CSE 484: Computer Security and Privacy

Anonymity

Spring 2024

David Kohlbrenner
dkohlbre@cs

Thanks to Franzi Roesner, Dan Boneh, Dieter Gollmann, Dan Halperin, David Kohlbrenner, Yoshi Kohno, Ada Lerner, John Manferdelli, John Mitchell, Vitaly Shmatikov, Bennet Yee, and many others for sample slides and materials ...
Logistics

• **Final Project Part A due Wednesday**
  • Make sure your patch passes the gradescope autograder
  • Think about what your patch does with valid and invalid Host: headers
  • Consider what the range of valid Host: headers is
    • Note: you don’t need to make tinyserv better than it was, just prevent exploitation
  • Please make your forks private!

• **Sploit2**
  • We’ve had requests for RCA+Patch
    • No patch release
  • Maybe a sploit1 patch release _after_ all are in
Part 1: Anonymity in Datasets
How to release an anonymous dataset?

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.; Saul Hansell contributed reporting for this article. Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "nunc fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shawd lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.
How to release an anonymous dataset?

• Possible approach: remove identifying information from datasets?

Massachusetts medical+voter data [Sweeney 1997]

Figure 1 Linking to re-identify data
k-Anonymity

- Each person contained in the dataset cannot be distinguished from at least \( k-1 \) others in the data.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>State of domicile</th>
<th>Religion</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramsha</td>
<td>29</td>
<td>Female</td>
<td>Tamil Nadu</td>
<td>Hindu</td>
<td>Cancer</td>
</tr>
<tr>
<td>Yadu</td>
<td>24</td>
<td>Female</td>
<td>Kerala</td>
<td>Hindu</td>
<td>Viral infection</td>
</tr>
<tr>
<td>Salima</td>
<td>28</td>
<td>Female</td>
<td>Tamil Nadu</td>
<td>Muslim</td>
<td>TB</td>
</tr>
<tr>
<td>Kaker</td>
<td>27</td>
<td>Male</td>
<td>Karnataka</td>
<td>Parsi</td>
<td>No illness</td>
</tr>
<tr>
<td>Joan</td>
<td>24</td>
<td>Female</td>
<td>Karnataka</td>
<td>Christian</td>
<td>Heart-related</td>
</tr>
<tr>
<td>Bahuksana</td>
<td>23</td>
<td>Male</td>
<td>Karnataka</td>
<td>Buddhist</td>
<td>TB</td>
</tr>
<tr>
<td>Rambha</td>
<td>19</td>
<td>Male</td>
<td>Kerala</td>
<td>Hindu</td>
<td>Cancer</td>
</tr>
<tr>
<td>Kishor</td>
<td>29</td>
<td>Male</td>
<td>Karnataka</td>
<td>Hindu</td>
<td>Heart-related</td>
</tr>
<tr>
<td>John</td>
<td>17</td>
<td>Male</td>
<td>Kerala</td>
<td>Christian</td>
<td>Heart-related</td>
</tr>
<tr>
<td>John</td>
<td>19</td>
<td>Male</td>
<td>Kerala</td>
<td>Christian</td>
<td>Viral infection</td>
</tr>
</tbody>
</table>
k-Anonymity

- Each person contained in the dataset cannot be distinguished from at least k-1 others in the data.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>State of domicile</th>
<th>Religion</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Female</td>
<td>Tamil Nadu</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Female</td>
<td>Kerala</td>
<td>*</td>
<td>Viral infection</td>
</tr>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Female</td>
<td>Tamil Nadu</td>
<td>*</td>
<td>TB</td>
</tr>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Male</td>
<td>Karnataka</td>
<td>*</td>
<td>No illness</td>
</tr>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Female</td>
<td>Kerala</td>
<td>*</td>
<td>Heart-related</td>
</tr>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>Age ≤ 20</td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>20 &lt; Age ≤ 30</td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>Age ≤ 20</td>
<td>Male</td>
<td>Kerala</td>
<td>*</td>
<td>Viral infection</td>
</tr>
</tbody>
</table>

Doesn’t work for high-dimensional datasets (which tend to be sparse)

[Robust De-anonymization of Large Sparse Datasets]

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin
Netflix Challenge:

• Netflix released a (non-uniform) random sample of user’s movie ratings
• Challenge was to build a better recommendation system
• Data was ‘anonymous’
  • ID # only
  • Random selection of a given user’s ratings
  • “noise” added (appears that there was no noise)
Result: No real anonymity

• Cross-correlate with IMBD ratings

• A handful (6 or fewer) ratings of non-top 500 movies is enough!
Part 2: Anonymity in Communication
Chaum’s Mix

• Early proposal for anonymous email

• Modern anonymity systems use Mix as the basic building block
Basic Mix Design

Adversary knows all senders and all receivers, but cannot link a sent message with a received message
Anonymous Return Addresses

M includes $\{K_1, A\}_{pk(mix)}$, $K_2$ where $K_2$ is a fresh public key

Secrecy without authentication (good for an online confession service 😊)
Mix Cascades and Mixnets

- Messages are sent through a sequence of mixes
  - Can also form an arbitrary network of mixes ("mixnet")
- Some of the mixes may be controlled by attacker, but even a single good mix ensures anonymity
- Pad and buffer traffic to foil correlation attacks
Disadvantages of Basic Mixnets

• Public-key encryption and decryption at each mix are **computationally expensive**
• Basic mixnets have **high latency**
  • OK for email, not OK for anonymous Web browsing
• Challenge: **low-latency anonymity network**
Another Idea: Randomized Routing
e.g., Onion Routing

- Sender chooses a random sequence of routers
  - Some routers are honest, some controlled by attacker
  - Sender controls the length of the path
Onion Routing

- Routing info for each link encrypted with router’s public key
- Each router learns only the identity of the next router
Tor

• Second-generation onion routing network
  • http://tor.eff.org
  • Developed by Roger Dingledine, Nick Mathewson and Paul Syverson
  • Specifically designed for low-latency anonymous Internet communications

• Running since October 2003

• “Easy-to-use” client proxy
  • Freely available, can use it for anonymous browsing
Tor Circuit Setup (1)

- Client proxy establishes a symmetric session key and circuit with Onion Router #1
Tor Circuit Setup (2)

- Client proxy extends the circuit by establishing a symmetric session key with Onion Router #2
  - Tunnel through Onion Router #1
Tor Circuit Setup (3)

- Client proxy extends the circuit by establishing a symmetric session key with Onion Router #3
  - Tunnel through Onion Routers #1 and #2
Using a Tor Circuit

• Client applications connect and communicate over the established Tor circuit.
How do you know who to talk to?

• Directory servers
  • Maintain lists of active onion routers, their locations, current public keys, etc.
  • Control how new routers join the network
    • “Sybil attack”: attacker creates a large number of routers
  • Directory servers’ keys ship with Tor code
Location Hidden Service

- **Goal:** deploy a server on the Internet that anyone can connect to without knowing where it is or who runs it
- Accessible from anywhere
- Resistant to censorship
- Can survive a full-blown DoS attack
- Resistant to physical attack
  - Can’t find the physical server!
Creating a Location Hidden Server

Server creates circuits to “introduction points”

Server gives intro points’ descriptors and addresses to service lookup directory

Client obtains service descriptor and intro point address from directory

Introduction Points

Client Alice

Server Bob

Service Lookup Server

Bob’s Service
Using a Location Hidden Server

Client creates a circuit to a “rendezvous point”

Rendezvous point splices the circuits from client & server

If server chooses to talk to client, connect to rendezvous point

Client sends address of the rendezvous point and any authorization, if needed, to server through intro point

Client Alice

Rendezvous Point

Introduction Points

Server Bob
Issues and Notes of Caution

• Passive traffic analysis
  • Infer from network traffic who is talking to whom
  • To hide your traffic, must carry other people’s traffic!

• Active traffic analysis
  • Inject packets or put a timing signature on packet flow

• Compromise of network nodes
  • Attacker may compromise some routers
    • Powerful adversaries may compromise “too many”
  • It is not obvious which nodes have been compromised
    • Attacker may be passively logging traffic
  • Better not to trust any individual router
    • Assume that some fraction of routers is good, don’t know which
Issues and Notes of Caution

- Tor isn’t completely effective by itself
  - Tracking cookies, fingerprinting, etc.
  - Exit nodes can see everything!
Issues and Notes of Caution

• The simple act of using Tor could make one a target for additional surveillance

• Hosting an exit node could result in illegal activity coming from your machine

• Tor not designed to protect against adversaries with the capabilities of a state (public statement by designers, at least in the past)
Aside -- HDCP
Problem: People like copying movies!

• Solution: DRM (Digital Rights Management)
  • DVD players, Streaming service plugins, etc
  • Encrypt video in-transit, decrypt on device
Problem: People like copying movies!

• Solution: DRM (Digital Rights Management)
  • DVD players, Streaming service plugins, etc
  • Encrypt video in-transit, decrypt on device

• Problem: The analog hole – You have to display the context eventually
Problem: Analog Hole

• Solution: … The same thing again – DRM
  • HDCP -- High-bandwidth Digital Content Protection
  • Encrypt data on the wire between the computer output and the monitor