CrowdLogging:

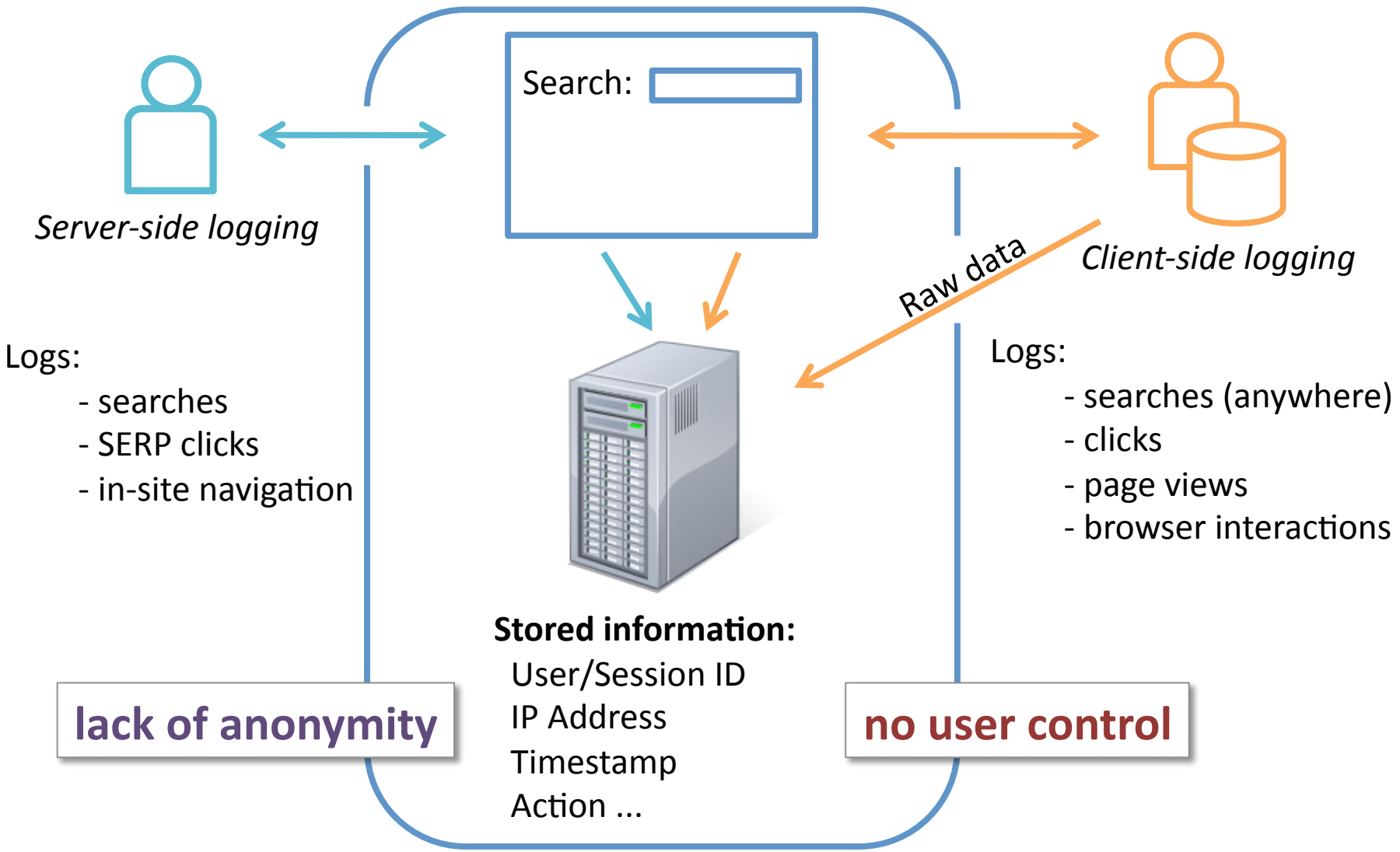
Distributed, private, and anonymous search logging

Henry Feild
James Allan
Joshua Glatt

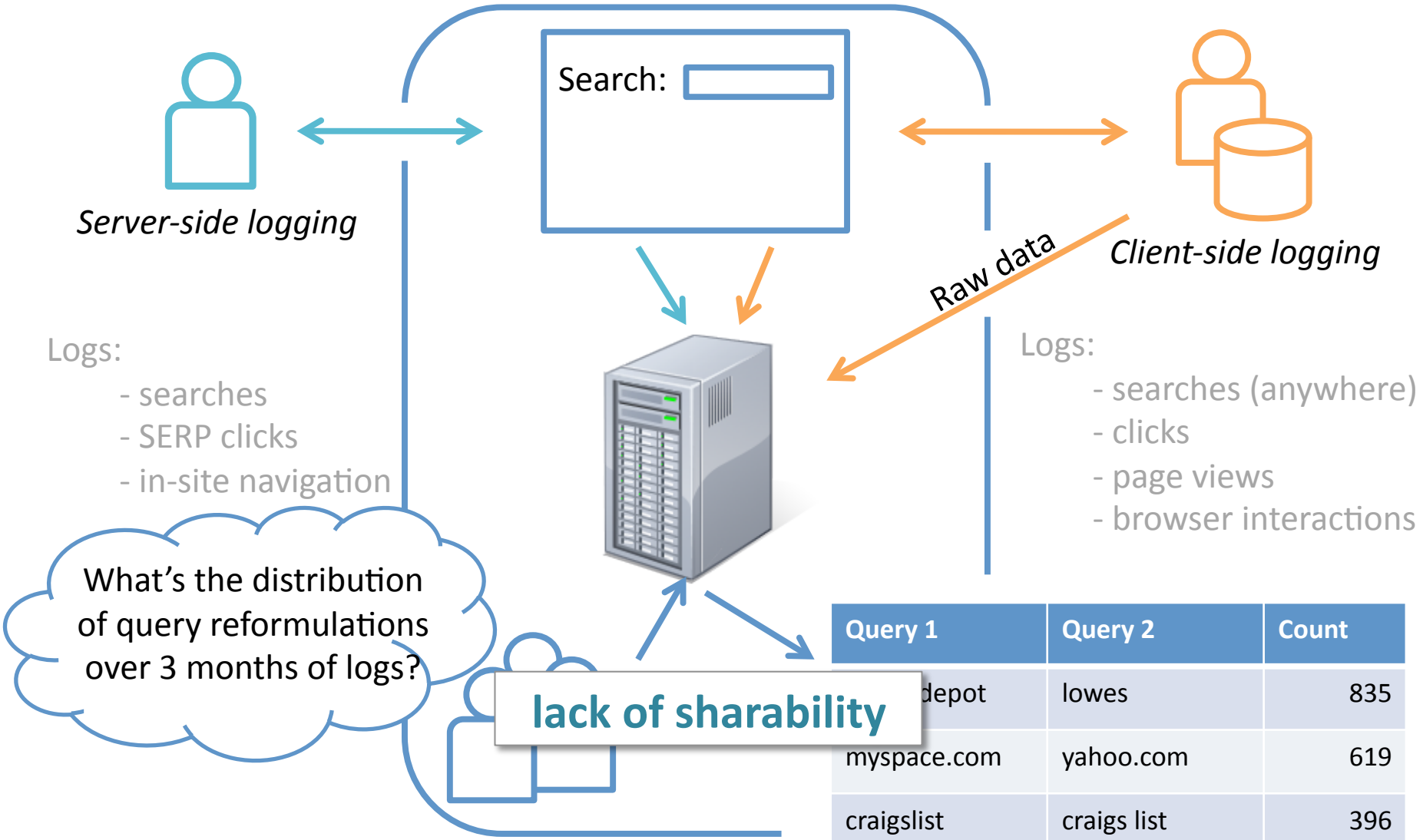
Center for Intelligent Information Retrieval
University of Massachusetts Amherst

July 26, 2011

Centralized search logging and mining



Centralized search logging and mining

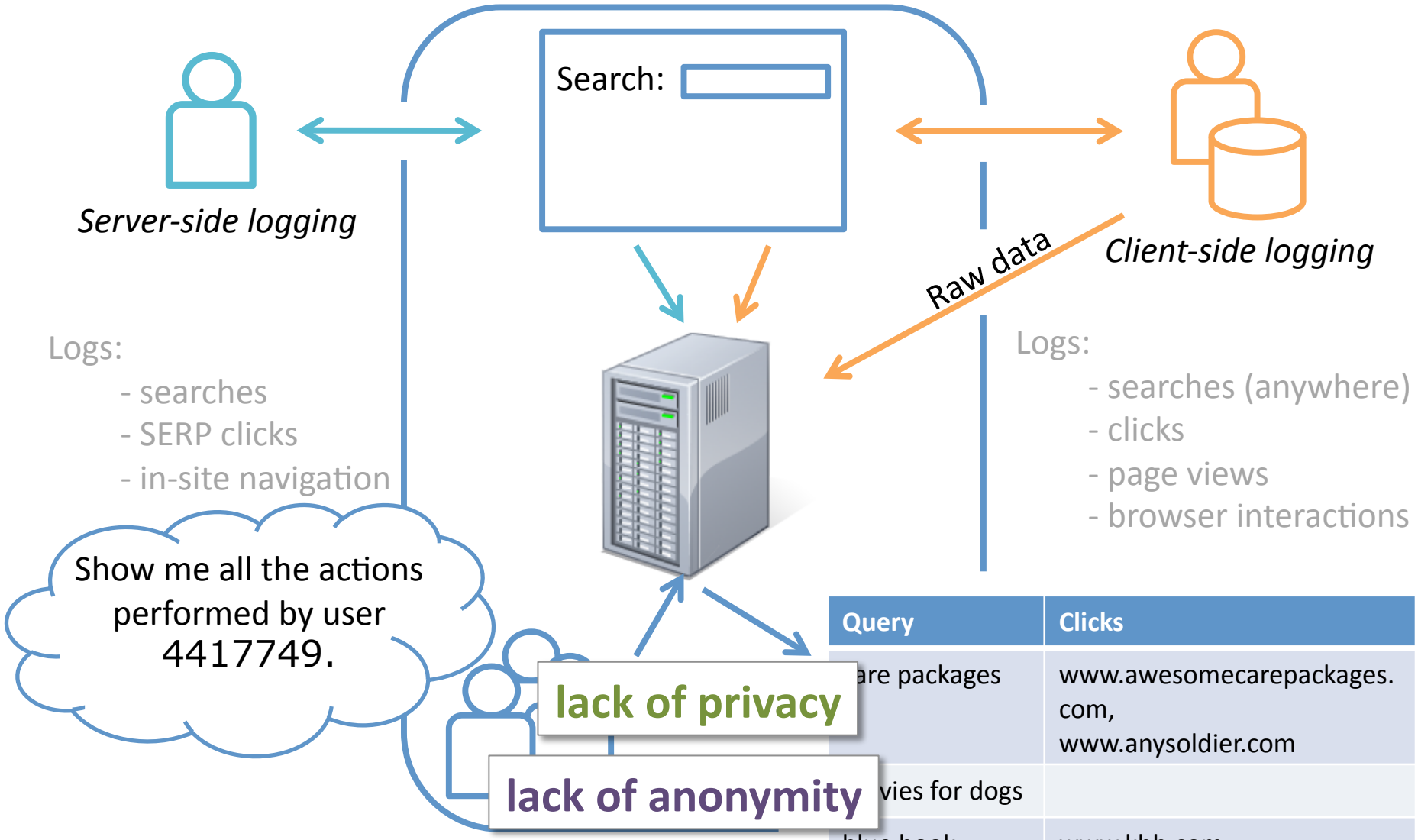


Query 1	Query 2	Count
depot	lowes	835
myspace.com	yahoo.com	619
craigslist	craigs list	396

...

Query reformulations from the AOL 2006 log.

Centralized search logging and mining



Query	Clicks
are packages	www.awesomecarepackages.com, www.anysoldier.com
vies for dogs	
blue book	www.kbb.com
...	

From the AOL 2006 log.

Drawbacks of the centralized model for *users* and *researchers*

- **lack of user control**
  - raw search data is stored out of reach of users
- **lack of privacy**
  - raw data *could* contain personally identifiable information
  - multiple user actions with common identifier
- **lack of anonymity**
  - source information logged (e.g., IP address)
- **lack of sharability**
  - logs not shared (privacy, legal, and competition issues)
  - cannot reproducible research results
  - stifles scientific process

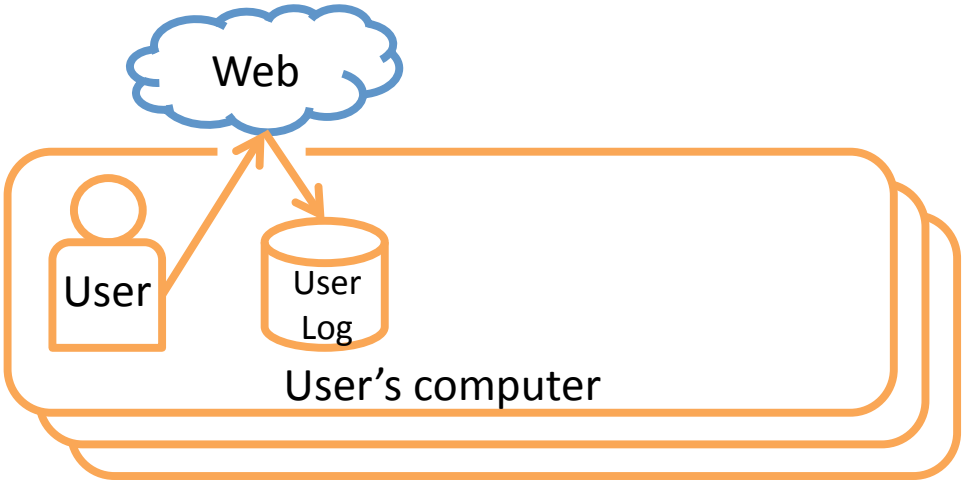
Outline

- *Centralized search logging and mining*
- CrowdLogging
  - logging, mining, and releasing data
  - advantages
  - comparison with centralized model
- The CrowdLogger browser extension
  - overview
  - collected data
- ~~Technical stuff~~
  - ~~secret sharing~~
  - ~~privacy policies (e.g., differential privacy)~~

See the paper for details

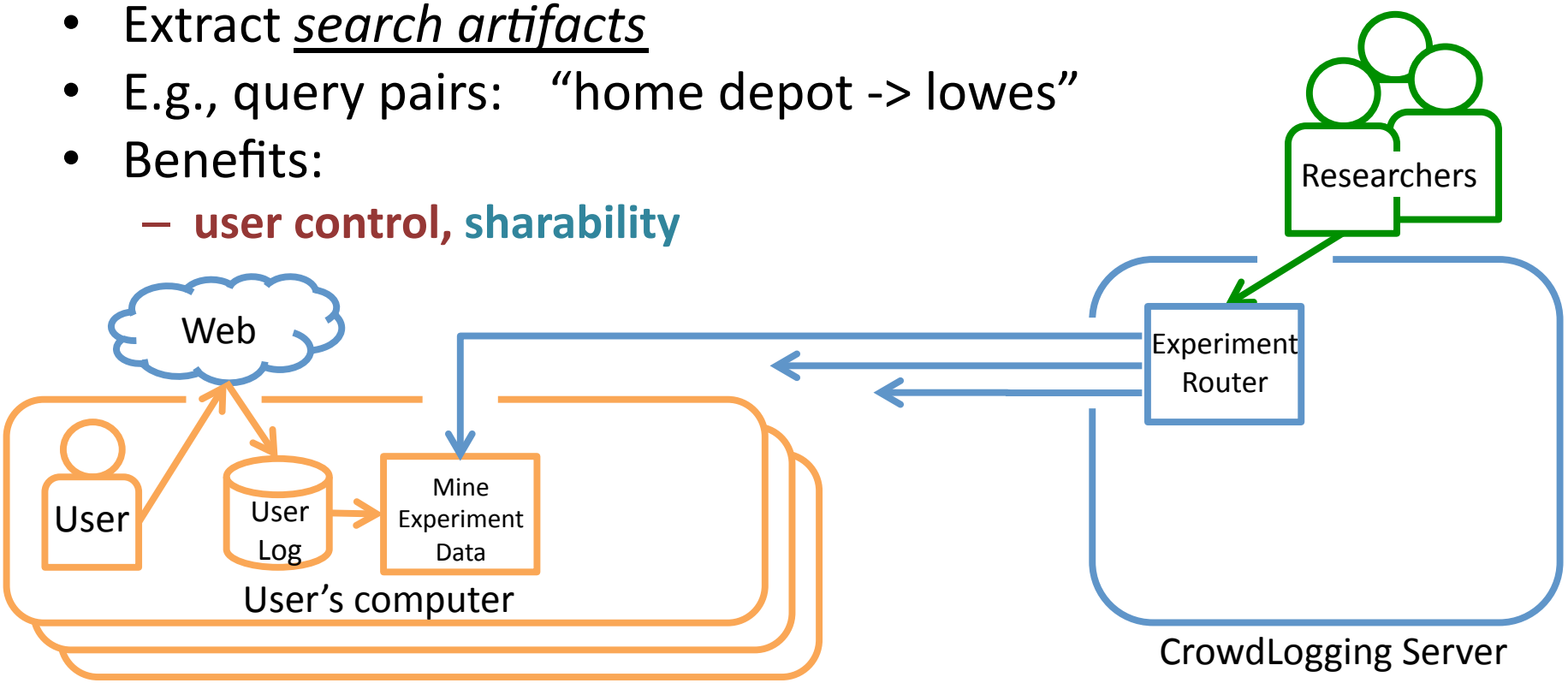
CrowdLogging: how data is logged

- User downloads browser extension or proxy
- User's web interactions logged locally
  - can be examined and deleted at any time
- Benefits:
  - **user control**



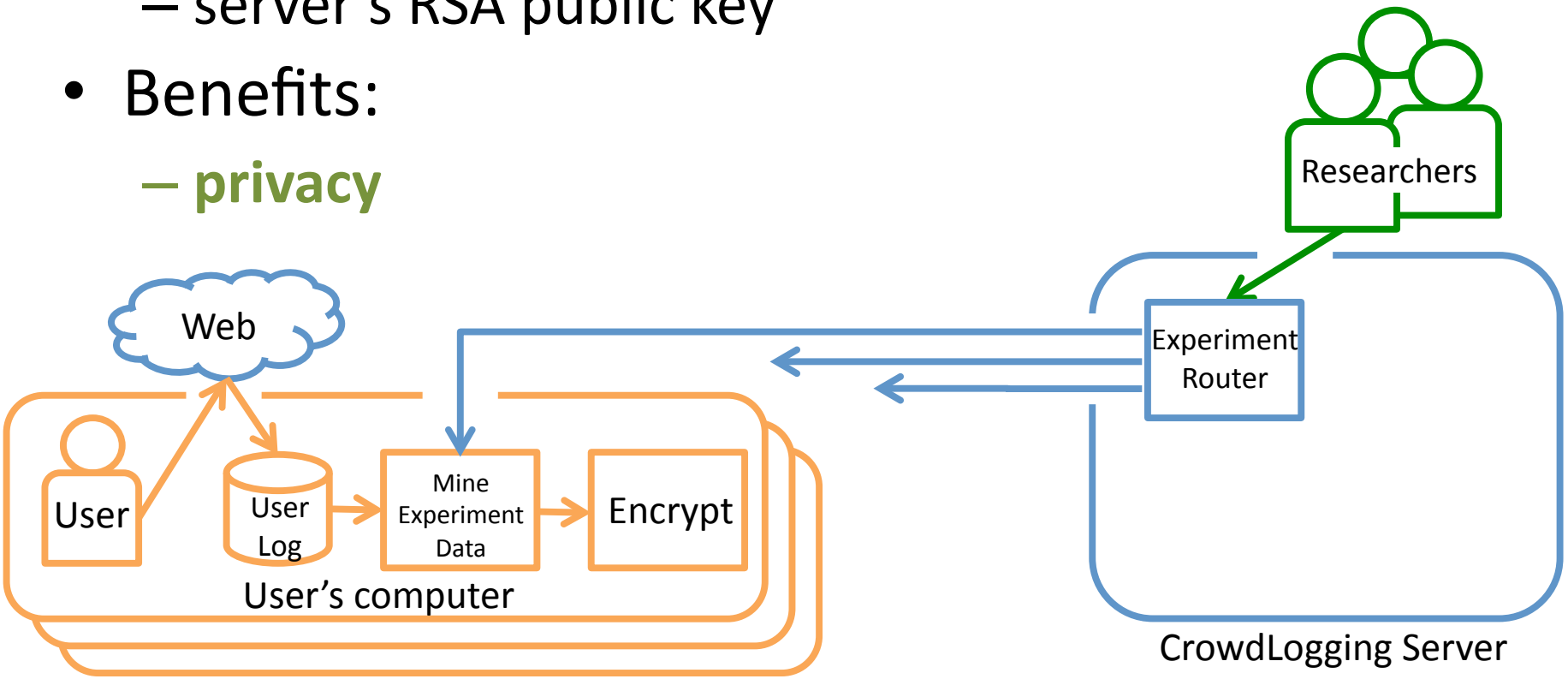
CrowdLogging: how data is mined

- Researchers request a mining experiment
- User software pulls experiment request
- *User approves experiment*
- Extract *search artifacts*
- E.g., query pairs: "home depot -> lowes"
- Benefits:
  - **user control, sharability**



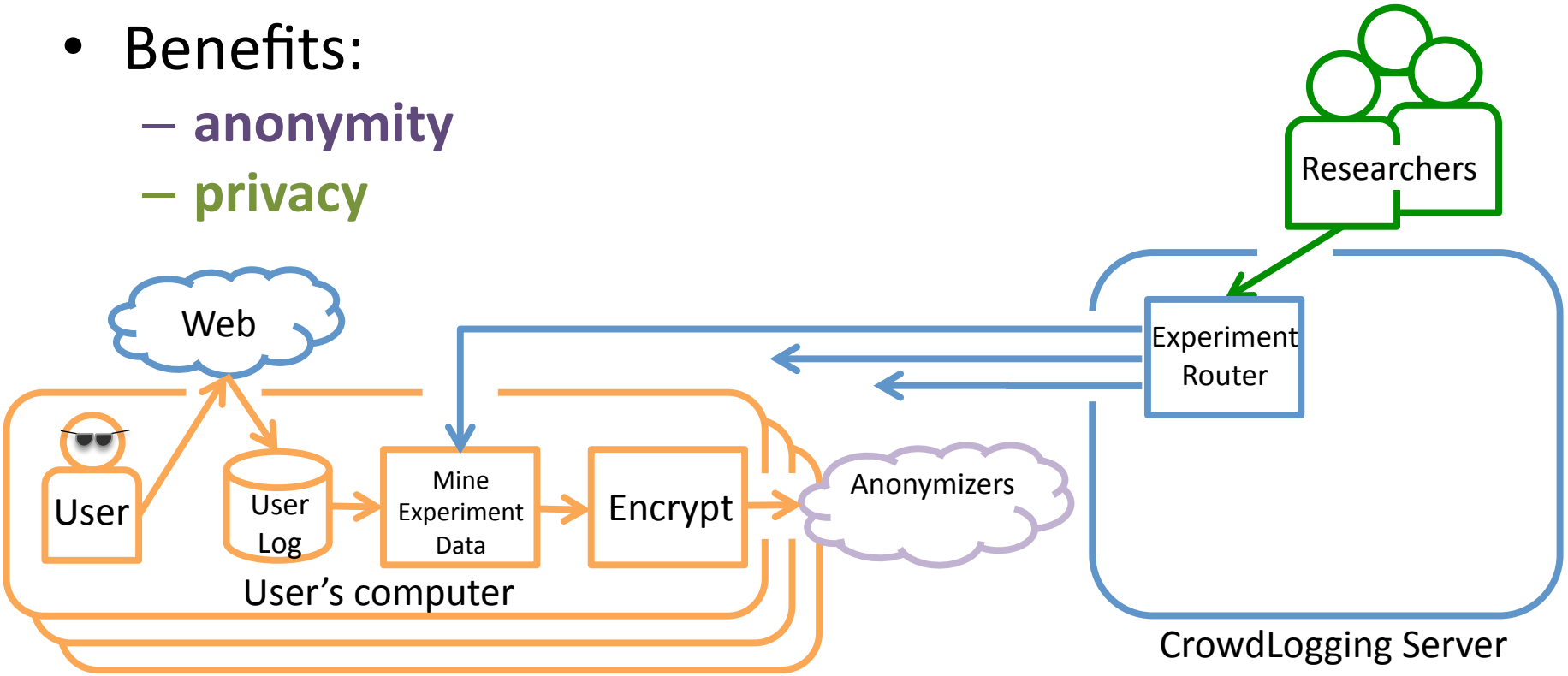
CrowdLogging: how data is encrypted

- Each artifact is encrypted with:
  - secret sharing scheme
  - server's RSA public key
- Benefits:
  - **privacy**



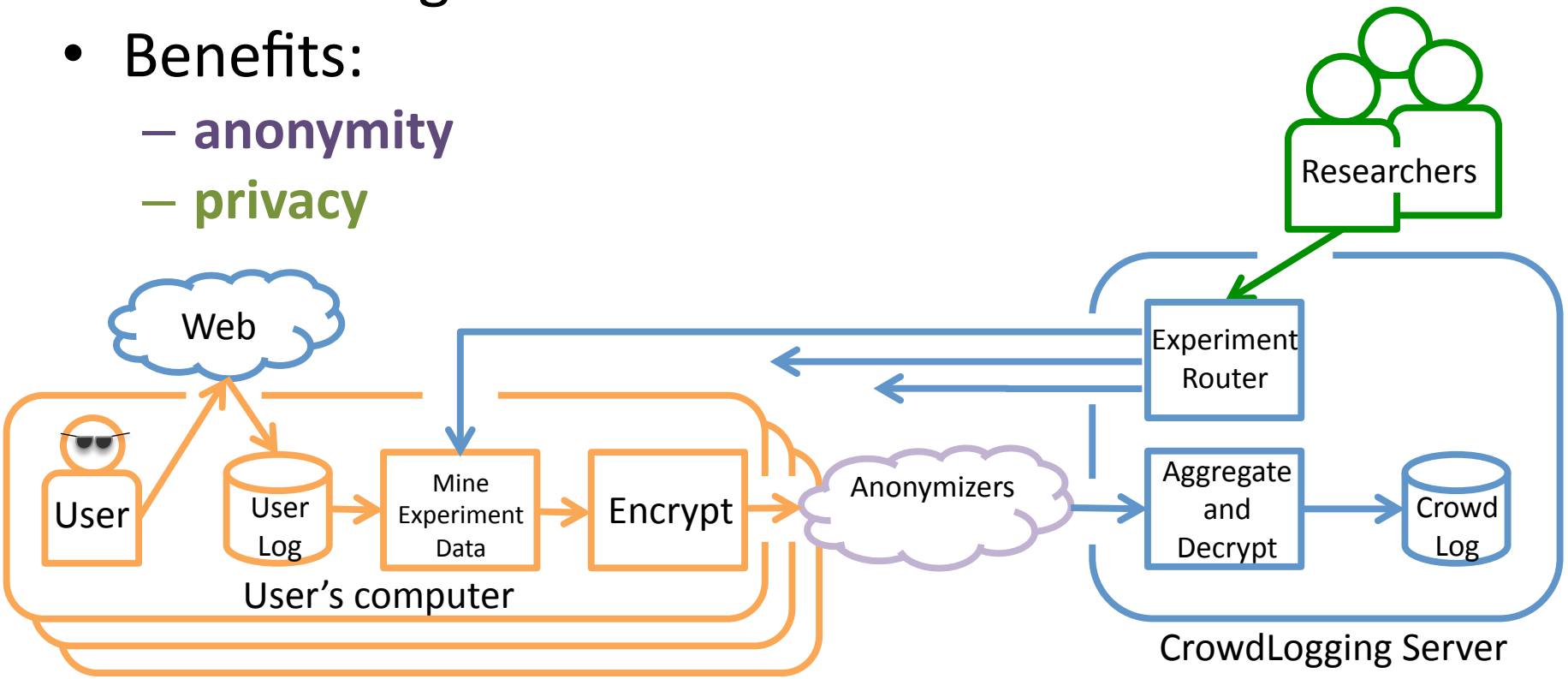
CrowdLogging: how data is uploaded

- Uploaded via an anonymization network
- Prevents server from knowing the source of an encrypted artifact
- Benefits:
  - **anonymity**
  - **privacy**



CrowdLogging: how data is aggregated

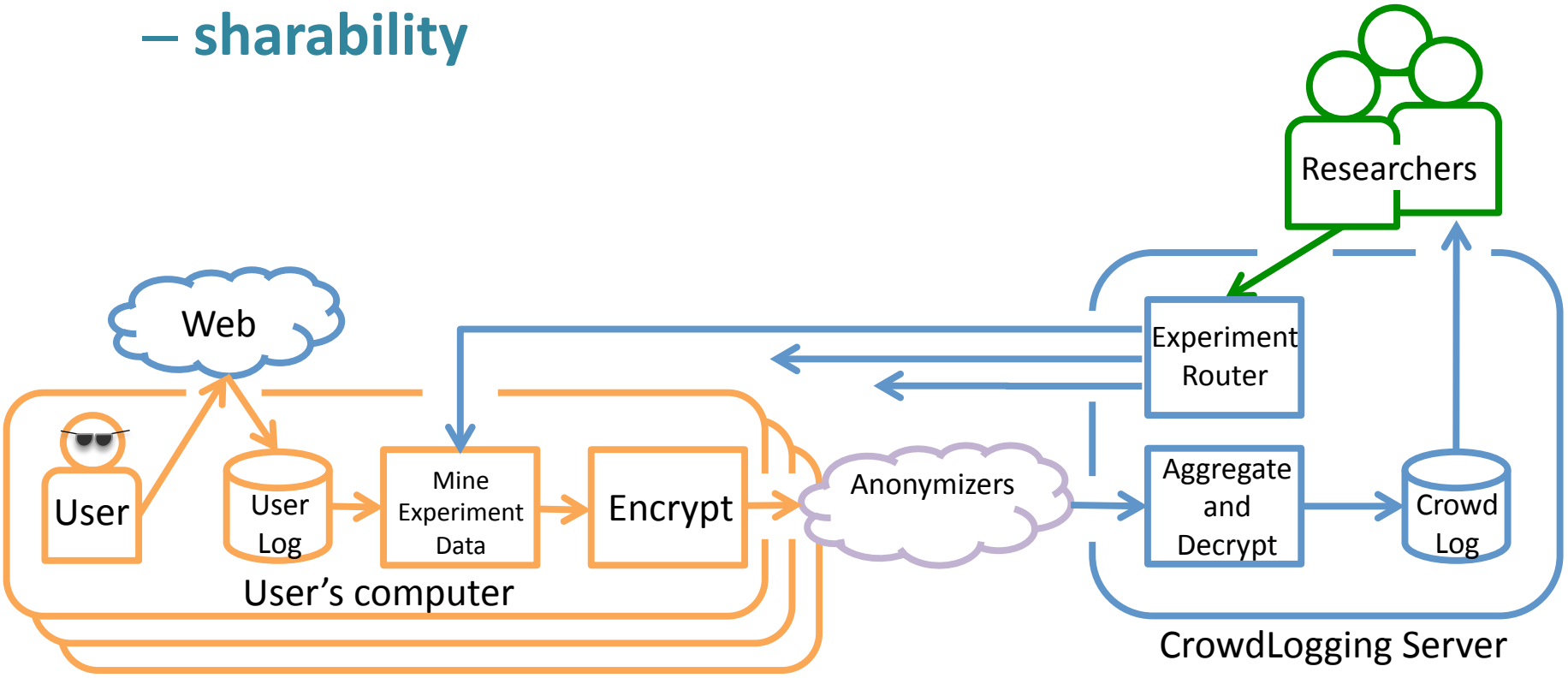
- Artifacts aggregated & decrypted
  - artifacts must be shared by many *different* users*
- A CrowdLog is born
- Benefits:
  - **anonymity**
  - **privacy**



* This can be made more or less strict according to the privacy protocol in use

CrowdLogging: how data is released

- Researchers can access the CrowdLog
- Benefits:
  - **sharability**



CrowdLogging advantages

- *now have* **user control**
  - search data is logged and mined on users' computers
- *now have* **privacy**
  - mined data does not expose PII
- *now have* **anonymity**
  - mined data is uploaded via an anonymization network
- *now have* **sharability**
  - created with the idea of open access search data

CrowdLog examples on AOL

Query CrowdLog (sample)

Query	User Count	Query Count
cheap tickets	1 696	2 438
member rewards	1 626	1 753
florida lottery	1 596	3 410
free games	1 392	1 869
chat	1 391	1 996
jokes	1 391	1 932
lottery	1 360	3 076
dogs	1 330	1 639
...		

Query Click Pair CrowdLog (sample)

Query	Clicked URL	User Count	Query Count
dictionary	dictionary.reference.com	4316	5629
lyrics	www.azlyrics.com	1409	2135
www.yahoo.com	mail.yahoo.com	1173	2056
dictionary	www.m-w.com	1013	1415
myrtle beach	www.mbchamber.com	99	106
song lyrics	www.musicsonglyrics.com	95	103
...			

Decryptable (user count > 5)

Distinct Queries	Total Queries
248 030 (2.5%)	8 620 013 (41.0%)

Decryptable (user count > 5)

Distinct Query Click Pairs	Total Query Click Pairs
106 510 (1.9%)	2 898 912 (31.6%)

Undecryptable

Users	Distinct Queries	Total Queries
4	85 908	423 303
3	171 429	631 246
2	510 602	1 241 115
1	9 138 773	10 097 419

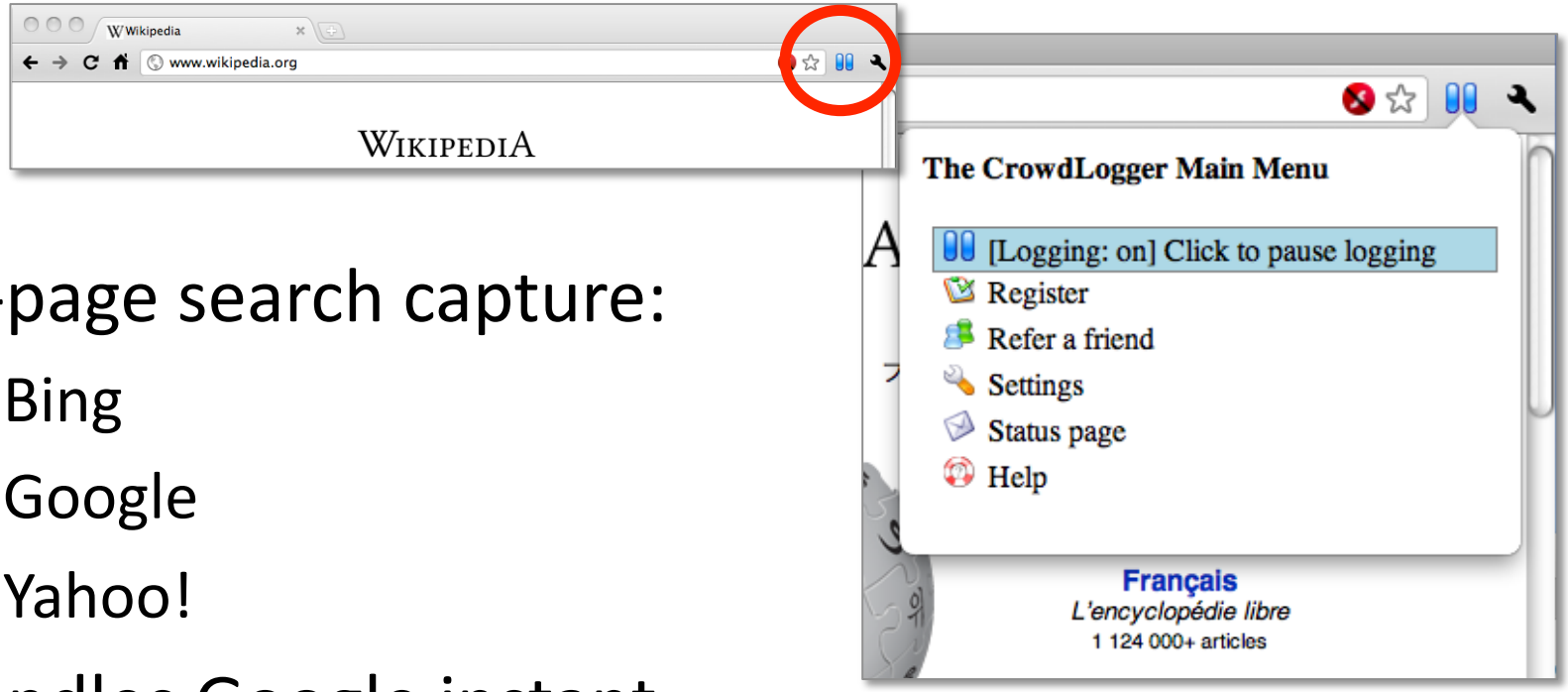
Undecryptable

Users	Distinct Query Click Pairs	Total Query Click Pairs
4	40 906	197 944
3	84 080	304 326
2	259 517	613 674
1	4 910 665	5 169 520

Outline

- *Centralized search logging and mining*
- *CrowdLogging*
  - *logging, mining, and releasing data*
  - *advantages*
  - *comparison with centralized model*
- The CrowdLogger browser extension
  - overview
  - collected data

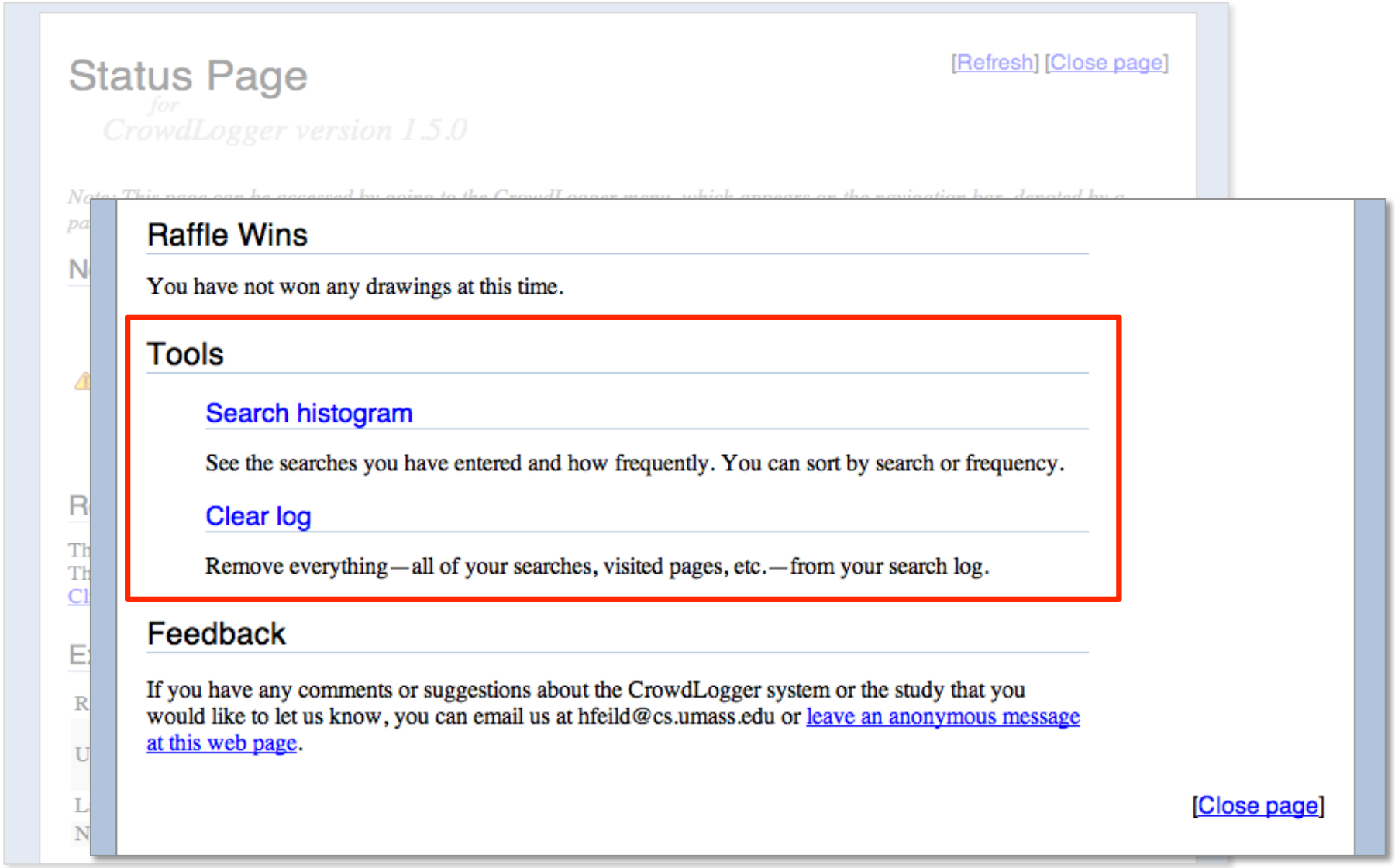
CrowdLogger



- In-page search capture:
  - Bing
  - Google
  - Yahoo!
- Handles Google instant
- Ignores HTTPS URL parameters
- Automatic removal of SSN/phone number patterns
- No logging while in “Privacy” or “Incognito” modes

CrowdLogger

CrowdLogger



CrowdLogger data

- 63 downloads
- 34 distinct registered
- currently cannot rele

- Queries:
  - **sigir 2011**, **cikm 2011**, **wsdm 2012**
- Query click pairs:
  - **cikm 2011 -> www.cikm2011.org**
  - **wsdm 2012 -> wsdm2012.org**

Summary

- **CrowdLogging**
  - a new way to collect and mine search data
  - it's private, distributed, and anonymous
  - less useful, more practical then centralized data
- **CrowdLogger**
  - an implementation for Chrome and Firefox
  - join the study and download: http://crowdlogger.org
  - questions/suggestions? email: info@crowdlogger.org

Thanks

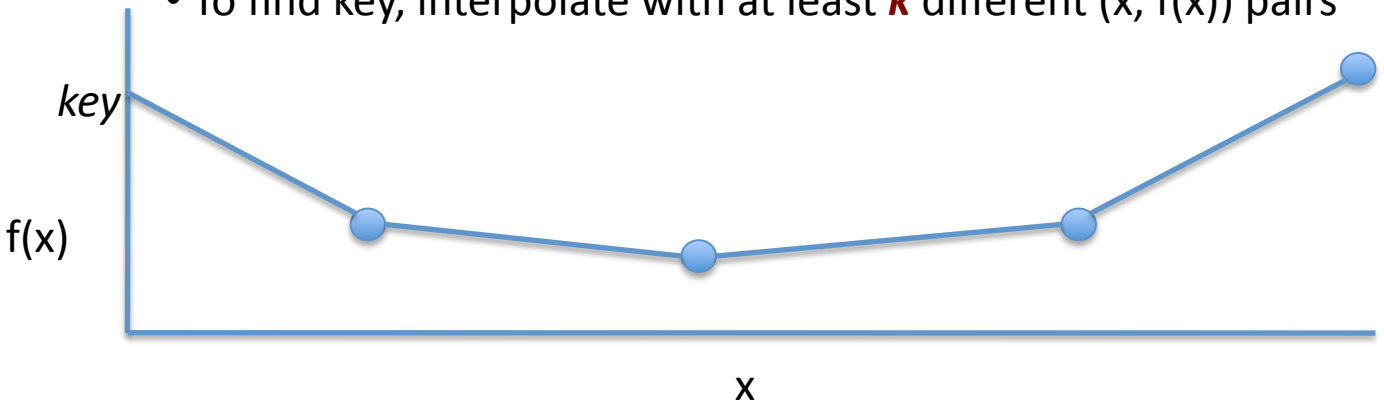
Secret Sharing

- Start with: artifact, ***k***, user's pass phrase, experiment ID
- Deterministically pick some key = genKey(artifact + experiment ID)
  - Range(genKey) = [0, very large prime]
- Deterministically pick ***k*** numbers *n* given artifact + experiment ID
- Create a polynomial $f(x) = y + n_1*x + n_2*x^2 + ... + n_k*x^k$
- Set x = genX(artifact + pass phrase)
  - Range(genX) = R+
- Symmetrically encrypt artifact using key
- Send off with: **[enc(artifact, key), x, f(x)]**

...

- To find key, interpolate with at least ***k*** different (x, f(x)) pairs

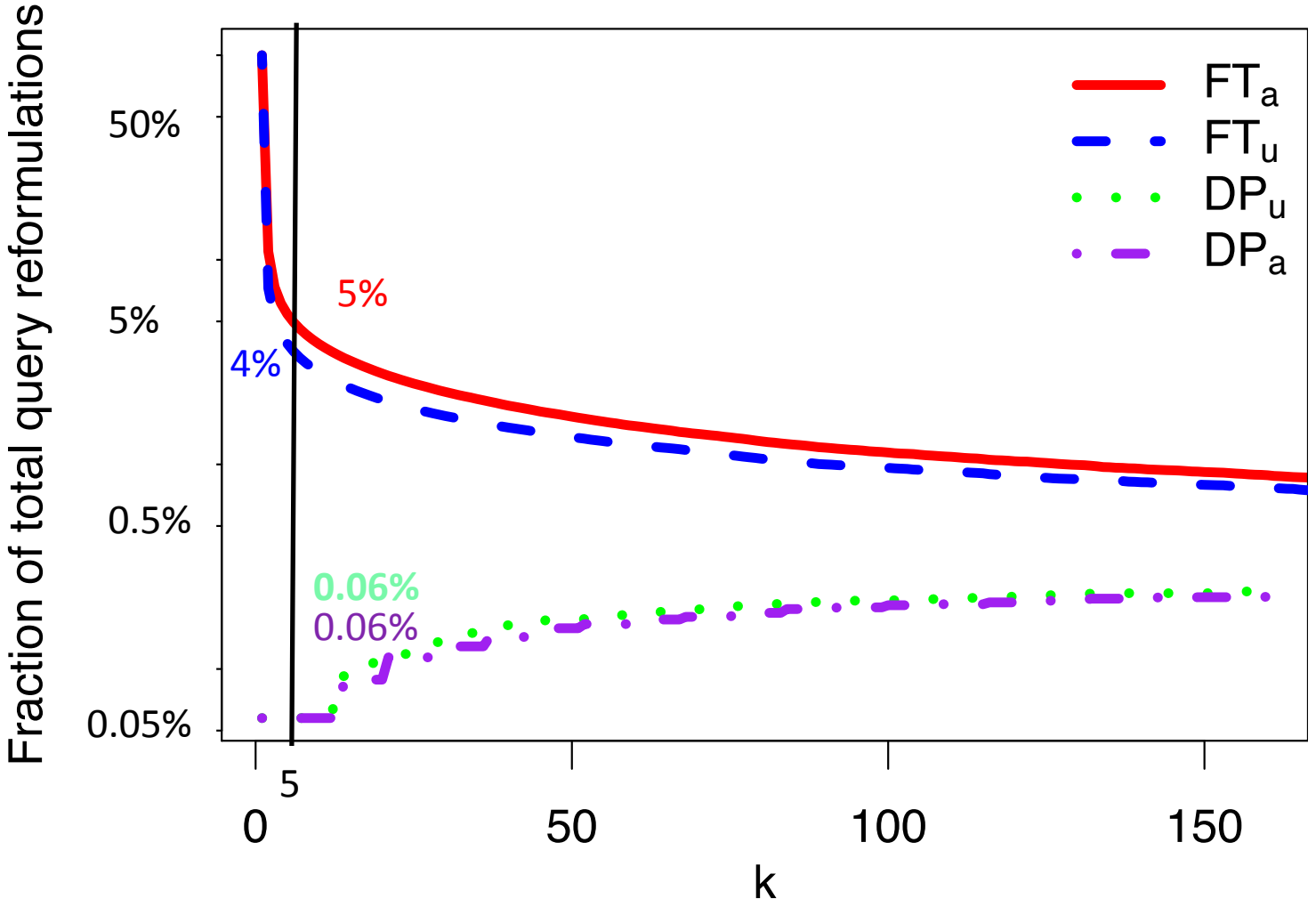
Demo:
http://ciir.cs.umass.edu/~hfeild/ssss



Interpolated polynomial for some given artifact + experiment ID combination.

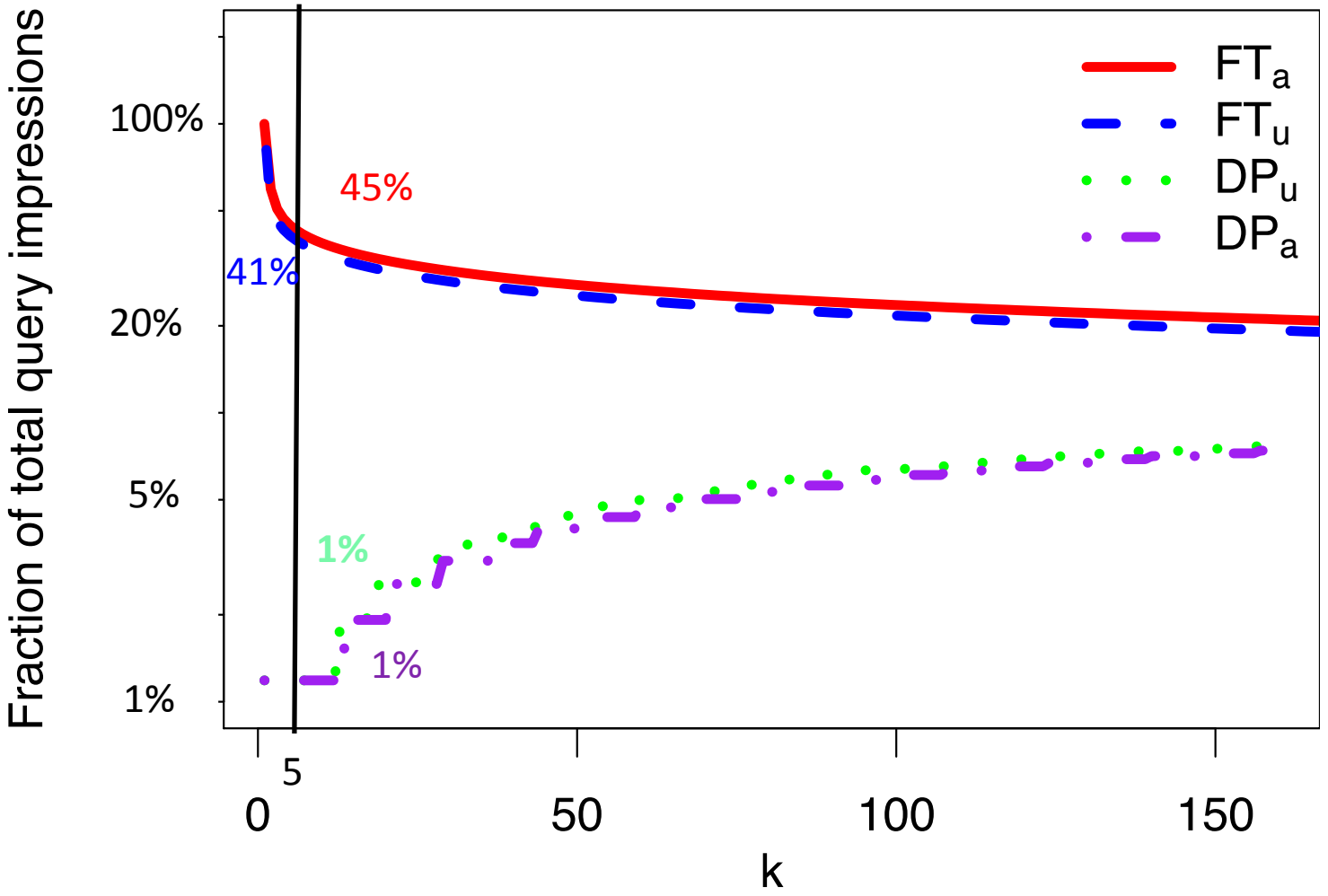
CrowdLogging vs. Centralized logging

Query Reformulations on AOL



CrowdLogging vs. Centralized logging

Query Counts on AOL



CrowdLog examples on AOL

Query CrowdLog (sample)

Query	User Count	Query Count
cheap tickets	1 696	2 438
member rewards	1 626	1 753
florida lottery	1 596	3 410
free games	1 392	1 869
chat	1 391	1 996
jokes	1 391	1 932
lottery	1 360	3 076
dogs	1 330	1 639
...		

Query Pair CrowdLog (sample)

QueryA	QueryB	User Count	Query Count
weather	wheather	70	73
ups	usps	64	81
greyhound	amtrak	63	65
american idol results	american idol	62	63
internet	webunlock	54	55
fredericks of hollywood	fredricks of hollywood	53	60
mycl.cravelyrics.com	bad day lyrics	53	62
...			

Decryptable @ k = 5

Distinct Queries	Total Queries
248 030	8 620 013

Decryptable @ k = 5

Distinct Query Pairs	Total Query Pairs
46 267	792 864

Undecryptable @ k = 5

Users (k)	Distinct Queries	Total Queries
4	85 908	423 303
3	171 429	631 246
2	510 602	1 241 115
1	9 138 773	10 097 419

Undecryptable @ k = 5

Users (k)	Distinct Query Pairs	Total Query Pairs
4	21 228	95 469
3	48 380	163 696
2	186 721	425 921
1	18 380 942	18 877 722

CrowdLog examples on AOL

Query Click Pair CrowdLog (sample)

Query	Clicked URL	User Count	Query Count
dictionary	http://dictionary.reference.com	4316	5629
lyrics	http://www.azlyrics.com	1409	2135
www.yahoo.com	http://mail.yahoo.com	1173	2056
dictionary	http://www.m-w.com	1013	1415
myrtle beach	http://www.mbchamber.com	99	106
song lyrics	http://www.musicsonglyrics.com	95	103

...

Decryptable @ k = 5

Distinct Query Click Pairs	Total Query Click Pairs
106 510	2 898 912

Undecryptable @ k = 5

Users (k)	Distinct Query Click Pairs	Total Query Click Pairs
4	40 906	197 944
3	84 080	304 326
2	259 517	613 674
1	4 910 665	5 169 520