Introduction to Bibliometrics

Applied Bibliometrics
KAN Min-Yen

What is Bibliometrics?

- Statistical and other forms of quantitative analysis
- Used to discover and chart the growth patterns of information
    - Production
    - Use

Outline

- What is bibliometrics? √
- Bibliometric laws
- Properties of information and its production

Properties of Academic Literature

- Growth
- Fragmentation
- Obsolescence
- Linkage

Growth

- Exponential rate for several centuries: "information overload"
- 1st known scientific journal: ~1600
- Today:
  - LINC has about 15,000 in all libraries

- Factors:
  - Ease of publication
  - Ease of use and increased availability
  - Known reputation

Zipf-Yule-Pareto Law

$P_n \approx 1/n^a$

where P_n is the frequency of occurrence of the n^{th} ranked item and $a \approx 1$.

"The probability of occurrence of a value of some variable starts high and tapers off. Thus, a few values occur very often while many others occur rarely."

- Pareto – for land ownership in the 1800's
- Zipf – for word frequency
- Also known as the 80/20 rule and as Zipf-Mandelbrot
- Used to measure of citings per paper:

  \# of papers cited n times is about $1/n^a$ of those being cited once, where $a \approx 1$

Random processes and Zipfian behavior

- Some random processes can also result in Zipfian behavior:
  - At the beginning there is one “seminal" paper.
  - Every sequential paper makes at most ten citations (or cites all preceding papers if their number does not exceed ten).
  - All preceding papers have an equal probability to be cited.

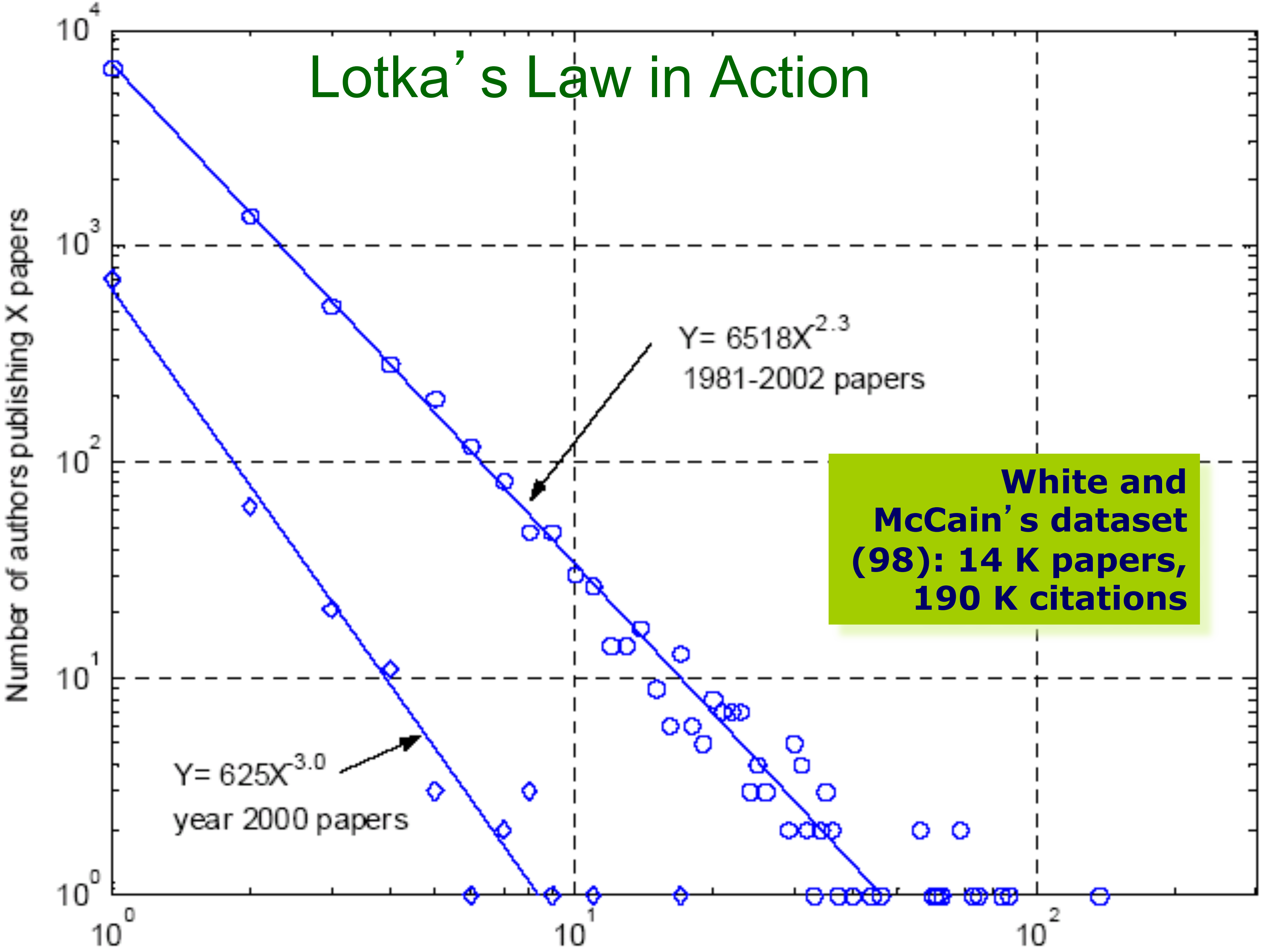
- Result: A Zipfian curve, with a≈1.
  What’s your conclusion?

Lotka's Law

The number of authors making n contributions is about $1/n^a$ of those making one contribution, where a ≈ **2**.

- Implications:
  - A small number of authors produce large number of papers, e.g., 10% of authors produce half of literature in a field
  - Those who achieve success in writing papers are likely continue having it

Lotka's Law in Action



White and McCain's dataset (98): 14 K papers, 190 K citations

Bradford's Law of Scattering

Journals in a field can be divided into three parts, each with about one-third of all articles:

1) a core of a few journals,
2) a second zone, with more journals, and
3) a third zone, with the bulk of journals.

The number of journals is $1:n:n^2$

To think about: Why is this true?

Fragmentation

- Influenced by scientific method
  - Information is continuous, but discretized into standard chunks

    (e.g., conference papers, journal article, surveys, texts, Ph.D. thesis)

- One paper reports one experiment
- Scientists aim to publish in diverse places

Fragmentation

- Motivation from academia
  - The “popularity contest”
  - Getting others to use your intellectual property and credit you with it
    - Spread your knowledge wide across disciplines
  - Academic yardstick for tenure (and for hiring)
    - The more the better – fragment your results
    - The higher quality the better – chase best journals

To think about: what is fragmentation’s relation to the aforementioned bibliometric laws?

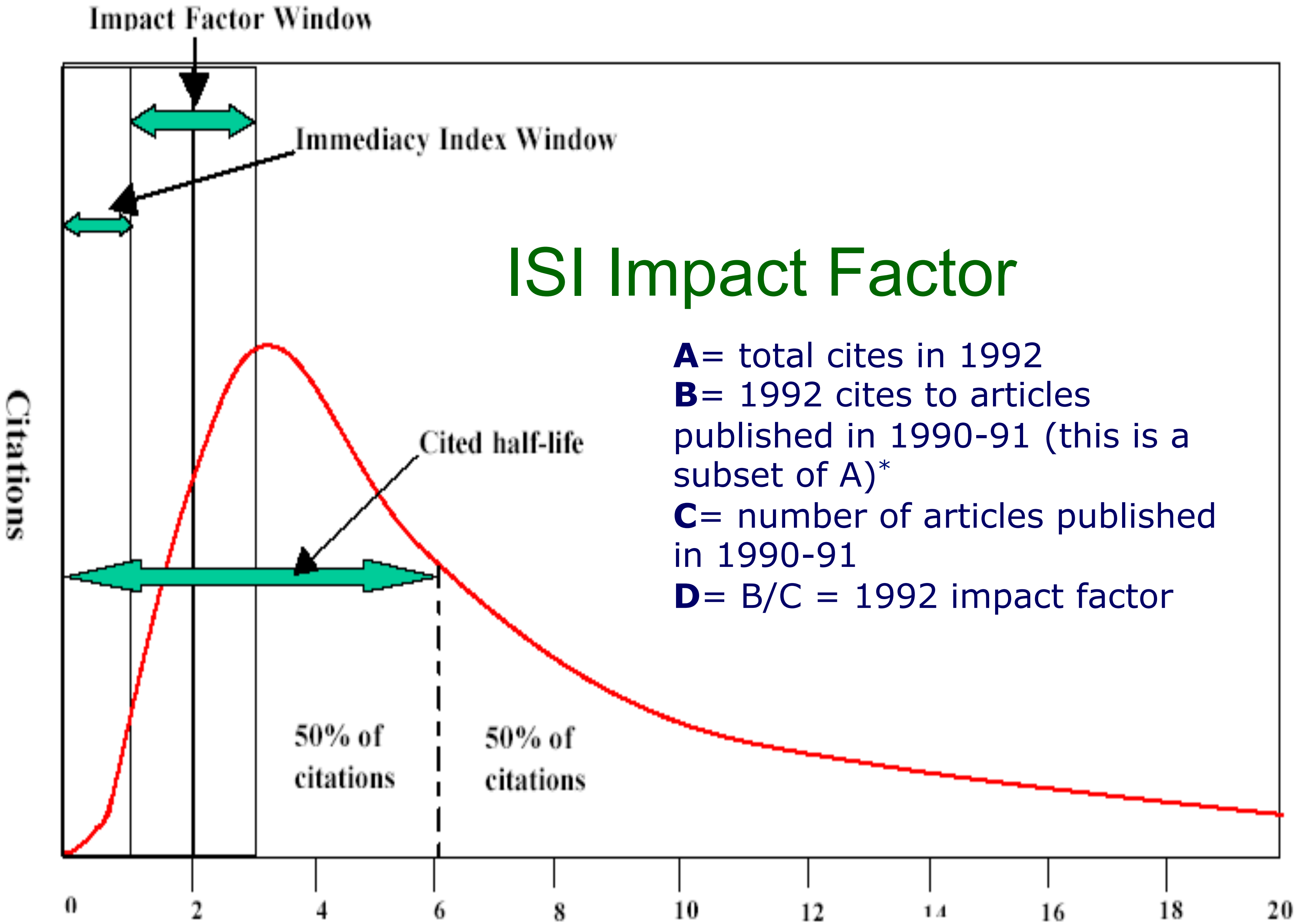
Obsolescence

Literature gets outdated fast!

- ½ references < 8 yrs. Chemistry
- ½ references < 5 yrs. Physics

- Textbooks out dated when published
- Practical implications in the digital library
- What about computer science?

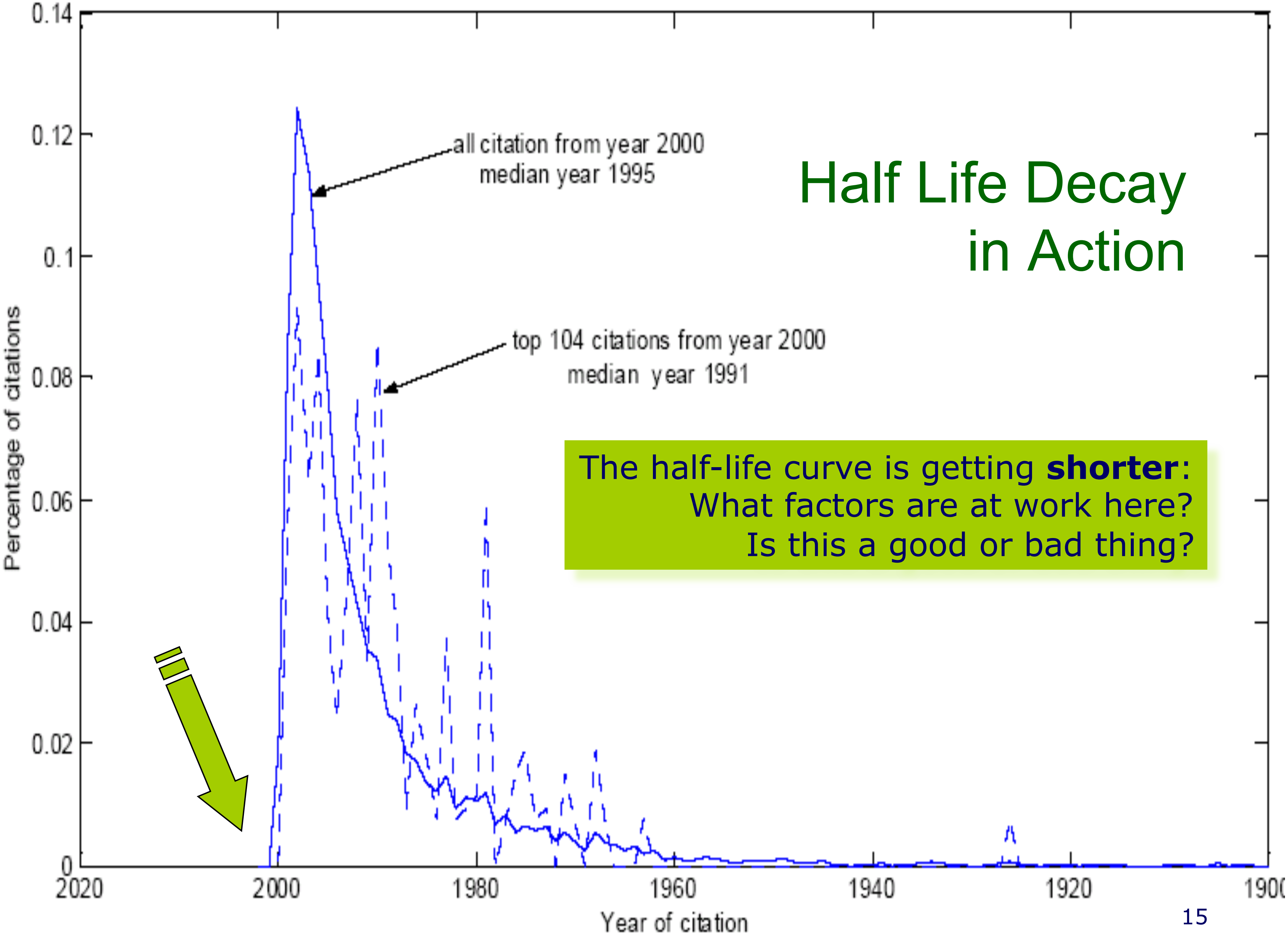
To think about: Is it really outdated-ness that is measured or something else?



ISI Impact Factor

A= total cites in 1992
B= 1992 cites to articles published in 1990-91 (this is a subset of A)*
C= number of articles published in 1990-91
D= B/C = 1992 impact factor

Half Life Decay in Action



The half-life curve is getting **shorter**:
What factors are at work here?
Is this a good or bad thing?

Expected Citation Rates

- From a large sample can calculate expected rates of citations
  - For journals vs. conferences
  - For specific journals vs. other ones

- Can find a researcher's productivities against this specific rate
  - Basis for promotion

To think about: what types of papers are cited most often?
(Hint: what types of papers dominate the top ten in Citeseer?)

All Years | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010

Next 100 →

1. ***5204*** A Dempster, N Laird, Rubin. *D.: Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, , 1977.

2. ***2900*** S Brookes, C Hoare, A Roscoe. *A theory of communicating sequential processes*. J. ACM, , 1984.

3. ***2668*** J R Quinlan. *Induction of decision trees*. Machine Learning, , 0.

4. ***2644*** I Stoica, R Morris, D Karger, M Kaashoek, H Balakrishnan. *Chord: A scalable peer-to-peer lookup service for Internet applications*. in Proceedings of the ACM SIGCOMM 2001 Conference, , 2001.

5. ***2611*** L R Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, , 1989.

6. ***2503*** S Kirkpatrick, C D Gelatt, M P Vecchi. *Optimization by Simulated Annealing*. Science, Number, , 1983.

7. ***2471*** R E Bryant. *Graph-based algorithms for boolean function manipulation*. IEEE Transactions on Computers, , 1986.

8. ***2357*** C L Lui, J W Layland. *Scheduling Algorithms for multiprogramming in a hard realtime environment*. Journal of the ACM, , 1973.

9. ***2265*** R Sutton, Barto. *A.: Reinforcement Learning: An Introduction*. , 1998.

10. ***2230*** R L Rivest, A Shamir, L Adleman. *A method for obtaining digital signatures and public-key cryptosystems, Comm*. MR 83m:94003 [Sag08] Sage, Free Open Source Mathematical Software (Version 3.0.4), 2008, http://www.sagemath.org. [Sch85] [Sho97, , 1978.

11. ***2195*** M Kass, A Witkin, D Terzopolous. *Snakes: Active contour models*. Inter. J. Computer Vision, , 1988.

12. ***2114*** S Ratnasamy, P Francis, M Handley, R Karp, S Shenker. *A scalable content-addressable network*. in Proceedings of the ACM SIGCOMM 2001 Conference, , 2001.

13. ***2092*** S Brin, L Page. *The anatomy of a large-scale hypertextual web search engine*. In WWW7: Proceedings of the 7th International Conference on World Wide Web 7, , 1998.

14. ***2008*** W Diffie, M E Hellman. *New directions in cryptography*. IEEE Transactions on Information Theory, , 1976.

15. ***1955*** M Turk, A Pentland. *Eigenfaces for recognition*. Journal of Cognitive Neuroscience, , 1991.

Linkage

- Citations in scientific papers are important:
  - Demonstrate awareness of background
  - Prior work being built upon
  - Substantiate claims
  - Contrast to competing work

Any other reasons?

One of the main reasons # of citations by themselves not a good rationale for evaluation.

Non-trivial to unify citations

- Citations have different styles:

Rosenblatt F (1961). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington, D.C.
[97] Rosenblatt, F. (1962). Principles of Neurodynamics. Washington, DC: Spartan
[Ros62] F. Rosenblatt. Principles of Neurodynamics. Spartan Books, 1962.

- Citeseer tried edit distance, structured field recognition
  - Settled on word (unigram) + section n-gram matching after normalization
  - More work to be done here: OpCit GPL code

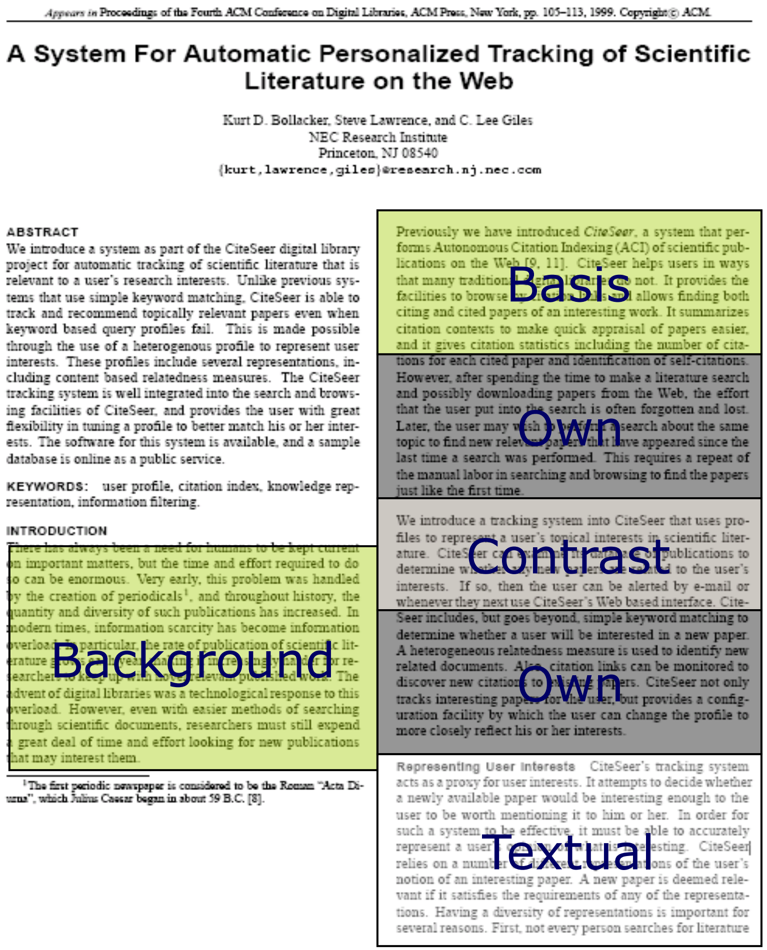
Non-trivial even for the web: Think URL redirects, domain names

Computational Analysis of Links

- If we know what type of citations/links exist, that can help:
  - In scientific articles:
    - In calculating impact
    - In relevance judgment (browsing → survey paper)
    - Checking whether paper author's are informed
  - In DL item retrieval:
    - In classifying items pointed by a link
    - In calculating an item's importance (removal of self-citations)

Calculating citation types

- Teufel (00): creates Rhetorical Document Profiles
  - Capitalizes on fixed structure and argumentative goals in scientific articles (e.g. Related Work)
  - Uses discourse cue phrases and position of citation to classify (e.g., In constrast to [1], we …) a zone



Appears in Proceedings of the Fourth ACM Conference on Digital Libraries, ACM Press, New York, pp. 105–113, 1999. Copyright © ACM.

A System For Automatic Personalized Tracking of Scientific Literature on the Web

Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles
NEC Research Institute
Princeton, NJ 08540
{kurt,lawrence,giles}@research.nj.nec.com

ABSTRACT
We introduce a system as part of the CiteSeer digital library project for automatic tracking of scientific literature that is relevant to a user's research interests. Unlike previous systems that use simple keyword matching, CiteSeer is able to track and recommend topically relevant papers even when keyword based query profiles fail. This is made possible through the use of a heterogenous profile to represent user interests. These profiles include several representations, including content based relatedness measures. The CiteSeer tracking system is well integrated into the search and browsing facilities of CiteSeer, and provides the user with great flexibility in tuning a profile to better match his or her interests. The software for this system is available, and a sample database is online as a public service.

KEYWORDS: user profile, citation index, knowledge representation, information filtering.

INTRODUCTION
There has always been a need for humans to be kept current on important matters, but the time and effort required to do so can be enormous. Very early, this problem was handled by the creation of periodicals[1], and throughout history, the quantity and diversity of such publications has increased. In modern times, information scarcity has become information overload. In particular, the rate of publication of scientific literature [illegible] for researchers [illegible] The advent of digital libraries was a technological response to this overload. However, even with easier methods of searching through scientific documents, researchers must still expend a great deal of time and effort looking for new publications that may interest them.

[1] The first periodic newspaper is considered to be the Roman "Acta Diurna", which Julius Caesar began in about 59 B.C. [8].

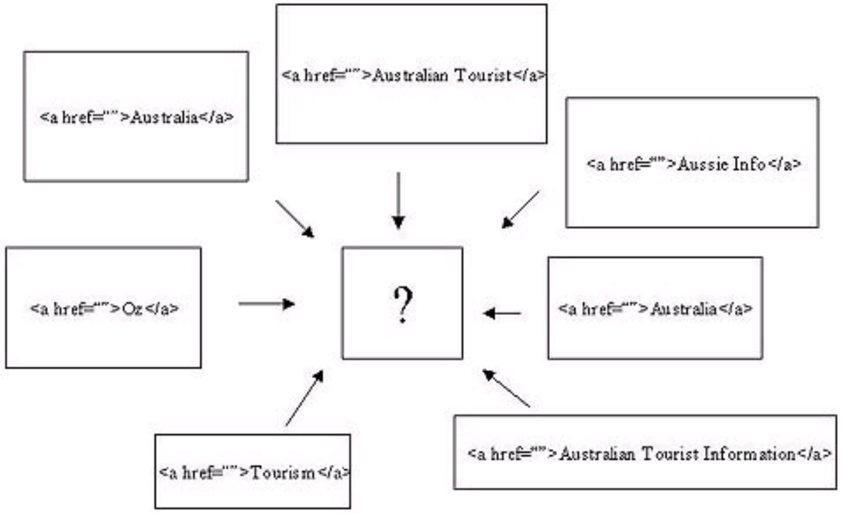
Previously we have introduced *CiteSeer*, a system that performs Autonomous Citation Indexing (ACI) of scientific publications on the Web [9, 11]. CiteSeer helps users in ways that many tradition[illegible] not. It provides the facilities to browse [illegible] and allows finding both citing and cited papers of an interesting work. It summarizes citation contexts to make quick appraisal of papers easier, and it gives citation statistics including the number of citations for each cited paper and identification of self-citations. However, after spending the time to make a literature search and possibly downloading papers from the Web, the effort that the user put into the search is often forgotten and lost. Later, the user may [illegible] a search about the same topic to find new [illegible] that have appeared since the last time a search was performed. This requires a repeat of the manual labor in searching and browsing to find the papers just like the first time.

We introduce a tracking system into CiteSeer that uses profiles to represent a user's topical interests in scientific literature. CiteSeer can [illegible] publications to determine [illegible] to the user's interests. If so, then the user can be alerted by e-mail or whenever they next use CiteSeer's Web based interface. CiteSeer includes, but goes beyond, simple keyword matching to determine whether a user will be interested in a new paper. A heterogeneous relatedness measure is used to identify new related documents. Also, citation links can be monitored to discover new citations to [illegible] papers. CiteSeer not only tracks interesting papers for the user, but provides a configuration facility by which the user can change the profile to more closely reflect his or her interests.

Representing User Interests CiteSeer's tracking system acts as a proxy for user interests. It attempts to decide whether a newly available paper would be interesting enough to the user to be worth mentioning it to him or her. In order for such a system to be effective, it must be able to accurately represent a user[illegible]esting. CiteSeer relies on a numb[illegible]ons of the user's notion of an interesting paper. A new paper is deemed relevant if it satisfies the requirements of any of the representations. Having a diversity of representations is important for several reasons. First, not every person searches for literature

Using link text for classification

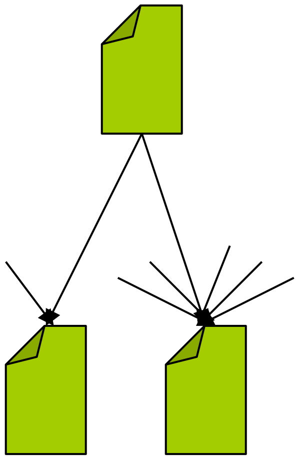
- The link text that describes a page in another page can be used for classification.



- Amitay (98) extended this concept by ranking nearby text fragments using (among other things) positional information.
  - **XXXX**: …. … .. … ..
  - … … … .. …. **XXX**, …. … .. … …
  - … **XXXX**[…] […] [….]

Ranking related papers in retrieval

- Citeseer uses two forms of relatedness to recommend “related articles”:



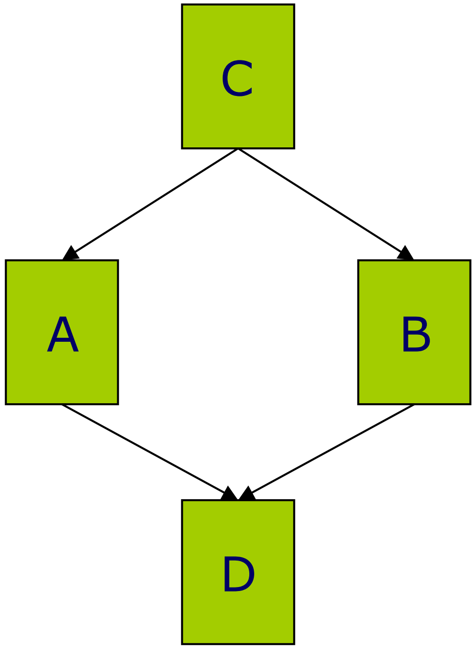
- TF × IDF
  - If above a threshold, report it
- CC (Common Citation) × IDF
  - CC = Bibliographic Coupling
  - If two papers share a rare citation, this is more important than if they share a common one.

Citation Analysis

Deciding which (web sites, authors) are most prominent

Citation Analysis

- Despite shortcomings, still useful
- Citation links viewed as a DAG
- Incoming and outgoing links have different treatments



Analysis types

- **Co-citation** analysis – A and B both cited by C
- **Bibliographic coupling** – A and B both have similar citations (e.g., D)

Sociometric experiment types

- Ego-centered: focal person and its alters
  (Wasserman and Faust, pg. 53)
- Small World: how many actors a respondent is away from a target

Prominence

Consider a node prominent if its ties make it particularly visible to other nodes in the network
(adapted from WF, pg 172)

- Centrality – no distinction on incoming or outgoing edges (thus directionality doesn't matter. How *involved* is the node in the graph.

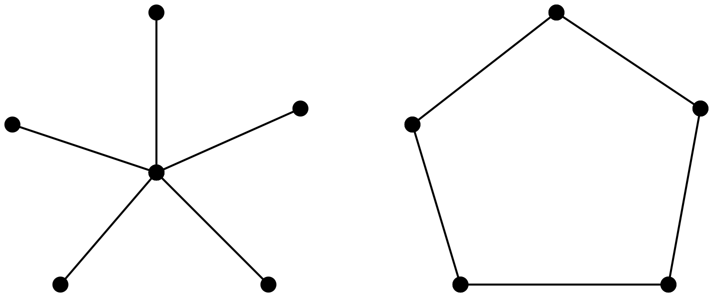
- Prestige – "Status". Ranking the prestige of nodes among other nodes. In degree counts towards prestige.

Centrality

- How central is a particular
  - Graph?
  - Node?
- Graph-wide measures assist in comparing graphs, subgraphs

Node Degree Centrality

- Degree (In + Out)
- Normalized Degree (In+Out/Possible)
  - What's max possible?
- Variance of Degrees



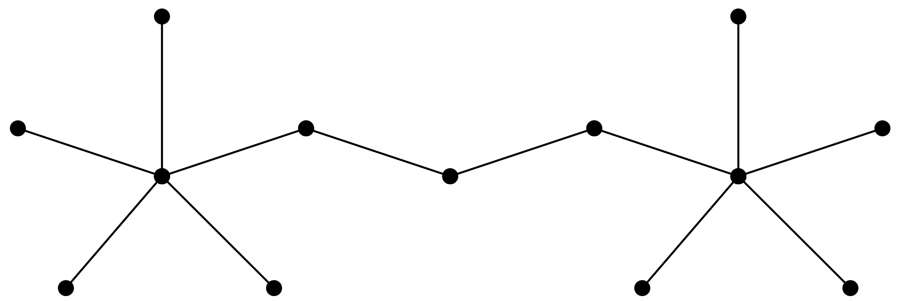
Distance Centrality

- Closeness = minimal distance
- Sum of shortest paths should be minimal in a central graph
- (Jordan) Center = subset of nodes that have minimal sum distance to all nodes.

What about disconnected components?

Betweenness Centrality

- A node is central iff it lies between other nodes on their shortest path.
- If there is more than one shortest path,
  - Treat each with equal weight
  - Use some weighting scheme
    - Inverse of path length



References (besides readings)

- Bollen and Luce (02) *Evaluation of Digital Library Impact and User Communities by Analysis of Usage Patterns*
  http://www.dlib.org/dlib/june02/bollen/06bollen.html
- Kaplan and Nelson (00) *Determining the publication impact of a digital library*
  http://download.interscience.wiley.com/cgi-bin/fulltext?ID=69503874&PLACEBO=IE.pdf&mode=pdf
- Wasserman and Faust (94) *Social Network Analysis* (on reserve)

Things to think about

- What's the relationship between these three laws (Bradford, Zipf-Yule-Pareto and Lotka)?
- How would you define the three zones in Bradford's law?

Pagerank and HITS*

Applied Bibliometrics

KAN Min-Yen

*Part of these lecture notes come from Manning, Raghavan and Schütze @ Stanford CS

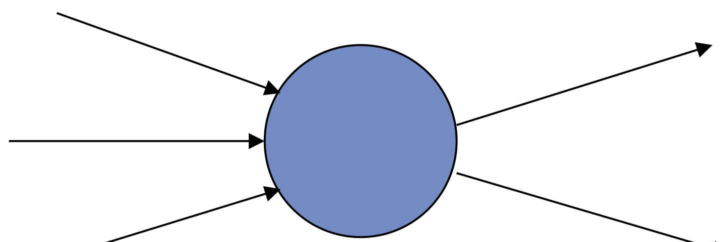
Connectivity analysis

- Idea: mine hyperlink information in the Web
- Assumptions:
  - Links often connect related pages
  - A link between pages is a recommendation
    - “people vote with their links”

Query-independent ordering

- Using link counts as simple measures of popularity

- Two basic suggestions:
  - Undirected popularity: **Centrality**
    - in-links plus out-links (3+2=5)
  - Directed popularity: **Prestige**
    - number of its in-links (3)



Algorithm

1. Retrieve all pages meeting the text query (say ***venture capital***), perhaps by using Boolean model

2. Order these by link popularity (either variant on the previous page)

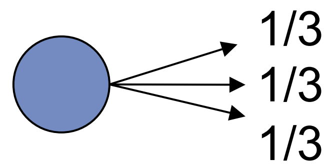
Exercise: How do you spam each of the following heuristics so your page gets a high score?

- score = # in-links plus # out-links

- score = # in-links

Pagerank scoring

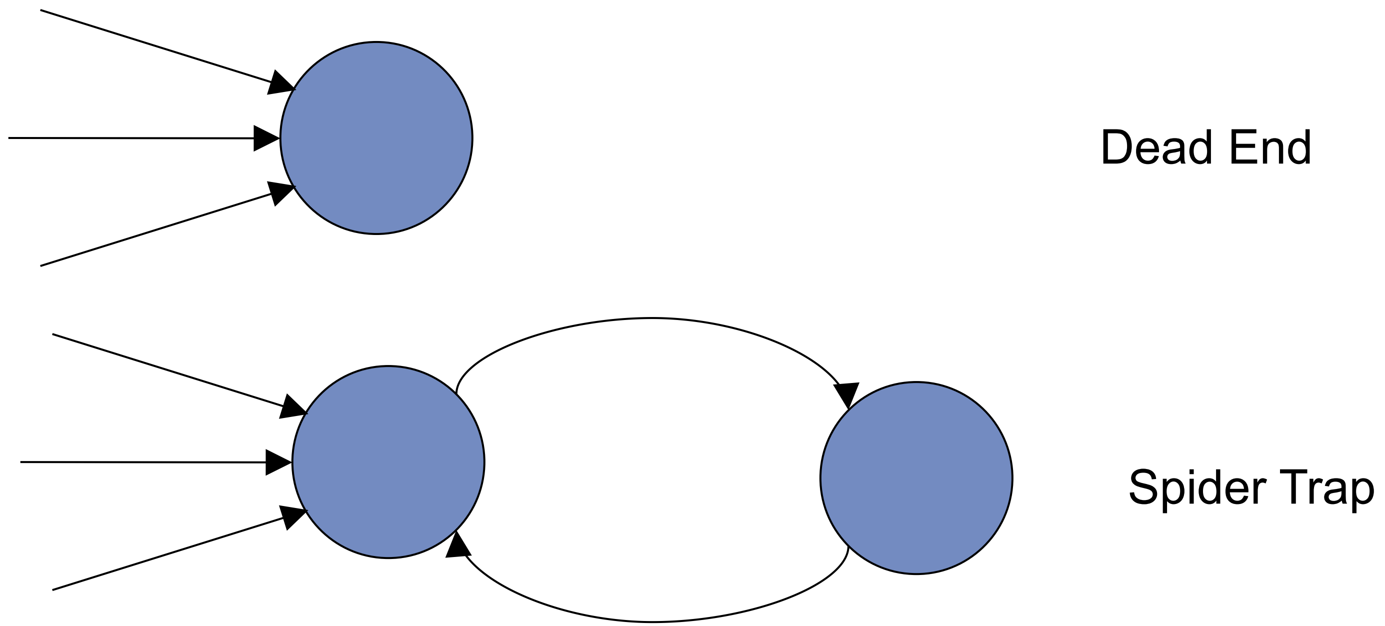
- Imagine a browser doing a random walk on web pages:
  - Start at a random page
  - At each step, follow one of the *n* links on that page, each with $1/n$ probability



- Do this repeatedly. Use the "long-term visit rate" as the page's score

Not quite enough

- The web is full of dead ends.
  - What sites have dead ends?
  - Our random walk can get stuck.



Teleporting

- At each step, with probability 10%, teleport to a random web page
- With remaining probability (90%), follow a random link on the page
  - If a dead-end, stay put in this case

This is lay explanation of the "damping factor" (1-a) in the rank propagation algorithm

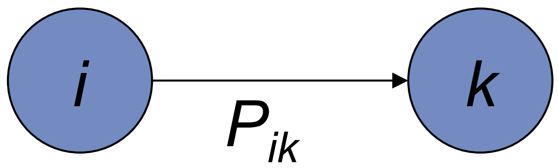
Result of teleporting

- Now we cannot get stuck locally
- There is a long-term rate at which any page is visited (not obvious, will show this)
  - How do we compute this visit rate?

Markov chains

A Markov chain consists of n states, plus an $n \times n$ transition probability matrix P.

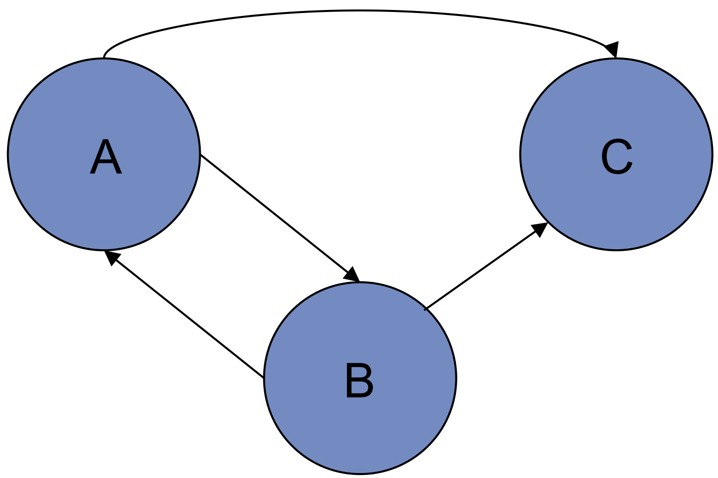
- At each step, we are in exactly one of the states.
- For $1 \leq i,k \leq n$, the matrix entry P_{ik} tells us the probability of k being the next state, given we are currently in state i.



$P_{ik} > 0$ is OK.

Markov chains

- Clearly, for all i, $\sum_{k=1}^{n} P_{ik} = 1.$
- Markov chains are abstractions of random walks



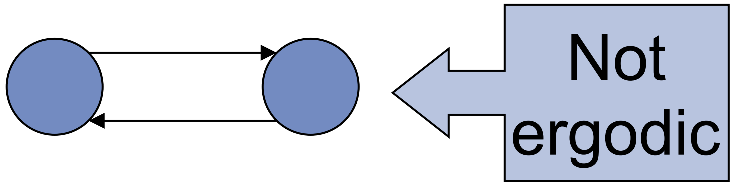
Try this: Calculate the matrix P_{ik} using a 10% probability of uniform teleportation

P_{ik}:

	A	**B**	**C**
A	.03	.48	.48
B	.48	.03	,48
C	.03	.03	.93

Ergodic Markov chains

- A Markov chain is ergodic if
  - you have a path from any state to any other
  - you can be in any state at every time step, with non-zero probability



  - With teleportation, our Markov chain is ergodic

Steady State

- For any ergodic Markov chain, there is a unique long-term visit rate for each state
  - Over a long period, we'll visit each state in proportion to this rate
  - It doesn't matter where we start

Probability vectors

- A probability (row) vector **x** = $(x_1, \ldots x_n)$ tells us where the walk is at any point
- E.g., (000…1…000) means we're in state i.

  $1 \quad i \quad n$

More generally, the vector **x** = $(x_1, \ldots x_n)$ means the walk is in state i with probability x_i.

$$\sum_{i=1}^{n} x_i = 1.$$

Change in probability vector

- If the probability vector is **x** = $(x_1, \ldots x_n)$ at this step, what is it at the next step?
- Recall that row *i* of the transition prob. Matrix **P** tells us where we go next from state *i*.
- So from **x**, our next state is distributed as **xP**.

Pagerank algorithm

- Regardless of where we start, we eventually reach the steady state **a**
  - Start with any distribution (say $\mathbf{x}=(10\ldots0)$)
  - After one step, we' re at $\mathbf{xP}$
  - After two steps at $\mathbf{xP}^2$, then $\mathbf{xP}^3$ and so on.
  - "Eventually" means for "large" k, $\mathbf{xP}^k = \mathbf{a}$
- Algorithm: multiply **x** by increasing powers of **P** until the product looks stable

Pagerank summary

- Pre-processing:
  - Given graph of links, build matrix **P**
  - From it compute **a**
  - The pagerank a_i is a scaled number between 0 and 1
- Query processing:
  - Retrieve pages meeting query
  - Rank them by their pagerank
  - Order is query-*independent*

Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find two sets of inter-related pages:
  - Hub pages are good lists of links on a subject.
    - e.g., “Bob’s list of cancer-related links.”
  - Authority pages occur recurrently on good hubs for the subject.
- Best suited for “broad topic” browsing queries rather than for known-item queries.
- Gets at a broader slice of common opinion.

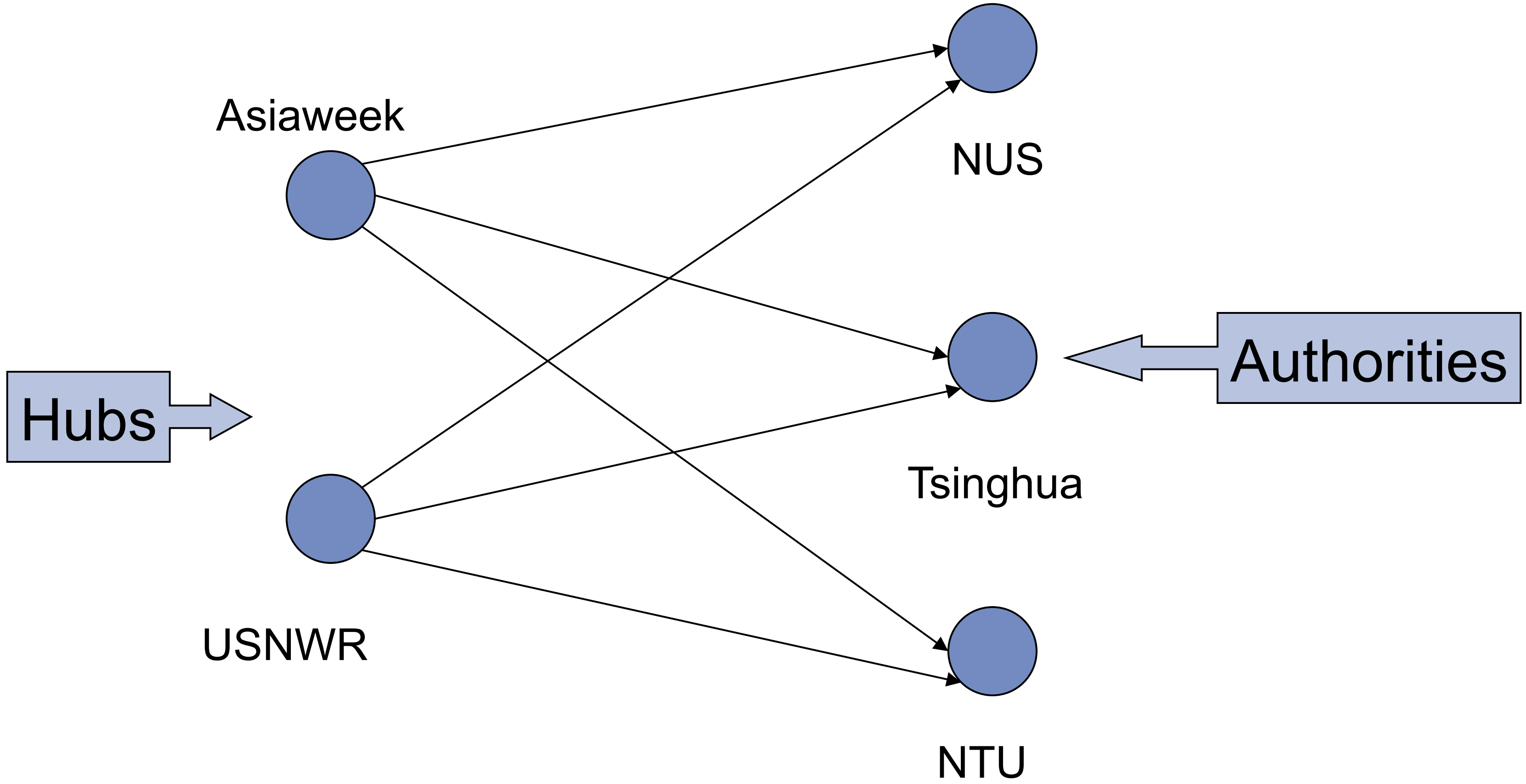
Hubs and Authorities

- Thus, a good hub page for a topic *points* to many authoritative pages for that topic.

- A good authority page for a topic is *pointed* to by many good hubs for that topic.

- Circular definition - will turn this into an iterative computation.

Hubs and Authorities

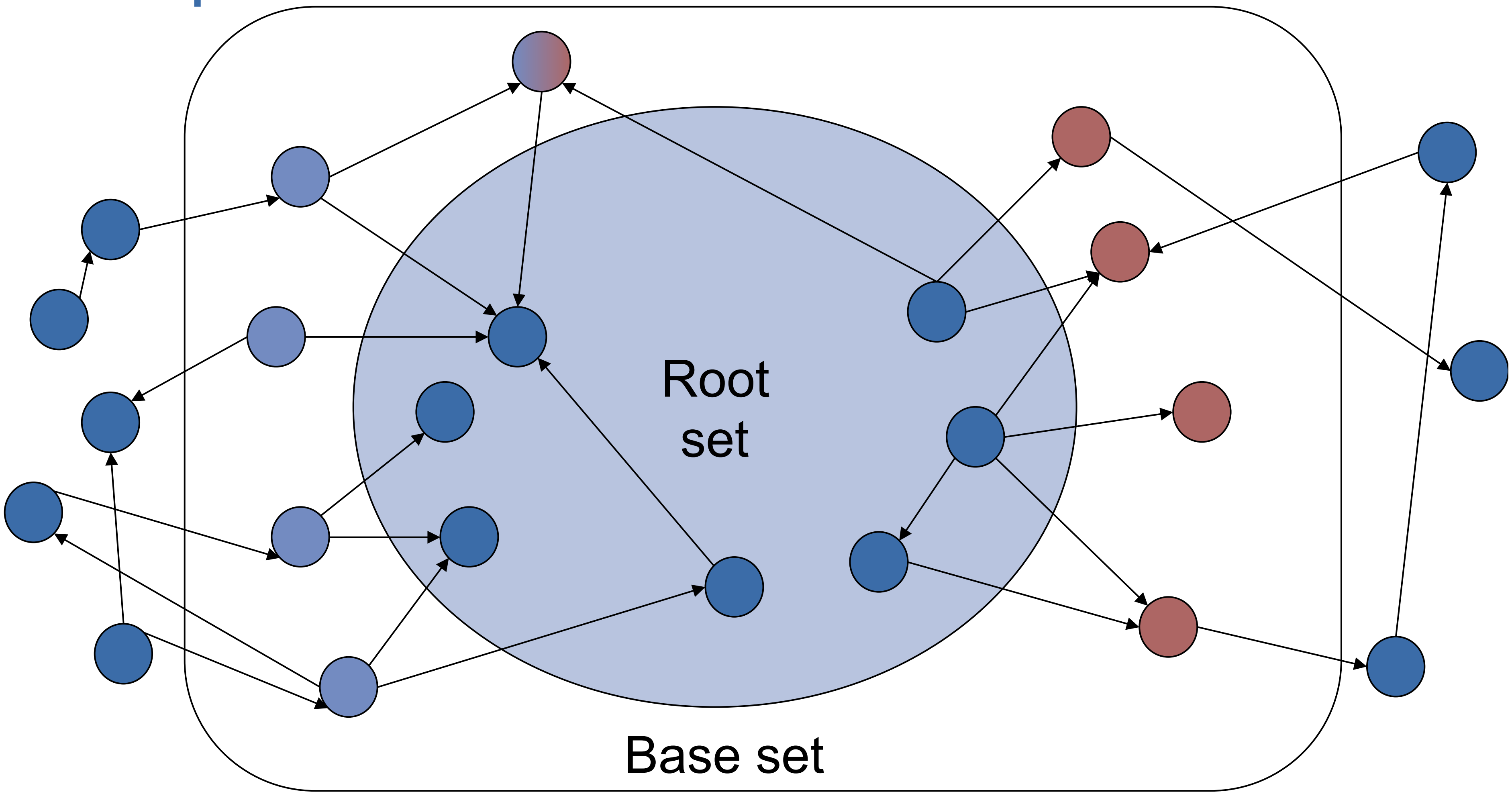


High-level scheme

- Extract from the web a base set of pages that *could* be good hubs or authorities.
- From these, identify a small set of top hub and authority pages
  - → iterative algorithm

Base set

1. Given text query (say **university**), use a text index to get all pages containing **university**.
   - Call this the <u>root set</u> of pages
2. Add in any page that either:
   - points to a page in the root set, or
   - is pointed to by a page in the root set
3. Call this the <u>base set</u>


Root
set
Base set

Assembling the base set

- Root set typically 200-1000 nodes.
- Base set may have up to 5000 nodes.
- How do you find the base set nodes?
  - Follow out-links by parsing root set pages.
  - Get in-links (and out-links) from a *connectivity server.*

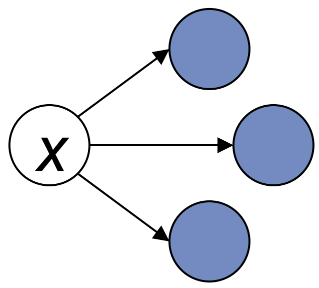
Distilling hubs and authorities

1. Compute, for each page *x* in the base set, a <u>hub score</u> *h(x)* and an <u>authority score</u> *a(x)*.
2. Initialize: for all *x*, $h(x) \leftarrow 1$; $a(x) \leftarrow 1$;
3. Iteratively update all *h(x)*, *a(x)*; ⇐ Key
4. After iterations:
   - highest *h()* scores are hubs
   - highest *a()* scores are authorities

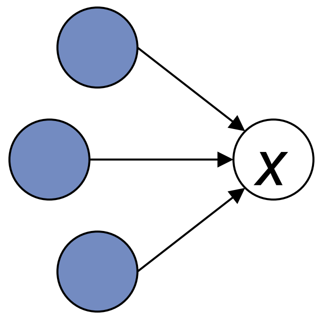
Iterative update

- Repeat the following updates, for all *x*:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$



$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$



How many iterations?

- Relative values of scores will converge after a few iterations
- We only require the <u>relative order</u> of the *h()* and *a()* scores - not their absolute values
- In practice, ~5 iterations needed

Things to think about

- Use *only* link analysis after base set assembled
  - iterative scoring is query-independent
- Iterative computation after text index retrieval - significant overhead

Things to think about

- How does the selection of the base set influence computation of H & As?
- Can we embed the computation of H & A during the standard VS retrieval algorithm?
- A pagerank score is a global score. Can there be a fusion between H&A (which are query sensitive) and pagerank? How would you do it?
- How do you relate CCIDF in Citeseer to Pagerank?