Trends in Web Content and Web Search

Andrew Tomkins

Chief Scientist, Yahoo! Search

Outline

- Content trends
  - Growth
  - Fragmentation and complexity
- New problems in search
- Academic research
- Sharing query logs

Content Growth

Content trends

Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	$\sim$ 2 GB
User generated content	8-10 GB
Private text content	$\sim$ 3 TB (300x more)
Upper bound on typed content	$\sim$700 TB ($\sim$200x more)

[Ramakrishnan and Tomkins 2007]

Metadata trends

Metadata type	Amount of metadata produced per day
Anchortext	100 MB
Tags	40 MB
Pageviews	180 GB
Reviews	Around 10 MB

[Ramakrishnan and Tomkins 2007]

Content becoming more complex

Content ownership

- Content ownership is fragmenting
- Y! represents ~10% of WW consumption, and dropping
- Much smaller fraction of total content
- No single place will own all the content
- Best of breed processing will operate on the web version (?)

Content Consumption is fragmenting

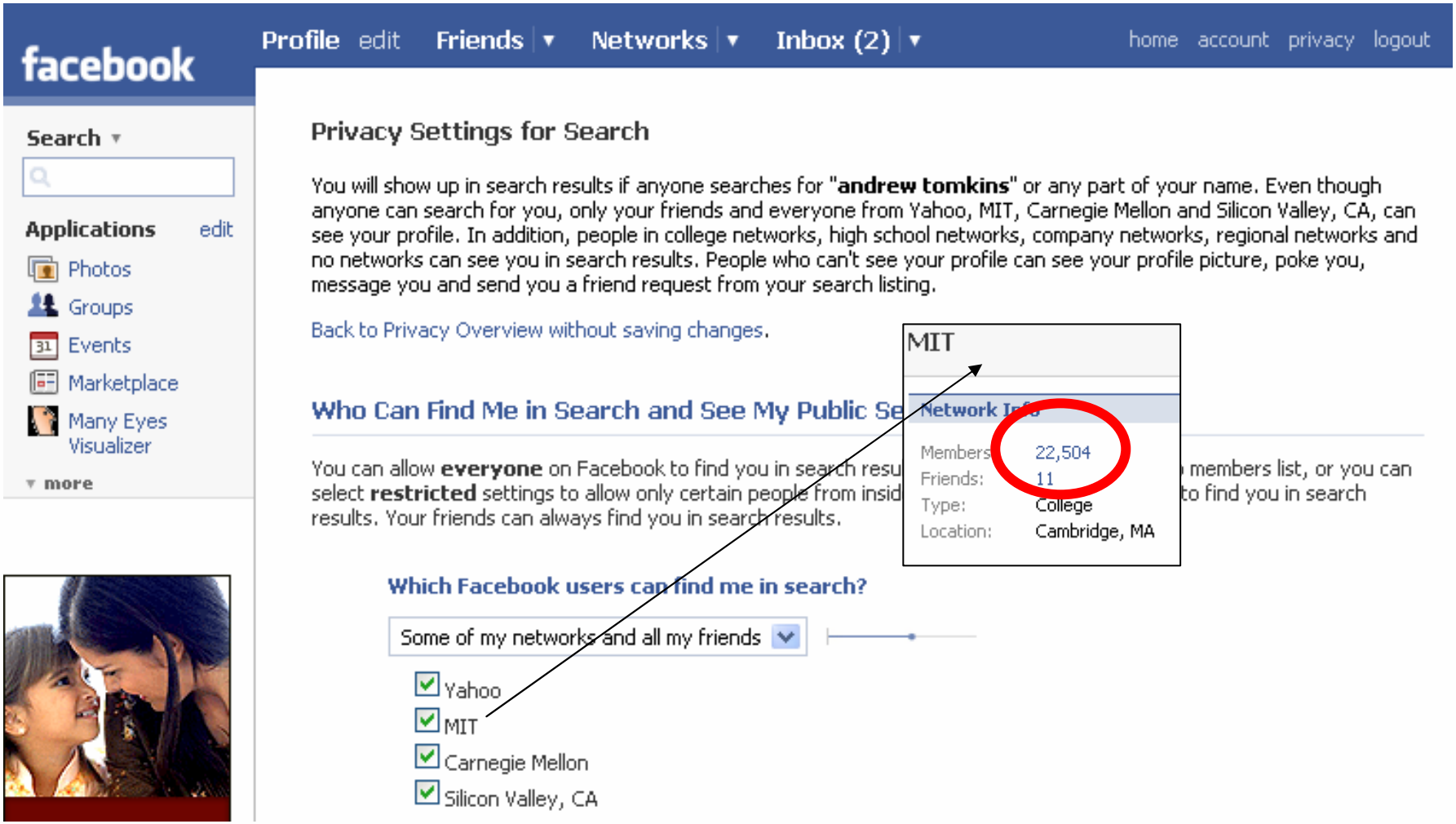
By topic

Content Consumption is fragmenting

By age…

1 to 3	0.5	treats, catnips, daddy, mommy, purring, mice, playing, napping, scratching, milk
13 to 15	3.5	webdesigning, Jeremy Sumpter, Chris Wilson, Emma Watson, T. V., Tom Felton, FUSE, Adam Carson, Guyz, Pac Sun, mall, going online
16 to 18	25.2	198{6,7,8}, class of 200{4,5}, dream street, drama club, band trips, 16, Brave New Girl, drum major, talkin on the phone, highschool, JROTC
19 to 21	32.8	198{3,5}, class of 2003, dorm life, frat parties, college life, my tattoo, pre-med
22 to 24	18.7	198{1,2}, Dumbledore's army, Midori sours, Long island iced tea, Liquid Television, bar hopping, disco house, Sam Adams, fraternity, He-Man, She-Ra
25 to 27	8.4	1979, Catherine Wheel, dive bars, grad school, preacher, Garth Ennis, good beer, public radio
28 to 30	4.4	Hal Hartley, geocaching, Camarilla, Amtgard, Tivo, Concrete Blonde, motherhood, SQL, TRON
31 to 33	2.4	my kids, parenting, my daughter, my wife, Bloom County, Doctor Who, geocaching, the prisoner, good eats, herbalism
34 to 36	1.5	Cross Stitch, Thelema, Tivo, parenting, cubs, role-playing games, bicycling, shamanism, Burning Man
37 to 45	1.6	SCA, Babylon 5, pagan, gardening, Star Trek, Hogwarts, Macintosh, Kate Bush, Zen, tarot
46 to 57	0.5	science fiction, wine, walking, travel, cooking, politics, history, poetry, jazz, writing, reading, hiking
> 57	0.2	death, cheese, photography, cats, poetry

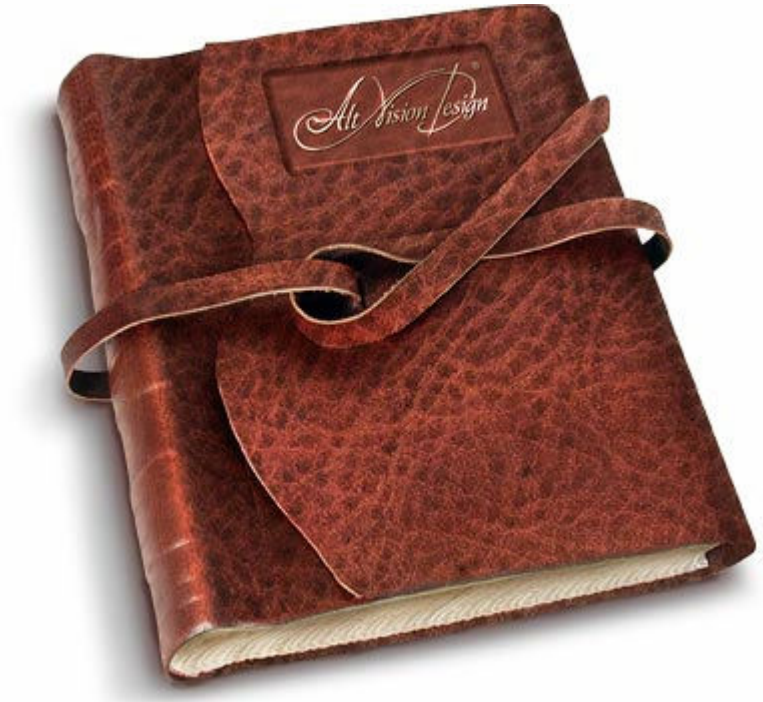
Content access is fragmenting



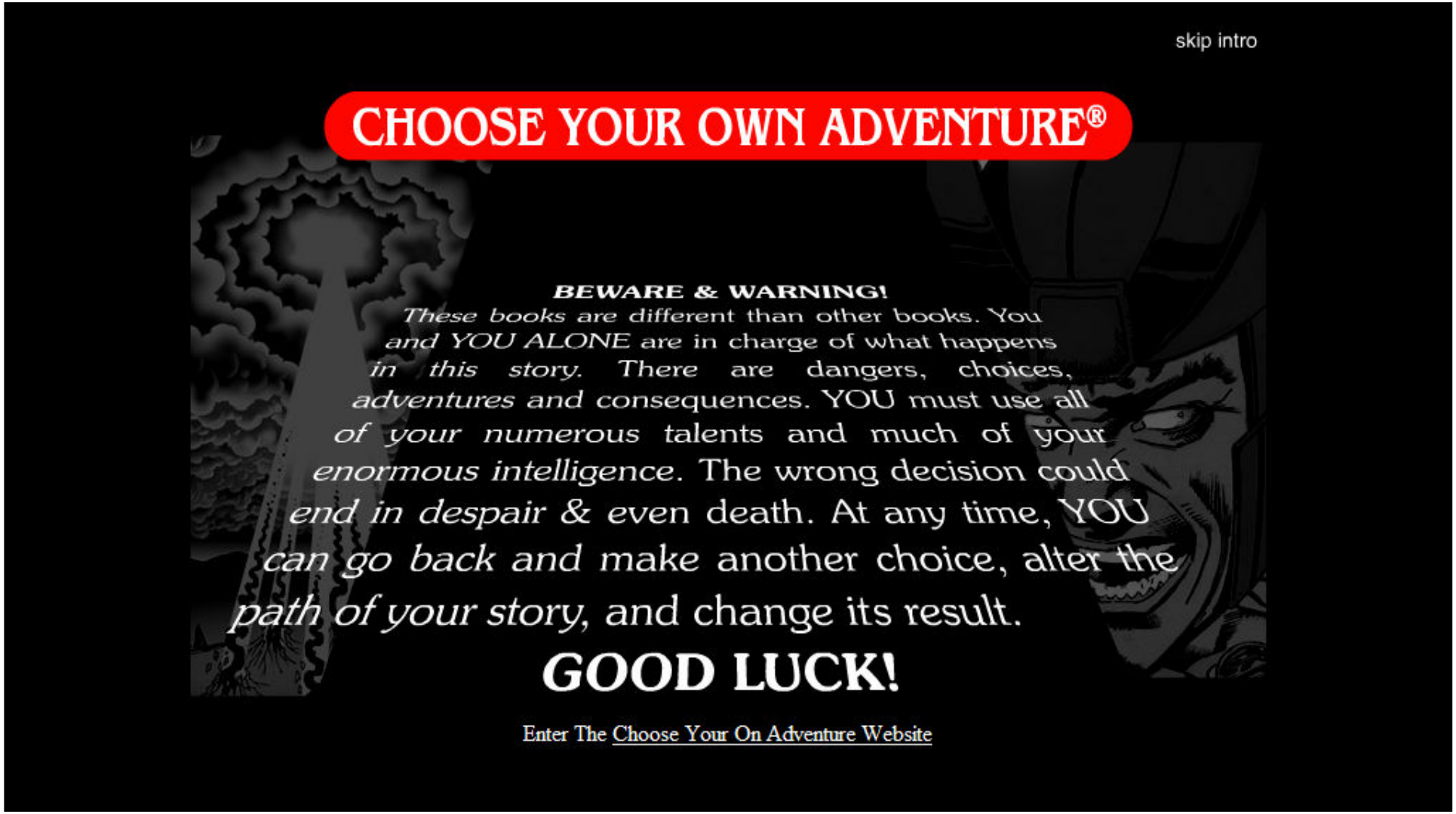
Content itself is fragmenting

- Content types are fragmenting
- Dawn of time: everything delivered as HTML, so treat HTML as the “base type”
- Need a more nuanced view of content types

Web pages yesterday….

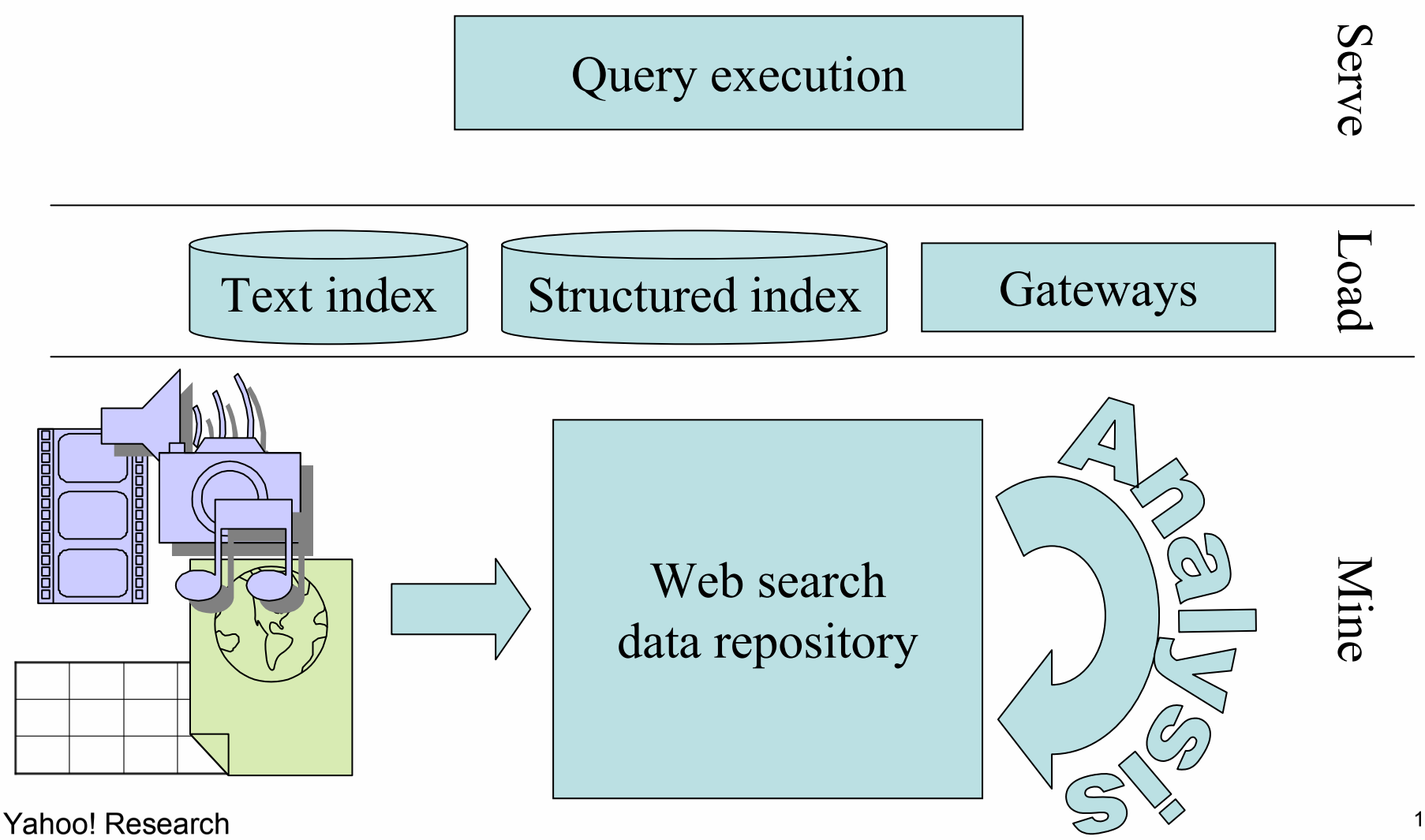


Web pages tomorrow….

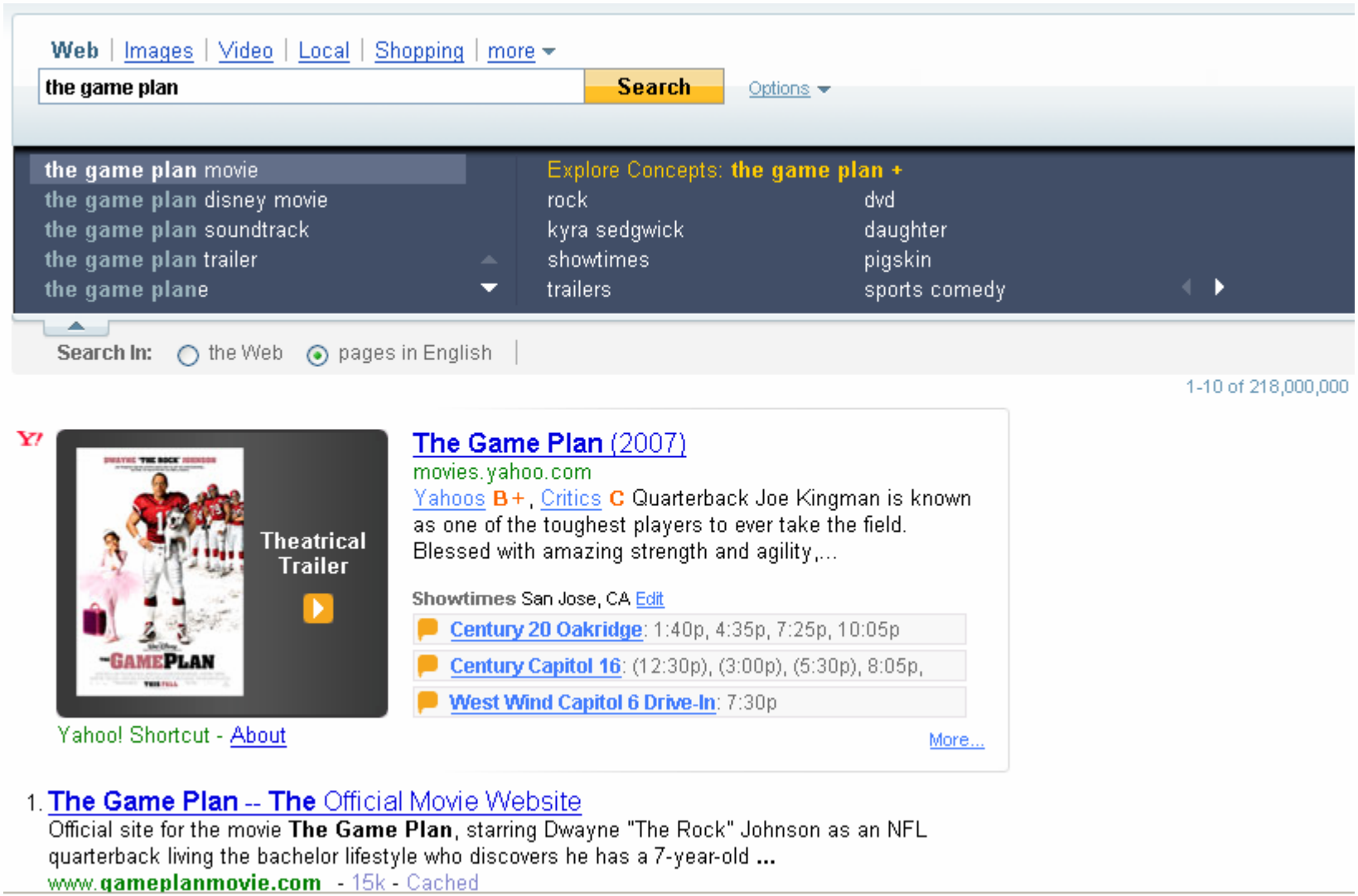


More complex content raises new issues in search

The emerging picture of search



Rich media and search assistance



Structured aggregation

Apple iPod nano digital player | Search

Apple iPod nano digital player

Is this useful?

$79.99 - $169.99 · Buy on MSN Shopping

User reviews

Take everything you love about iPod and shrink it - then shrink it again. Now meet iPod nano, the pencil-thin marvel featuring a color display, up to 14 hours of battery life and space for skip-free songs, audiobooks and... More...

User reviews at a glance What's this?

User reviews

General Comments (47 comments) 79% positive

Ease Of Use (21 comments) 90% positive

Size (15 comments) 87% positive

Price (14 comments) 64% positive

Ease Of Use

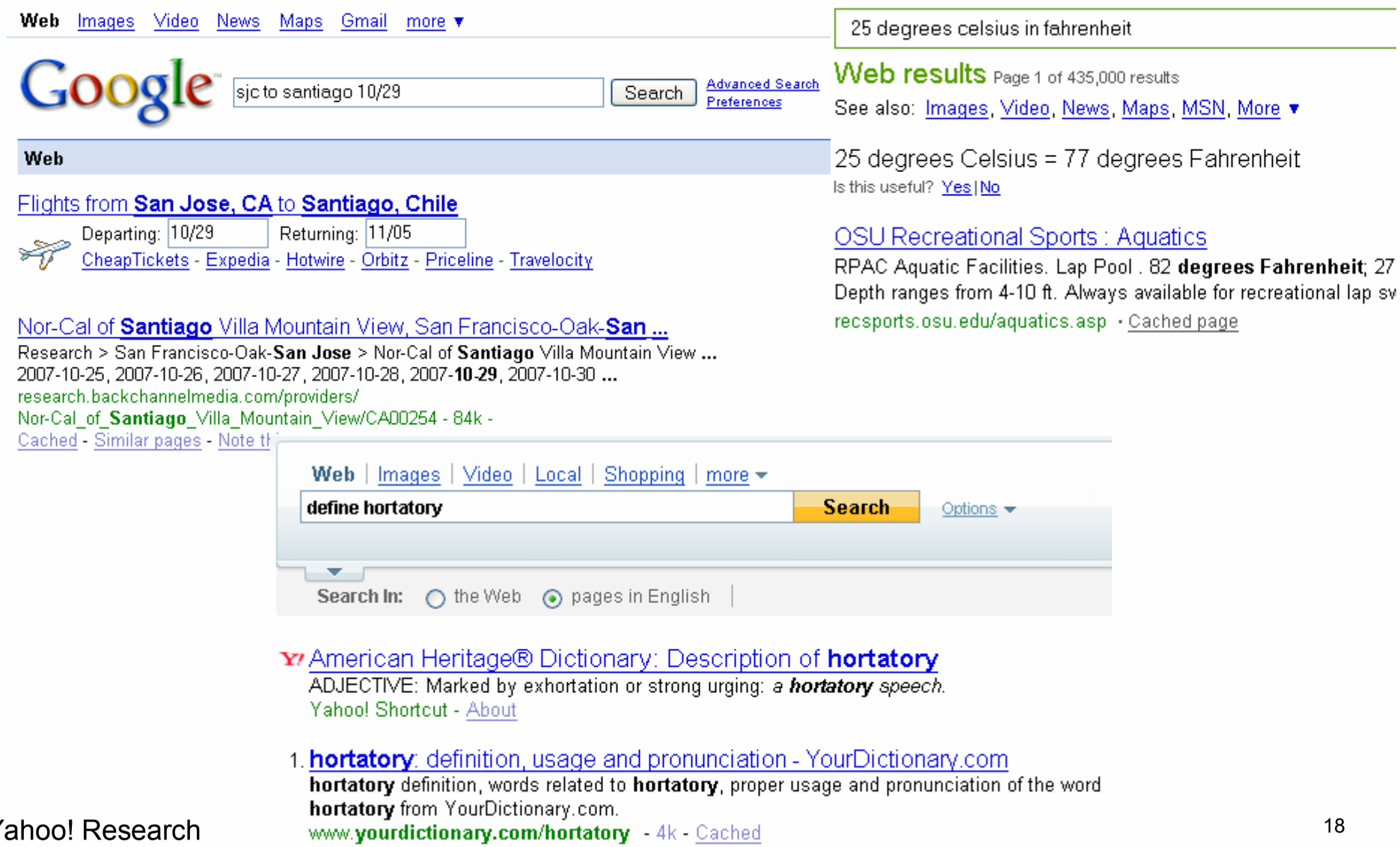
Positive comments (90%) | **Negative comments** (10%)

Pros: Silky smooth and easy to use. More...
www.epinions.com

I love how the Nano is so small and easy to use, you just plug it into your computers and it automatically loads your songs up on to it. More...
search.reviews.ebay.com

It is very easy to use. More...
search.reviews.ebay.com

Simple task-focused queries



Questions

- How can we move from targeted exploitation of key verticals to broad exploitation of available structure?
- How do we represent the affordances of web services with more meaningful semantics than, say, XML schema?
- How do we represent user tasks in a way that maps to content sources and services?
- What are appropriate models to extract complex hierarchical structures, and expose them to users?
- How do we represent and reason about a user's information need?
  - Long-running proclivities/preferences
  - Longitudinal tasks
  - Session-level models
  - Current query
- How do we model the interactions between data sources in a way that allows us to produce diverse results sets that maximize satisfaction over a distribution of users?

And more questions

- Given detailed user data, what are the appropriate model-based diversity measures we should be optimizing?
- How should we create lists of blended results optimizing relevance by exploiting different sources?
- How do we generate proxy measures for user satisfaction with rich result pages and interactive experiences?
  - Editorial metrics
  - Behavioral metrics
- How do we optimize the wide range of federation, blending, and presentation decisions automatically?
- What's the basic unit of relevance for multi-level annotated content?
- What user models are most effective for learning to rank?
- How can judgment-resistant features be incorporated into ranking approaches optimized for traditional metrics
  - Personalized ranking
  - Social information

Academic research and the cost of web search

Generated text content

- Storage: 52PB/yr
- Cost: $25M/yr
- In another 5 years, this looks about like the cost of having 10 people on your payroll
- Conclusion 1: any company with a multiple of ten people can afford to store every bit of text produced by every human on the planet
- Conclusion 2: no scale-based differentiation around text content

(of course, not all content is text…)

Some implications

- Technical platform for hosting the content not the differentiating factor
- Two things are key:
  - Gathering the content
    - Making deals
    - Working with users
  - Understanding the content
    - Strongest signal comes from user interactions

      (and has for ten years now)
    - Need to acquire and analyze content type-specific metadata

The cost story

- Storage is cheap: currently, think $500 for 1T of storage
- A credible web search engine can be built on a quality crawl of 1B pages
- With some simple compression, can store at <5K/page
- Total: $12.5K

Should academics work on core web search?

- Great problem!
- Major dollars behind it!
- But…
  - Hard to compete without user data
  - Very high capital requirements to serve
  - Hundreds of person-years of pieces not yet in the public domain

Academics cannot address core web search on a level playing field today

(can we help?)

Sharing query logs

Rosie Jones
Ravi Kumar
Bo Pang
Andrew Tomkins
Jasmine Novak

Why share query logs?

- Help academia contribute to web search
- Untold insights into human behavior exist within logs

Privacy entering public perception

HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Your Life as an Open Book

Illustration by The New York Times

By TOM ZELLER Jr.
Published: August 12, 2006

SIGN IN TO E-MAIL THIS

Quick overview of 3 pieces of work

- Goal for this work: try to find the boundaries of what seems breakable
- Approaches:
  1. Attack a specific (reasonable?) anonymization scheme
  2. Approximately extract age/gender/ location to leak bits about the user

Ba Nabalzvmvat Dhrel Ybtf ivn Gbxra-onfrq Unfuvat

Eniv Xhzne, Wnfzvar Abinx, *Ob Cnat*, Naqerj Gbzxvaf
Lnubb! Erfrnepu, Fhaalinyr, PN.
obcnat@lnubb-vap.pbz

Znl, 2007

On Anonymizing Query Logs via Token-based Hashing

Ravi Kumar, Jasmine Novak, *Bo Pang*, Andrew Tomkins
Yahoo! Research, Sunnyvale, CA.
bopang@yahoo-inc.com

May, 2007

Data

random samples from six-hour query logs from a week apart

	token side	hash side
number of queries	3,849,916	3,187,228
vocabulary size (n)	freq. of the least freq. term	
1000	1406	1181
2000	737	598
4000	358	296
8000	159	131
16000	63	52

Summary of results

Vocabulary (n)	matchable set	score greedy
1000	918	0.99
2000	1851	0.96
4000	3648	0.92
8000	7182	0.83

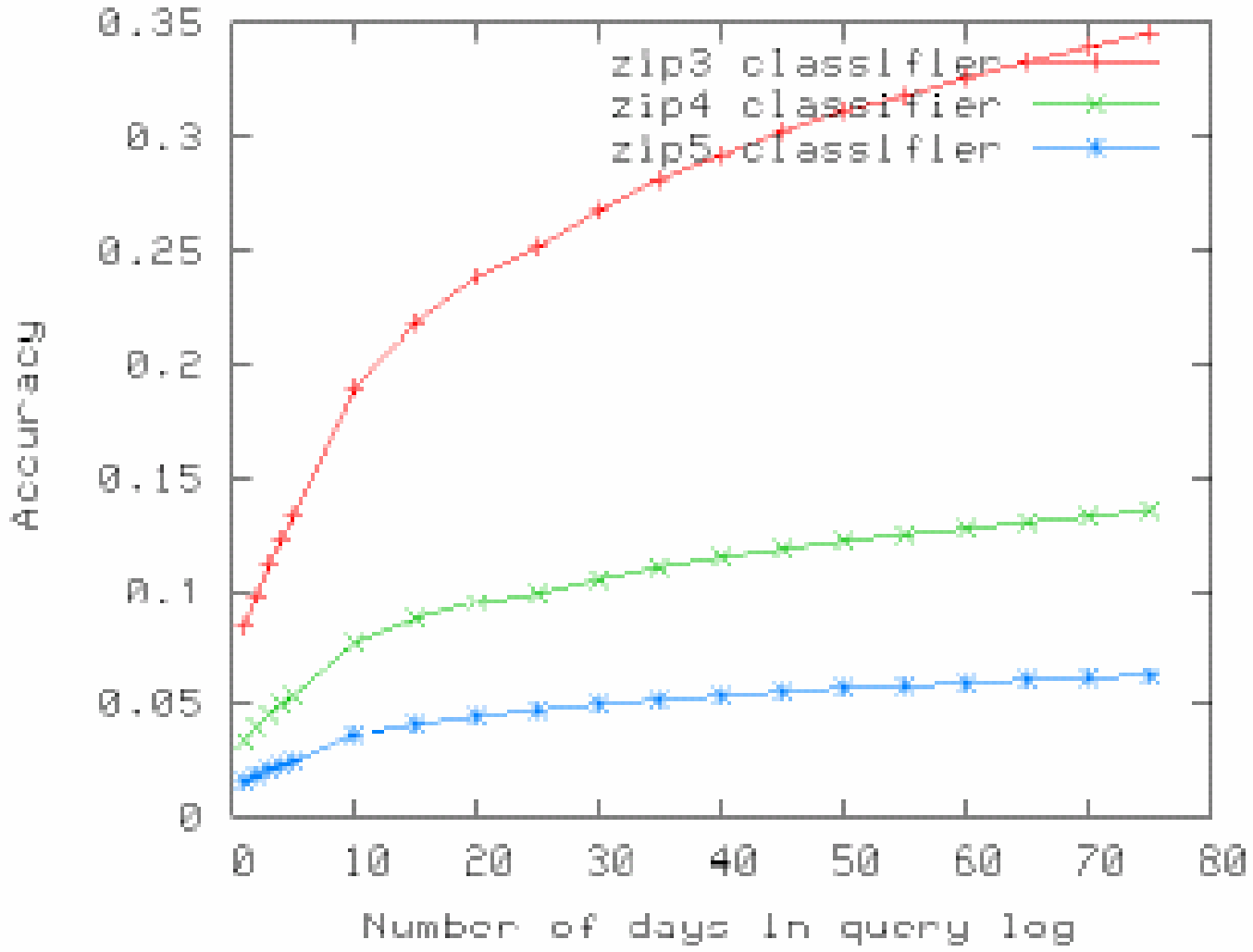
What leaks?

with session information, in the top 1k most frequent terms:

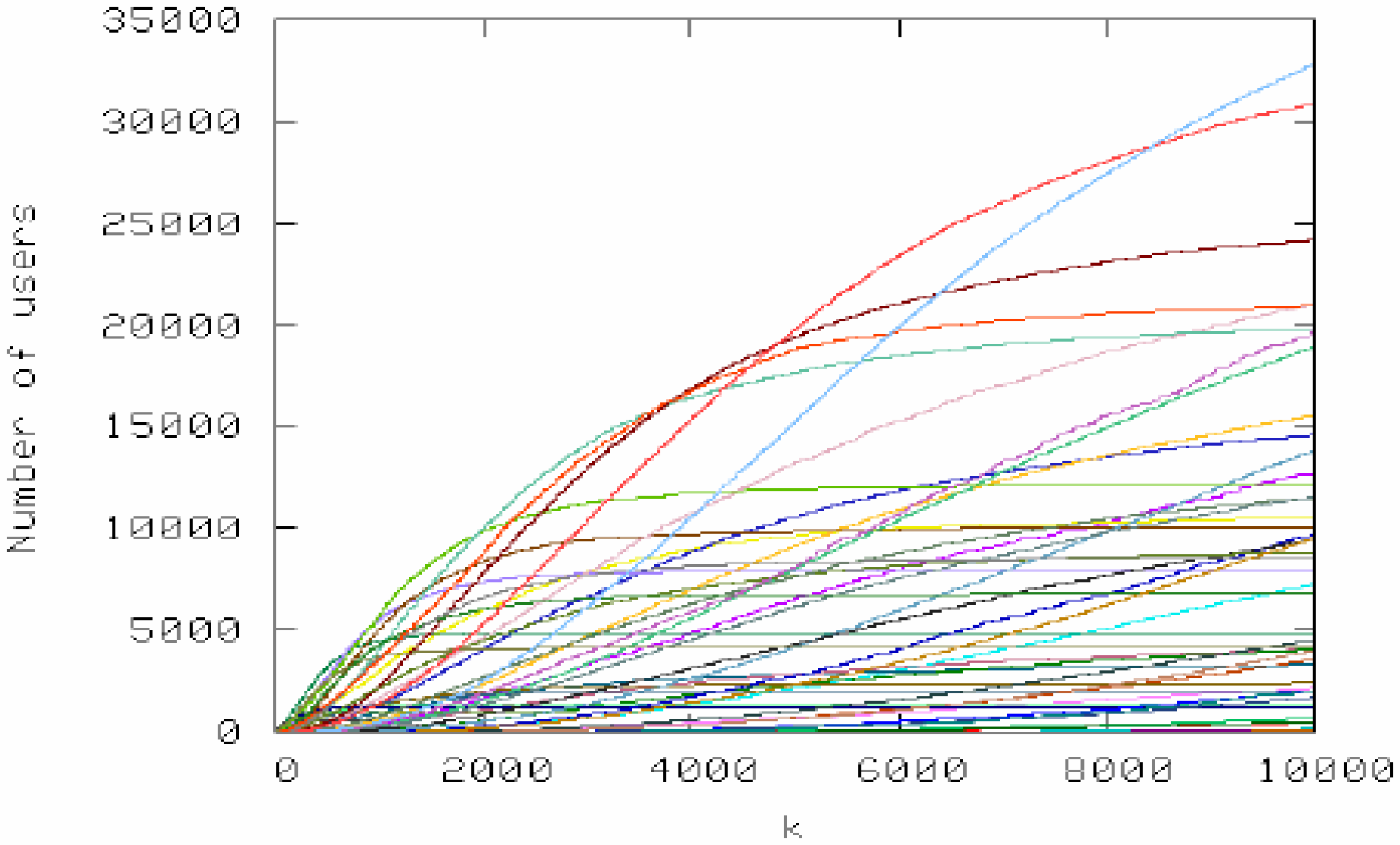
Session count	Entity type
7417	person name
83801	company name
7769	place name
2960	non-star name
83	non-star name and a place name
169	non-star name and a company name
12	non-star name and an adult term
14	non-star name and a revealing term

Using metadata to breach privacy

Classification accuracy



Accuracy:
400-1000X better than random



Information leakage (100M users)

Number of users	Bits leaked
61	23.3–26.5
1428	20–23.3
16260	16.6–20
56288	13.3–16.6
90027	10–13.3
89898	< 10
496038	0

“Person” and “Trace” attacks

- Trace attack: given a trace, identify the person
- Person attack: the dual
- Adversaries in person attack:
  - “Neighborly” knowledge
  - Query knowledge
  - Browser compromise

Person attack (750K users)

Query set	Bin size
harry potter, pizza	4855
football, skiing	2430
italian restaurant, pizza	1441
harry potter, volkswagen beetle	27
honda odyssey, italian restaurant	20
football, skiing, toyota prius	9
football, triumph tr3	4
football, harry potter, volkswagen beetle	3
pizza, triumph tr3	2
danielle steele, volkswagen beetle	1
brie, holly lisle, pizza	1

Conclusions

- The only technical proposal in which we feel there is some hope of success is
  - removal of all session information
  - obscuring of timestamps to avoid re-identification of sessions (ie, by adding a random value in [0,60 minutes] to all timestamps)
  - removing certain highly-sensitive individual searches (credit cards and the like)
- It's possible such a proposal could be weakened slightly to allow reformulations to pass through, but we don't have confidence that much beyond this is possible without some significant leakage

The End