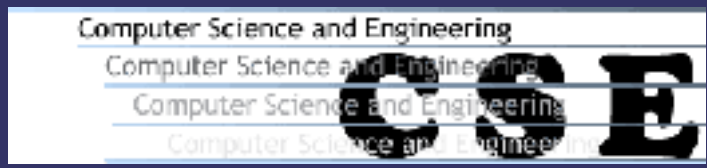
OVERVIEW:

WWW Search Engines

Prof. Brian D. Davison
davison@cse.lehigh.edu
http://www.cse.lehigh.edu/~brian/



Two Advertisements :-)

- This talk has been brought to you today by....
  - CSE Department Seminar Thursday 4pm PL466
    - Craig Nevill-Manning
    - Senior Research Scientist, Google
  - WWW Search Engines course
    - (at least a year from now)
    - http://www.cse.lehigh.edu/~brian/course/searchengines/

WWW Search Engines

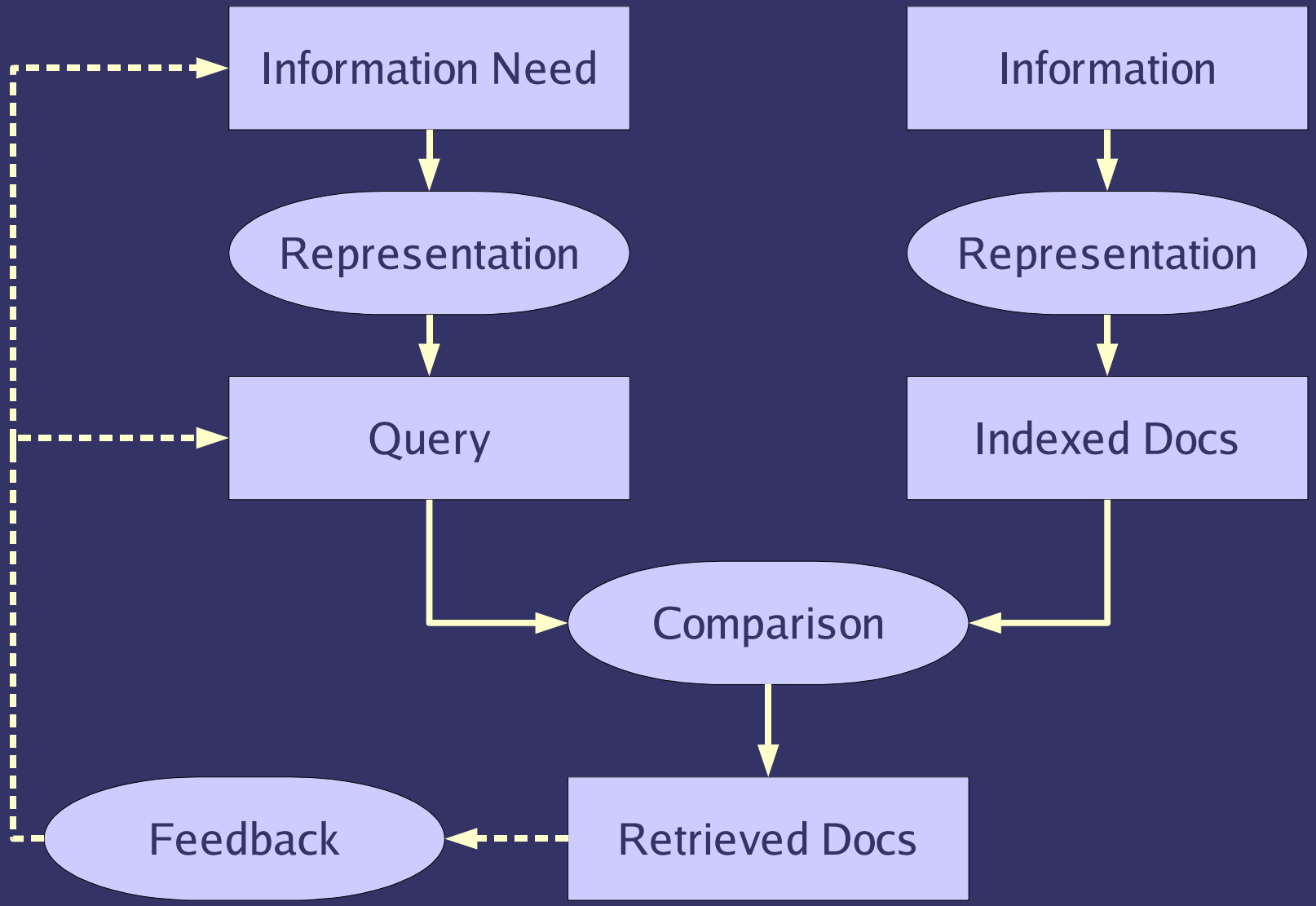
- What are they?
- Examples? Which ones do you use?
- How do they work?
- How can they be improved?
- How did they start?
- What are they really supposed to do?

Commercial Web Search Engines

What do SE designers consider?

- Information Retrieval fundamentals
  - Indexing, retrieval and ranking methods
  - Performance evaluation
- What does the searcher really want?
- Web-specific content
  - Properties of hypertext
  - Web crawling
  - Link analysis
- Scaling to large num. of docs, queries (cheaply)

Overview of IR Task



Motivation

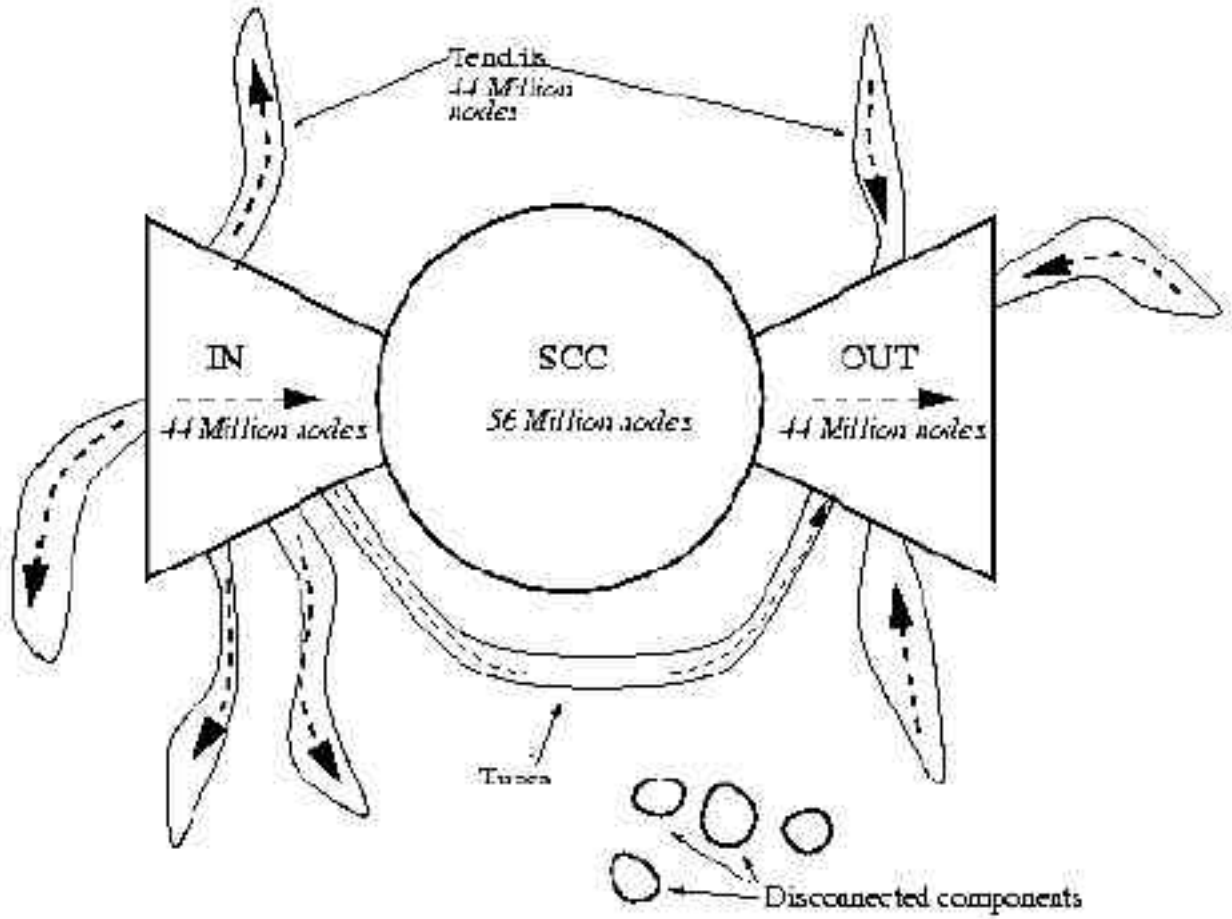
- Data retrieval (i.e., in databases)
  - Which docs contain a set of keywords?
  - Problem is well-defined.
  - A single mistake (including the wrong doc or excluding the right doc) constitutes an error!
- Information retrieval
  - Semantics of what docs are needed are often not well-specified.
  - Small errors are tolerated.
- IR System
  - Interprets the contents of info. objects (queries and docs).
  - Generates a ranking to reflect expected relevance.

Searching on the WWW

- How is the Web different from other IR systems?
  - Content
    - Universal data repository
      - Large number of documents
    - No central editorial control
      - Varied kind and quality of documents
      - May try to manipulate search engine rankings
  - Searchers
    - Low-cost universal access
      - Varied searchers with varying interests
    - Lots of queries, some very popular
    - Queries very short

Abundance and authority crisis

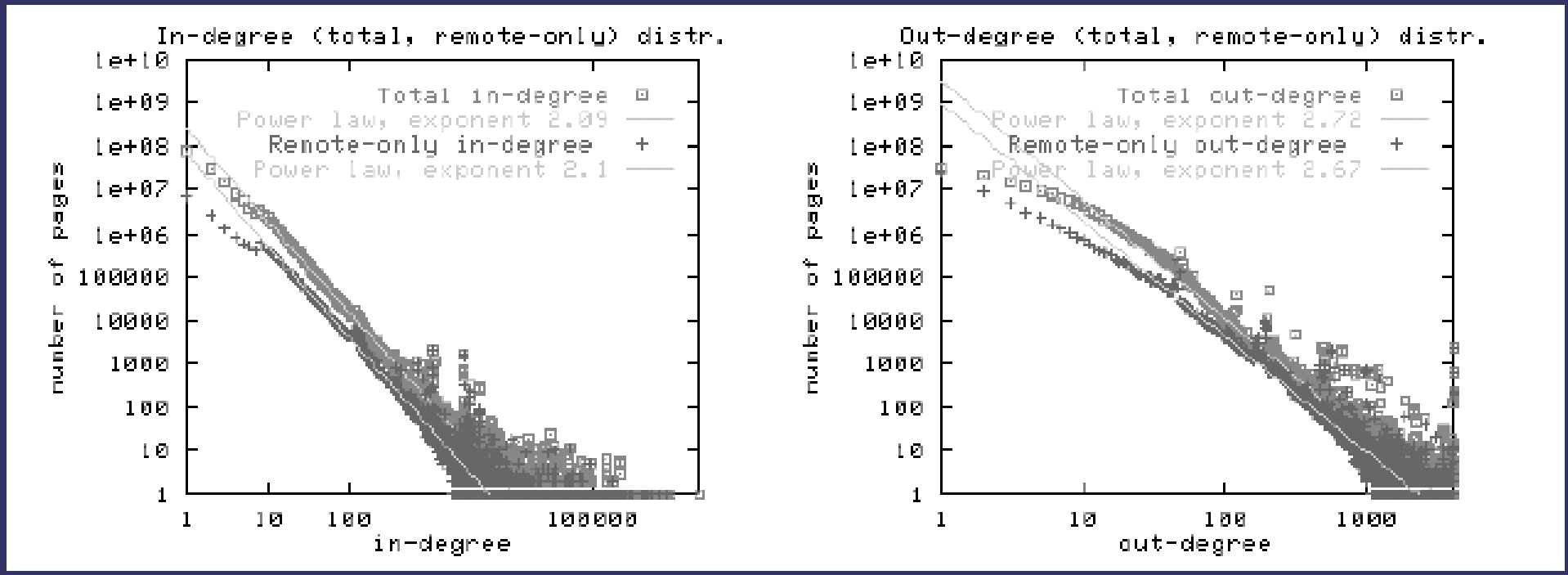
- Liberal and informal culture of content generation and dissemination.
- Redundancy and non-standard form and content.
- Millions of qualifying pages for most broad queries.
  - Example: java or kayaking
- No authoritative information about the reliability of a site.



Region:	SCC	IN	OUT	TENDRILS	Disconnected	TOTAL
Size:	56,463,993	43,343,168	43,166,185	43,797,944	16,777,756	203,549,046

The Web is a bow-tie

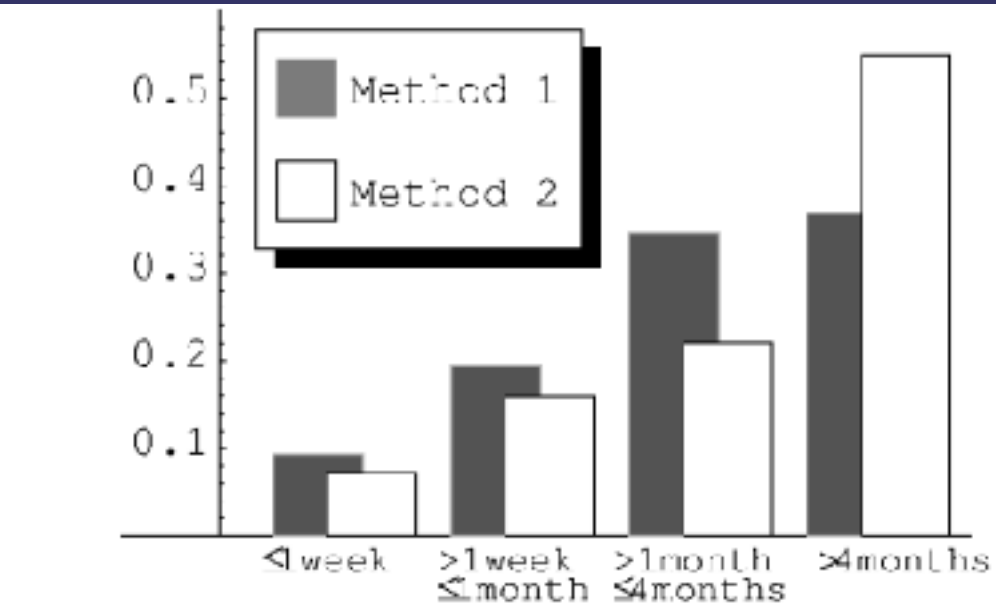
Link Distribution



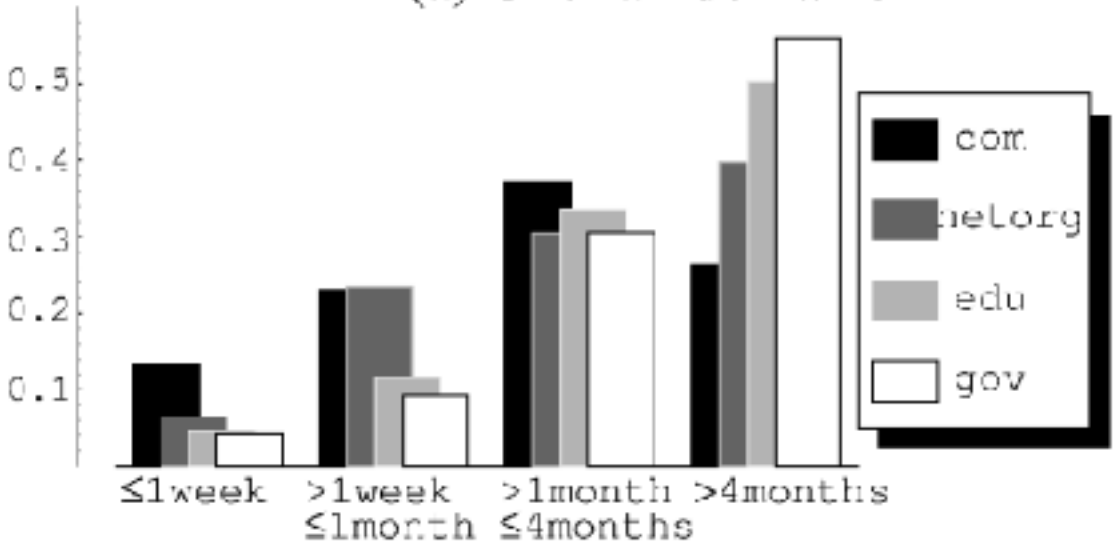
The in- and out-degree of Web nodes closely follow power-law distributions.

Web Pages Change:

- Typical lifespan is a few months.
- A significant number of pages are accessible for a long period.
- More than 50% of EDU/GOV pages were accessible for more than 4 months.

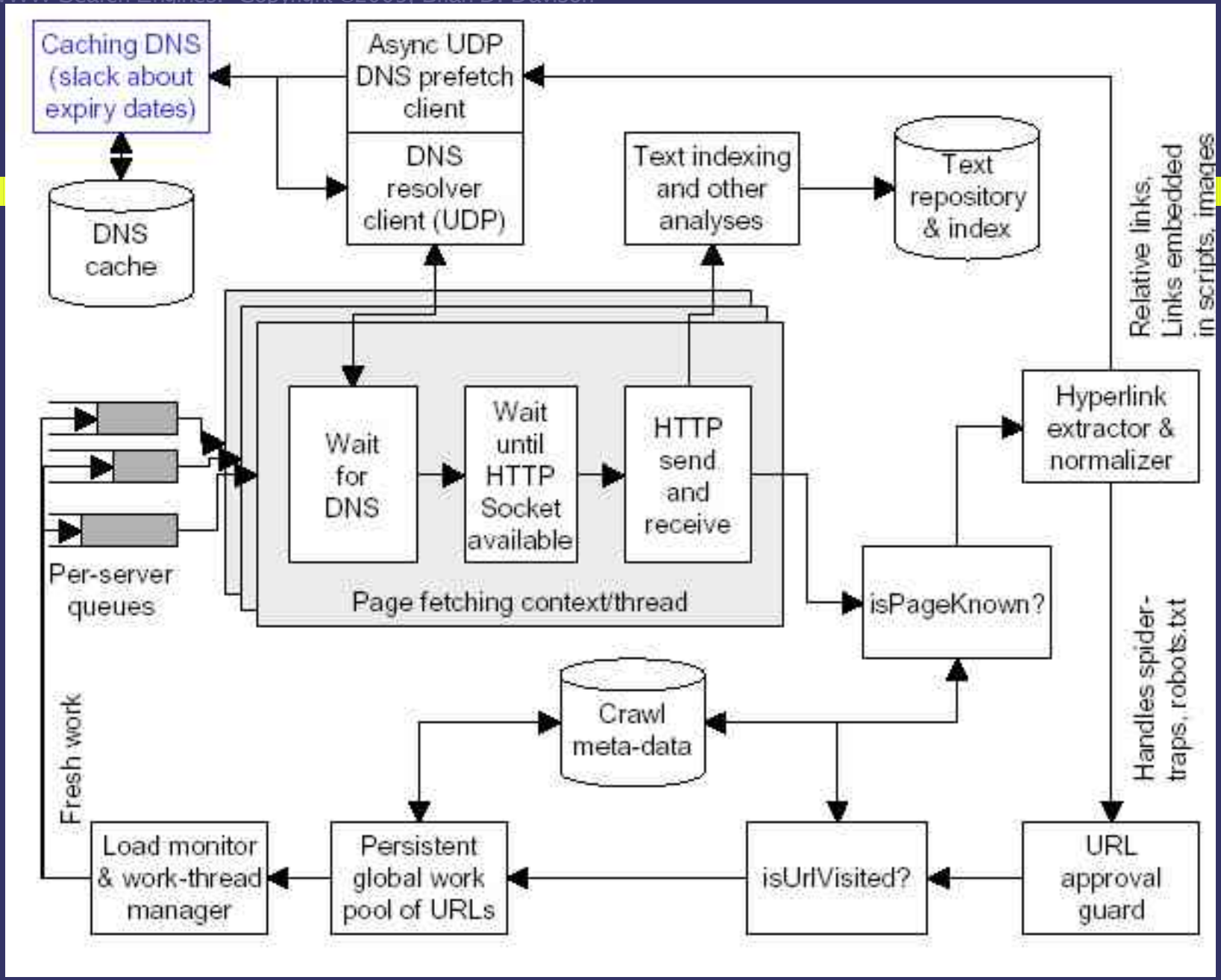


(a) Over all domains



(b) For each domain

Figure 4: Percentage of pages with given visible lifespan



Typical anatomy of a large-scale crawler.

Relationships

- There are many examples of linkage relationships in various domains, including social networks and term co-occurence.
  - Six degrees of separation
    - "Small world phenomena"
    - Six degrees of Kevin Bacon
- Scientific documents, like WWW pages, do not exist alone – they cite other documents.
  - Unlike co-occurence, these links are directed.

Bibliometrics

- Citation analysis in scientific documents.
- Papers cite other papers that have previously been published.
- Highly cited papers are deemed important.
  - In-degree is the number of citation to a paper.
- Broad survey papers cite many other related works.
- Cocitation can find related documents.
  - If two documents are cited the by same papers, the two docs are likely related.
- So can bibliographic coupling (when two documents cite the same papers).

WWW Link Analysis

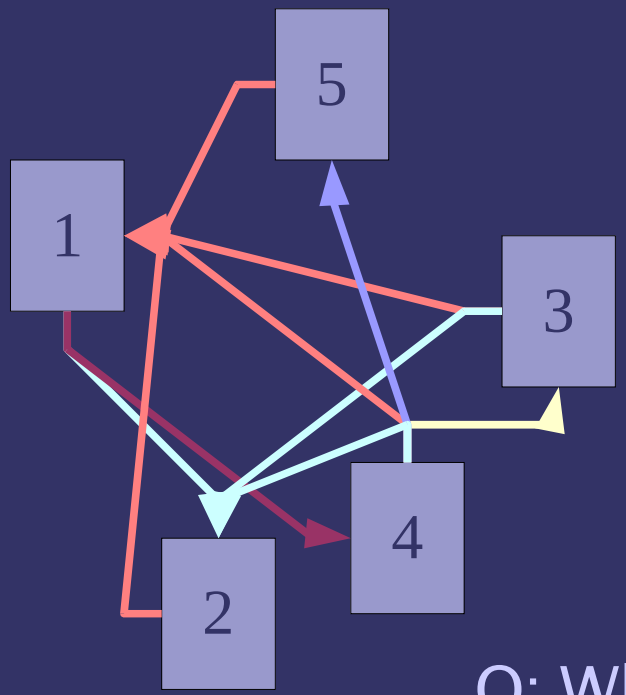
- Links on the Web can be interpreted as more than just navigation.
  - Often, as in document citation, a link points to more information or the source of presented facts.
  - In general, a link connotes authority or quality to the destination (e.g., an endorsement).
- A page with n outgoing links has out-degree n.
- A page with m incoming links has in-degree m.
- The simplest form of link analysis is just to rank pages based on their indegree == popularity.

Web as a Graph/Matrix

- Typically the Web is thought of as a directed graph:
  - The nodes represent pages.
  - Directed edges represent links between pages.
- But represented in a page-page adjacency matrix.
  - Given a graph of n nodes, the adjacency matrix A is an nxn matrix with A(p,q) = 1 iff p->q

Example Matrix

- In this sample 5x5 matrix, document 1 has high in-degree (4), compared to the others which have in-degrees of only 1 or 2.



Docs	To 1	2	3	4	5
From 1	0	1	0	1	0
2	1	0	0	0	0
3	1	1	0	0	0
4	1	1	1	0	1
5	1	0	0	0	0

Q: Why is this a Web graph and not a citation graph?

Authority/Quality/Importance

- Popularity is a very simple measure of authority or quality or importance or standing or influence.
- Imagine a highly cited (important) paper A cites another paper B. Similarly, C cites D.
- If in-degree(B) == in-degree(D), should they have the same measure of importance?
- Most link analysis techniques would give a higher measure of importance to B.
  - However, some form of popularity may be sufficient to find (just) the most important pages.

Link-based Ranking Strategies

- Leverage the
  - "Abundance problems" inherent in broad queries
- Google's PageRank [Brin and Page WWW7]
  - Measure of prestige with every page on web
- HITS: Hyperlink Induced Topic Search [Jon Kleinberg '98]
  - Use query to select a sub-graph from the Web.
  - Identify "hubs" and "authorities" in the sub-graph

PageRank

"The PageRank Citation Ranking: Bringing Order to the Web," Page *et al.*, 1998.

- PageRank provides a query-independent ranking of the pages in a graph.
- Links from more important pages are more valuable than links from less important pages.
  - E.g. a link from Yahoo is better than your home page.
- PageRank can calculate "importance" from the link structure.
- A form of PageRank is the basis of the ranking used by Google in conjunction with other factors.

PageRank continued

- Simple iterative formulation:
  - Let Fu=set of forward links from u, Bu=backlink set.
  - The rank R(u) = c * sum of R(v)/|Fv| for all v->u
    - [c is a normalization factor so that the total rank of all web pages is constant.]
  - Note that the rank of u is evenly divided among each of its outgoing links.
  - In this case, the matrix A(p,q) = 1/|Fp] when p->q
  - The simple equation R(u) is recursively defined, but will converge (with almost any starting rank assignments).

Problems

- Some pages have no forward links
  - especially those that you have not yet visited
  - links to such pages are called 'dangling links'
- Pages can reference each other, forming loops
  - Loops act as 'sinks' for the rank values
- Solution is to have a decay factor and a rank source involved
  - R' = c(A + Ex1)R' where 1 is the vector of all ones
  - R' is an eigenvector of (A+Ex1)

PageRank Surfing

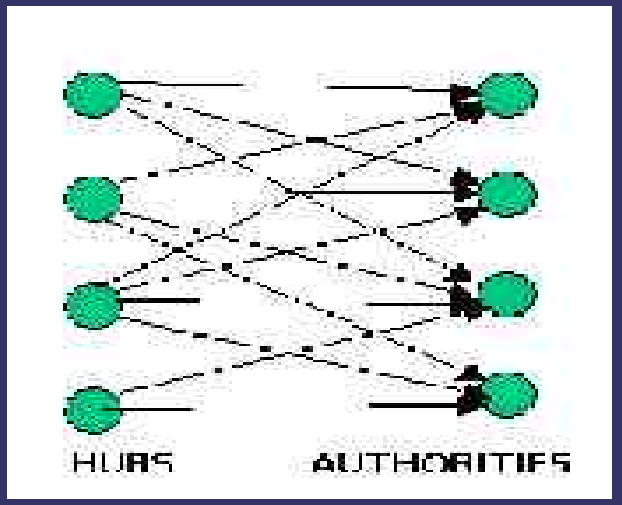
- PageRank measures the likelihood of a random surfer to be on a particular page.
  - R(u) = d + (1-d)(sum of R(v)/|Fv|) for all v->u
  - With probability d, a surfer jumps to a random page; with probability (1-d), surfer follows a random outlink.
  - [Dangling links are removed.]
- PageRank is equiv. to the principal eigenvector of the normalized link matrix of the Web.
  - Alternatively, it is the stationary probability of the transition matrix of the Markov chain.

PageRank architecture at Google

- Ranking of pages more important than exact values
- Convergence of page ranks in 52 iterations for a crawl with 322 million links.
- Pre-compute and store the PageRank of each page.
  - PageRank independent of any query or textual content.
- Ranking scheme combines PageRank with textual match
  - Unpublished
  - Many empirical parameters, human effort and regression testing.
  - Criticism: Ad-hoc coupling and decoupling between relevance and prestige

Hubs and Authorities

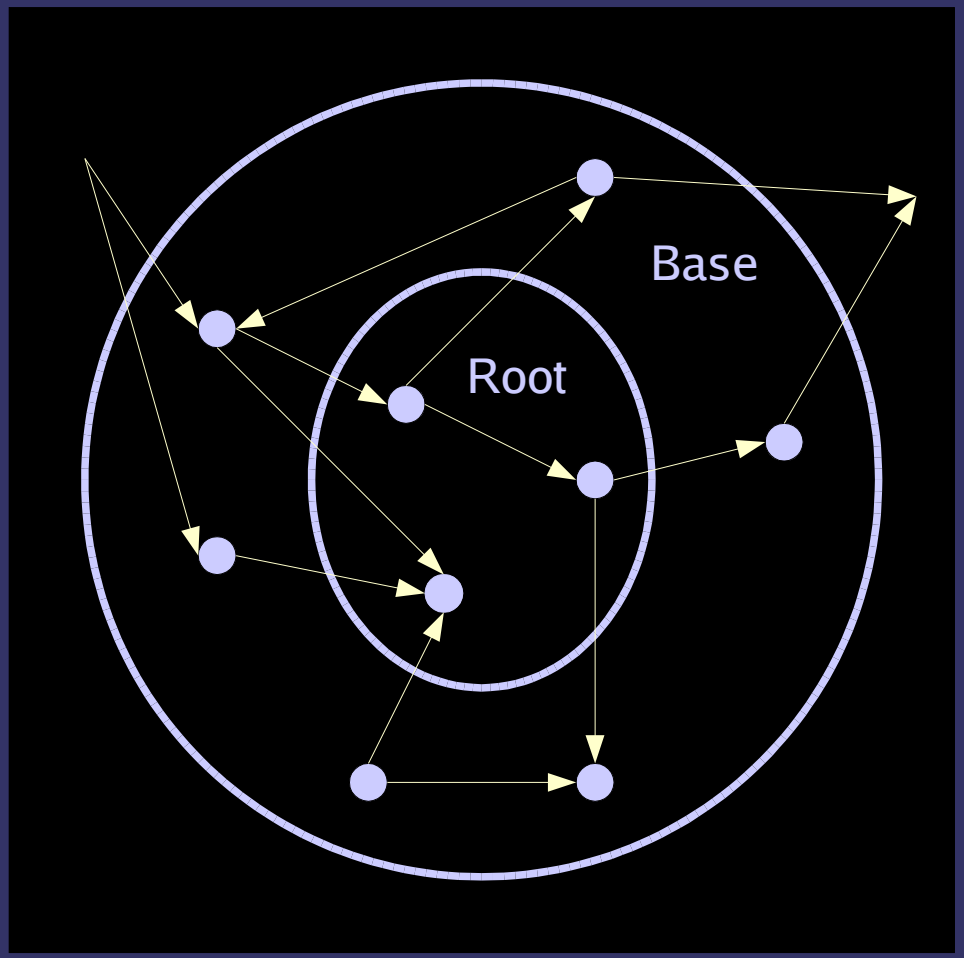
- In PageRank, authorities are based on which and how many pages point to you.
  - This corresponds to one kind of value.
- Another kind is a page containing good outlinks.
- In Kleinberg's HITS (Hyperlinked Induced Topic Search) algorithm, both types of pages are used:
  - Good hubs are pages that point to good (related) authorities.
  - Good authorities are pages pointed to by good hubs.



Calculating HITS

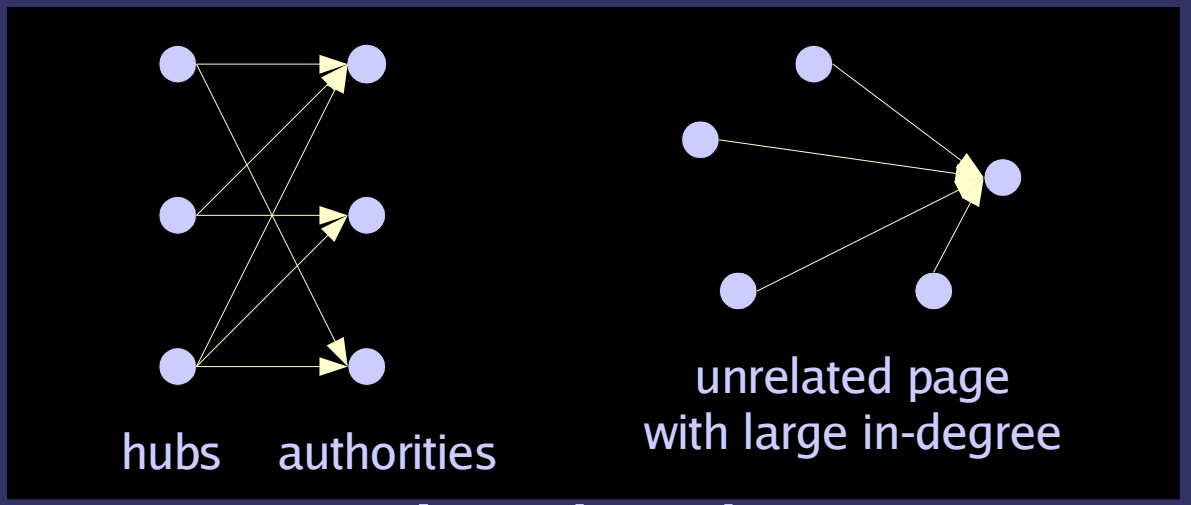
"Authoritative Sources in a Hyperlinked Environment," J. Kleinberg, JACM, 1999.

- HITS is also query-specific.
  - Use query to collect a root set of pages from a (text-based) search engine.
  - Expand the root set into a base set by adding pages linked to and from the root set.



Calculating HITS cont.

- Avoids unrelated pages with large in-degree.



- Each page is initialized with a non-negative authority weight and hub weight.
  - H(u) = sum of A(v) where u->v
  - A(u) = sum of H(v) where v->u
  - Normalize, repeat process until weights converge.

Matrix View of HITS

- Update rules translate to
  - $a=A^T h$ and $h=Aa$
  - $a=(A^T A)a$ and $h=(AA^T)h$
  - (again, where A is the page adjacency matrix.)
- Thus, authority and hub vectors are principal eigenvectors of the $A^T A$ and AA^T matrices, repectively.
- Additional nonprincipal eigenvectors can be calculated and interpreted.

Additional Eigenvectors

- Nonprincipal eigenvectors will have both positive and negative entries.
- Often, the highly positive entries will correspond to a cluster of pages.
- Likewise, the highly negative entries will correspond to a different cluster.
- Typically the two clusters will not be tightly intertwined.

Jaguar Example

- Authority principal eigenvector is primarily about the Atari product.
- In the positive end of the 2nd nonprincipal eigenvector, the pages are primarily about the Jacksonville Jaguars.
- In the positive end of the 3rd nonprincipal eigenvector, the pages are primarily about the car.

Abortion Example

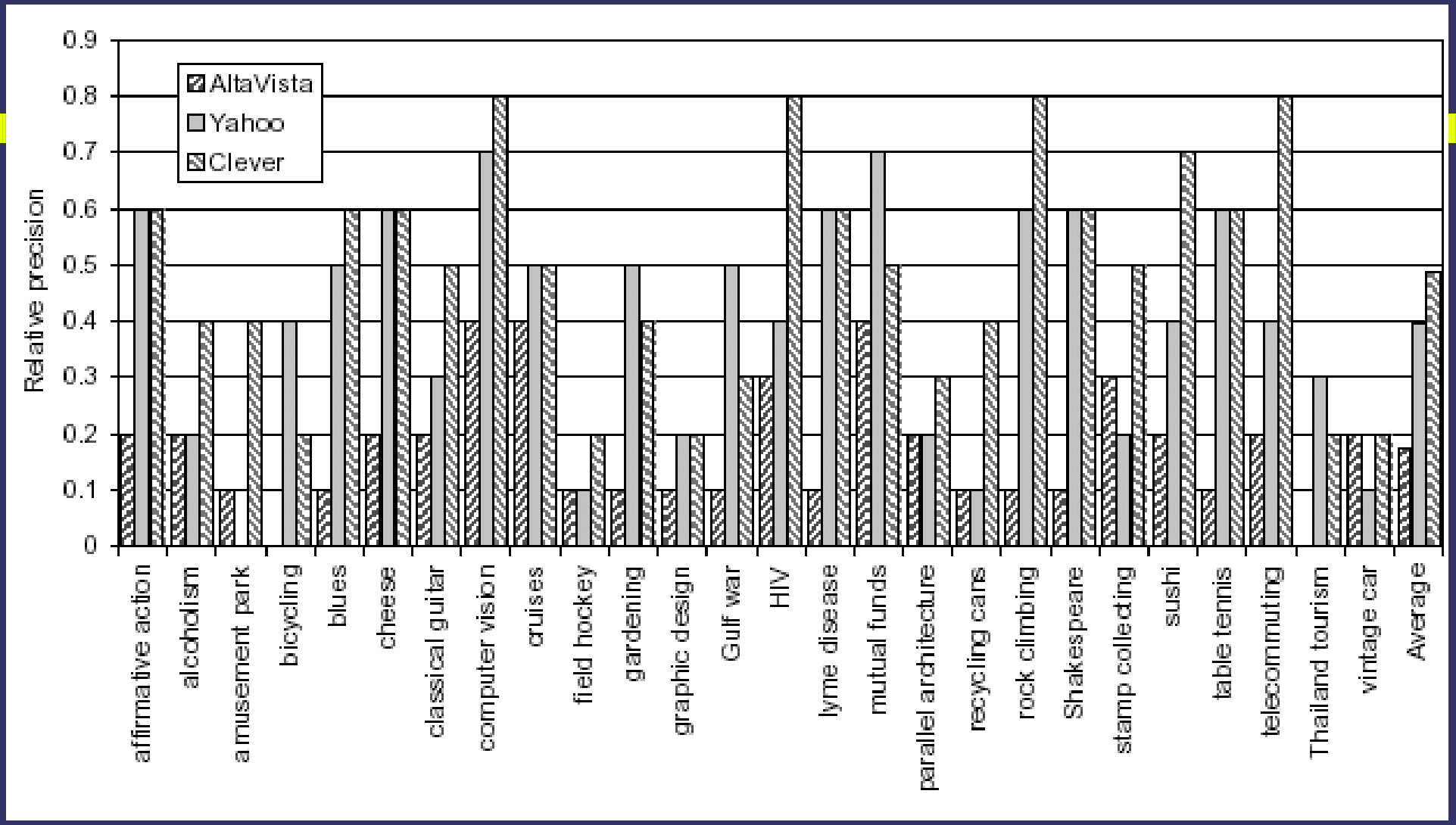
- Authority 2nd nonprincipal eigenvector postive end contains pages primarily supporting pro-choice.
- The negative end contains pages primarily supporting pro-life.
- Other eigenvectors might have other clusters (e.g., religious viewpoints).

Difficulties

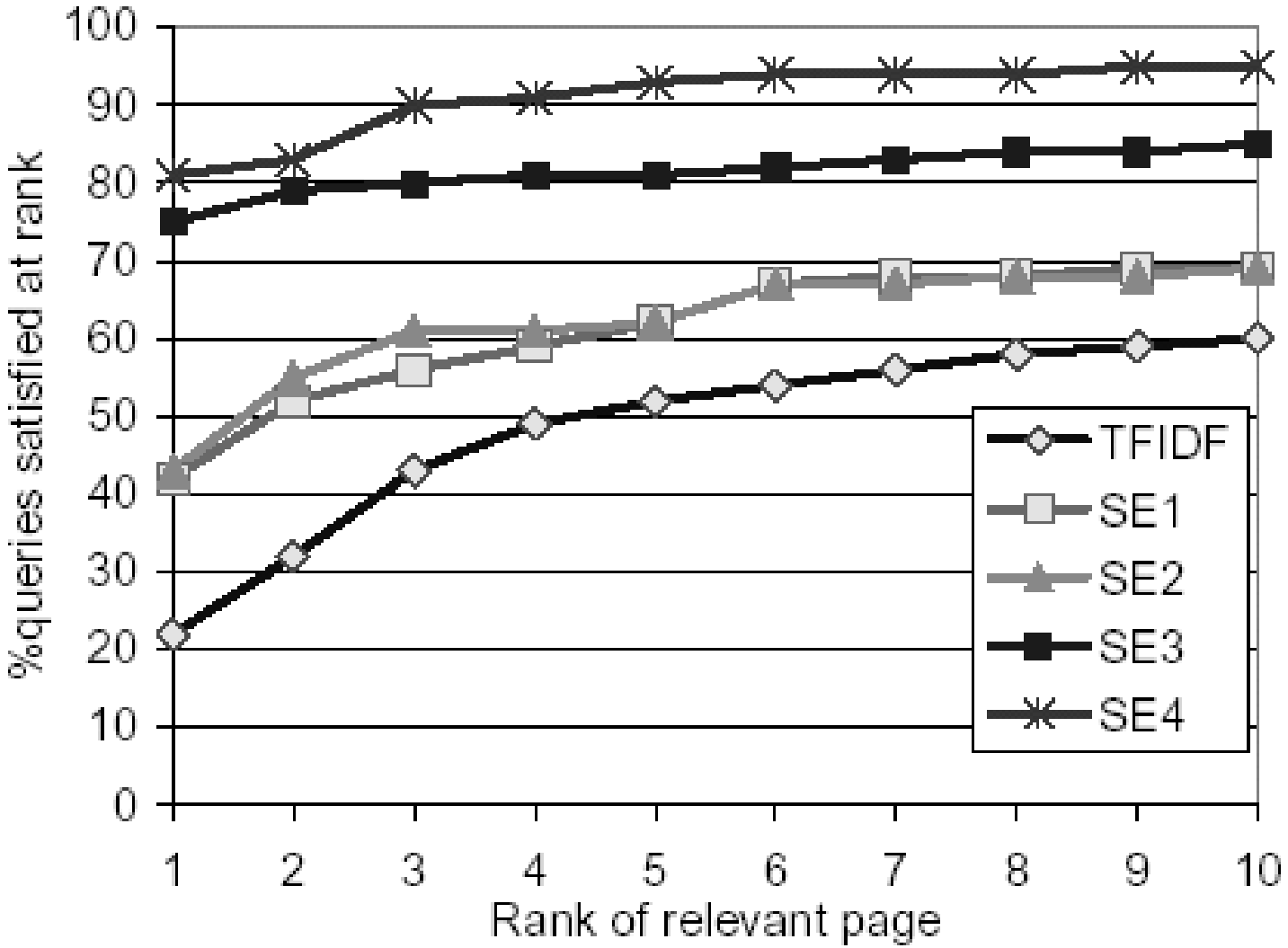
- The query may not be sufficiently "broad."
  - In this case there will not be enough highly relevant pages in the base set to extract a sufficiently dense subgraph of relevant hubs and authorities.
- When this occurs, the collection will often represent a broader topic, and the results will reflect a *diffused* version of the initial query.
- Example: "WWW conferences" -> WWW resource pages.

Calculating Similar Pages

- If, instead of a root set from the results of a search engine query, we use a single page, we can find related pages.
- Starting with the home page for Honda, we also find the home pages for Toyota, Ford, BMW, Volvo, Saturn, Nissan, Audi, Dodge, Chrysler, etc.



In studies conducted in 1998 over 26 queries and 37 volunteers, Clever reported better authorities than Yahoo!, which in turn was better than Alta Vista. Since then most search engines have incorporated some notion of link-based ranking.



Link-based ranking beats a traditional text-based IR system by a clear margin for Web workloads. 100 queries were evaluated. The x-axis shows the smallest rank where a relevant page was found and the y-axis shows how many out of the 100 queries were satisfied at that rank.
A standard TFIDF ranking engine is compared with four well-known Web search engines (AltaVista/Raging, Lycos, Google, and Excite). Their identities have been withheld in this chart by [Singhal et al].

Relation between HITS, PageRank and LSI

- HITS algorithm = running SVD on the hyperlink relation (source,target)
- LSI algorithm = running SVD on the relation (term,document).
- PageRank on root set R gives same ranking as the ranking of authorities as given by HITS

PageRank vs HITS

- PageRank advantage over HITS
  - Query-time cost is low
    - HITS: computes an eigenvector for every query
  - Less susceptible to localized link-spam
- HITS advantage over PageRank
  - HITS ranking is sensitive to query
  - HITS has notion of hubs and authorities
- Topic-sensitive PageRanking [Haveliwala WWW11]
  - Attempt to make PageRanking query sensitive

  [Aside: Kaltix was a company formed a few months ago with people who had expertise in generating personalized and topic-sensitive PageRanking systems. Google bought Kaltix this month.]