

# Google Changes the World

(with a little help from its friends)

How?

Why?

What's Next?

# Early Search Engines

1. *Crawl the Web* (follow links from page to page, finding and copying as many pages as they could).
2. Index pages by the words they contained.
3. Respond to *search queries* (lists of words) with the pages containing those words.

# Early Page Ranking

- ◆ Attempt to order pages matching a search query by “importance.”
- ◆ First search engines considered:
  1. Number of times query words appeared.
  2. Prominence of word position, e.g. title, header.

# The First Spammers

- ◆ As people began to use search engines to find things on the Web, those with commercial interests tried to exploit search engines to bring people to their own site – whether they wanted to be there or not.
- ◆ **Example:** shirt-seller might pretend to be about “movies.”

# The First Spammers – (2)

- ◆ How do you make your page appear to be about movies?
- ◆ Add the word `movie` 1000 times to your page.
- ◆ Set its color to the background color, so only search engines would see it.

# The First Spammers – (3)

- ◆ Or, run the query `movie` on your target search engine.
- ◆ See what page came first in the listings.
- ◆ Copy it into your page, invisibly.
- ◆ These and similar techniques are *term spam*.

# The First Spammers – (4)

- ◆ Rapidly, the promise of search engines disappeared.
- ◆ Spam dominated the listings to the extent that responses to search queries were useless.

# The Google Solution to Term Spam

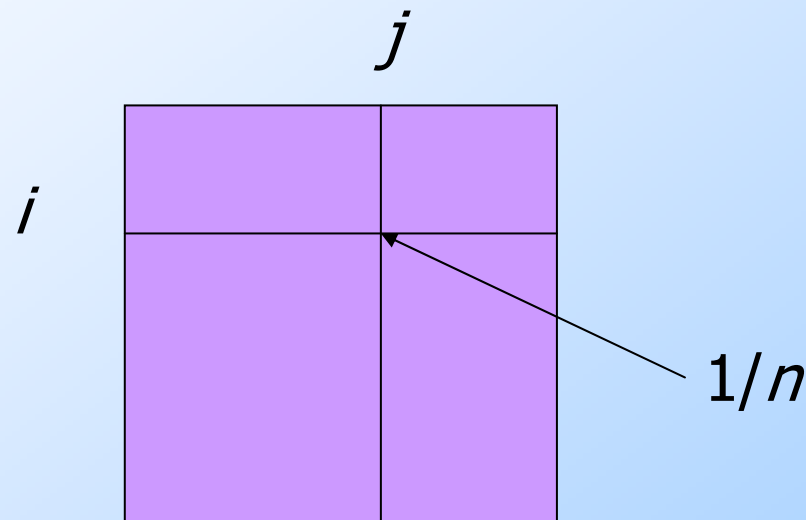
1. Believe what people say about you, rather than what you say about yourself.
  - ◆ Consider words in the *anchor text* (words that appear underlined to represent the link) and its surrounding text.
2. PageRank as a tool to measure the “importance” of Web pages.

# PageRank

- ◆ **Intuition**: solve the recursive equation:  
“a page is important if important pages link to it.”
  - ◆ Let the world vote, by their links, on what is important.
  - ◆ But you can't “stuff the ballot box.”
- ◆ In high-falutin' terms: *importance* = the principal eigenvector of the transition matrix of the Web.

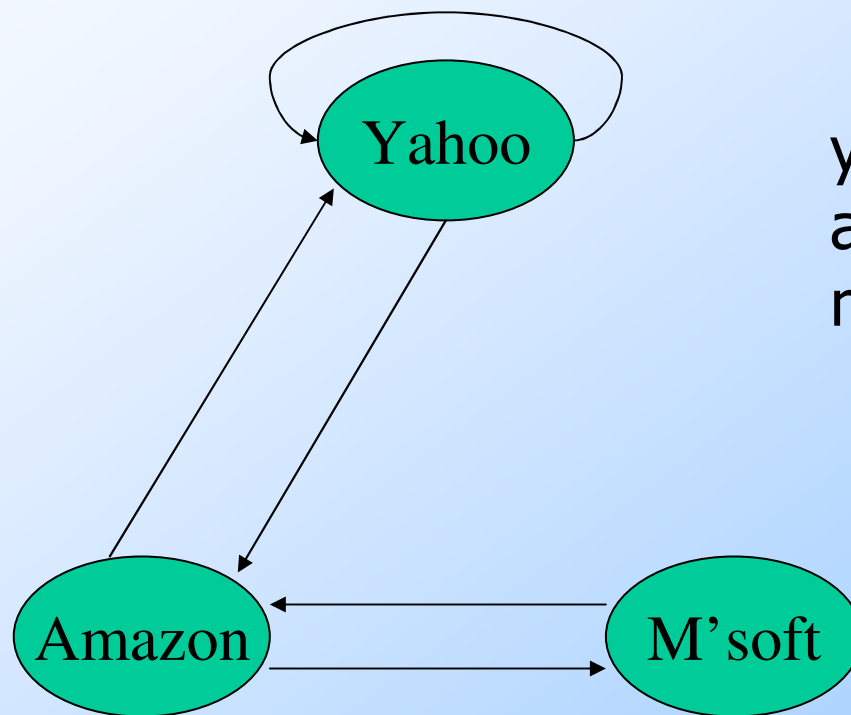
# Transition Matrix $M$ of the Web

Suppose page  $j$  links to  $n$  pages, including  $i$



Expresses how “importance” flows around the Web. Equivalent to following “random walkers.”

# Example: The Web in 1839



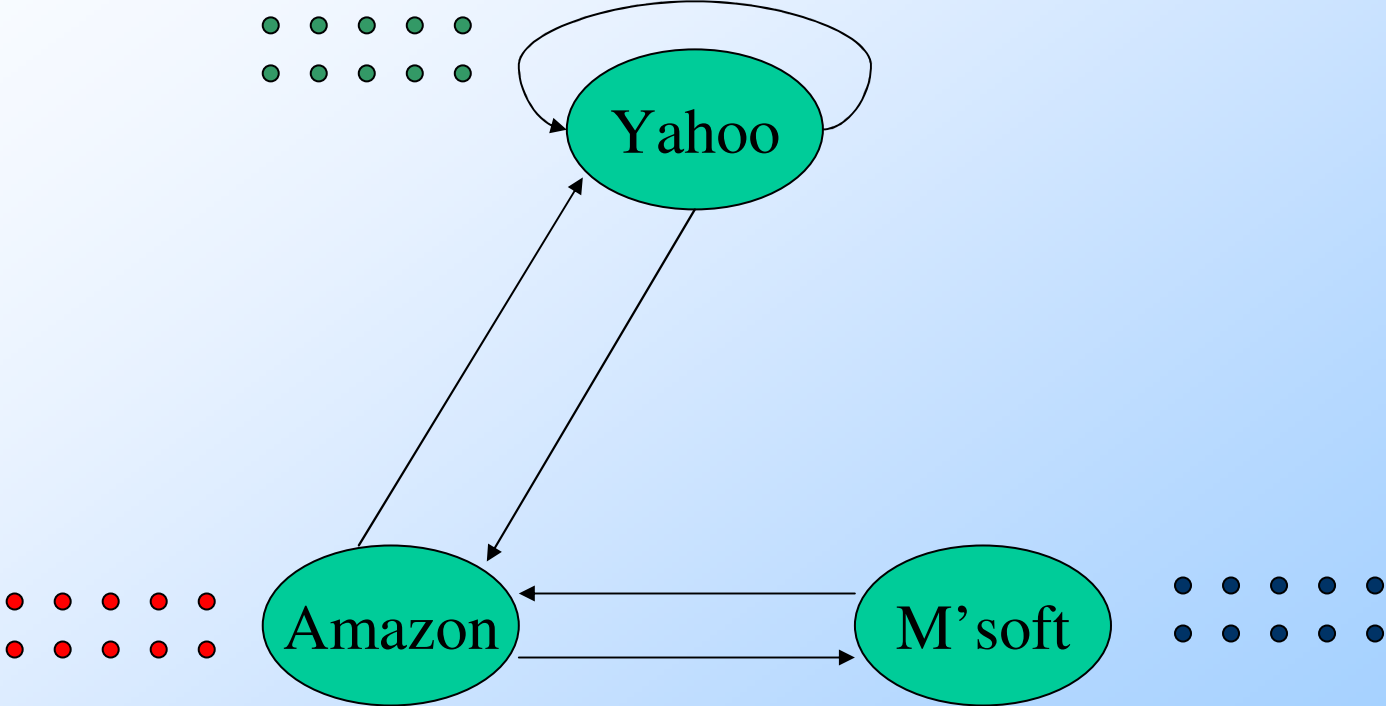
	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

*M*

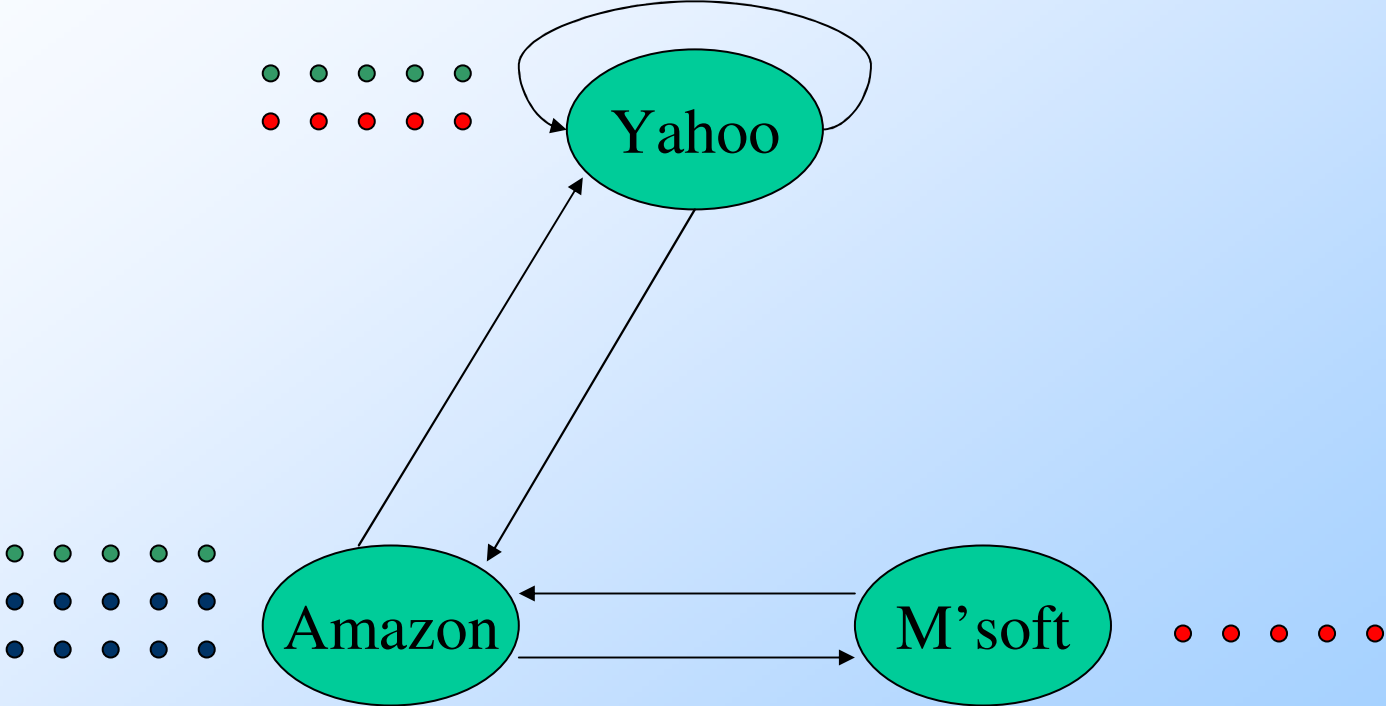
# The Idea Behind PageRank

- ◆ Imagine many random walkers on the Web.
- ◆ At each “tick,” each walker picks an out-link at random and follows it.
- ◆ Distribution of walkers  $\mathbf{v}$  becomes  $M\mathbf{v}$  after one tick.
- ◆ Compute  $M^{50}\mathbf{v}$  (approximately 50).

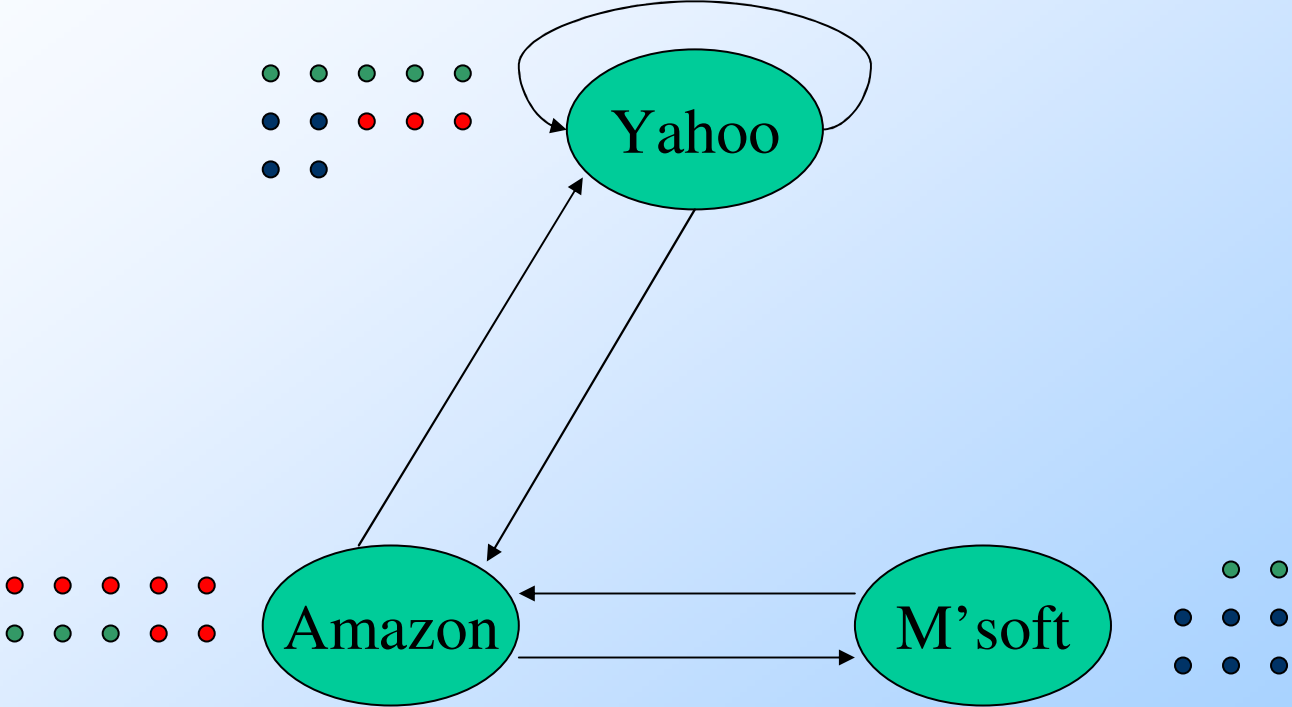
# The Walkers



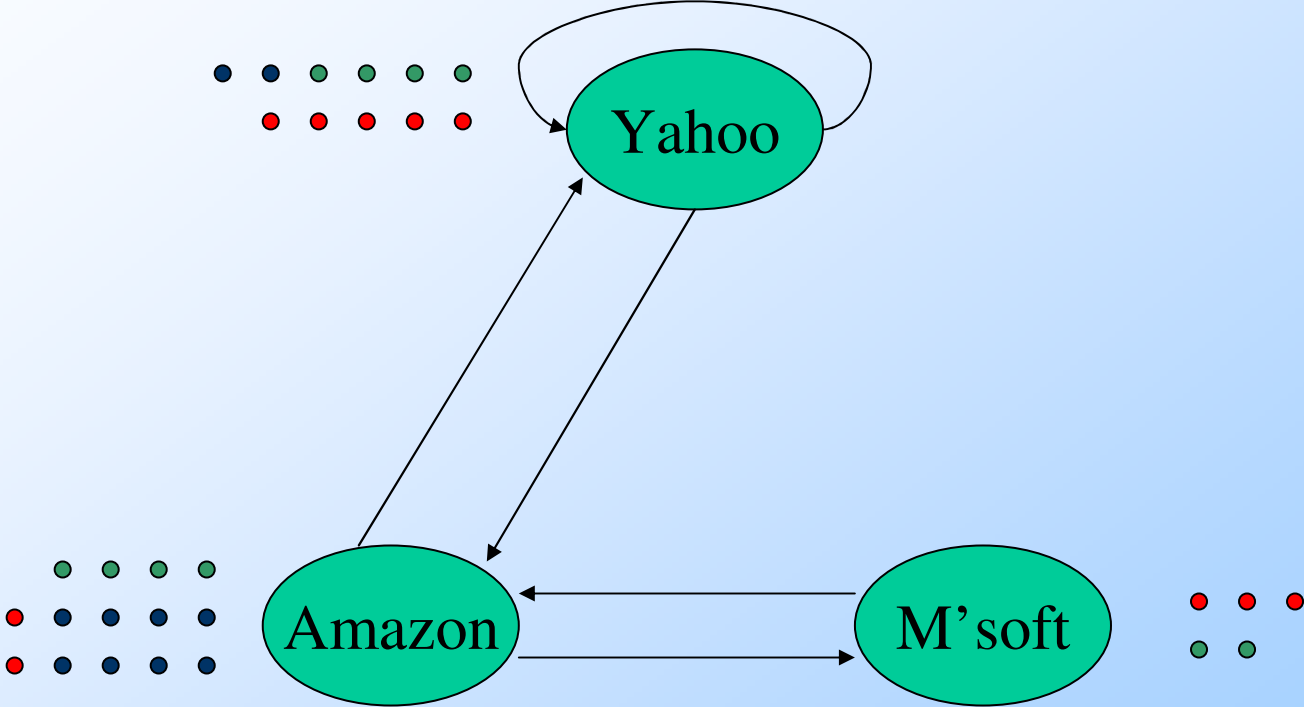
# The Walkers



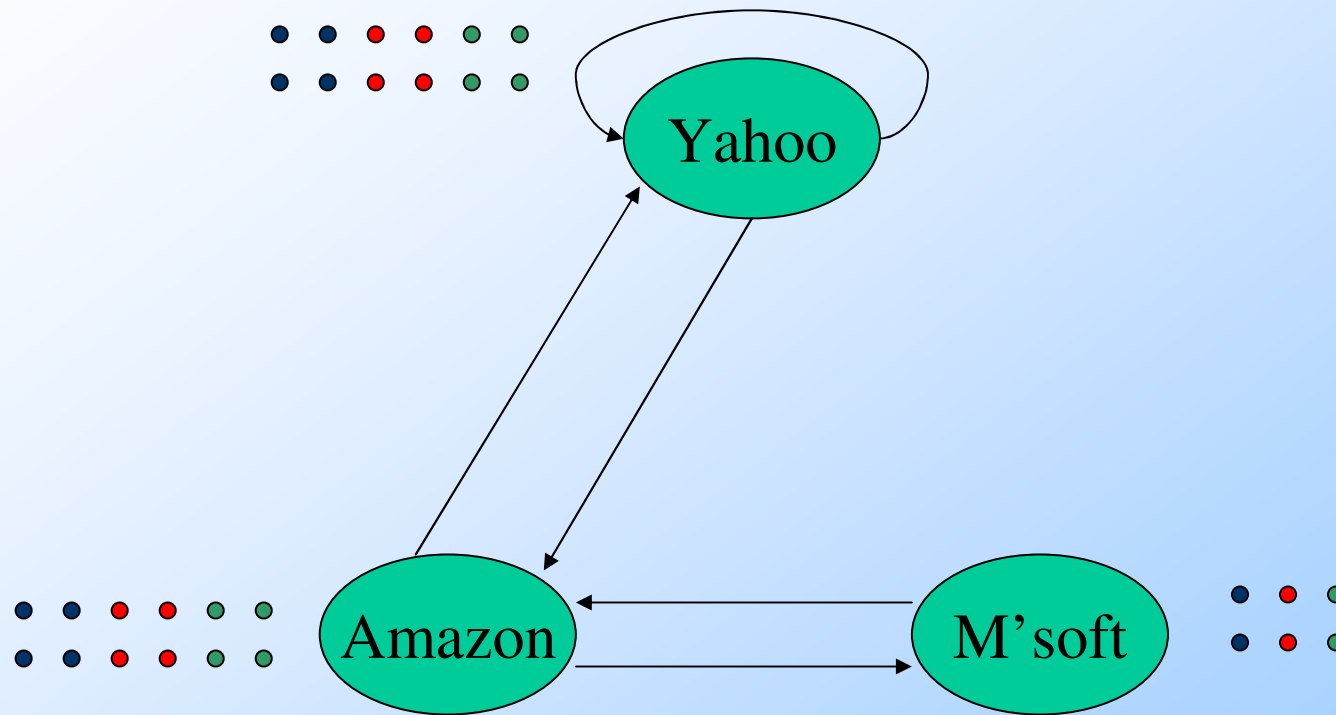
# The Walkers



# The Walkers



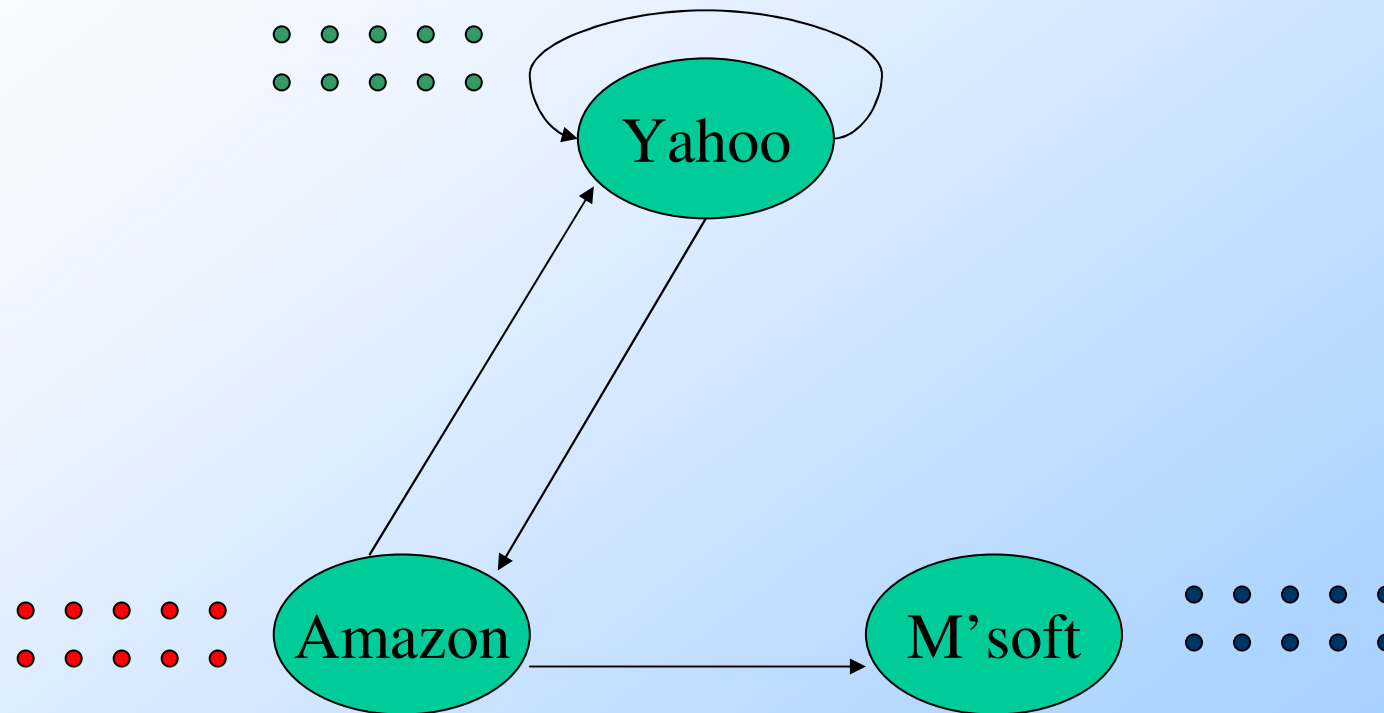
# In the Limit ...



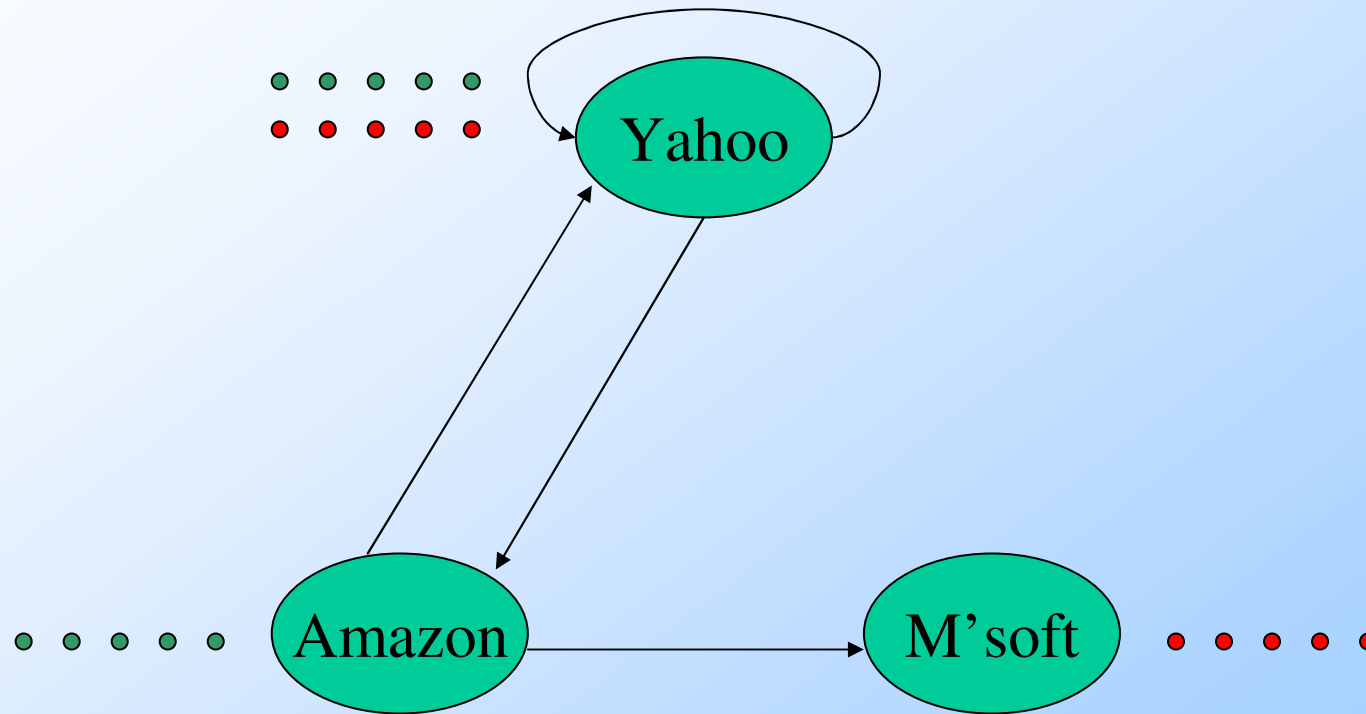
# Real-World Problems

- ◆ Some pages are “dead ends” (have no links out).
  - ◆ Such a page causes importance to leak out.
- ◆ Other (groups of) pages are *spider traps* (all out-links are within the group).
  - ◆ Eventually spider traps absorb all importance.

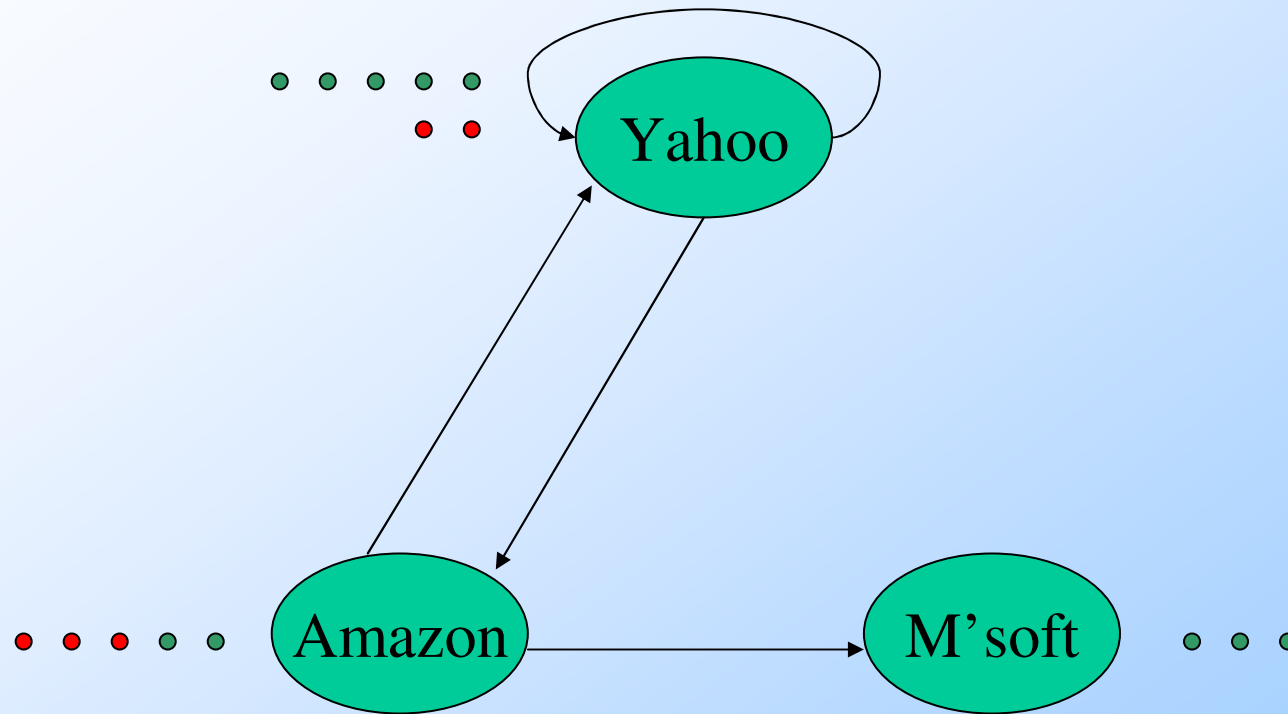
# Microsoft Becomes a Dead End



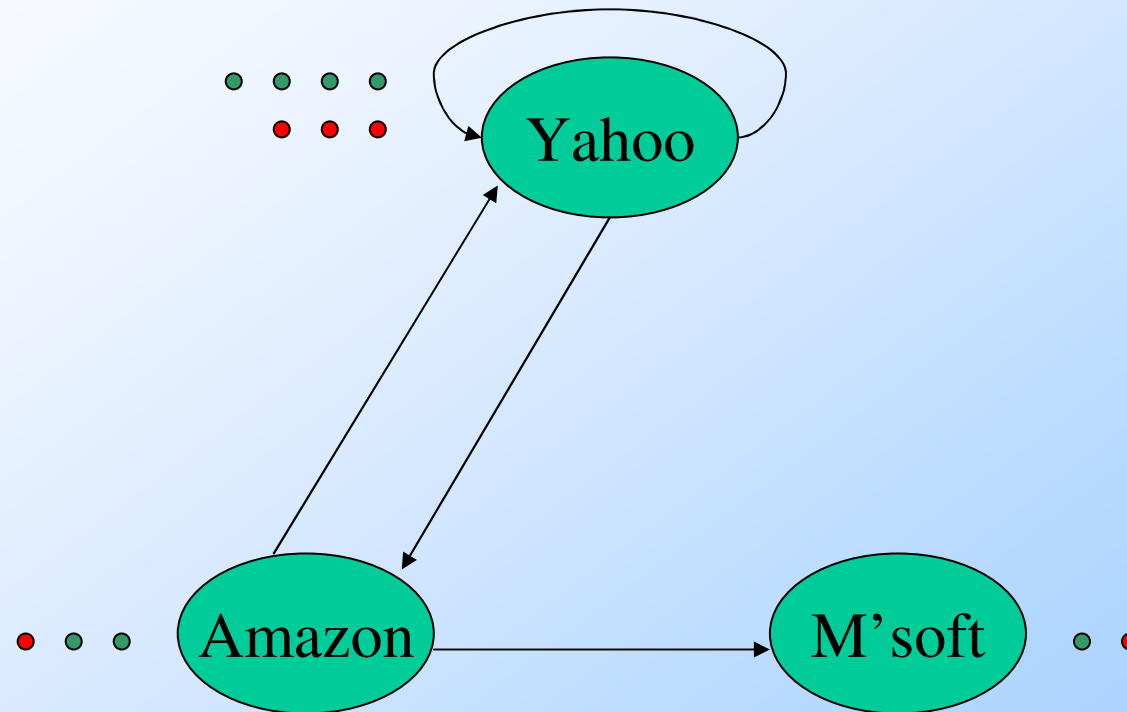
# Microsoft Becomes a Dead End



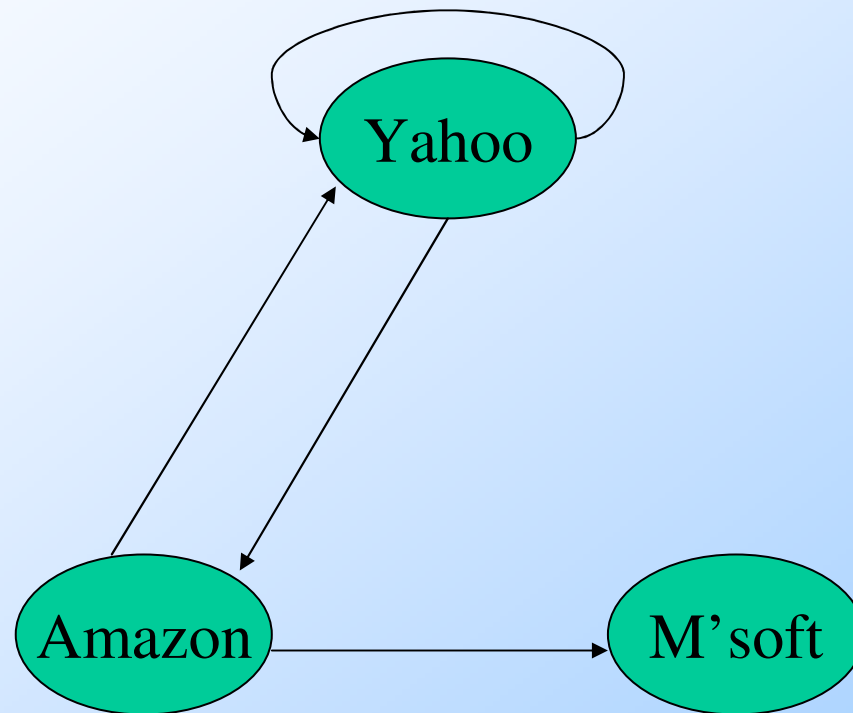
# Microsoft Becomes a Dead End



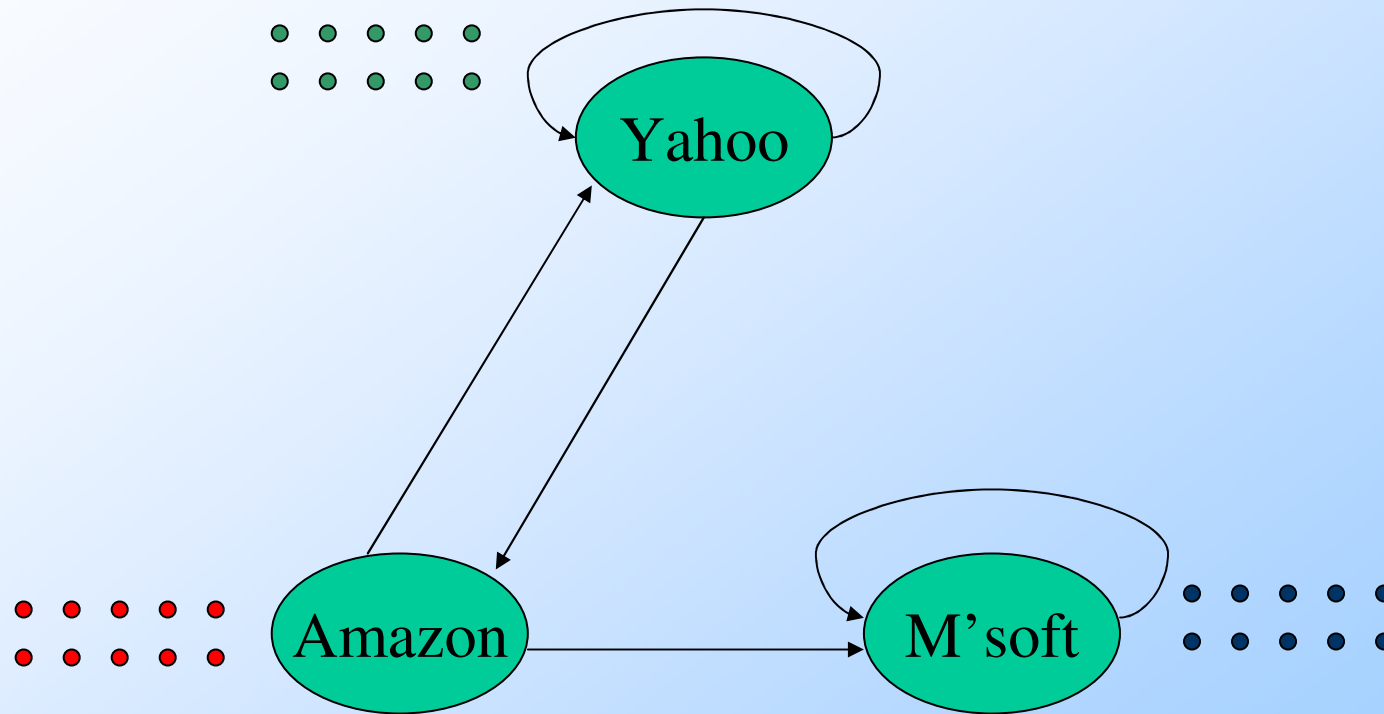
# Microsoft Becomes a Dead End



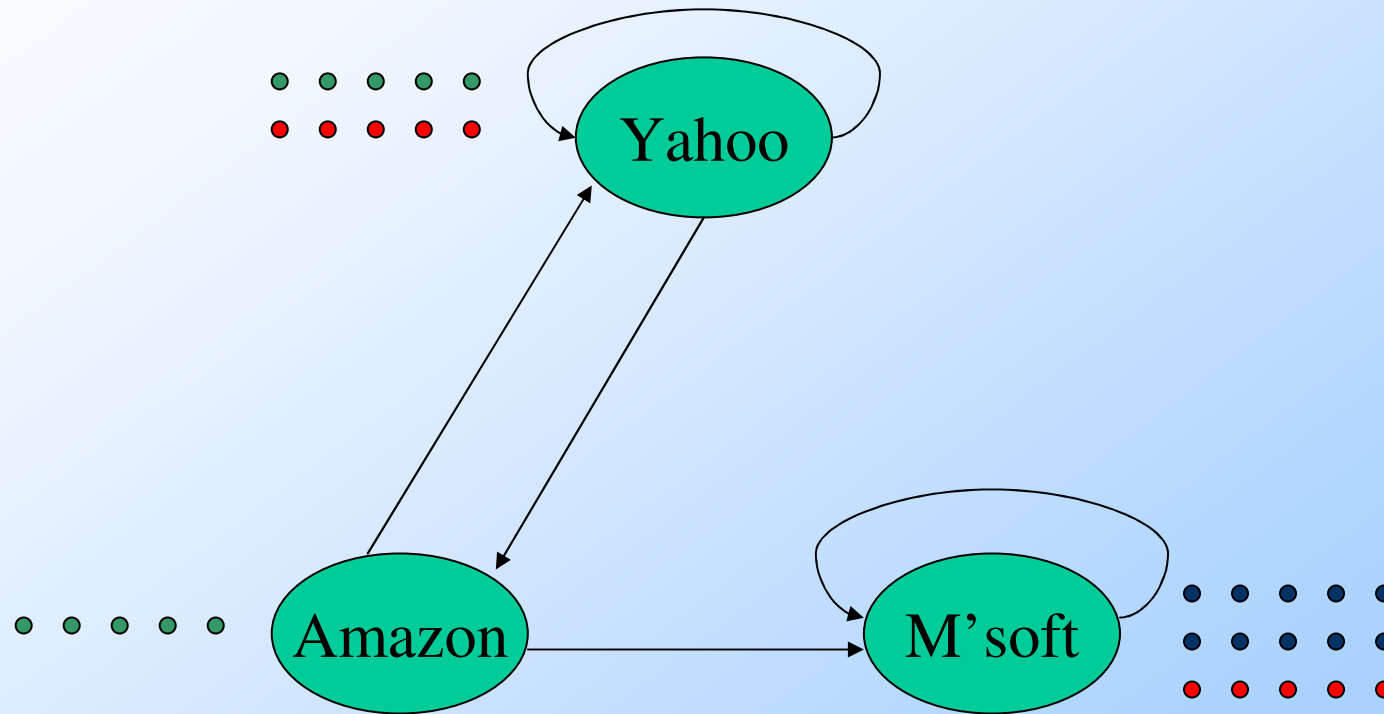
# In the Limit ...



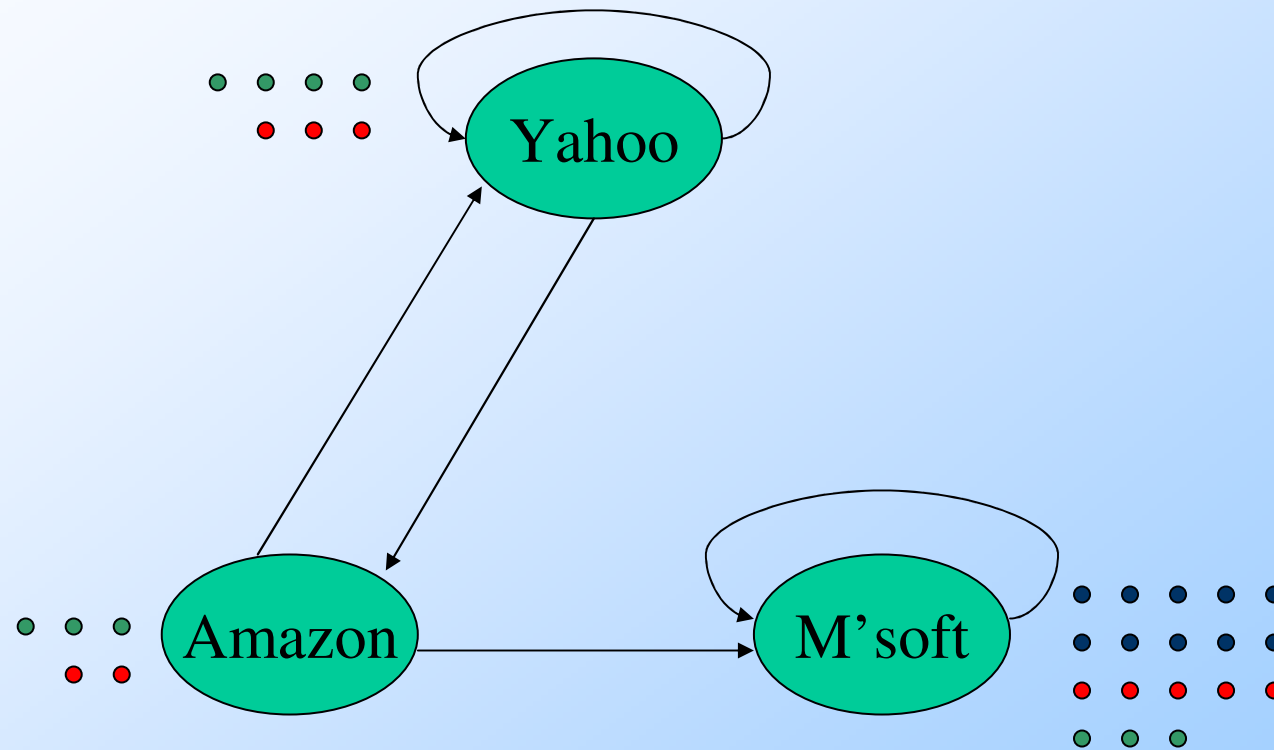
# Microsoft Becomes a Spider Trap



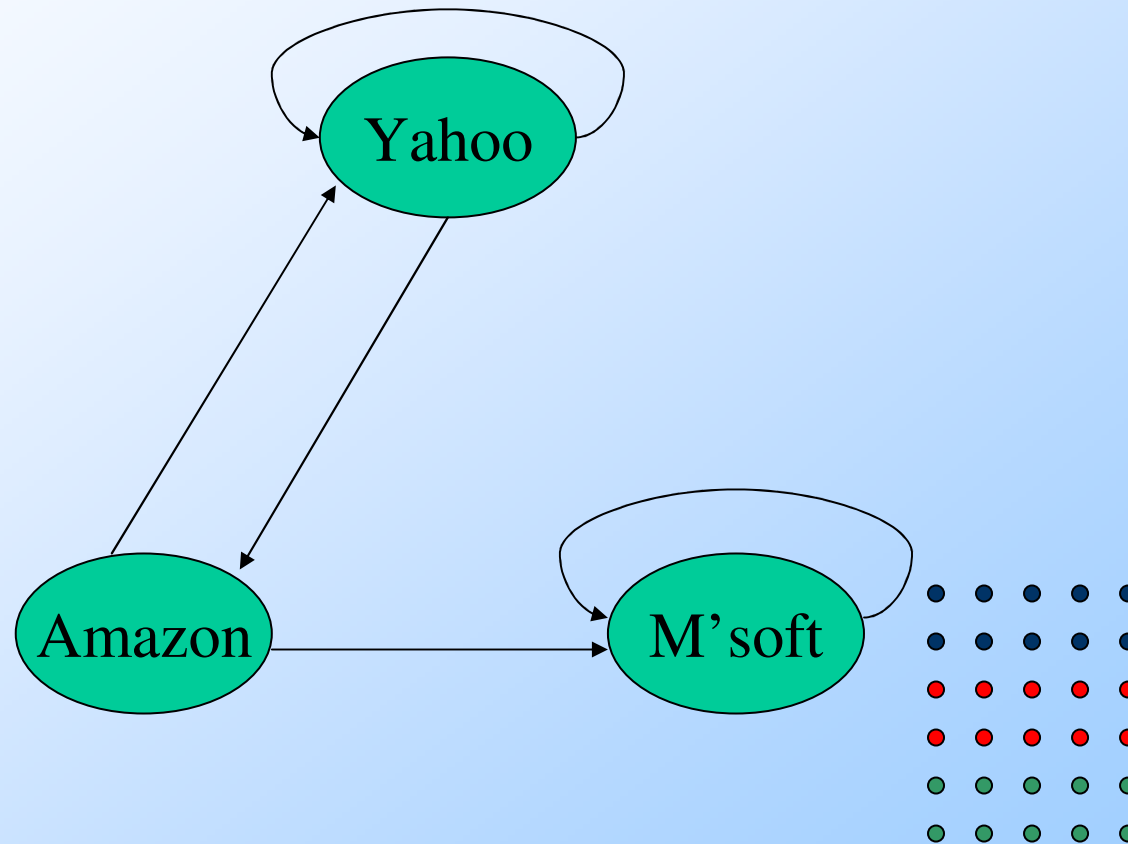
# Microsoft Becomes a Spider Trap



# Microsoft Becomes a Spider Trap



# In the Limit ...



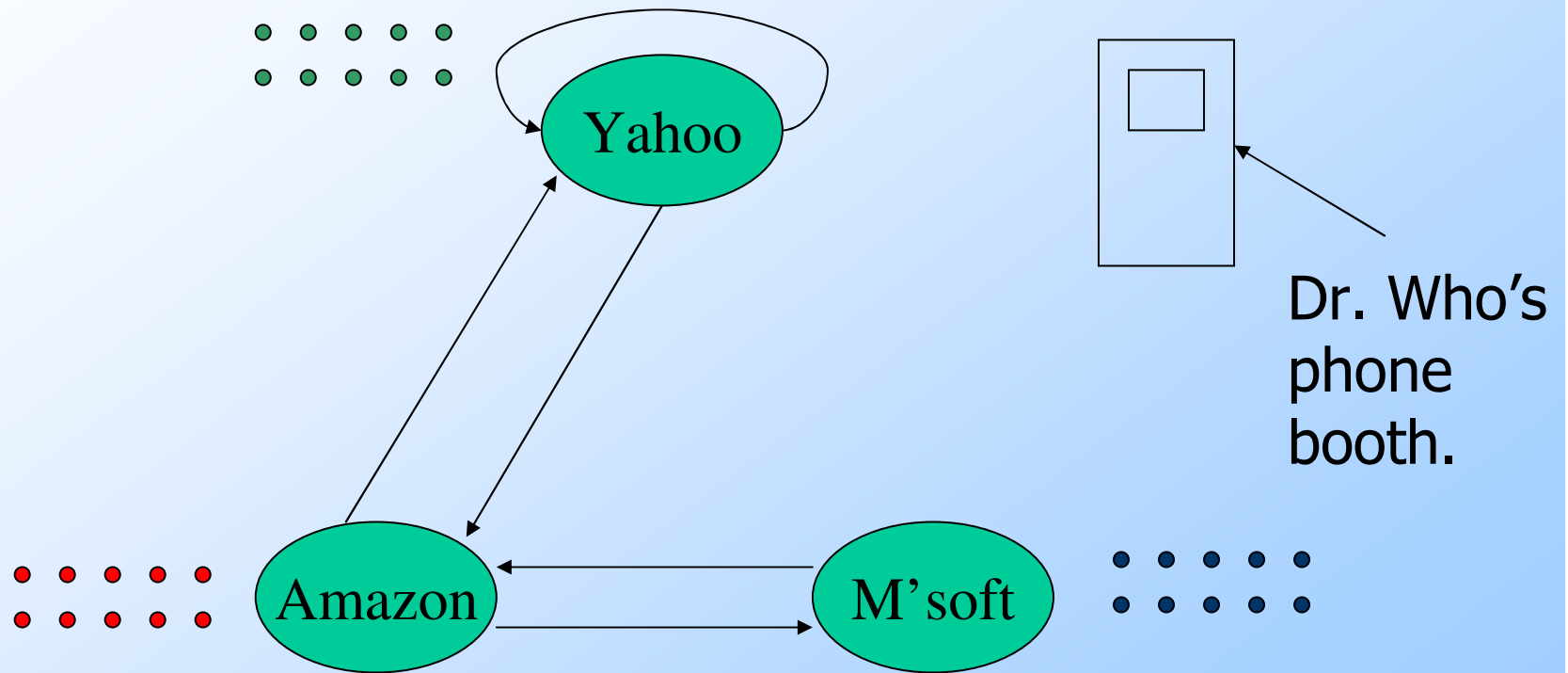
# Teleport Sets

- ◆ Assume each walker has a small probability of “teleporting” at any tick.
- ◆ Teleport can go to:
  1. Any page with equal probability.
    - ◆ To avoid dead-end and spider-trap problems.
  2. A topic-sensitive set of “relevant” pages (*teleport set*).
    - ◆ For *topic-sensitive* PageRank.

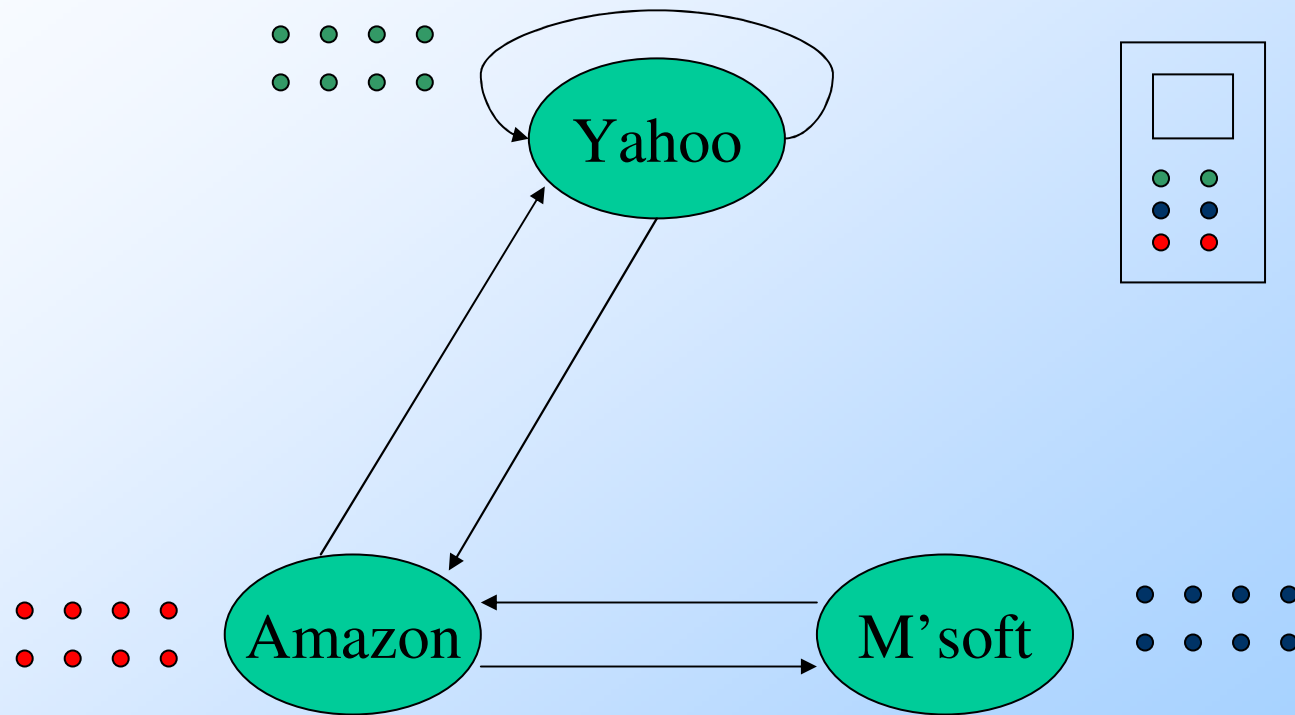
## Example: Topic = Software

- ◆ Only Microsoft is in the teleport set.
- ◆ Assume 20% "tax."

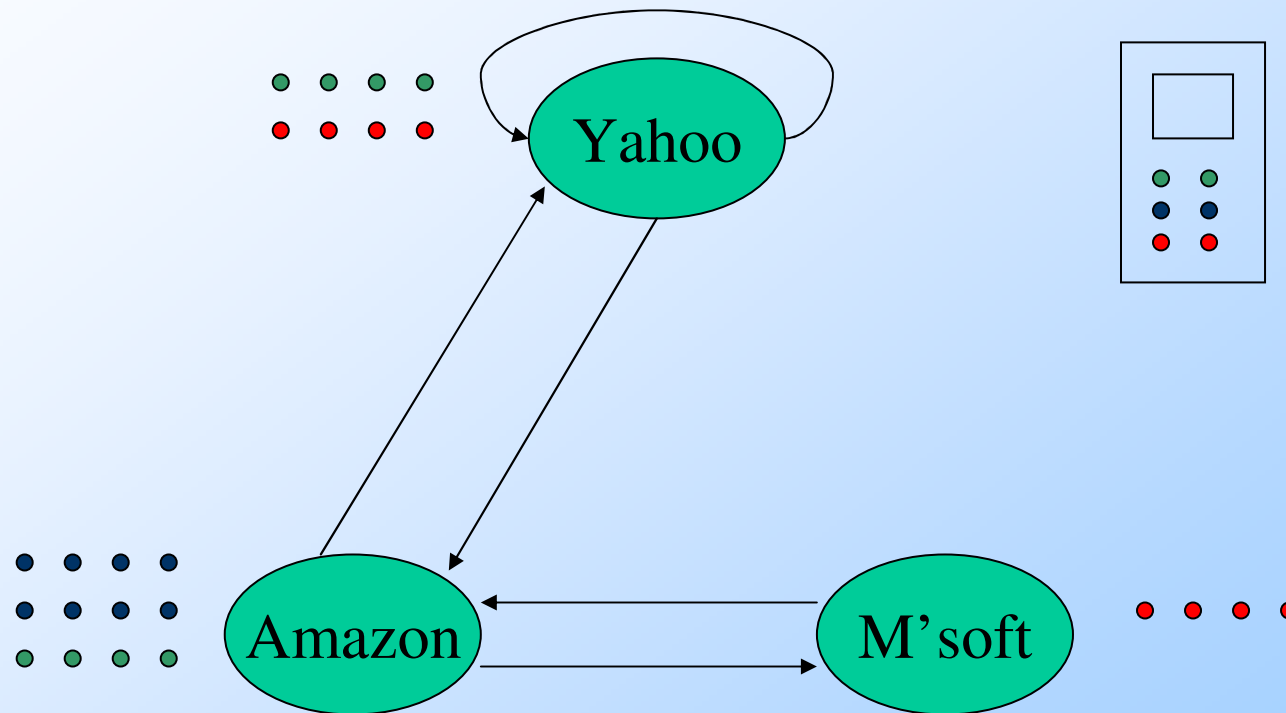
# Only Microsoft in Teleport Set



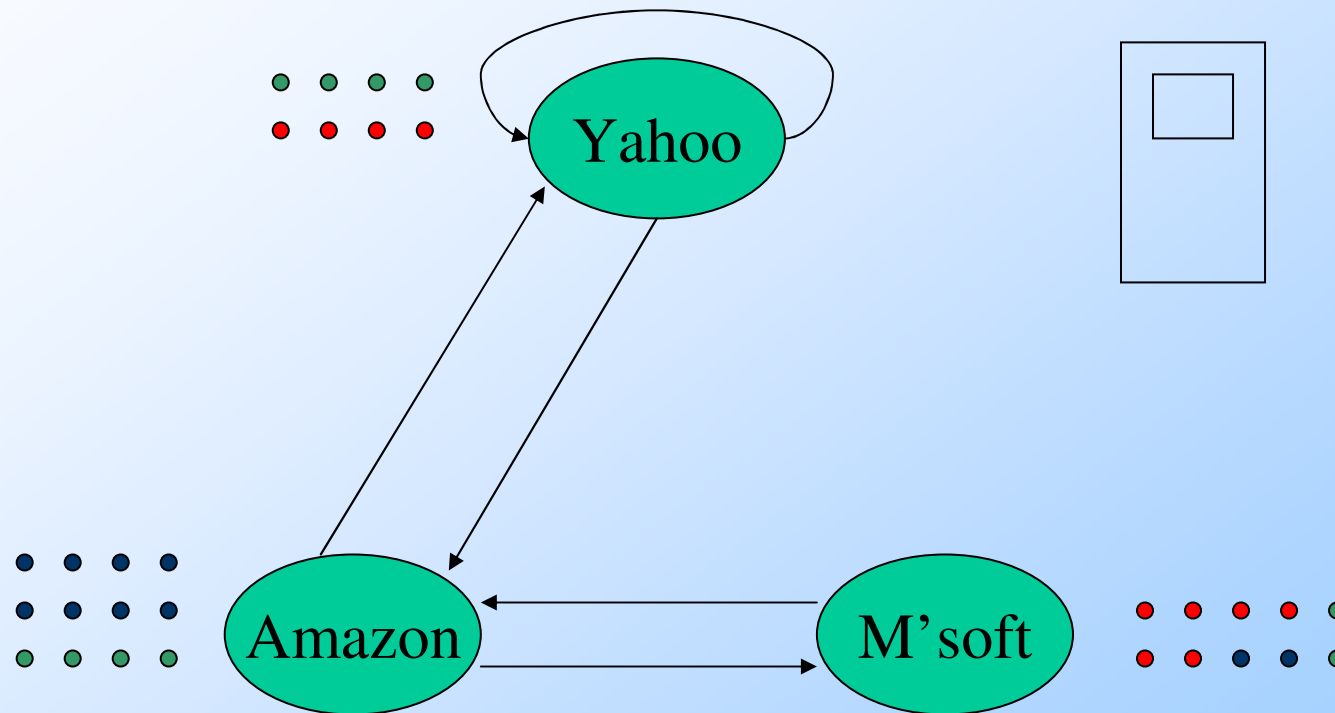
# Only Microsoft in Teleport Set



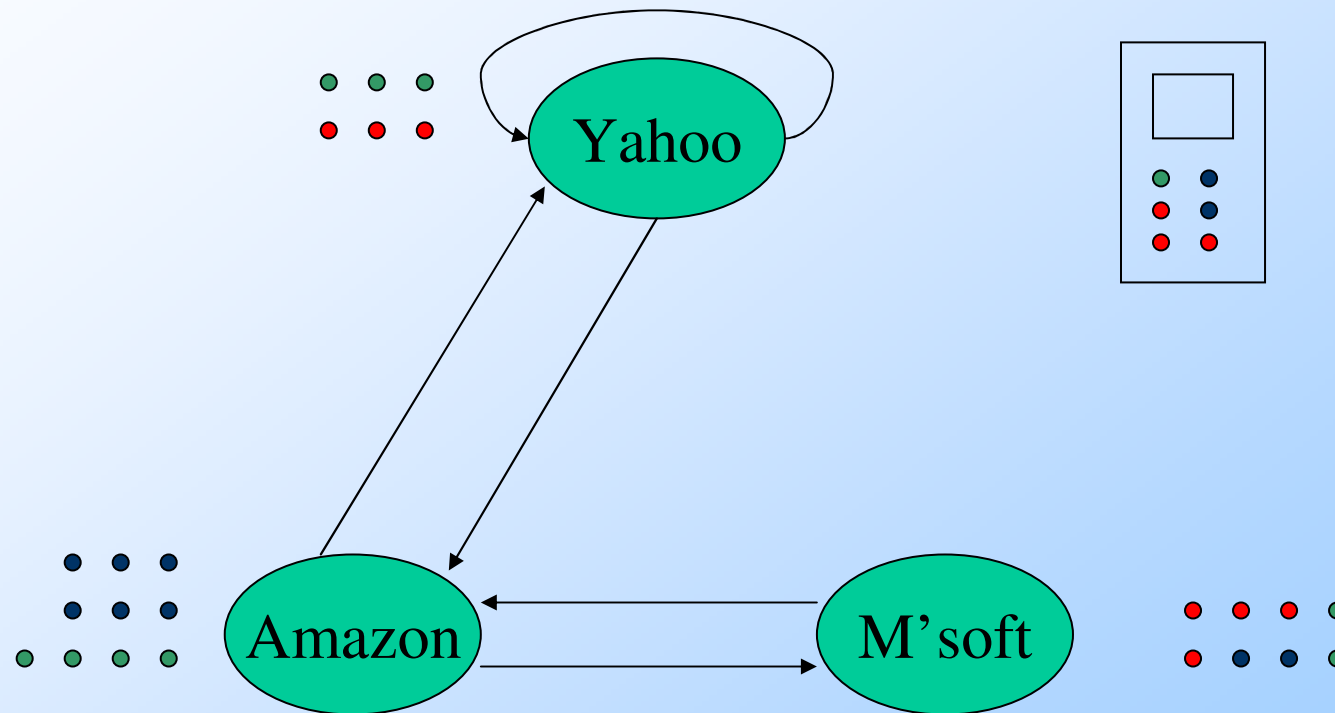
# Only Microsoft in Teleport Set



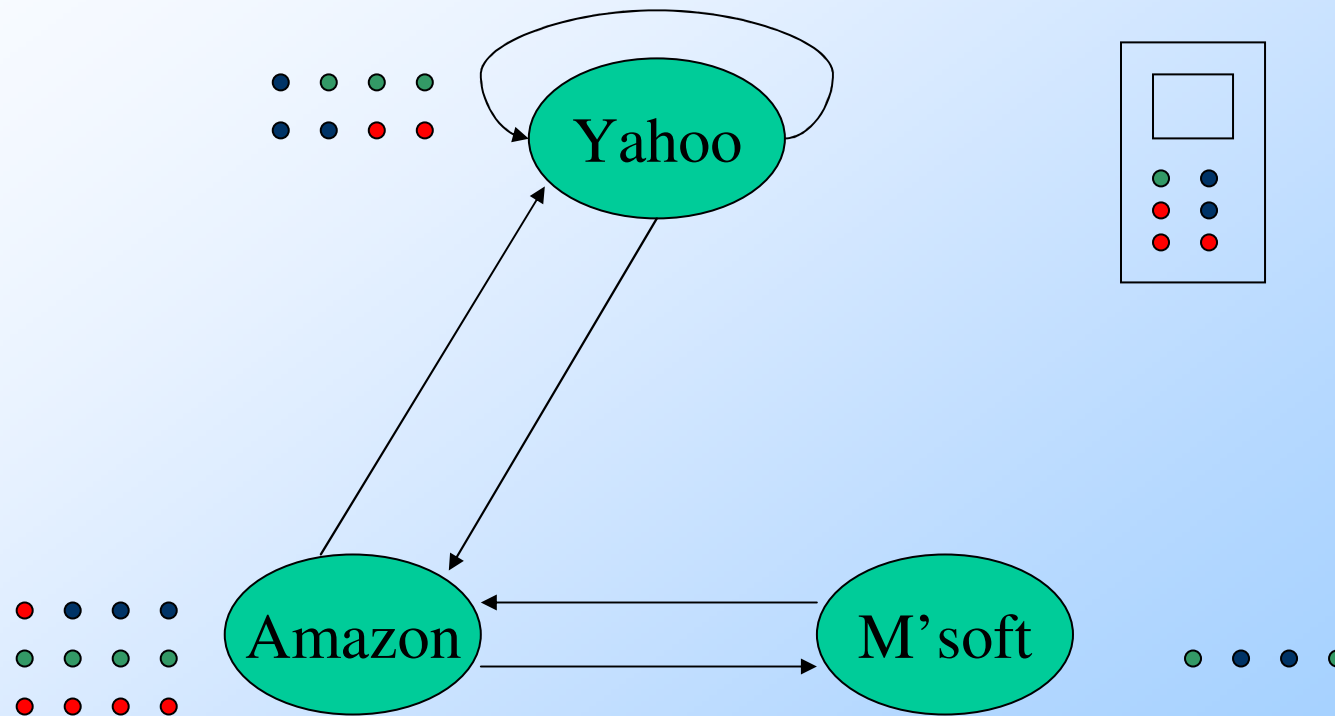
# Only Microsoft in Teleport Set



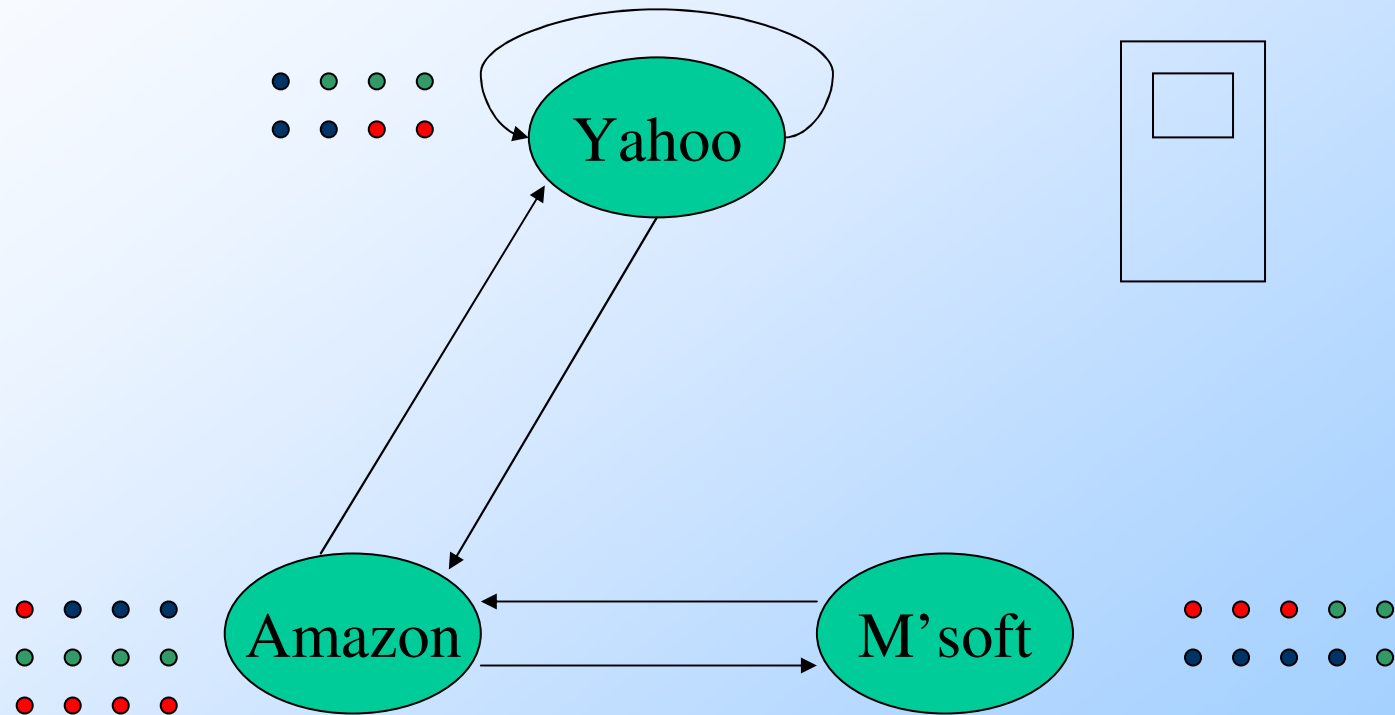
# Only Microsoft in Teleport Set



# Only Microsoft in Teleport Set



# Only Microsoft in Teleport Set



# Why Google Works

- ◆ Our hypothetical shirt-seller loses.
- ◆ His page isn't very important, so it won't be ranked high for shirts or movies.
- ◆ Saying he is about movies doesn't help, because others don't say he is about movies.

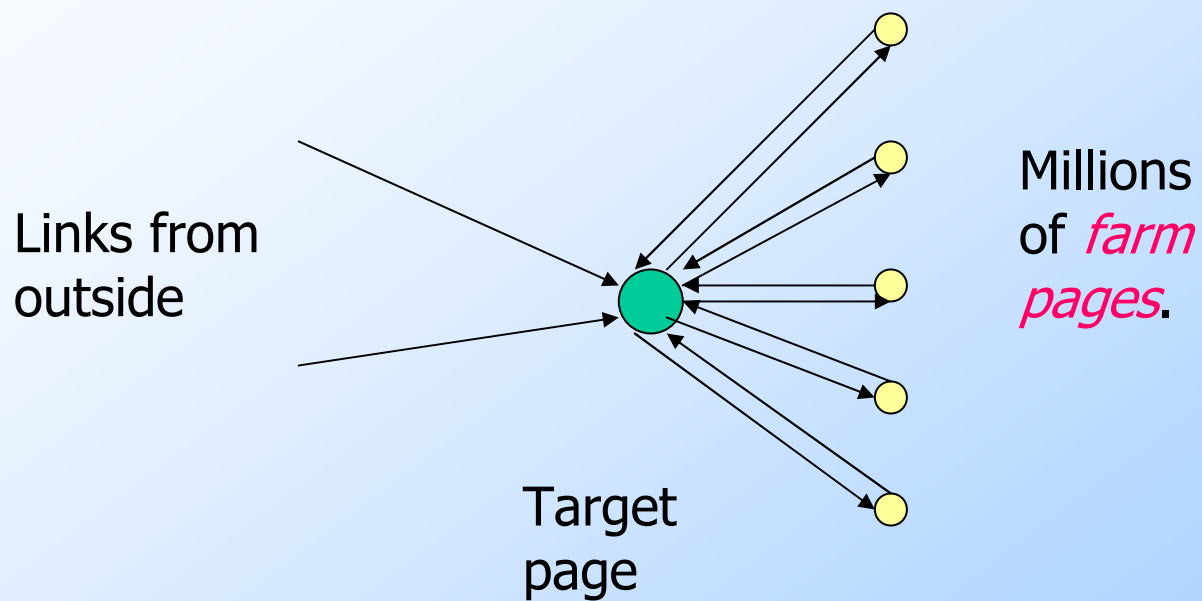
# Simple Spam Techniques Fail

- ◆ **Example:** shirt-seller creates 1000 pages, each of which links to his with `movie` in the anchor text.
- ◆ These pages have no links in, so they get little PageRank.
- ◆ So the shirt-seller can't beat truly important movie pages like OMDb.

## Round 2: *Link Spam*

- ◆ Once Google became the dominant search engine, spammers began to work out ways to fool Google.
- ◆ *Spam farms* were developed to concentrate PageRank on a single page.

# Structure of a Typical Spam Farm



# Farm Pages

- ◆ Even with taxation, farm pages can preserve most of the PageRank that the farm starts with.
- ◆ And it amplifies externally supplied PageRank by a significant factor.

# External Links

- ◆ Where do external links come from?
- ◆ Blog pages allow spammers to add comments, e.g., “I agree. See [www.mySpamFarm.com](http://www.mySpamFarm.com).”

# Combating Link Spam

1. Detection and blacklisting of structures that look like spam farms.
  - ◆ Leads to another war – hiding and detecting spam farms.
2. *TrustRank* = topic-specific PageRank with a teleport set of “trusted” pages.
  - ◆ **Example:** .edu domain, plus similar domains for non-US schools.

# Spam Mass

- ◆ Run ordinary PageRank and TrustRank.
- ◆ Pages whose TrustRank is much less than their PageRank are said to have high *spam mass* and are likely to be part of a spam farm.

# Future Consequences of Reliable Search

1. Advertising moving on-line.
2. Textbook market destroyed.
3. Newspapers destroyed.

# Advertising

- ◆ The original Brin/Page article on Google says “we do not believe advertising is a way to support search.”
  - ◆ True if “advertising” meant the DoubleClick display ad that took 10 seconds to load.
  - ◆ Took years for people to trust search enough that they would use it to find vendors.

# Why is Advertising Moving On-Line?

- ◆ Pay-per-click model is “measurable.”
  - ◆ But so is newspaper advertising – run the ad in one city, and not in a similar city.
- ◆ Ability to target.
  - ◆ Raises privacy issues.
    - My position: OK as long as done by machines.
    - Do you care if your toaster sees you naked?

# Textbooks

- ◆ Internet has made resale feasible; reliable search makes it easy.
- ◆ Leads to lower sales, annoying tricks by publishers.
  - ◆ **Example:** I am asked to reorder exercises so old editions cannot be used.

# Textbooks – (2)

- ◆ Trips to the library are replaced by search queries.
  - ◆ Academics are happy to put their slides and course notes on-line for free.
  - ◆ PageRank elevates the best of these to the top of the list.

# Textbooks – (3)

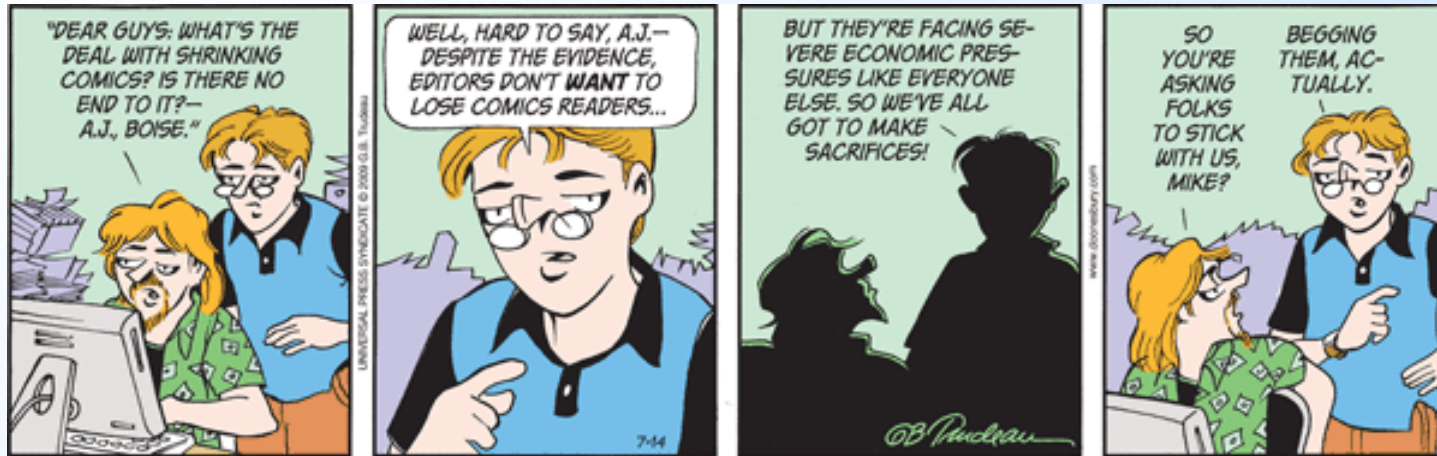
- ◆ Royalties for books are a relatively modern invention.
- ◆ The Internet may take us back to situation where you wrote for the glory.
- ◆ **Example:** Jokes were never copyrighted, even though they are intellectual property.
  - ◆ Why? Easy transmission was possible without the Internet.

# Who Killed Newspapers?

- ◆ It wasn't Google, exactly, although sucking advertising to search doesn't help.
- ◆ Newspapers always made their money from **classified** ads, not display ads.
- ◆ So blame Craig's List and similar sites for stealing the classified business.

# How Do We Know They're Dead?

- ◆ I read it in the newspaper.
  - ◆ Well the on-line newspaper, anyway.



From the NY Times, April 27, 2009:

## Fall in Newspaper Sales Accelerates to Pass 7%

By [TIM ARANGO](#)

The rate of decline in print circulation at the nation's newspapers has accelerated since last fall, as industry figures released Monday show a more than 7 percent drop compared with the previous year, while another recent analysis showed that newspaper Web site audiences had increased 10.5 percent in the first quarter.

# From the Newspaper Association of America:

Year	Morn., Aft., Total Newspapers			Morn., Aft., Total Circulation			Sunday Newspapers, Circulation	
2000	766	727	1,480	46,772	9,000	55,773	917	59,421
2001	776	704	1,468	46,821	8,756	55,578	913	59,090
2002	777	692	1,457	46,617	8,568	55,186	913	58,780
2003	787	680	1,456	46,930	8,255	55,185	917	58,495
2004	814	653	1,457	46,887	7,738	54,626	915	57,754
2005	817	645	1,452	46,122	7,222	53,345	914	55,270
2006	833	614	1,437	45,441	6,888	52,329	907	53,179
2007	867	565	1,422	44,548	6,194	50,742	907	51,246
2008	872	546	1,408	42,757	5,840	48,597	902	49,115

Down 13%
Down 17%

# Benefits of On-Line News

- ◆ More economical delivery of news.
- ◆ News aggregators like Yahoo! News or Google News are a big win for the consumer.
  - ◆ Search by topic.
  - ◆ See differing viewpoints on the same story.

# Dark Side of On-Line News

- ◆ Unlike many occupations being eliminated by the Internet, news reporting serves a vital function in a democracy.
- ◆ Who is going to pay for true journalism?
- ◆ Are bloggers a substitute? Is “crowdsourcing”?

# Some Interesting Research Questions

1. *Collaborative filtering* : suggest news to read based on what people like you are reading.
2. Eliminate near-identical versions of the same basic article.
3. Cluster articles by their content.
  - ◆ **Example:** "Steve Jobs has heart attack" vs. "Rumors Steve Jobs has heart attack are false."

# Summary

- ◆ Reliable search requires constant defense against those who would subvert it for their own purposes.
- ◆ Only institutions that can move on-line can be supported by advertising.
- ◆ Newspapers are particularly threatened (sigh!).
- ◆ So are textbooks (yaay!).