How Anonymous Is My Anonymous Data?

Matt Bishop
Dept. of Computer Science
University of California at Davis

The Problem

We need to share data

"We believe that **data sharing** is **essential** for expedited translation of research results into knowledge, products, and procedures to improve human health."
Final NIH Statement on Sharing Research Data, National Institutes of Health (2003)

"**[N]on-reproducible** single occurrences are of **no significance** to science"
The Logic of Scientific Discovery, Karl Popper (1959)

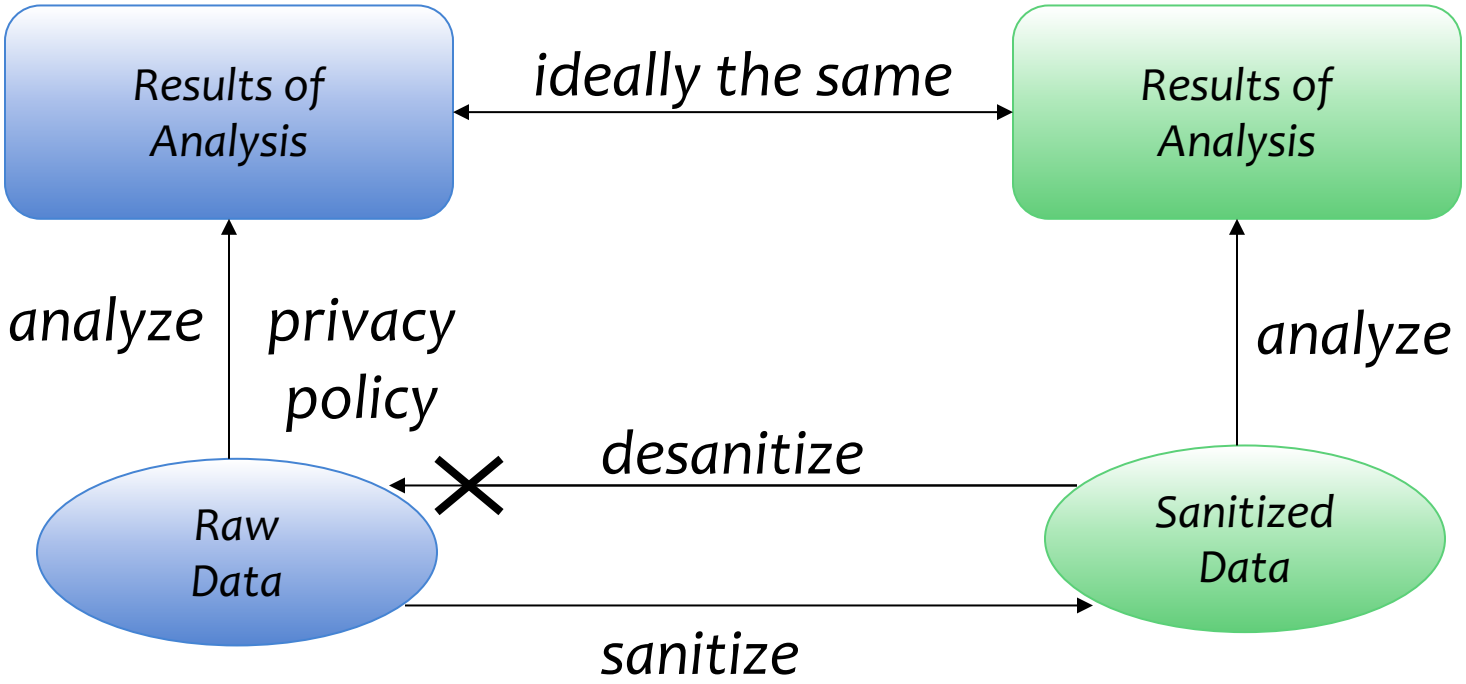
"**Data-driven innovation** is a **key enabler** of growth and jobs in Europe. The importance of data collected online and generated by the Internet of Things (IoT) objects, and the **availability of big data analytics tools** and **artificial intelligence applications** are key technical drivers."
Guidance on Private Sector Data Sharing, European Commission (2019)

Key Point to Take Away

Core message:

Anonymization is a problem of threats, risks, and relationships, with some of each being unknown.

The System Model



The Contradiction

* Exposing personal or other confidential information can cause problems
  * Data may be private
  * Data may allow *inferences* about private matters
* But other parts of the information must be exposed for analysis
  * Network traces, for Internet attacks
  * Medical data, for research or public health analysis

Two Goals

A statistical disclosure takes place when the release of the statistics S makes it possible to determine the value of an attribute more accurately than is possible without access to S

* Dalenius (1977)

The risk to privacy should not substantially increase as a result of participating in a statistical database

* Dwork (2006)

Differential Privacy!

* K randomized function; $S \subseteq ran(K)$; D_1, D_2 data sets differing in at most 1 element
* Second goal is met if, for some small ε:

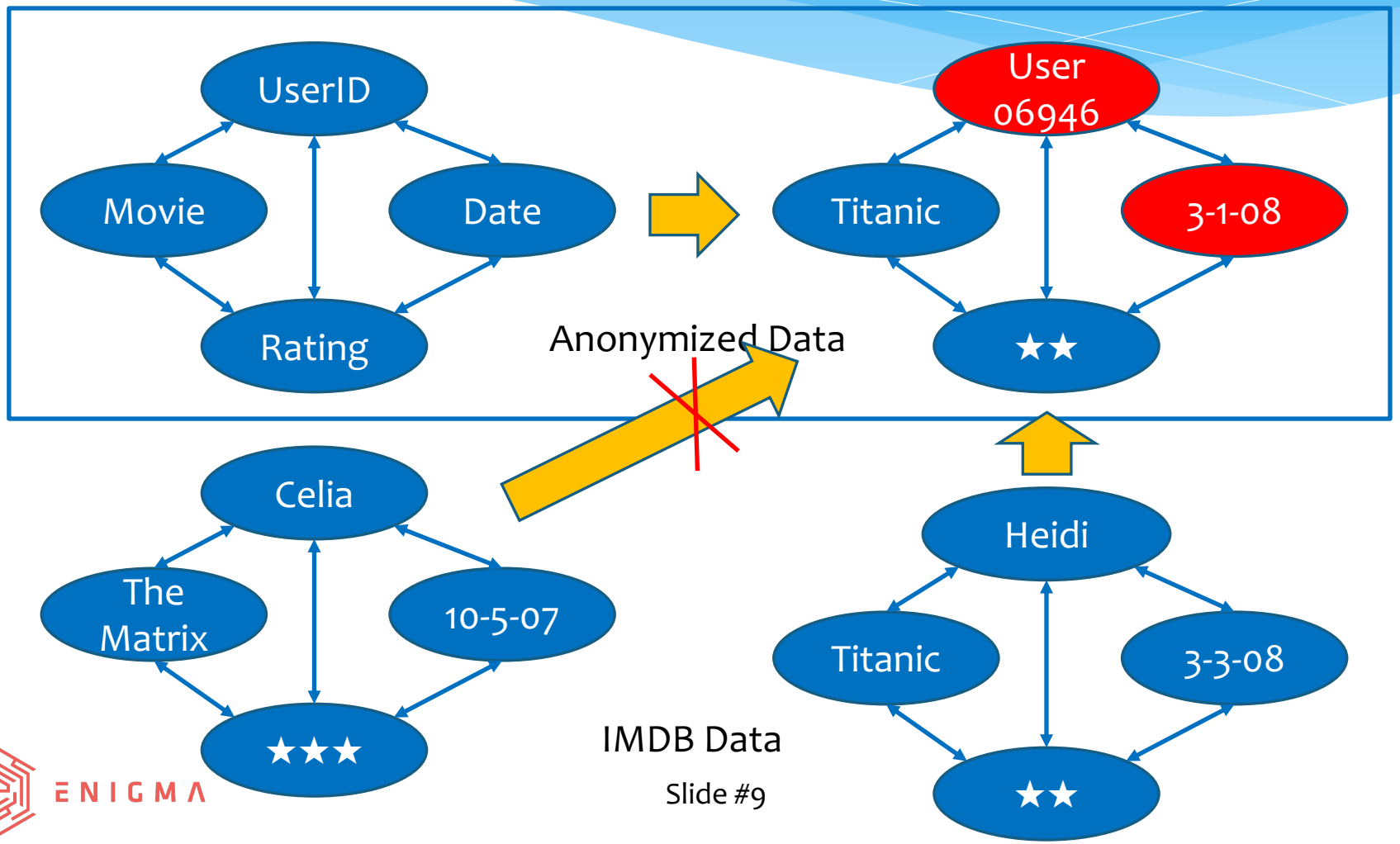
$$\Pr[K(D_1) \in S] = e^{\varepsilon} \Pr[K(D_2) \in S]$$

* Here, K anonymizes

About That First One . . .

* Dwork: for non-interactive release of S, there is always external information that, combined with the released S, will enable one to deduce information that could not be learned without access
  * Question is, how to determine what that information is, and how an adversary can get it

Netflix Data



ENIGMA

Threat Model

* What access to the raw data do adversaries have?
  * Can they inject "markers" that elude sanitization but that help with desanitization?
* What access to the sanitized data do adversaries have?
  * ***Assume the same as analysts***

Threat Model

* What auxiliary data do adversaries have access to?
  * Adversary can't desanitize data set based on information in that (sanitized) set
* But adversary knows Paul works in late evening
  * Given that datum, adversary can figure out which entries in data set apply to Paul

Temporal

* In the USA:
  * During Great Depression, many people joined Communist Party; not seen as a threat
  * During 1950s, past membership seen as threat
    * HUAC, McCarthy, others
  * Now, not seen as a threat

Situational

* Account names, passwords
  * (Friendly) attacker cracked many of these, sent owners of accounts email warning them
  * Also noticed many of the passwords of male account holders were variants of female name, and vice versa
    * So now we know who is involved with whom

Non-Obvious Relations

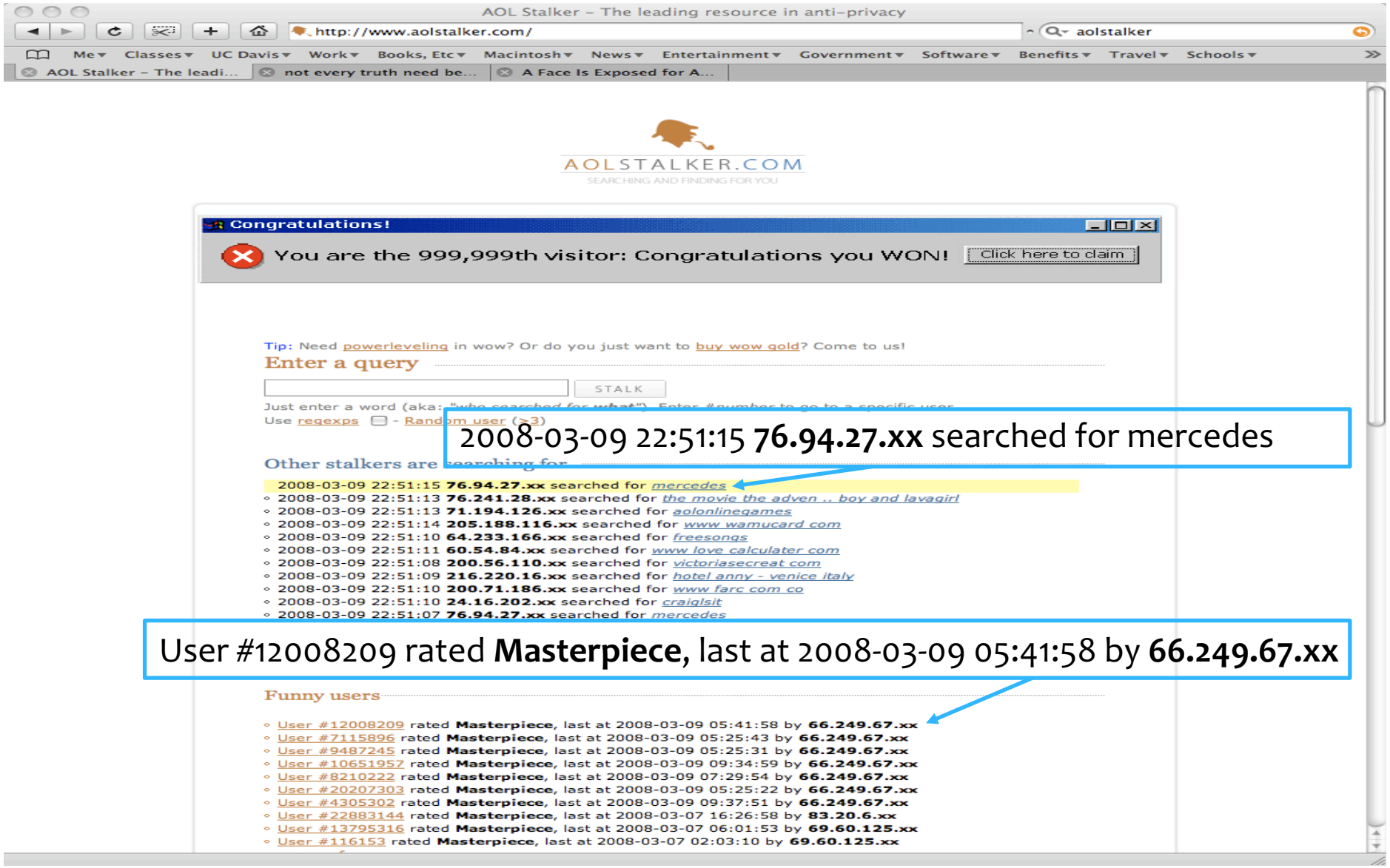
* Goal: hide name
  * Phone number
  * Social security number
  * (ZIP code, date of birth, gender)
* Goal: social security number
  * Date of birth

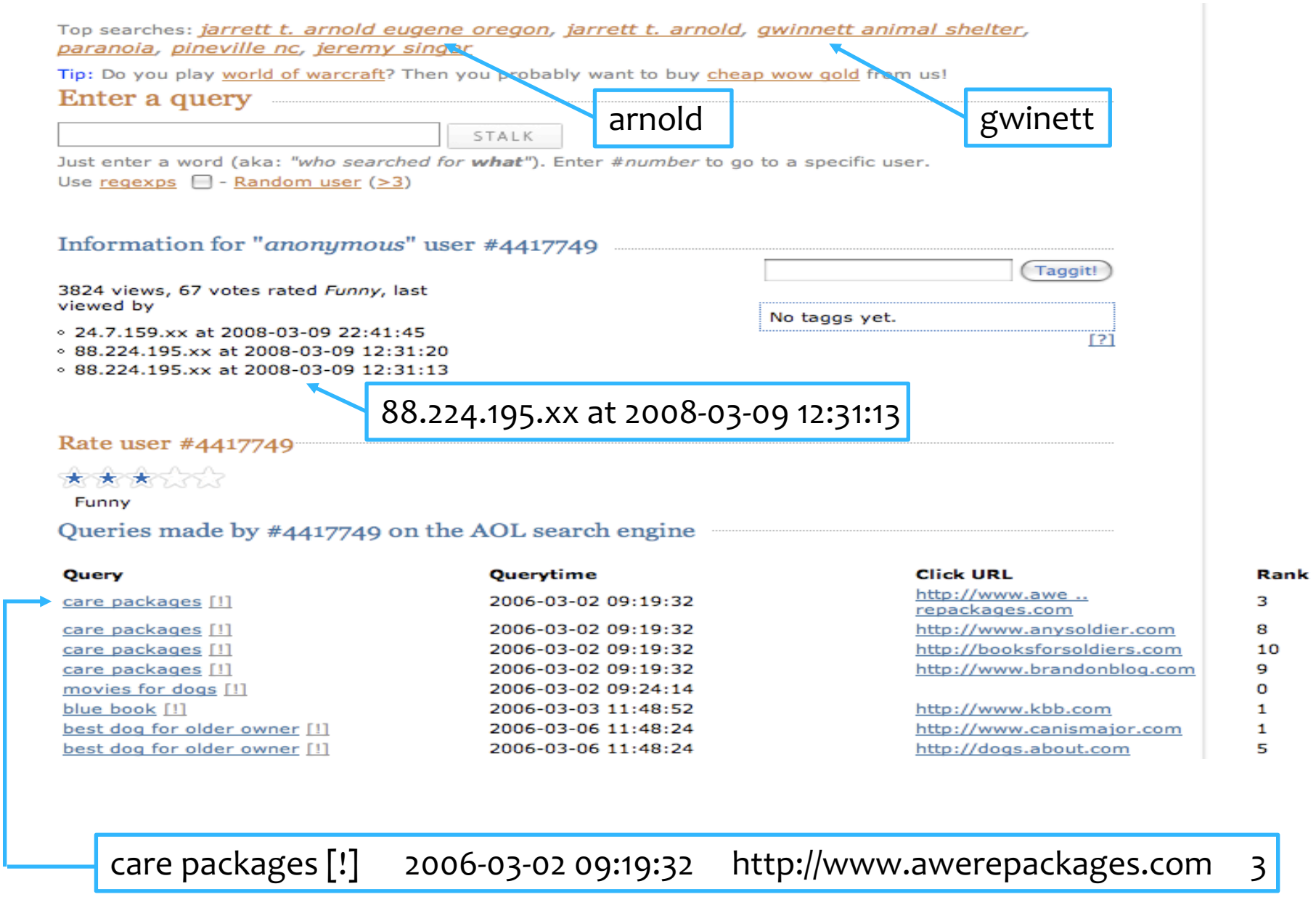
Adventures with AOL

* AOL released 21,011,340 search queries involving 657,426 users for March-May, 2006
* Data set has:
  * Anonymous user id
  * Query
  * Time of query
  * If click through, rank of item clicked on
  * If click through, URL clicked on
* Data posted August 3, 2006
* Data taken down August 7, 2006

First Aftermath

* *New York Times*, August 9, 2006:
  "A Face Is Exposed for AOL Searcher No. 4417749"
* User #4417749: Thelma Arnold of Lilburn, GA
* Found by following queries such as
  * landscapers in Lilburn, GA
  * several people with the last name Arnold
  * homes sold in shadow lake subdivision gwinnett county georgia


AOL Stalker – The leading resource in anti-privacy
http://www.aolstalker.com/
aolstalker
AOLSTALKER.COM
SEARCHING AND FINDING FOR YOU
Congratulations!
You are the 999,999th visitor: Congratulations you WON!
Click here to claim
Tip: Need powerleveling in wow? Or do you just want to buy wow gold? Come to us!
Enter a query
STALK
Use regexps - Random user (<3)
Other stalkers are searching for
2008-03-09 22:51:15 76.94.27.xx searched for mercedes
2008-03-09 22:51:13 76.241.28.xx searched for the movie the adven .. boy and lavagirl
2008-03-09 22:51:13 71.194.126.xx searched for aolonlinegames
2008-03-09 22:51:14 205.188.116.xx searched for www wamucard com
2008-03-09 22:51:10 64.233.166.xx searched for freesongs
2008-03-09 22:51:11 60.54.84.xx searched for www love calculater com
2008-03-09 22:51:08 200.56.110.xx searched for victoriasecreat com
2008-03-09 22:51:09 216.220.16.xx searched for hotel anny - venice italy
2008-03-09 22:51:10 200.71.186.xx searched for www farc com co
2008-03-09 22:51:10 24.16.202.xx searched for craigslsit
2008-03-09 22:51:07 76.94.27.xx searched for mercedes
Funny users
User #12008209 rated Masterpiece, last at 2008-03-09 05:41:58 by 66.249.67.xx
User #7115896 rated Masterpiece, last at 2008-03-09 05:25:43 by 66.249.67.xx
User #9487245 rated Masterpiece, last at 2008-03-09 05:25:31 by 66.249.67.xx
User #10651957 rated Masterpiece, last at 2008-03-09 09:34:59 by 66.249.67.xx
User #8210222 rated Masterpiece, last at 2008-03-09 07:29:54 by 66.249.67.xx
User #20207303 rated Masterpiece, last at 2008-03-09 05:25:22 by 66.249.67.xx
User #4305302 rated Masterpiece, last at 2008-03-09 09:37:51 by 66.249.67.xx
User #22883144 rated Masterpiece, last at 2008-03-07 16:26:58 by 83.20.6.xx
User #13795316 rated Masterpiece, last at 2008-03-07 06:01:53 by 69.60.125.xx
User #116153 rated Masterpiece, last at 2008-03-07 02:03:10 by 69.60.125.xx
2008-03-09 22:51:15 76.94.27.xx searched for mercedes
User #12008209 rated Masterpiece, last at 2008-03-09 05:41:58 by 66.249.67.xx


Top searches: jarrett t. arnold eugene oregon, jarrett t. arnold, gwinnett animal shelter, paranoia, pineville nc, jeremy singer
Tip: Do you play world of warcraft? Then you probably want to buy cheap wow gold from us!
Enter a query
STALK
arnold
gwinett
Just enter a word (aka: "who searched for what"). Enter #number to go to a specific user.
Use regexps - Random user (>3)
Information for "anonymous" user #4417749
Taggit!
3824 views, 67 votes rated Funny, last viewed by
24.7.159.xx at 2008-03-09 22:41:45
88.224.195.xx at 2008-03-09 12:31:20
88.224.195.xx at 2008-03-09 12:31:13
No taggs yet.
[?]
88.224.195.xx at 2008-03-09 12:31:13
Rate user #4417749
Funny
Queries made by #4417749 on the AOL search engine
Query
Querytime
Click URL
Rank
care packages [!] 2006-03-02 09:19:32 http://www.awe .. repackages.com 3
care packages [!] 2006-03-02 09:19:32 http://www.anysoldier.com 8
care packages [!] 2006-03-02 09:19:32 http://booksforsoldiers.com 10
care packages [!] 2006-03-02 09:19:32 http://www.brandonblog.com 9
movies for dogs [!] 2006-03-02 09:24:14 0
blue book [!] 2006-03-03 11:48:52 http://www.kbb.com 1
best dog for older owner [!] 2006-03-06 11:48:24 http://www.canismajor.com 1
best dog for older owner [!] 2006-03-06 11:48:24 http://dogs.about.com 5
care packages [!] 2006-03-02 09:19:32 http://www.awerepackages.com 3

What This Means

* Privacy protection requires sanitization based on accurate, complete threat model
  * Exclusive focus on single, isolated aspect of risk *without adequate, complete threat model* can result in privacy policies and sanitization methods that amplify, not reduce, detrimental consequence

Tying Votes to People

* Poll books record when people enter polling station to vote
* DREs print timestamp on paper representing ballots
* Canvass audits vote counting to ensure the ballots and votes are counted correctly
  * Does not tie names to ballots!
* Someone requested poll books, paper ballots and correlated times
  * This showed how many people voted

Conclusion

* Sanitization problem depends not only on what must be kept secret now, but what might have to be kept secret in future
* Threat modeling aspect of data sanitization critical to effective sanitization
* Environment (laws, customs, etc.) affect both the problem and its solution
* Need to understand the trade-offs is critical, not just from a scientific and engineering point of view, but also from a social point of view

A Parting Thought

The personal life of every individual is based on secrecy, and perhaps it is partly for that reason that civilized man is so nervously anxious that personal privacy should be respected.

— *Anton Chekhov*

Contact Information

Matt Bishop
Department of Computer Science
University of California, Davis
Davis, CA 95616-8562
USA

phone: +1 (530) 752-8060
email: mabishop@ucdavis.edu
www: http://seclab.cs.ucdavis.edu/~bishop