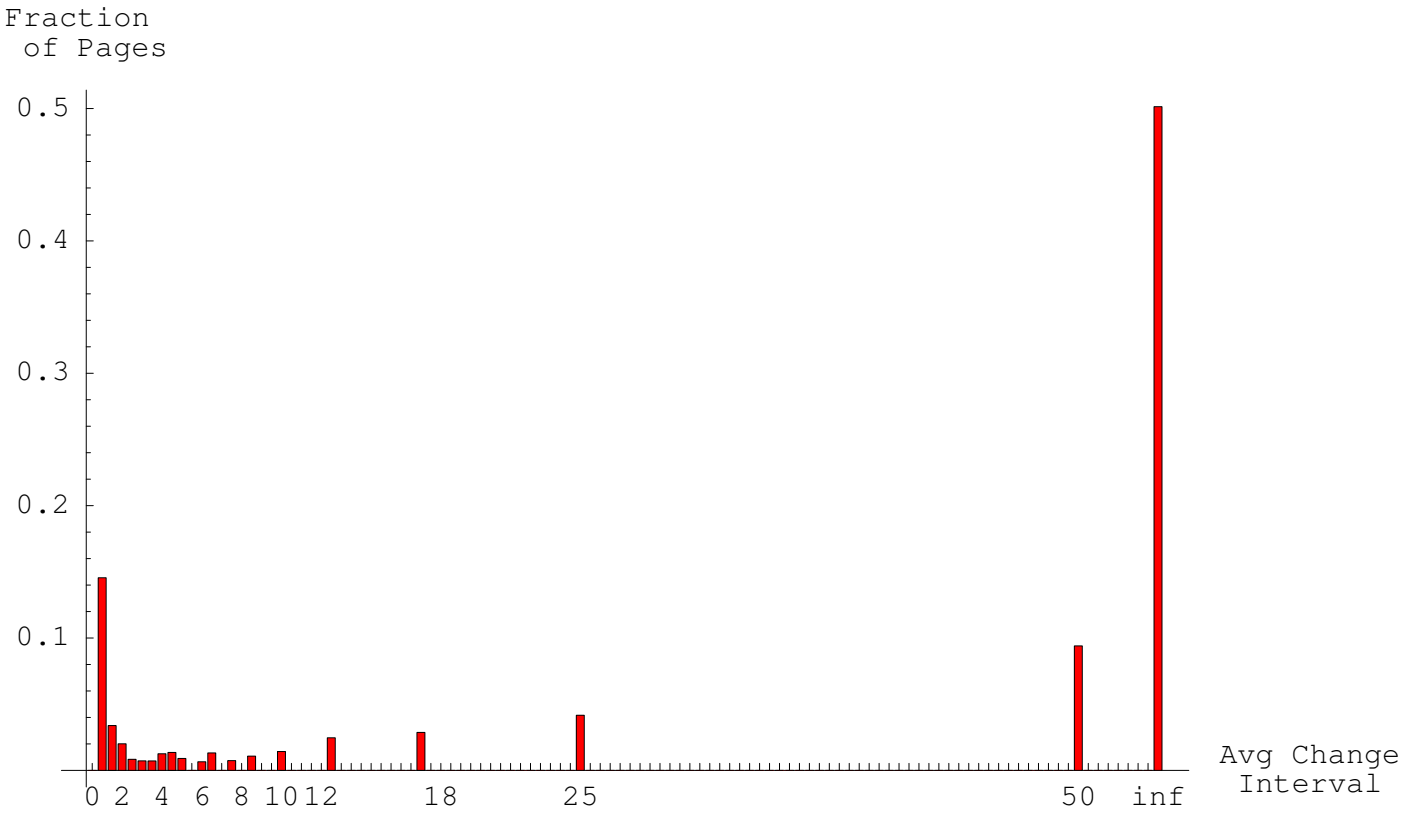
Word Distance



Amount of change

- Most of the changes are minor under both metrics
- A significant number of changes is due to low-weight words

Frequency of change



- Measured change frequency based on "simple" definition of change
- Most pages either change often or not at all