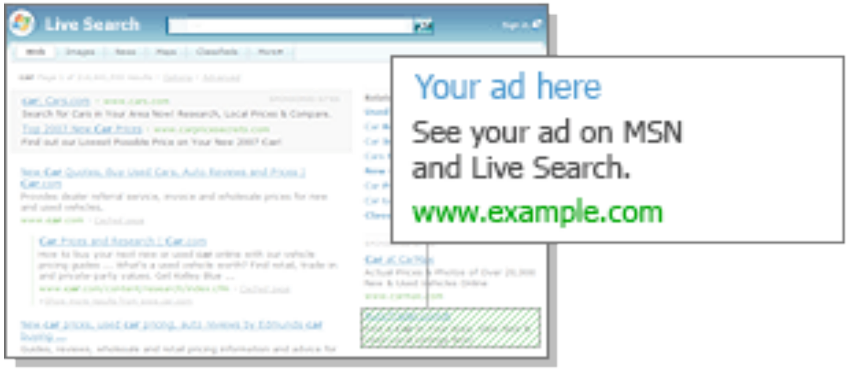
Releasing Search Queries and Clicks Privately

Aleksandra Korolova	Stanford University
Krishnaram Kenthapadi	Microsoft Research – Search Labs
Nina Mishra	Microsoft Research – Search Labs
Alexandros Ntoulas	Microsoft Research – Search Labs

All examples are fictitious

Why Release Search Logs

Online Ad Campaign

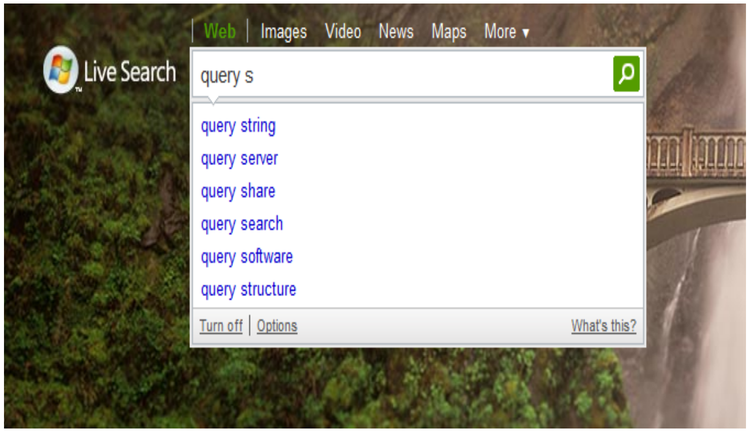


Your ad here

See your ad on MSN and Live Search.

www.example.com

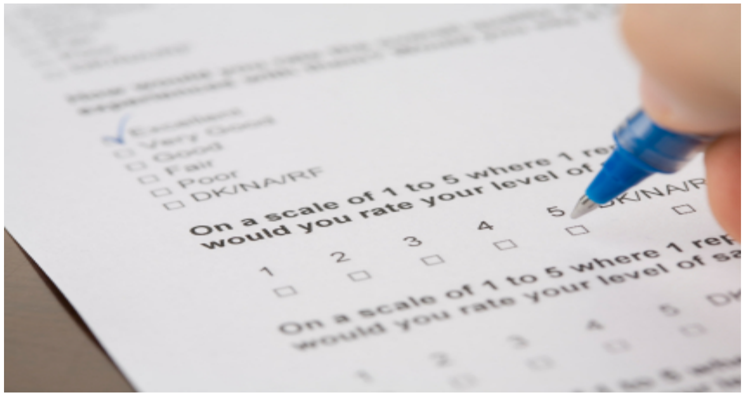
Query Suggestions



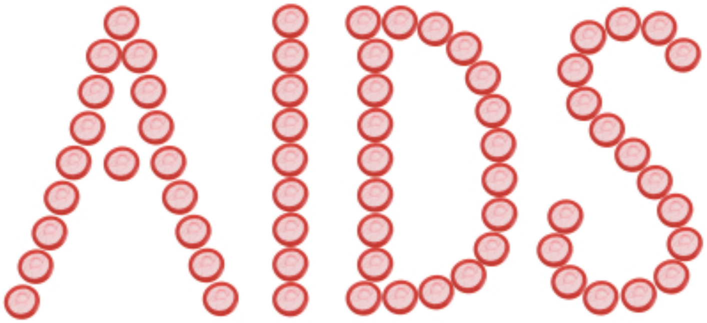
query s

Mining Search Data

Social Science



Why Search Logs are Private



Previous Approaches

Anonymize Usernames/Omit IP Addresses

AOL data release, 2006

- •CTO resigned, 2 employees fired
- •Class action law suit pending
- •CNN Money:
“101dumbest moments in business”

Searches by user 4417749

Thelma Arnold, 62
from Lilburn, Georgia

landscapers in lilburn ga.	3/6/2006	18:37:26
effects of nicotine	3/7/2006	19:17:19
jarrett t. arnold eugene oregon	3/23/2006	21:48:01
plastic surgeons in gwinnett county	3/28/2006	15:04:23
60 single men	3/29/2006	20:11:52
clothes for 60 plus age	4/19/2006	12:44:03
lactose intolerant	4/21/2006	20:53:51
dog who urinate on everything	4/28/2006	13:24:07

Ad-hoc Techniques do Not Work

- Remove names, dates, numbers, locations
  - "MIT math major with multiple sclerosis"
- Token-based hashing fails
  - [Kumar, Novak, Pang, Tomkins WWW'07]
- Release only frequent queries
  - What's sufficiently frequent?
- Combining data from multiple sources
  - Previous/future releases useful to break privacy

Our Goal

Can we release search logs with

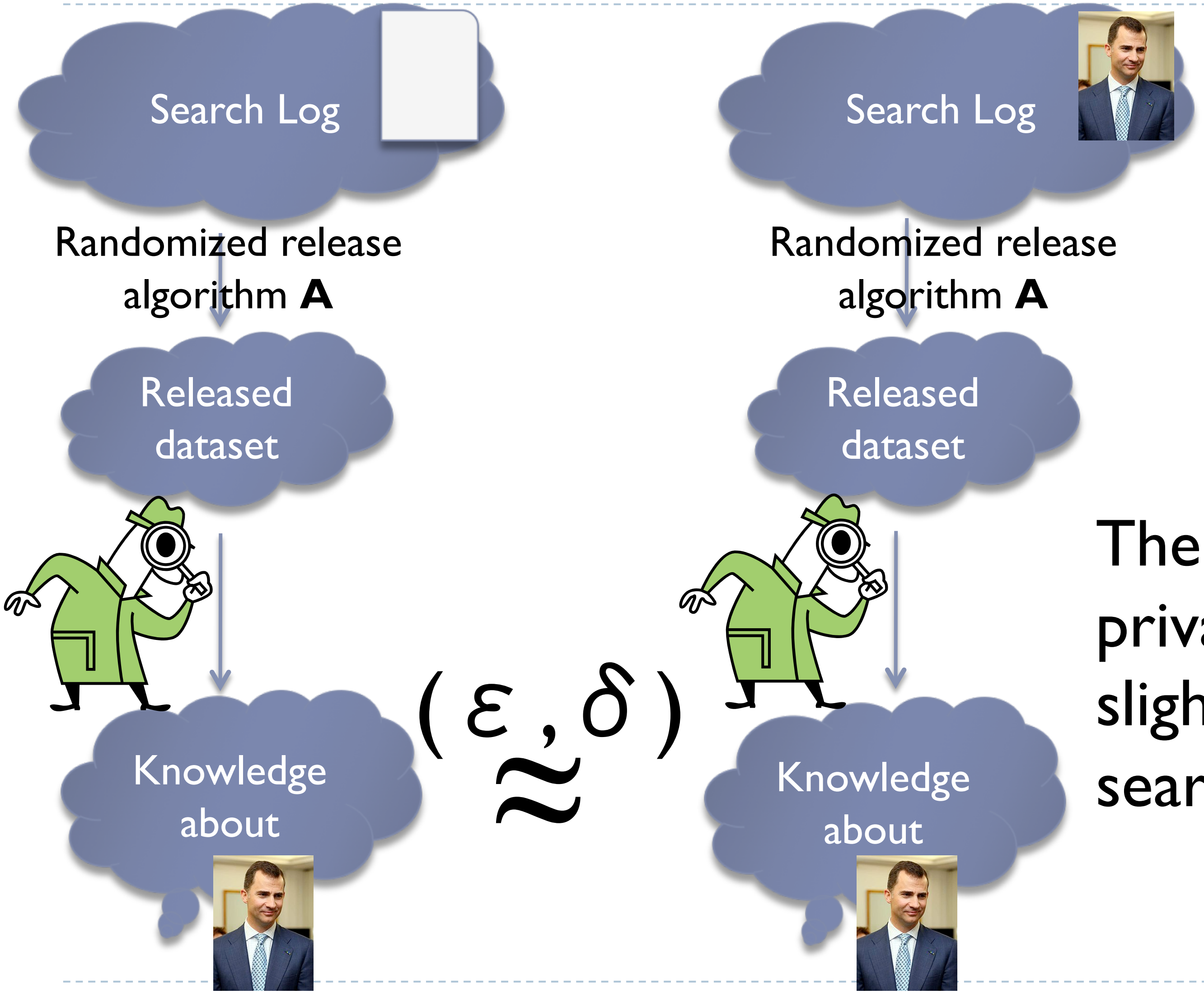
- provable privacy guarantees
- preserving usefulness

Rigorous Privacy Definition

Desired Features of Privacy Definition

- No assumptions on attacker’s
  - prior knowledge
  - computational powers
  - access to other datasets
- No assumptions on user’s
  - search patterns
  - what constitutes private information

Differential Privacy [Dwork et al, 2006]



Search Log

Randomized release algorithm **A**

Released dataset

Knowledge about

$\overset{(\varepsilon,\delta)}{\approx}$

Search Log

Randomized release algorithm **A**

Released dataset

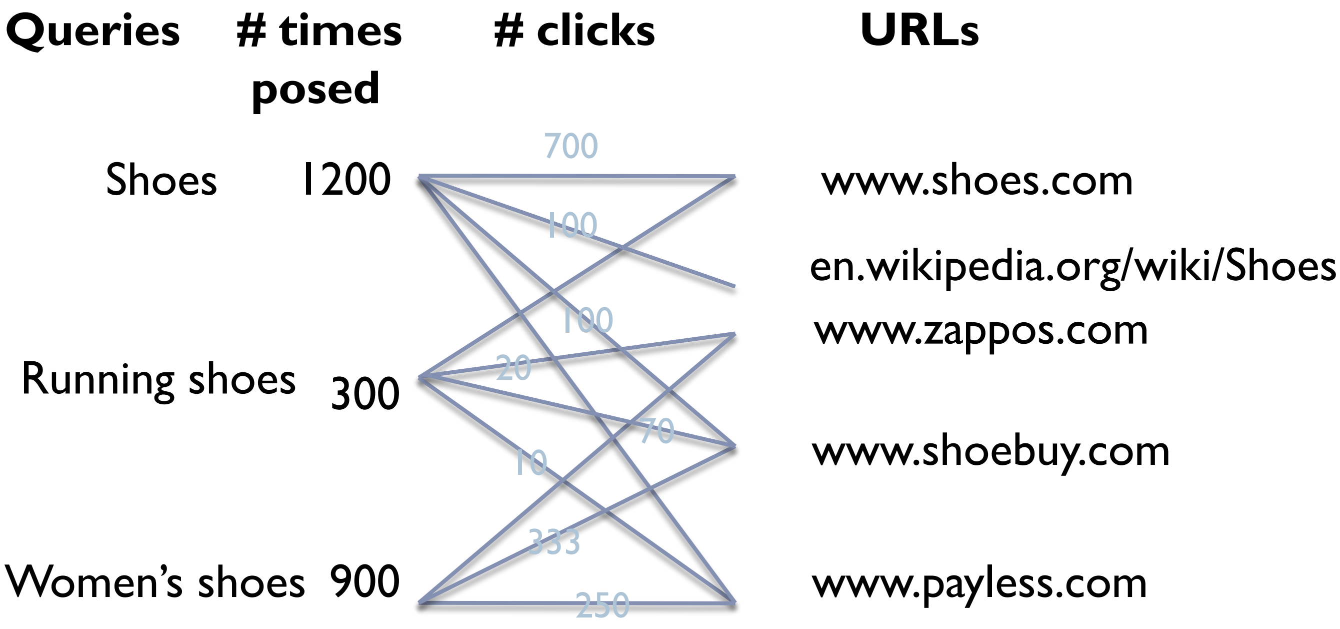
Knowledge about

The risk to one's privacy increases only slightly from using the search engine

Our Approach

Query-Click Graph
Data Release Algorithm
Privacy Guarantees

Query-Click Graph



Queries	# times posed
Shoes	1200
Running shoes	300
Women's shoes	900

# clicks

URLs

www.shoes.com

en.wikipedia.org/wiki/Shoes

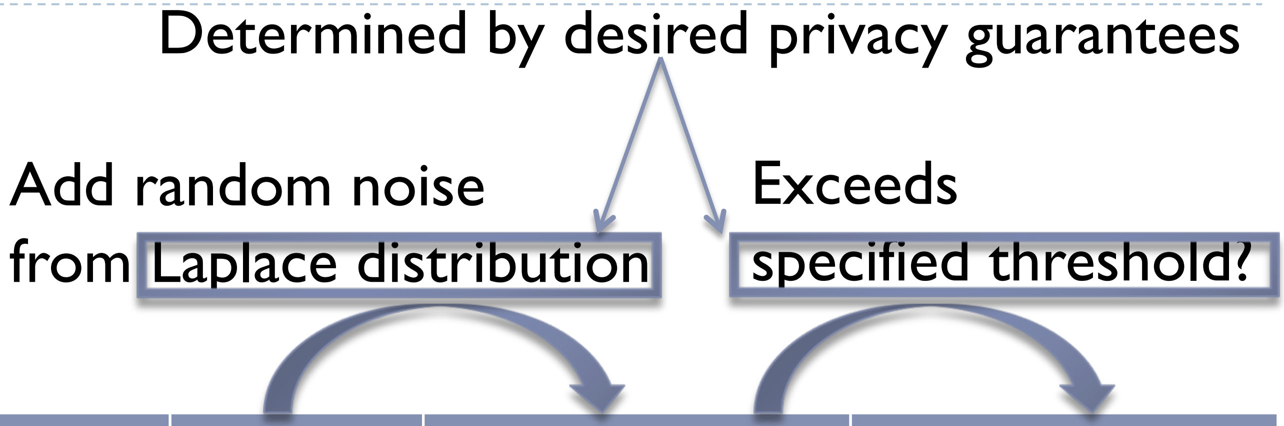
www.zappos.com

www.shoebuy.com

www.payless.com

- ✓ Useful for many applications
  - ✓ Related searches
  - ✓ Spell corrections
  - ✓ Expanding acronyms
  - ✓ Estimating CTRs
  - ✓ Computations on query-click graph

Releasing Queries Privately



Determined by desired privacy guarantees

Add random noise from Laplace distribution

Exceeds specified threshold?

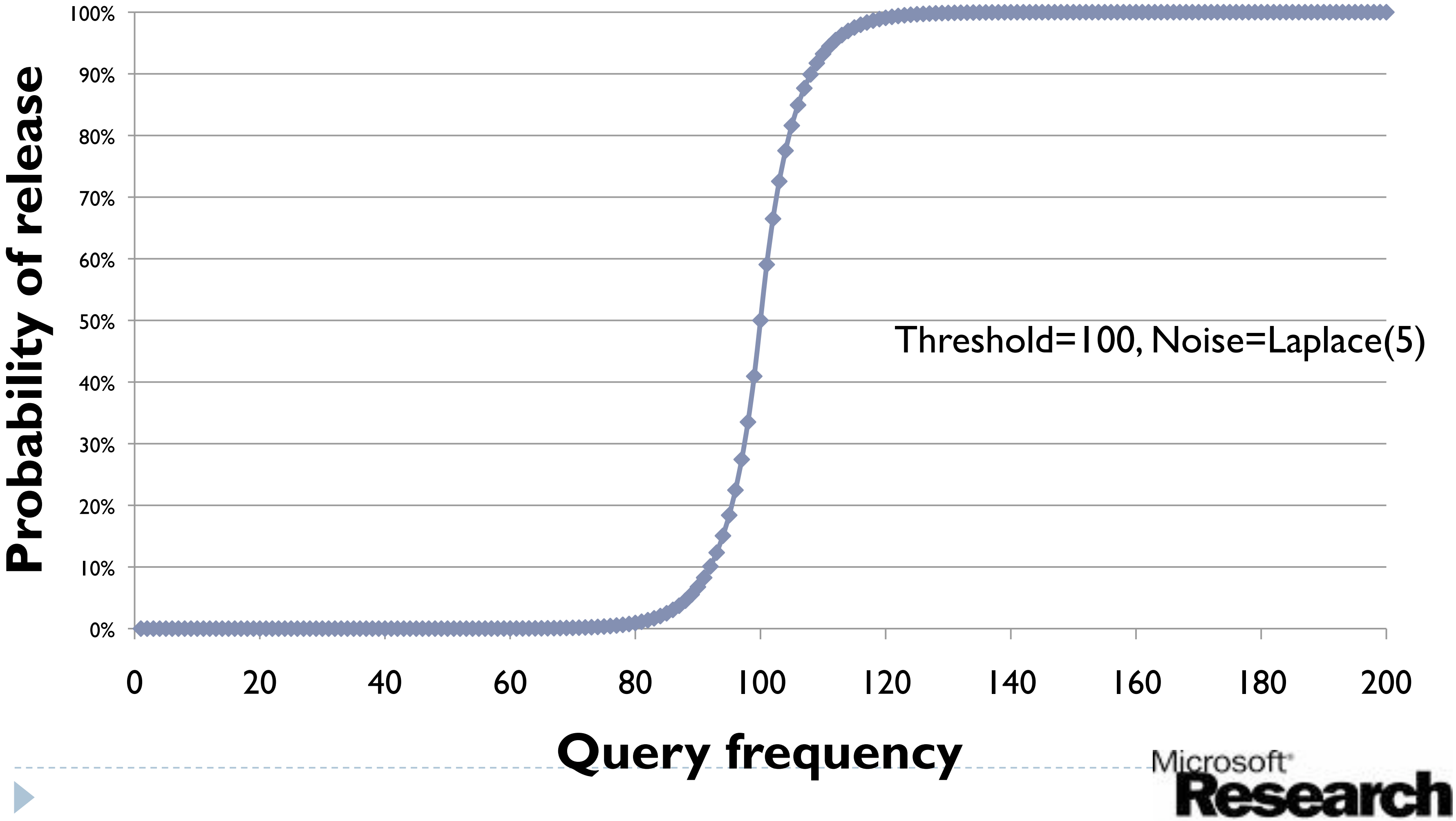
Query	Count	Noisy Count	Released?
Weather in Madrid	1150	1159	✓
WWW 2009	900	903	✓
Data-mining	710	698	✓
Report a stolen passport	20	19	✗
Aleksandra (650) 796-4536	2	7	✗

Understanding Private Query Release

- Why add random noise?
  - Suppose attacker has a guess for my SSN and poses the query containing the guess threshold-1 # of times

- What if one user disproportionally influences the log?
  - Solution: limit each user's activity to *d* queries and d_c clicks
  - Caveat: if using multiple computers, treated as two users

Probability of Release Depending on Frequency



Releasing Queries and Clicks Privately

Choose:

- Desired privacy guarantees (ε , δ)
- Limit on user activity d, d_c

→ Threshold
Noise Level

Release Queries:

- whose noisy frequency counts exceed the threshold

Release URL Click Counts:

- Given released query, top 10 URLs returned are public
- Release noisy click counts for top 10 URLs

Theorem: Algorithm Provably Private

✓ Satisfies (ε , δ)-differential privacy, when

- Threshold $= d\left(1 + \frac{\ln(\frac{d}{2\delta})}{\epsilon}\right)$
- Noise from Laplace distribution w/ scale $\frac{d}{\epsilon}$
- Keeping the first *d* queries per user

✓ Quantifies what constitutes "sufficiently frequent" queries

Utility

Released Data Characteristics
Social Science Research
Algorithmic Application

Quantity of Privately Releasable Data

Distinct Queries	Impressions
2.5 million	3.5 billion

Example queries releasable:

- How to tie a windsor knot
- Girl born with 8 limbs
- Cash register software
- Vintage aluminum Christmas trees

Utility: Studying Human Nature

[Tancer "Click" 2008]

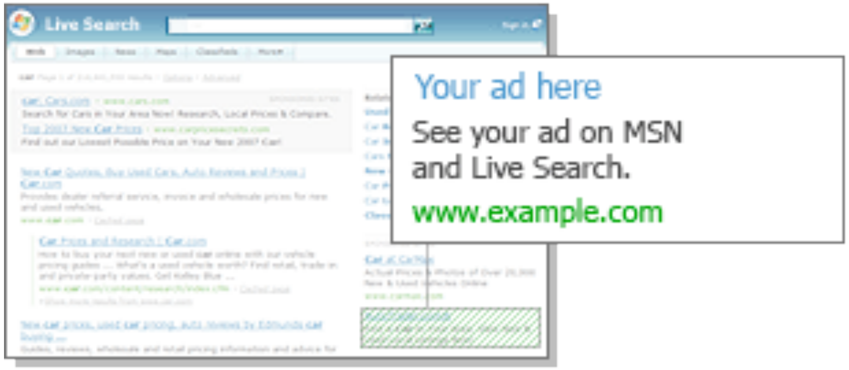
"Fear of …" queries

Rank	Phone Survey	Original Search Log	Released Queries
1	Bugs, mice, snakes	Flying	Flying
2	Heights	Heights	Heights
3	Water	Snakes, spiders	Public Speaking
4	Public transportation	Death	Snakes, spiders
5	Storms	Public speaking	Death
6	Closed spaces	Commitment	Commitment
7	Tunnels and bridges	Intimacy	Abandonment
8	Crowds	Abandonment	The dark
9	Speaking in public	The dark	Intimacy

Social Fears

Utility: Recommending Keywords to Online Advertisers

- Launch an online ad campaign around a concept



Your ad here

See your ad on MSN and Live Search.

www.example.com

travelocity

- Goal:
  - given a seed set of keywords/URLs, suggest relevant keywords.
- Solution:
  - Random walk on Query-Click Graph
  - [Fuxman, Tsaparas, Achan, Agrawal, WWW'08]

Recommending Keywords: travelocity

Original

flight travelocity
travalocity
travalosity
travel velocity
travelacity
travellocity
travellocity com
travelocity
travelocity air fares
travelocity ca
travelocity cheap flight
travelocity com
travelocity vacations
travelociy
travelosity
travelosity com
travilocity
travleocity
travlocity
travolicity
travolocity
trvelocity
ww travelocity com
www travellocity com
www travelocity
www travelocity co
www travelocity com
www travelosity com
wwwtravelocity com
aarp passport
air fares
airfare
airfares
cheap flights
cruises
flight travelocity
flights
last minute travel
last minute travel deals
vacation packages
vacations to go

Private (13% of Original)

travellosity
traveloscity
flights

Conclusions

Contributions

- Algorithm for releasing queries and clicks with provable privacy guarantees
  - Non-trivial amount of queries, impressions, clicks
  - Evidence that released data preserves utility
- Releasing frequent queries works
  - Quantify frequent
- Explored the trade-offs between privacy and utility

Future Work

- Grouping similar queries
- Choosing privacy parameters in practice
- Beyond privacy of users

Thank you!
Questions?