Exploiting social networks for Internet search

Alan Mislove[†‡] Krishna Gummadi[†] Peter Druschel[†]

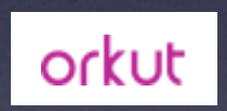
[†]Max Planck Institute for Software Systems
[‡]Rice University

HotNets 2006

Search in the Internet

- Web has transformed information exchange

- Social networking is now a popular way to share content
  - Photos, videos, blogs, music and profiles
  - MySpace (100 M users), Orkut (30 M users), ...



- Many studies examined Web: Web search well understood
  - Few looked at social networks

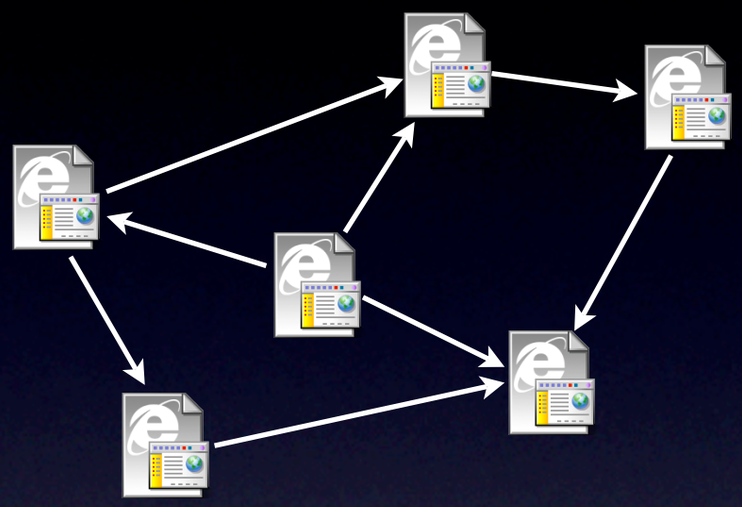
This talk

- Compares content sharing in the Web and social networks
  - Shows underlying mechanisms for publishing and locating differ
  - Examines implications for locating various types of content
- Investigates benefit of using social network search over Web

Web vs. social networks: Publishing

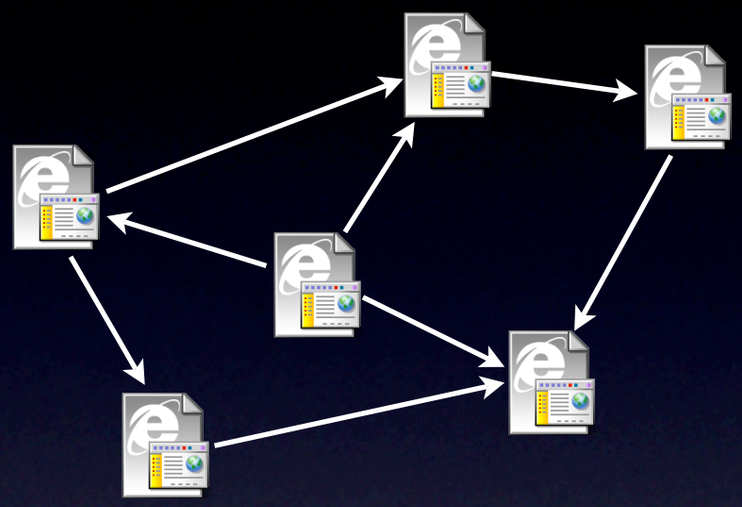
- In Web, links exist between content
  - Hyperlink is endorsement of relevance
- In social networks, no links between content
  - Links between users and content they create or endorse
  - Links between users with common interests or trust
- Different link structures affect how content is located

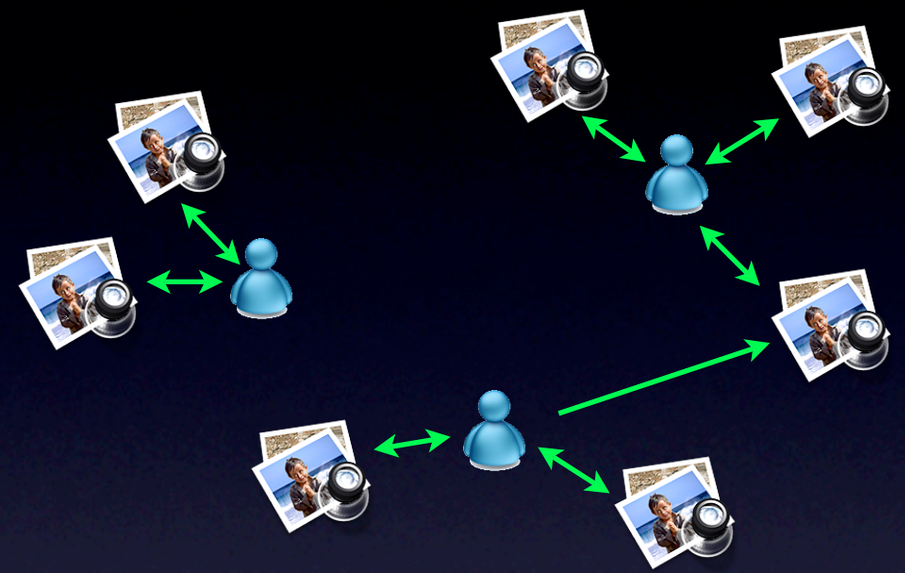
Web vs. social networks: Publishing



- In Web, links exist between content
  - Hyperlink is endorsement of relevance
- In social networks, no links between content
  - Links between users and content they create or endorse
  - Links between users with common interests or trust
- Different link structures affect how content is located

Web vs. social networks: Publishing



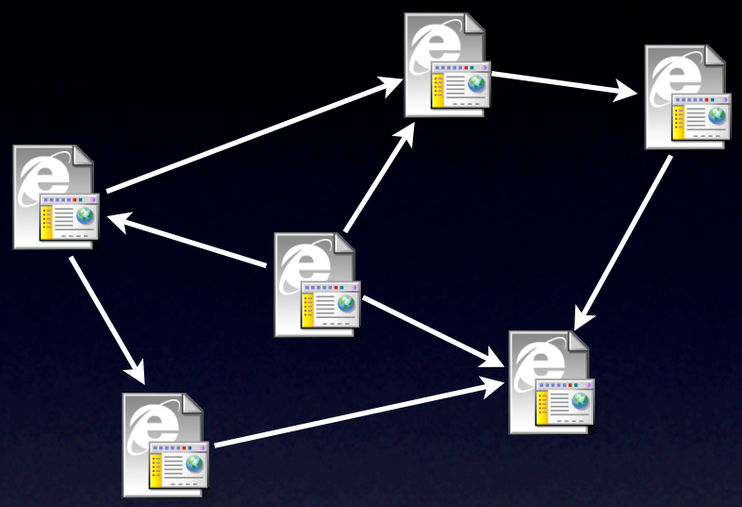


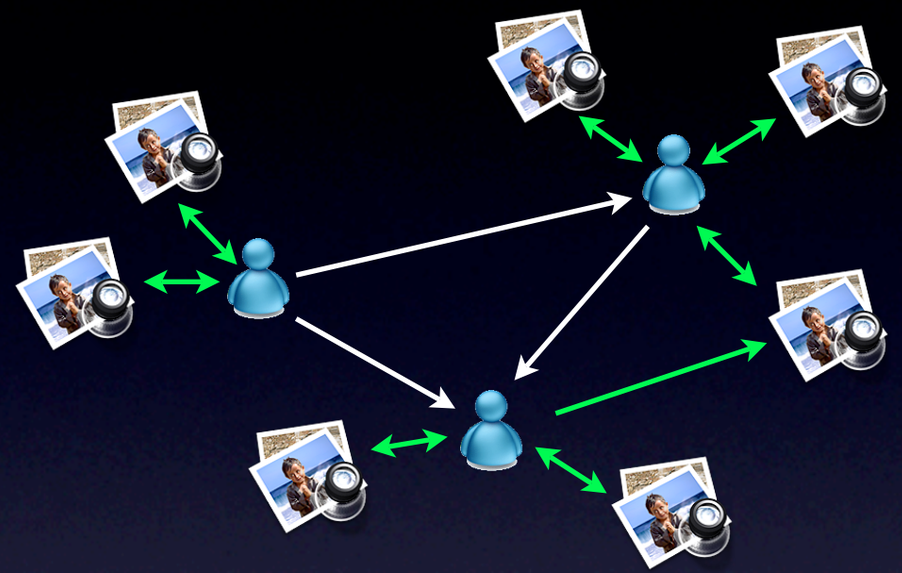
- In Web, links exist between content
  - Hyperlink is endorsement of relevance

- In social networks, no links between content
  - Links between users and content they create or endorse
  - Links between users with common interests or trust

- Different link structures affect how content is located

Web vs. social networks: Publishing





- In Web, links exist between content
  - Hyperlink is endorsement of relevance

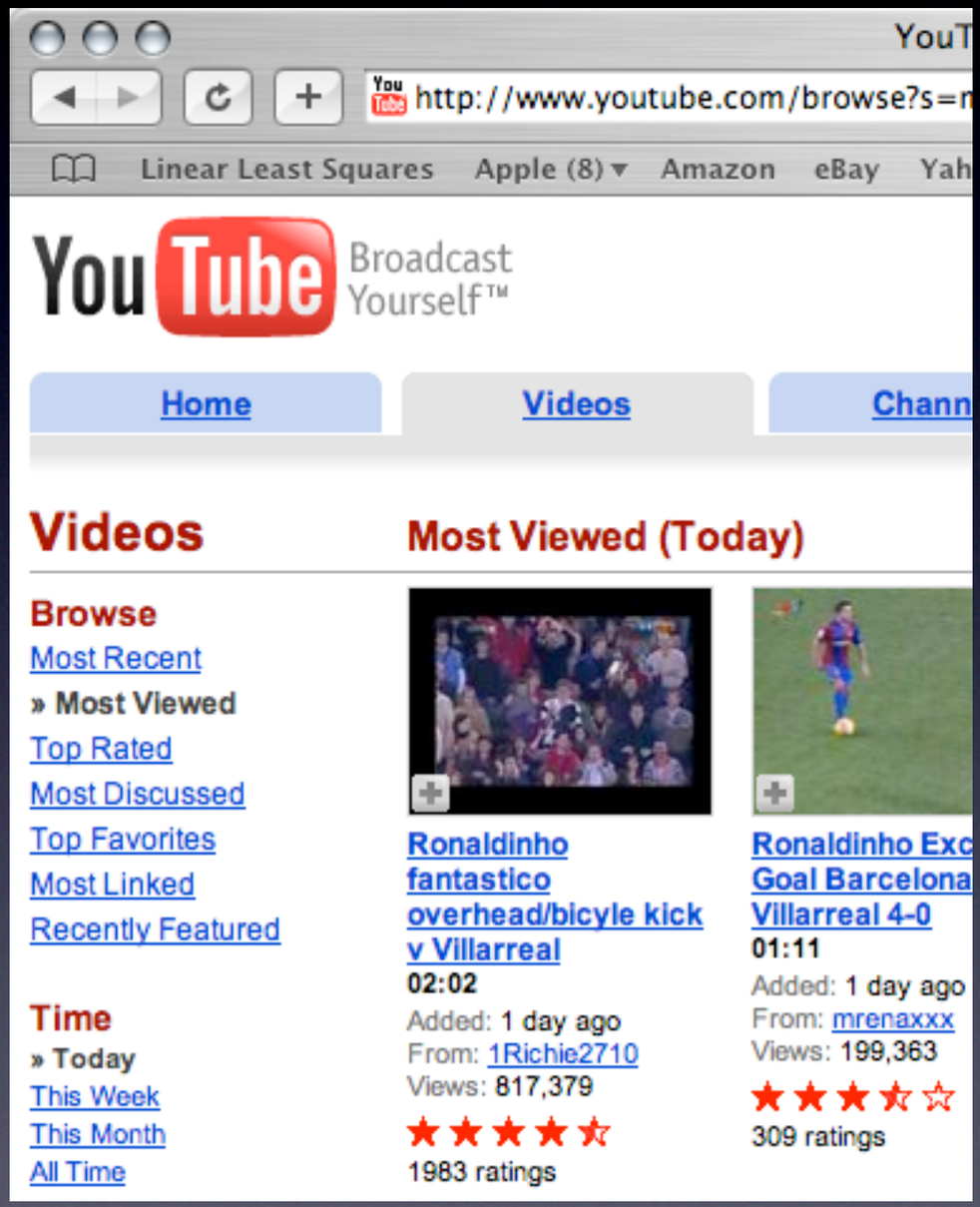
- In social networks, no links between content
  - Links between users and content they create or endorse
  - Links between users with common interests or trust

- Different link structures affect how content is located

Web vs. social networks: Locating

- Web search exploits hyperlink structure
  - More incoming links imply importance

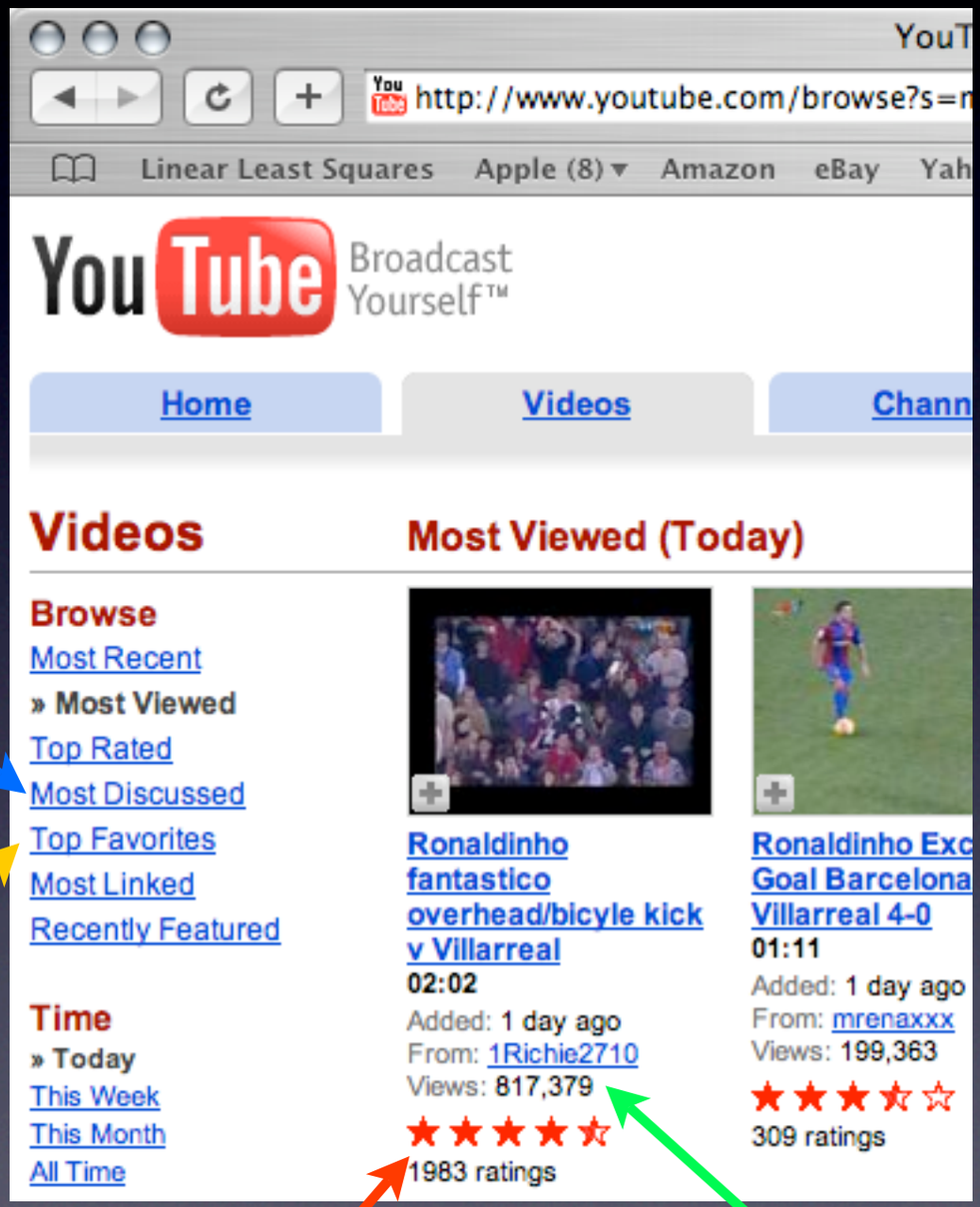
- Social networks use user feedback
  - Implicit (e.g. # of views)
  - Explicit (e.g. rating, # of comments, favorites)



Web vs. social networks: Locating

- Web search exploits hyperlink structure
  - More incoming links imply importance

- Social networks use user feedback
  - Implicit (e.g. # of views)
  - Explicit (e.g. rating, # of comments, favorites)



What content do social nets locate better?

- Recently added content
  - Creating Web links takes time, social nets rapidly rate content

- Information of interest to a specific community
  - Web ratings reflect interests of community at large
  - Web search misses deep web content

- Multimedia content
  - Hard to link content instances
  - Social network uses tags and comments

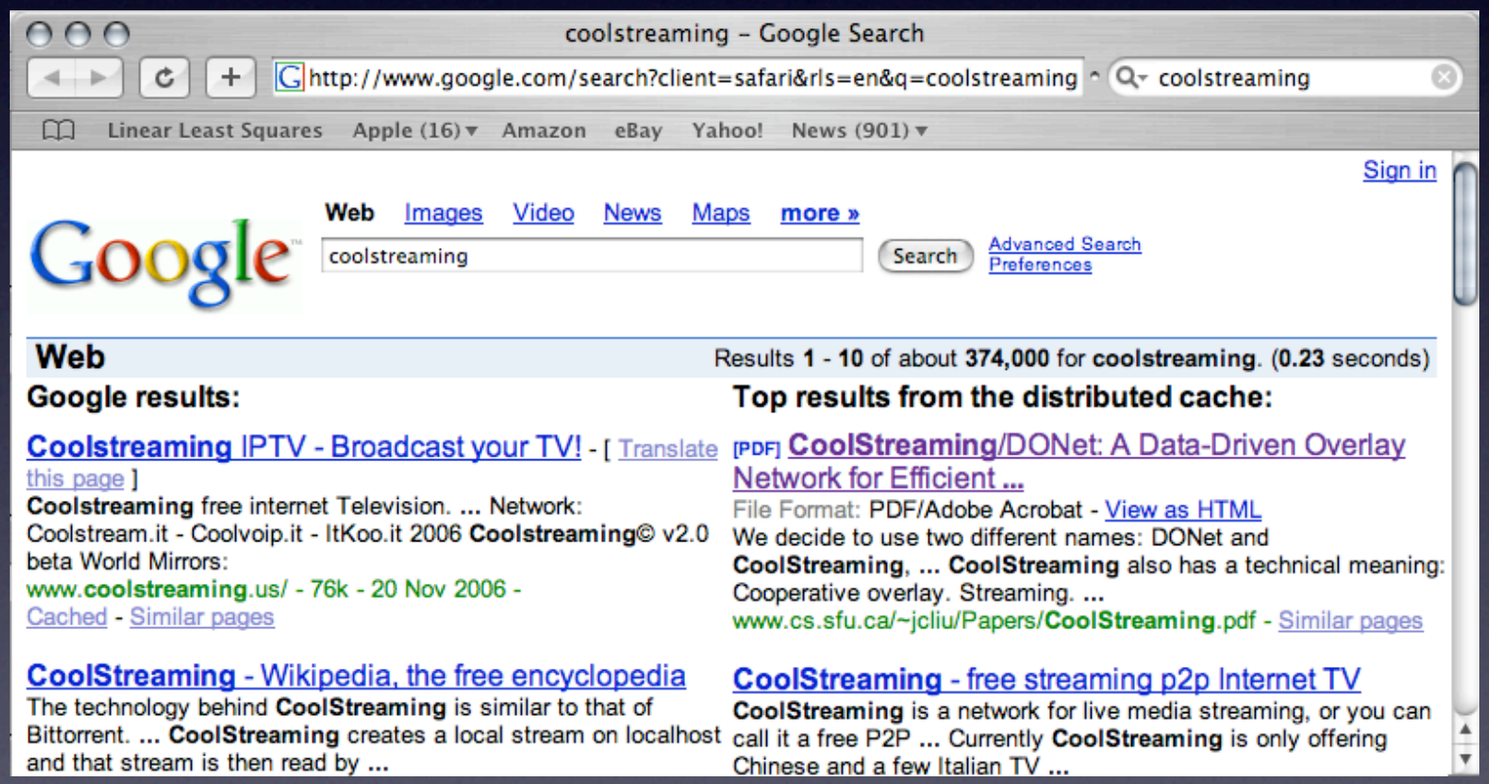
- Can this Web content be better located with social networks?

What content do social nets locate better?

- Recently added content
  - Creating Web links takes time, social nets rapidly rate content
- Information of interest to a specific community
  - Web ratings reflect interests of community at large
  - Web search misses deep web content
- Multimedia content
  - Hard to link content instances
  - Social network uses tags and comments
- Can this Web content be better located with social networks?

Applying social network search to Web

- PeerSpective experiment uses social nets to search the Web
- High level idea: users can query their friends' viewed pages



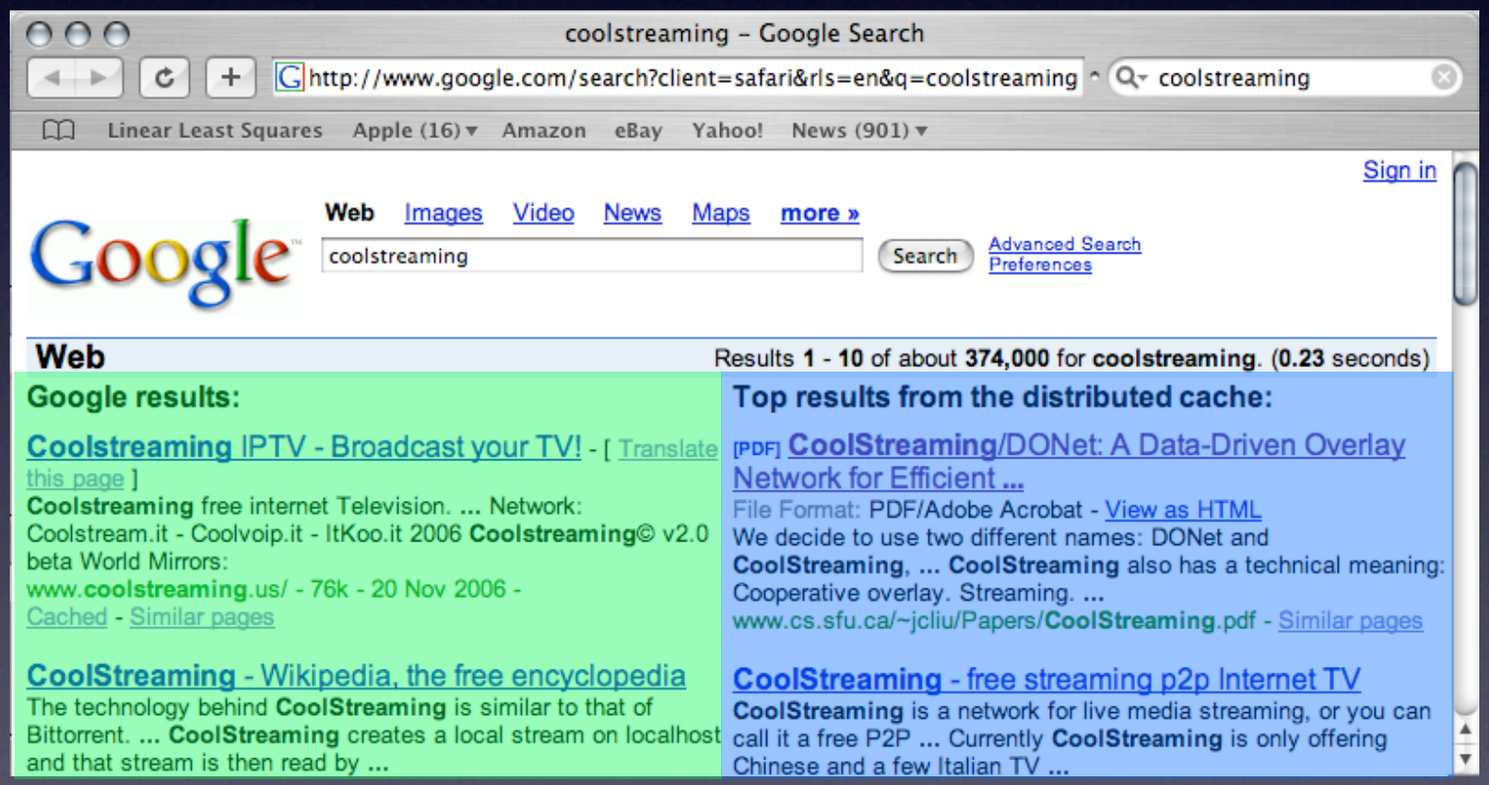
- Results from friends appear alongside Google results

Applying social network search to Web

- PeerSpective experiment uses social nets to search the Web
- High level idea: users can query their friends' viewed pages

Google

PeerSpective



- Results from friends appear alongside Google results

PeerSpective implementation

- Prototype is a lightweight HTTP proxy
  - Runs on users' desktop and indexes all browsed content
- When Google search is performed
  - Query other PeerSpective proxies in parallel with Google
  - Present results alongside each other

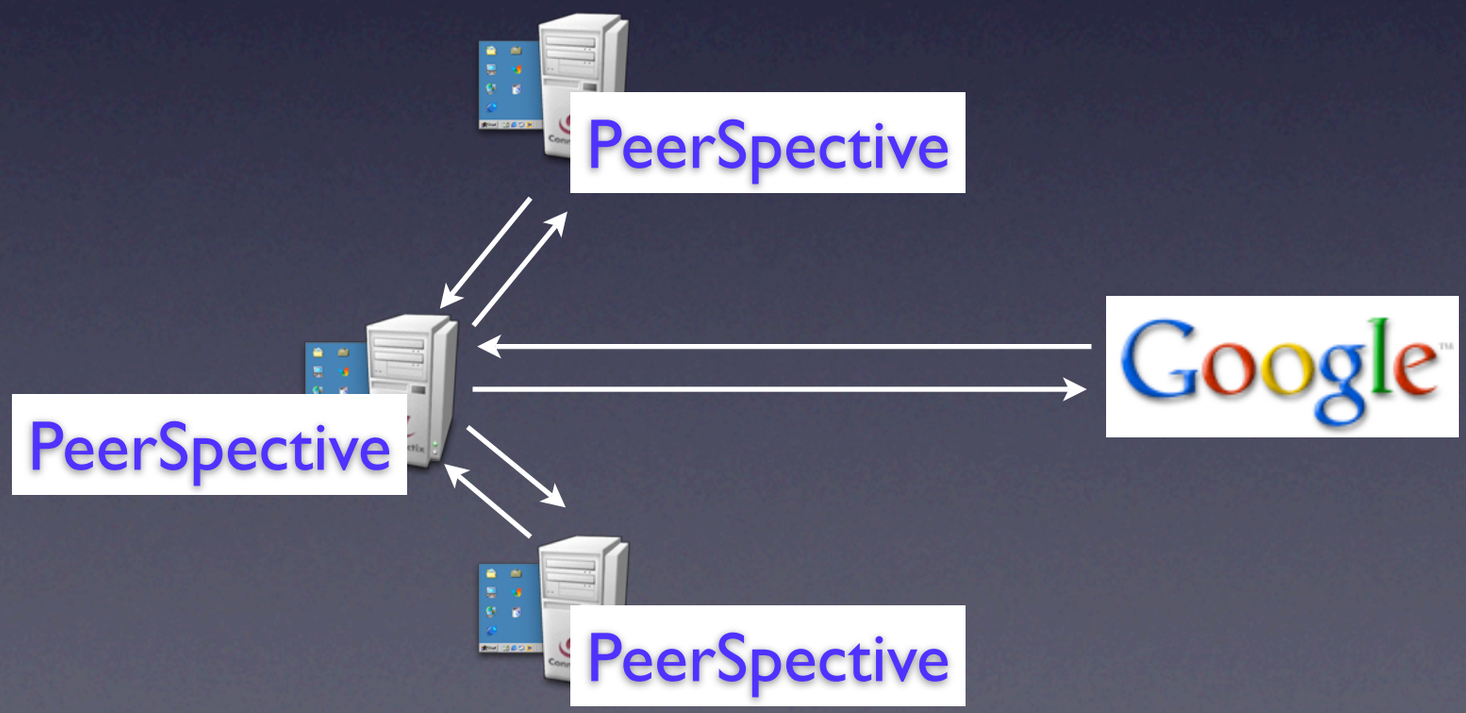



PeerSpective implementation

- Prototype is a lightweight HTTP proxy
  - Runs on users' desktop and indexes all browsed content
- When Google search is performed
  - Query other PeerSpective proxies in parallel with Google
  - Present results alongside each other

PeerSpective implementation

- Prototype is a lightweight HTTP proxy
  - Runs on users' desktop and indexes all browsed content
- When Google search is performed
  - Query other PeerSpective proxies in parallel with Google
- Present results alongside each other



Questions to answer

- Does PeerSpective improve coverage?
  - What is the coverage of Google's index for viewed pages?
  - What fraction of URLs already viewed by a friend?
- How good is PeerSpective at ranking results?
  - Do users click on PeerSpective or Google results?

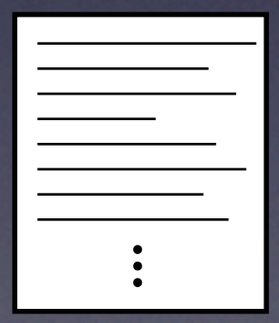
High-level results

- Ran PeerSpective with 10 users for one month
  - All users were researchers at MPI
  - 51,410 distinct URLs viewed
  - 1,730 Google searches

- Caveat: Small data set from group of computer scientists
  - User group includes authors
  - Results indicate potential, at least for special interest groups

What fraction of viewed URLs does Google index?

- Limited to static pages (`text/html` ending in `.html` or `.htm`)
- Queried Google's index for each URL
  - Using `about:URL` search request
- **Google contained only 62.5%** of URLs!
  - Representing 68.1% of HTTP requests



What fraction of viewed URLs does Google index?

- Limited to static pages (`text/html` ending in `.html` or `.htm`)
- Queried Google's index for each URL
  - Using `about:URL` search request
- Google contained only 62.5% of URLs!
  - Representing 68.1% of HTTP requests

Why are so many URLs not in Google?

- Examined URL list, found three reasons
  - Too new: Google has not had time to crawl this URL

    `http://edition.cnn.com/2006/ ... /italy.nesta/index.html`

  - Deep web: URL is not well-connected enough to crawl

    `http://www.mpi-sws.mpg.de/~pkouznet/ ... /pres0031.ht/pres0031.html`

  - Dark web: URL is not connected, or not visible

    `http://www.mpi-sws.org/intranet/index.htm`

What fraction of URLs viewed by a friend?

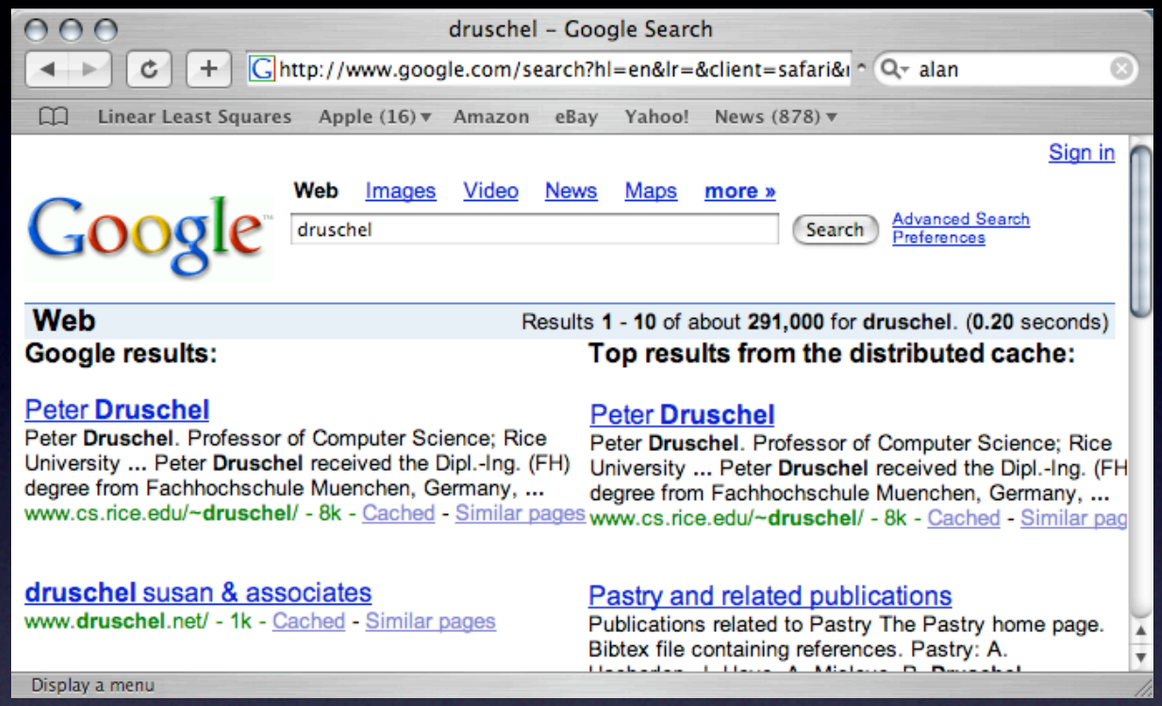
- Only static, `text/html` pages
  - Same methodology as Google coverage check

- 30.4% of URLs previously viewed by someone in network
  - Many previously viewed locally

- 13.3% of URLs previous viewed but not in Google!
  - Suggests social networks can extend index coverage
  - With comparatively small index

Did users click on PeerSpective results?

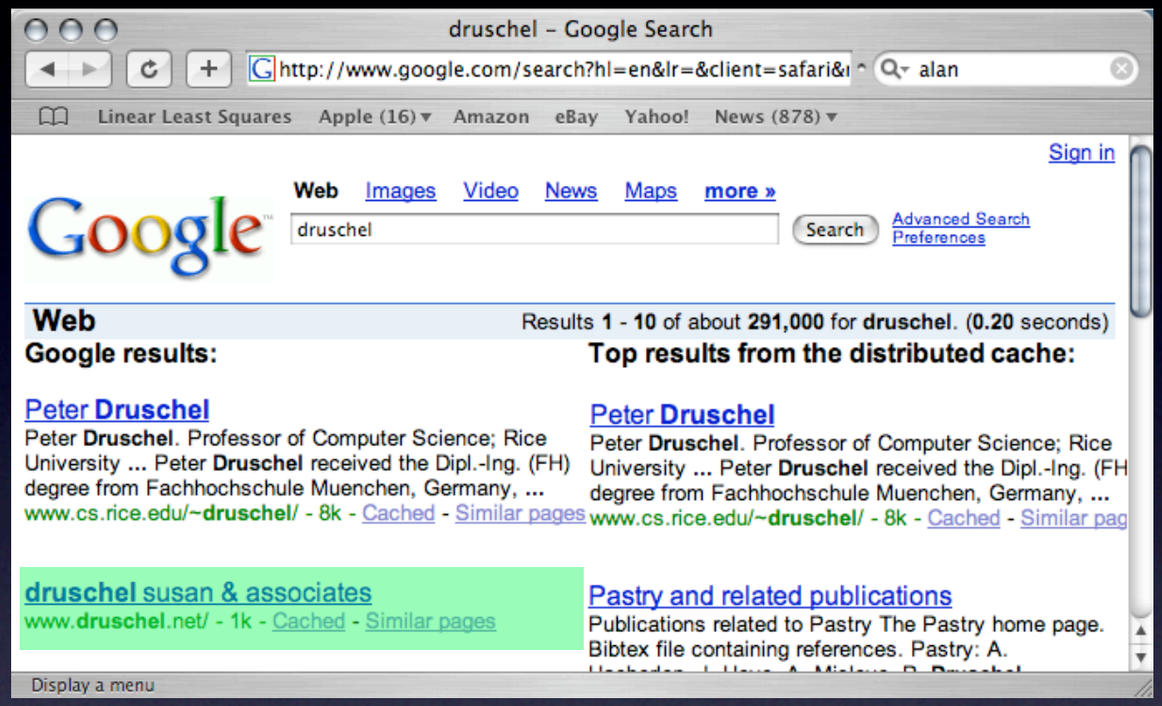
- For each result click, we ask
  - Only in Google's top-10?
  - Only in PeerSpective's top-10?
  - In top-10 from both?



- 7.7% of result clicks were on PeerSpective-only results!
  - Shows potential of social network search

Did users click on PeerSpective results?

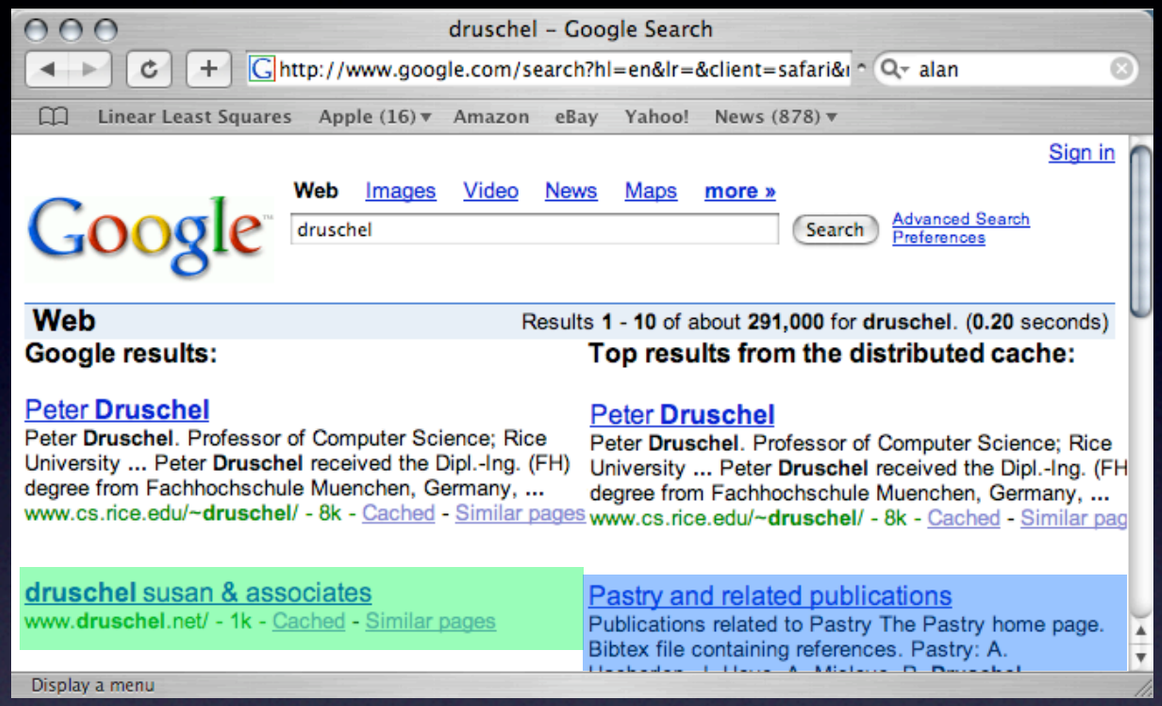
- For each result click, we ask
  - Only in Google's top-10?
  - Only in PeerSpective's top-10?
  - In top-10 from both?



- 7.7% of result clicks were on PeerSpective-only results!
  - Shows potential of social network search

Did users click on PeerSpective results?

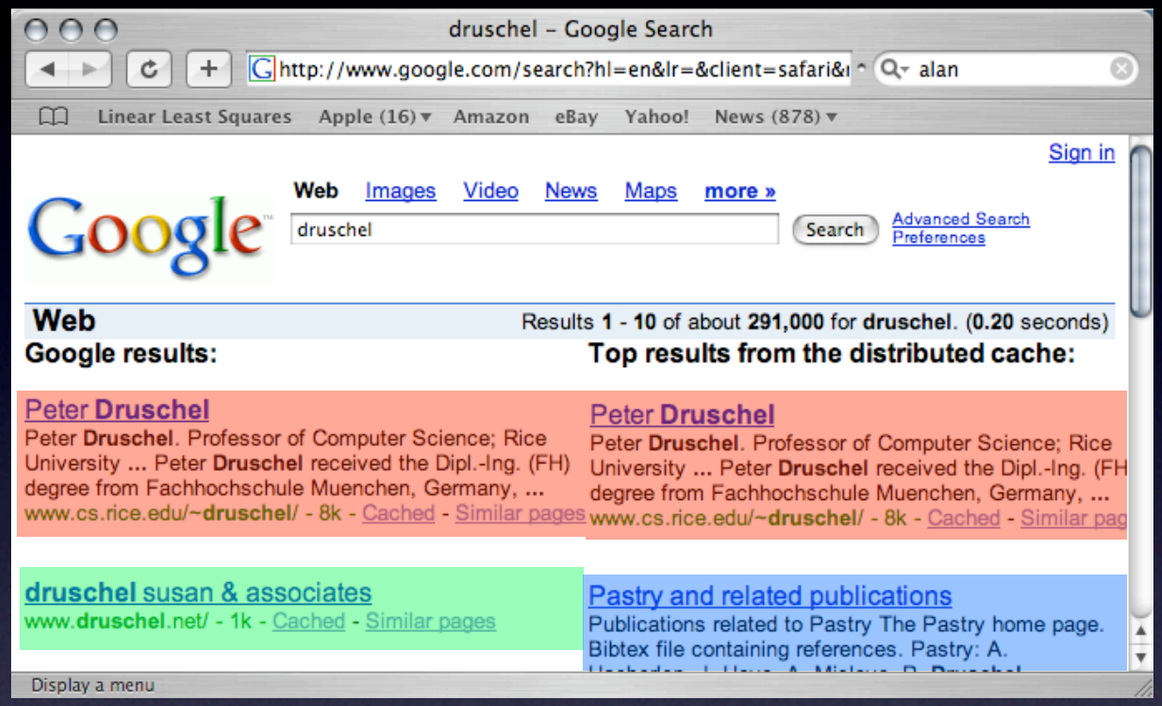
- For each result click, we ask
  - Only in Google's top-10?
  - Only in PeerSpective's top-10?
  - In top-10 from both?



- 7.7% of result clicks were on PeerSpective-only results!
  - Shows potential of social network search

Did users click on PeerSpective results?

- For each result click, we ask
  - Only in Google's top-10?
  - Only in PeerSpective's top-10?
  - In top-10 from both?



- 7.7% of result clicks were on PeerSpective-only results!
  - Shows potential of social network search

Why are PeerSpective-only URLs clicked on?

- Disambiguation: determining appropriate meaning of term
- Search engines currently pick most popular definition

MPI ?

Message Passing Interface

Max Planck Institute

Manitoba Public Insurance

Meeting Professionals International

- PeerSpective can leverage meaning relevant to friends

Why are PeerSpective-only URLs clicked on?

- Disambiguation: determining appropriate meaning of term
- Search engines currently pick most popular definition

MPI ?

Message Passing Interface

Max Planck Institute

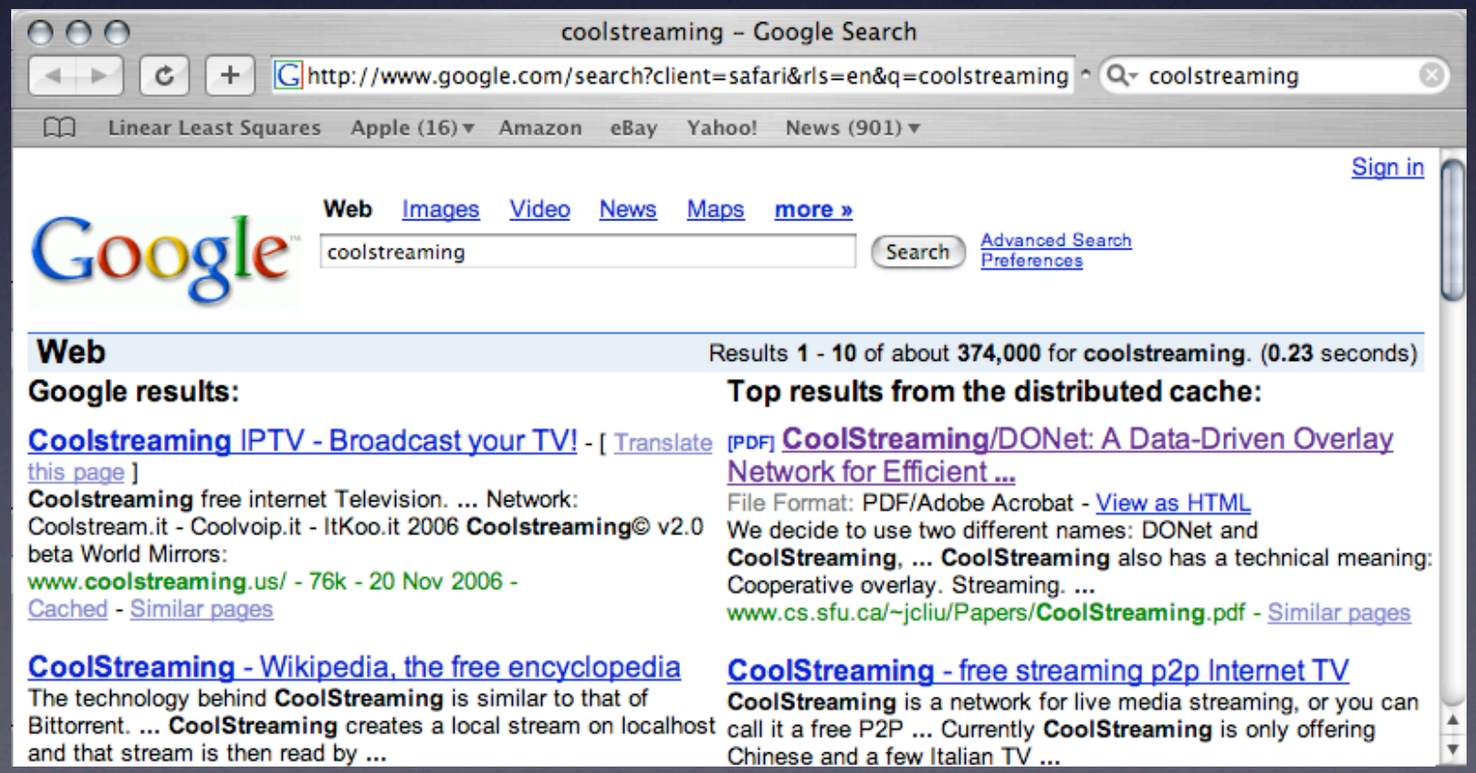
Manitoba Public Insurance

Meeting Professionals International

- PeerSpective can leverage meaning relevant to friends

Why are PeerSpective-only URLs clicked on?

- Relevance: picking best among matching documents
- Example: search for ‘coolstreaming’ leads to paper
- PeerSpective can use shared interests of friends



Why are PeerSpective-only URLs clicked on?

- Serendipity: finding interesting and unexpected content
  - Integral to web search experience
  - News sites are current examples of serendipitous sites

- Example: ‘Munich’ leads to co-worker’s homepage

- Serendipitous discoveries occur frequently in PeerSpective
  - Users often find pages viewed by friends interesting

Results summary

- PeerSpective explored potential of integrating Web and social network search

- Found that PeerSpective aided web search
  - Provided additional coverage for viewed sites
  - Improved ranking of results
  - Aided finding serendipitous content
  - Changed usage pattern of our users

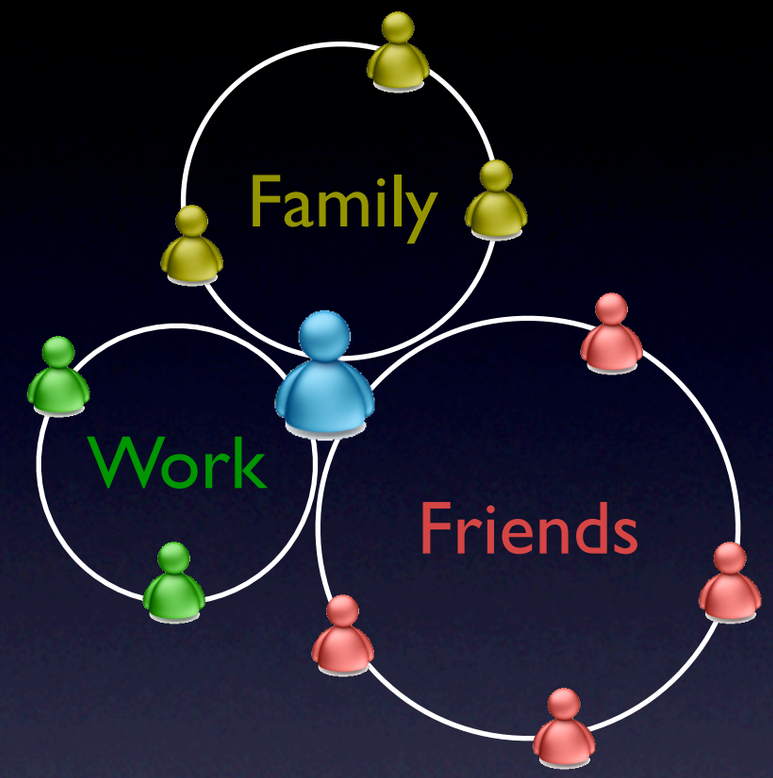
- However, just an experiment
  - Many challenges and opportunities to actual system

Opportunities and challenges

- Privacy
  - Users disclose someone in their group has viewed a URL
    - Subject to *k*-anonymity
  - In PeerSpective, currently
    - No HTTPS indexed
    - Allowed users to turn off indexing and purge pages
    - Search queries not recorded
- Need ways to ensure anonymity and privacy
  - While providing incentives to contribute

Opportunities and challenges

- Clustering
  - Users often members of multiple social groups
  - Necessary to route query to most useful users?



- Architecture
  - Centralized vs. decentralized?
    - Rather share URL history with centralized organization or friends?

- Others in the paper

Conclusion

- Content sharing mechanisms in Web and social nets differ widely
- Social nets are naturally better suited for certain content
- Early experiments suggest social nets can improve Web search
  - Found noticeable improvement in coverage and ranking
- Will soon release PeerSpective to the PlanetLab community

Questions?

What is the coverage of Google/PS?

In PeerSpective?

In Google?		Yes	No
	Yes	16.7%	45.8%
	No	13.3%	24.2%

What is the coverage of Google/PS?

In PeerSpective?

In Google?	Yes	No	
Yes	16.7%	45.8%	62.5%
No	13.3%	24.2%	

What is the coverage of Google/PS?

In PeerSpective?

In Google?	Yes	No	
Yes	16.7%	45.8%	62.5%
No	13.3%	24.2%	
	30.4%		

What is the coverage of Google/PS?

In PeerSpective?

In Google?	Yes	No	
Yes	16.7%	45.8%	62.5%
No	13.3%	24.2%	
	30.4%		

What results do users click on?

Google result?	PeerSpective result? Yes	PeerSpective result? No
Yes	5.8%	86.5%
No	7.7%	--

What results do users click on?

Google result?	PeerSpective result?		
	Yes	No	
Yes	5.8%	86.5%	92.3%
No	7.7%	--	

What results do users click on?

Google result?	PeerSpective result? Yes	PeerSpective result? No	
Yes	5.8%	86.5%	92.3%
No	7.7%	--	
	13.5%		

What results do users click on?

Google result? \ PeerSpective result?	Yes	No	
Yes	5.8%	86.5%	92.3%
No	7.7%	--	
	13.5%		