

# Framework for Data Anonymization

Matt Bishop

Dept. of Computer Science  
University of California at Davis



# Joint Work With ...

- Bhume Bhumiratana
- Rick Crawford
- Karl Levitt
- Lisa Clark
- *Aided by the other members of the Computer Security Laboratory*
- Work funded by the U.S. National Science Foundation and Promia, Inc.



# The Importance of Privacy

- “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*



# Recent News Article

## **To Aim Ads, Web Is Keeping Closer Eye on You**

A famous New Yorker cartoon from 1993 showed two dogs at a computer, with one saying to the other, “On the Internet, nobody knows you’re a dog.” That may no longer be true.

A new analysis of online consumer data shows that large Web companies are learning more about people than ever from what they search for and do on the Internet, gathering clues about the tastes and preferences of a typical user several hundred times a month.

These companies use that information to predict what content and advertisements people most likely want to see. They can charge steep prices for carefully tailored ads because of their high response rates.

- L. Story, *New York Times* (March 10, 2008); available at <http://www.nytimes.com/2008/03/10/technology/10privacy.html?th&emc=th>



# 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



# Second Aftermath

The screenshot shows a web browser window titled "AOL Stalker - The leading resource in anti-privacy". The address bar shows "http://www.aolstalker.com/". The browser's menu bar includes "Me", "Classes", "UC Davis", "Work", "Books, Etc", "Macintosh", "News", "Entertainment", "Government", "Software", "Benefits", "Travel", and "Schools". The page content features the AOL Stalker logo and a "Congratulations!" message: "You are the 999,999th visitor: Congratulations you WON!". Below this is a search section with a "STALK" button and a list of "Other stalkers are searching for" with various IP addresses and search terms. A "Funny users" section lists several users who rated "Masterpiece".

**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.calculator.com](#)
- 2008-03-09 22:51:08 **200.56.110.xx** searched for [victoriasecret.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 [craigslist](#)
- 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**



# Example: User 4417749

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

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

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 tags yet.



## Rate user #4417749



Funny

## Queries made by #4417749 on the AOL search engine

Query	Querytime	Click URL	Rank
<a href="#">care packages</a> [!]	2006-03-02 09:19:32	<a href="http://www.awe..repackages.com">http://www.awe..repackages.com</a>	3
<a href="#">care packages</a> [!]	2006-03-02 09:19:32	<a href="http://www.anysoldier.com">http://www.anysoldier.com</a>	8
<a href="#">care packages</a> [!]	2006-03-02 09:19:32	<a href="http://booksforsoldiers.com">http://booksforsoldiers.com</a>	10
<a href="#">care packages</a> [!]	2006-03-02 09:19:32	<a href="http://www.brandonblog.com">http://www.brandonblog.com</a>	9
<a href="#">movies for dogs</a> [!]	2006-03-02 09:24:14		0
<a href="#">blue book</a> [!]	2006-03-03 11:48:52	<a href="http://www.kbb.com">http://www.kbb.com</a>	1
<a href="#">best dog for older owner</a> [!]	2006-03-06 11:48:24	<a href="http://www.canismajor.com">http://www.canismajor.com</a>	1
<a href="#">best dog for older owner</a> [!]	2006-03-06 11:48:24	<a href="http://dogs.about.com">http://dogs.about.com</a>	5



# The Problem

- 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



# Idea

- Hide the sensitive data
  - Data said to be *deidentified*, *anonymized*, or *sanitized*
  - Original data called *raw* or *unsanitized*
- Several ways
  - Delete the data
  - Perturb the data
  - Generalize the data



# Controversial

- “If you’re not doing something wrong, why would you want to hide anything?”
  - Short answer: identity theft
  - Longer answer: privacy, “the right to be let alone”
- “You have zero privacy anyway ... get over it”
  - Short answer: so?
  - Longer answer: lack of privacy arises from not recognizing what needs to be protected, or protecting it inadequately



# Theme of This Talk

- Data sanitization requires analysis of how data *might* be used
  - Not merely how it *is intended to* be used
- Threat analysis in data sanitization is critical
  - Examine not only the current context, but what future contexts may bring
  - Goals of bad folks may differ from yours!



# Problem Statement

- Provide access to data in such a way as to satisfy the following conditions:
  - *Privacy constraints* — What you cannot reveal
  - *Analysis requirements* — What you must reveal



# Framework

- *Collectors* gather data
  - Also *sanitize* it to meet privacy constraints
- *Analysts* analyze the data
  - Goal is to draw conclusions about raw data
- *Adversaries* try to recover raw data
  - May be happy with only portions of it

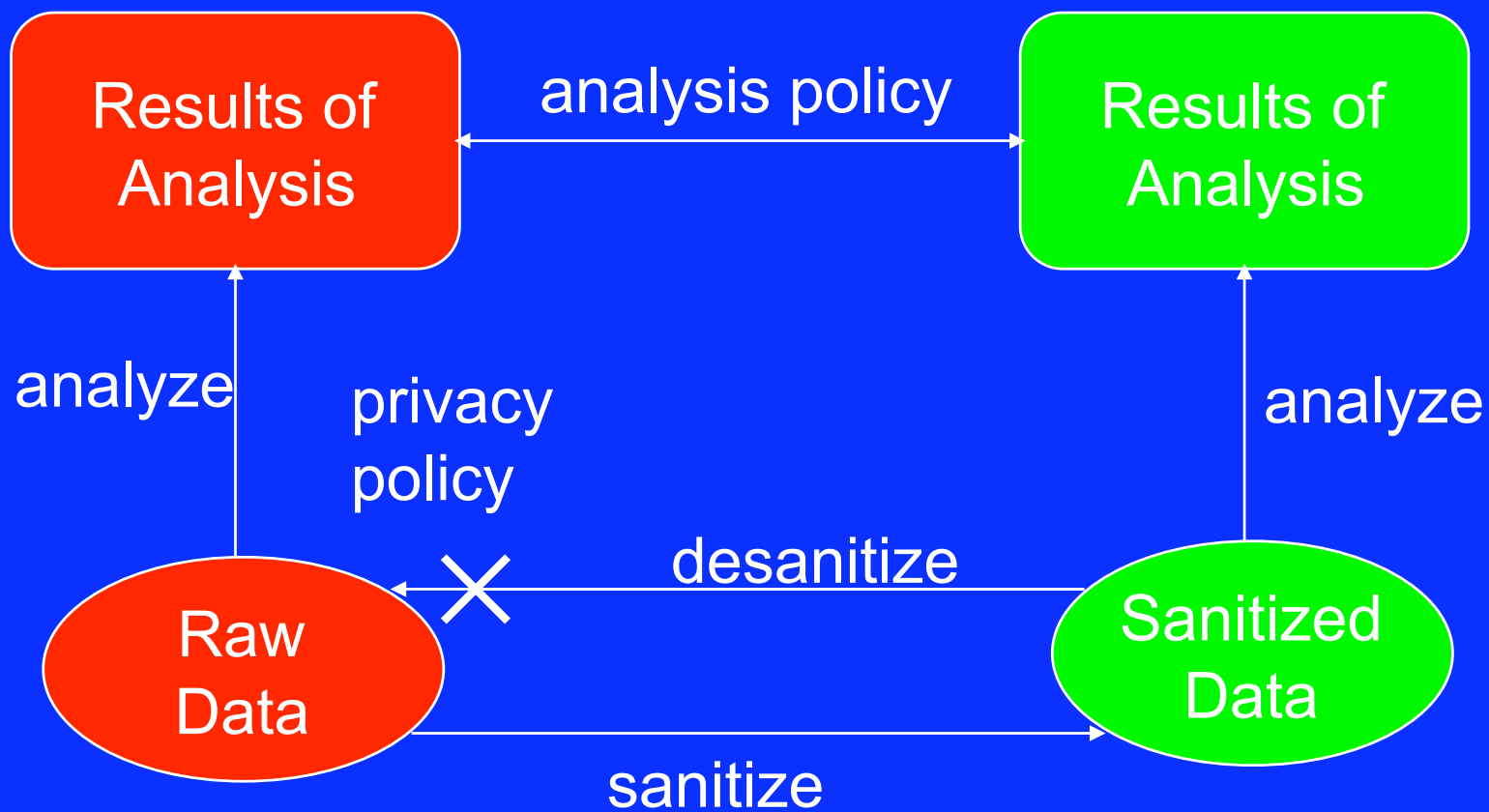


# Relationships

- Determine if particular host accessed `www.private.xz`; privacy policy requires origin of connections private
  - Conflict!
- Determine if particular host on given network accessed `www.private.xz`; same privacy policy
  - Only 1 host on network: conflict!
  - Multiple hosts on network: no conflict



# More Generally



# How to Sanitize

- Prevent adversary from tying sanitized data to an entity but consistently obscuring that entity

123.45.67.89	→	192.68.5.12
203.67.2.123		95.31.28.54
123.45.67.89	→	192.68.5.12
97.15.2.2		66.32.1.90
203.67.2.25		95.31.28.25



# Ways to Sanitize

- Prevent adversary from determining sanitized data refers to same entity

123.45.67.89	→	192.68.5.12
203.67.2.123		95.31.28.54
123.45.67.89	→	36.5.21.224
97.15.2.2		66.32.1.90
203.67.2.25		95.31.28.25



# Ways to Sanitize

- Prevent adversary from tying sanitized data to a set of entities
  - Called “*k*-anonymity”

123.45.67.89	→	192.68.5.12
203.67.2.123		95.31.28.54
123.45.67.89	→	192.19.5.70
97.15.2.2		66.32.1.90
203.67.2.25		95.31.28.25



# Threat Model

- What is the threat?
  - Reveal information about a *particular* person?
  - Reveal information about a *set of people*?
- When is the threat?
  - Is the threat for a limited time only?
  - Is the threat only for this data set, or will future sets be added to it?



# 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 set
  - Adversary knows Paul works in late evening
  - Given that datum, adversary can figure out which entries in data set apply to Paul



# Example Threat Evolution

- Consider credit cards
  - Identity thief trying to steal one card's information
    - Hide name, number, expiration date
  - Private investigator trying to determine if any cardholder having illicit affair
    - Also hide raw purchase transaction data



# Continued

- CC1's policy: PI not a threat
  - Sanitized data reveals 7% of cardholders seem to be having affairs
  - Basis: these folks suddenly start carrying balance near credit limit (“external knowledge”)
- CC2's policy: PI *is* threat
  - Now must sanitize balances or limits



# Morals

1. New data allows unexpected inferences
  - Here, revealing credit limits makes card holders vulnerable
2. Another person's data can make you vulnerable
  - Here, knowing people who have affairs have credit card balances near credit limits makes card holders having affairs vulnerable



# More On Threats

- Items of minor importance for current analysis may be of major importance for future analysis
  - Cell phone usage irrelevant for analysis of medical records for diseases
  - Cell phone usage relevant for marketing by phone companies



# What This Means

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



# Example

- Unsanitized diagnosis attribute
  - “Ms. D has cancer” (Ms. D a pseudonym)
- Initially predicate has undefined or default probability
  - But now we know the cancer attribute for Ms. D is 1
- Life insurance company: knows someone has cancer, but not who
  - Solves problem of identification risk



# But ...

- Life insurance company concludes “Ms. D” is “Ms. Diana James” (incorrectly)
  - Up go Ms. James’ rates
- Life insurance company concludes “Ms. D” is one of “Ms. Diana”, “Ms. Doris”, “Ms. Deirdre”, “Ms. Donna”, or “Ms. Delilah”
  - Up go all their rates



# Still More on Threats

- Information syntactically unrelated may be semantically related
- Example: AOL search data for user id #4417749
- Example: patient is pregnant
  - Inference: patient is female
- Example: patient has Tay-Sachs
  - Inference: patient is probably an Ashkenazi Jew



# Example

- Goal: correlate age, general residential area, gender, and income
- Data: contains name, address, city, state, ZIP code, gender, date of birth, income
- Sanitizer: deletes first 4 fields
  - Releases records with ZIP code, gender, date of birth, income
- Problem: you can uniquely identify 87% of the population of the United States by the first 3 fields



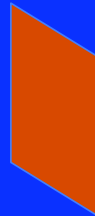
# Key Assumptions

- *Closed world assumption*
  - Attention restricted to items appearing explicitly in raw dataset
- Uniform analysis metric
  - All attributes equally valuable for analysis purposes

These are usually implicit, not explicit

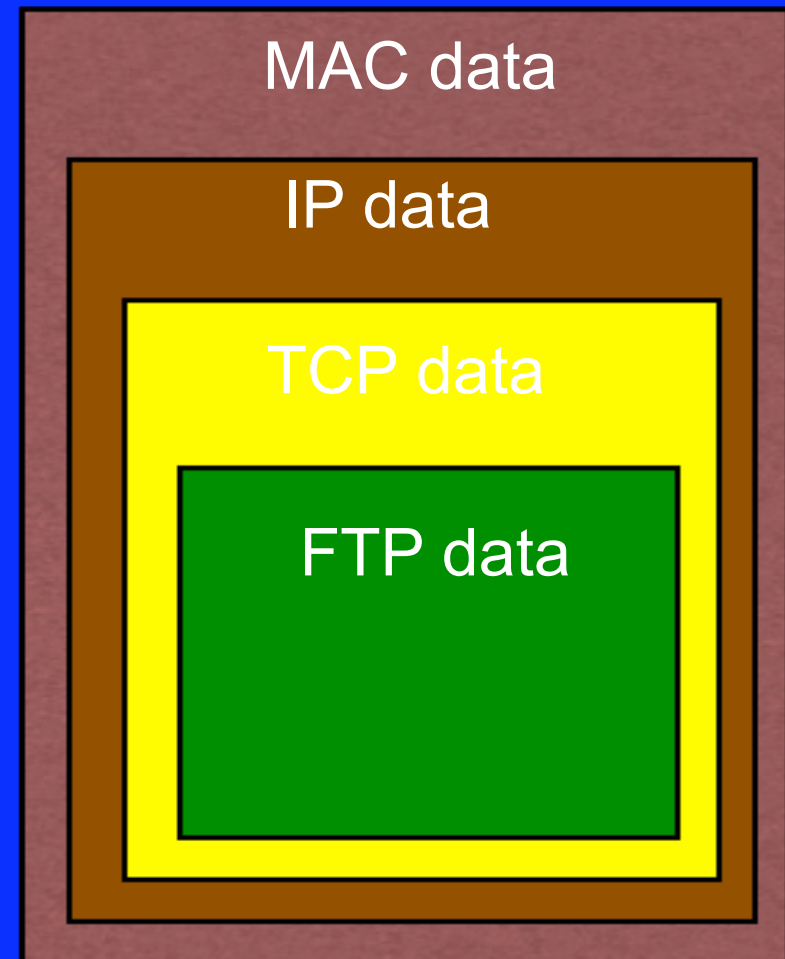


# One Approach



# XML Converter

- Input: raw data
- Output: structured representation
- Example: network packets
  - XML tree structure provides natural expression of nesting of layers



# Detailed Example

- FTP network traffic
  - Privacy policy: hide IP addresses
- Problem: IP addresses in headers
  - Converter must supply context for both IP, FTP
  - Note: IP addresses in former are binary, in latter are ASCII
    - Have converter, reverter handle it; hide from sanitizer



# Sanitizer

- Transform input data
  - Could derive structure from data itself, but this means one sanitizer for each type of input data
  - With XML, one sanitizer that keys on tags and attributes
- Example: FTP traffic
  - Sanitize addresses at session, IP layers
  - Can distinguish based on nesting in tree



# XML Reverter

- Restore sanitized data to original format
  - Allows tools that would work on original, unsanitized data to work on transformed, sanitized data



# Policy Languages

- Policy expressed in terms non-technical policy maker can understand
- Allow direct comparison between privacy, analysis policy to detect, identify conflicts
- Policy expression needs to lead to efficient sanitization function
- Expression must allow changes to threat model to update privacy policy automatically



# 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



# Contact Information

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

*phone:* +1 (530) 752-8060

*email:* [bishop@cs.ucdavis.edu](mailto:bishop@cs.ucdavis.edu)

*www:* <http://seclab.cs.ucdavis.edu/~bishop>

