Privacy

Presenter: Yinzhi Cao
Lehigh University
Some contents are borrowed from the following sources.


https://www.cs.utexas.edu/~shmat/courses/cs361s/webtrack.ppt
Outline

- Differential Privacy
- Web Privacy
- TrackingFree
General Setting

Medical data
Query logs
Social network data
...

Data mining
Statistical queries
General Setting

publish

Data mining
Statistical queries
How can you allow meaningful usage of such datasets while preserving individual privacy?
Blatant Non-Privacy
Blatant Non-Privacy

- Leak individual records
- Can link with public databases to re-identify individuals
- Allow adversary to reconstruct database with significant probability
Attempt 1: Crypto-ish Definitions

I am releasing some useful statistic $f(D)$, and nothing more will be revealed.

What kind of statistics are safe to publish?
How do you define privacy?
Attempt 2:

I am releasing researching findings showing that people who smoke are very likely to get cancer.

You cannot do that, since it will break my privacy. My insurance company happens to know that I am a smoker...
Attempt 2: Absolute Disclosure Prevention

“If the release of statistics $S$ makes it possible to determine the value [of private information] more accurately than is possible without access to $S$, a disclosure has taken place.”

[Dalenius]
An Impossibility Result

[informal] It is not possible to design any non-trivial mechanism that satisfies such strong notion of privacy. [Dalenius]
Attempt 3: “Blending into Crowd” or k-Anonymity

K people purchased A and B, and all of them also purchased C.
Attempt 3: “Blending into Crowd” or k-Anonymity

K people purchased A and B, and all of them also purchased C.

I know that Elaine bought A and B...
Attempt 4: Differential Privacy

\[ x' \text{ is a neighbor of } x \]  
if they differ in one row

From the released statistics, it is hard to tell which case it is.
Attempt 4: Differential Privacy

For all neighboring databases $x$ and $x'$
For all subsets of transcripts:

$$\Pr[A(x) \in S] \leq e^\epsilon \Pr[A(x') \in S]$$
Attempt 4: Differential Privacy

I am releasing researching findings showing that people who smoke are very likely to get cancer.

Please don’t blame me if your insurance company knows that you are a smoker, since I am doing the society a favor.

Oh, btw, please feel safe to participate in my survey, since you have nothing more to lose.

Since my mechanism is DP, whether or not you participate, your privacy loss would be roughly the same!
Notable Properties of DP

- Adversary knows arbitrary auxiliary information
  - No linkage attacks

- Oblivious to data distribution

- Sanitizer need not know the adversary’s prior distribution on the DB
Techniques for Achieving DP

- Output perturbation
- Input perturbation
- Perturbation of intermediate values
- Sample and aggregate
Method: Output Perturbation

- **Global Sensitivity:**
  \[ GS_f = \max_{x,x'} \| f(x) - f(x') \|_1 \]

Example: \[ GS_{avg} = \frac{1}{n} \]
Method: Output Perturbation

\[ A(x) = f(x) + \text{Lap}\left(\frac{G_{S_f}}{\epsilon}\right) \text{ is } \epsilon\text{-DP} \]

Laplace distribution \( \text{Lap}(\lambda) \) has density \( h(y) \propto e^{-\frac{\|y\|_1}{\lambda}} \)

Sliding property of \( \text{Lap}\left(\frac{G_{S_f}}{\epsilon}\right) \): \( \frac{h(y)}{h(y+\delta)} \leq e^{\epsilon \cdot \frac{\|\delta\|}{G_{S_f}}} \) for all \( y, \delta \)

Proof idea:

\( A(x) \): blue curve
\( A(x') \): red curve

\( \delta = f(x) - f(x') \leq G_{S_f} \)
Web Privacy
It’s the Internet! Of course they know you’re a dog. They also know your favorite brand of pet food and the name of the cute poodle at the park that you have a crush on!
Third-party cookies:
Disabled by default (Safari)
Can be disabled by user (many browsers)
Cannot be disabled (Android)
... but there are many other tracking technologies
Behavioral Targeting

Ad network

Publishers
The New York Times
okcupid
Google

Advertisers
Coca-Cola
Ford
<table>
<thead>
<tr>
<th>Partial List of Ad Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>24/7 Real Media</td>
</tr>
<tr>
<td>33Across</td>
</tr>
<tr>
<td>Adify</td>
</tr>
<tr>
<td>AdInterax (Yahoo!)</td>
</tr>
<tr>
<td>Aggregate Knowledge</td>
</tr>
<tr>
<td>Akamai</td>
</tr>
<tr>
<td>Blue Kai</td>
</tr>
<tr>
<td>BlueLithium (Yahoo!)</td>
</tr>
<tr>
<td>Casale Media</td>
</tr>
<tr>
<td>Chitika</td>
</tr>
<tr>
<td>Coremetrics</td>
</tr>
<tr>
<td>Cossette</td>
</tr>
<tr>
<td>eXelate</td>
</tr>
<tr>
<td>EyeWonder</td>
</tr>
<tr>
<td>Fox Audience Network</td>
</tr>
<tr>
<td>FreeWheel</td>
</tr>
<tr>
<td>Navegg</td>
</tr>
<tr>
<td>NextAction</td>
</tr>
<tr>
<td>Microsoft</td>
</tr>
<tr>
<td>MindSet Media</td>
</tr>
<tr>
<td>Outbrain</td>
</tr>
<tr>
<td>PointRoll</td>
</tr>
<tr>
<td>richrelevance</td>
</tr>
<tr>
<td>Right Media (Yahoo!)</td>
</tr>
<tr>
<td>Snoobi</td>
</tr>
<tr>
<td>Specific Media</td>
</tr>
<tr>
<td>Traffic Marketplace</td>
</tr>
<tr>
<td>Tribal Fusion / Exponential</td>
</tr>
<tr>
<td>ValueClick Media</td>
</tr>
<tr>
<td>Vizu</td>
</tr>
<tr>
<td>Acerno</td>
</tr>
<tr>
<td>AdJuggler</td>
</tr>
<tr>
<td>AlmondNet</td>
</tr>
<tr>
<td>Bluestreak</td>
</tr>
<tr>
<td>ChoiceStream</td>
</tr>
<tr>
<td>Criteo</td>
</tr>
<tr>
<td>e-planning</td>
</tr>
<tr>
<td>Google</td>
</tr>
<tr>
<td>NexTag</td>
</tr>
<tr>
<td>Nielsen Online</td>
</tr>
<tr>
<td>PrecisionClick</td>
</tr>
<tr>
<td>Rocket Fuel</td>
</tr>
<tr>
<td>TACODA (AOL)</td>
</tr>
<tr>
<td>Specific Media</td>
</tr>
<tr>
<td>Tribal Fusion / Exponential</td>
</tr>
<tr>
<td>Vizu</td>
</tr>
<tr>
<td>Axiom Relevance-X</td>
</tr>
<tr>
<td>AdShuffle</td>
</tr>
<tr>
<td>Atlas (Microsoft)</td>
</tr>
<tr>
<td>BrightRoll</td>
</tr>
<tr>
<td>ClickTale</td>
</tr>
<tr>
<td>Effective Measure</td>
</tr>
<tr>
<td>Facilitate Digital</td>
</tr>
<tr>
<td>Hurra</td>
</tr>
<tr>
<td>Mediaplex (ValueClick Media)</td>
</tr>
<tr>
<td>nugg.ad</td>
</tr>
<tr>
<td>Omniture</td>
</tr>
<tr>
<td>QuantaCast</td>
</tr>
<tr>
<td>Pulse 360</td>
</tr>
<tr>
<td>SafeCount</td>
</tr>
<tr>
<td>SafeCount *</td>
</tr>
<tr>
<td>Tatoo Media</td>
</tr>
<tr>
<td>Turn</td>
</tr>
<tr>
<td>Undertone Networks / Zedo</td>
</tr>
<tr>
<td>Yahoo!</td>
</tr>
<tr>
<td>[x+1]</td>
</tr>
</tbody>
</table>
Tracking Is Pervasive

64 independent tracking mechanisms in an average top-50 website
Sticky Tracking

Subverting same origin policy
(publisher also runs an ad network)

ad.hi5.com = ad.yieldmanager.com

Flash cookies

Browser fingerprinting

History sniffing
Tracking Technologies

- HTTP Cookies
- HTTP Auth
- HTTP Etags
- Content cache
- IE userData
- HTML5 protocol and content handlers
- HTML5 storage
- Flash cookies
- Silverlight storage
- TLS session ID & resume
- Browsing history
- window.name
- HTTP STS
- DNS cache
Everything Has a Fingerprint
Fingerprinting Web Browsers

- User agent
- HTTP ACCEPT headers
- Browser plug-ins
- MIME support
- Clock skew

- Installed fonts
- Cookies enabled?
- Browser add-ons
- Screen resolution
Your browser fingerprint **appears to be unique** among the 3,435,834 tested so far.

Only **anonymous data** will be collected by this site.

A paper reporting the statistical results of this experiment is now available: *How Unique Is Your Browser?*, Proceedings of the Privacy Enhancing Technologies Symposium (PETS 2010), Springer Lecture Notes in Computer Science.
Panopticlick Example

84% of browser fingerprints are unique
With Flash or Java, 94% are unique
How Websites Get Your Identity

Third party is sometimes the site itself

Leakage of identifiers

GET http://ad.doubleclick.net/adj/...
Referer: http://submit.SPORTS.com/...?email=jdoe@email.com
Cookie: id=35c192bcfe0000b1...

Security bugs

Remember XSUH (cross-site URL hijacking)?

Third party buys your identity
Introduction

Syphilis is a bacterial infection that is usually passed on through having sex with someone who is infected. It can also be passed from an infected mother to her unborn child and, in rare cases, can be caught through injecting drugs. It is extremely rare to catch syphilis through a blood transfusion in the UK as blood donors are carefully screened.

Three stages of disease

Stage 1 (primary syphilis). Symptoms of syphilis begin with a painless but highly infectious sore on the genitals or sometimes around the mouth. If somebody else comes into close contact with the sore, typically during sexual contact, they can also become infected. The sore lasts two to six weeks before disappearing.

Stage 2 (secondary syphilis). Secondary symptoms, such as a skin rash and sore throat, then develop. These symptoms may disappear within a few weeks, after which you experience a latent (hidden) phase with no symptoms, which can last for years. After this, syphilis can progress to its third, most dangerous stage.

Stage 3 (tertiary syphilis). At this stage, it can cause serious damage to the body.

The primary and secondary stages are when you are most infectious to other people. In the latent phase (and usually around two years after becoming infected), syphilis cannot be passed onto others but can still cause symptoms.
History Sniffing

How can a webpage figure out which sites you visited previously?

- Color of links
  - CSS :visited property
  - getComputedStyle()
- Cached Web content timing
- DNS timing
Do Not Track

Basics

HTTP header
- DNT: 1

Standardization

Browser support in FF4, IE9
Beginning to see adoption (AP, NAI) ... or not

Privacy protections

No tracking across sites
- Who is the “third” party?
  Can’t be based on domain
  Example: amazonaws.com, ad.hi5.com ...

No intrusive tracking

Limits on regular log data
 Exceptions for fraud prevention, etc.
“But the NAI code also recognizes that companies sometimes need to continue to collect data for operational reasons that are separate from ad targeting based on a user’s online behavior. For example, online advertising companies may need to gather data to prove to advertisers that an ad has been delivered and should be paid for; to limit the number of times a user sees the same ad; or to prevent fraud.”

Translation: we’re going to keep tracking you, but we’ll simply call it “operational reasons.”
TrackingFree

Goals and Challenges

- Anti-tracking Completeness
- Functionality/compatibility
- Performance

Core Idea: TrackingFree partitions client-side states into multiple isolation units so that the identifiers still exist but not unique any more!

Referer: http://online.wsj.com/
Cookie: id = 12345

Referer: http://www.cnn.com/
Cookie: id = 24578
Out-of-scope threats

TrackingFree doesn’t address following threats:

- Within-Site Tracking.
- Tracking by exploiting browser vulnerabilities
- Stateless tracking.
Architecture

Principal Kernel Interface

Principal

Profile

domain: a.com

Persistent Storage

tab: mail.a.com

Profile

domain: b.com

tab: online.b.com/n

Principal Backend

Message Policy Enforcer

Principal Manager

Public History Manager

Domain Data Manager

Preference Configure

Legend

user-activated flag  non-user-activated flag

navigation  cross-principal message

history update message  session data  user preference
Contents Allocation Mechanism

- Initial Contents Allocation
  - Handles those top frames that are navigated by users directly

- Derivative Contents Allocation
  - Handles those frames that are generated due to the contents on other frames, which we call child frame
Initial Contents Allocation

deals.ebay.com
books.ebay.com
toys.ebay.com

Persistent Storage
Cookie Store HTML5 Storage Cache Plugins User Config ...

news.sina.com.cn
sports.sina.com.cn
blogs.sina.com.cn

Persistent Storage
Cookie Store HTML5 Storage Cache Plugins User Config ...

44
Derivative Contents Allocation

- **Principal Switch**
  - Should we switch principle for child frame?

- **Principal Selection**
  - How to choose target principal?
Principal Switch

The deficiencies of two intuitive yet extreme policies:

- Not privacy-preserving (no switch)
- Unnecessary overhead (too much switch)

Our solution: switch principal only if the following two conditions are met:

- Cross-site
- User-triggered

Same principal

Different principal
Principal Selection

The deficiency of two intuitive yet extreme policies:

- Break compatibility (always create new principal)
- Break anti-tracking capacity (create at most one principal for each domain)

Our solution:

- Maintains an in-degree bounded graph for principals
- The in-degree of the graph is set to two

Principal Backend

(1) a.com
(2) b.com
(3) c.com
(4) d.com
(5) b.com
(6) c.com
(7) b.com
(10) b.com
(11) c.com

Gmail

Youtube
Explicit communication is widely used, but break the isolation mechanism.

Our solution:
we restrict the use of explicit communication as follows:

- Third-party elements in one principle can not explicitly communicate with other principles.
- First-party elements can only explicitly communicate with the first-party elements placed in its neighbor principles.
Principal Communication

Implicit Communication

- History Sharing
  - UI history manager
  - Accepts information from other managers
  
  Only UI manager gets associated with browser

Communication through navigation URL

send msg
Preference Configure

- User preference can be abused to store tracking identifier. (e.g. strict transport security)
- Completely isolating user preference affects user preference.
- Our solution:
  - Isolate user preference.
  - Apply user-initiated changes to all of the principals.
  - Monitor GUI message to determine user-initiated preference change.
Evaluation

- Anti-tracking capability
  - Formal proof
  - Experiments with real world websites

- Performance
  - Overhead (latency, memory, disk)

- Compatibility
Formal Proof

- Use Alloy to formally analyze TrackingFree’s anti-tracking ability.
  - Alloy is the most popular formal proof system

- Describe TrackingFree’s behaviors on an existing Alloy Web model [Akhwae et al. CSF 2010].

- Formally verified trackers can correlate TrackingFree user’s activities up to three principals without site collaboration.
Anti-tracking Capability with Real World Web Sites

- Gathered tracking tokens on Alexa Top web sites by following the tracker detection of [Roesner et al. NSDI 2012].
- Detection based on the observation that each tracking request must contain the user’s globally unique identifier.
- Some false negative, no false positive.
## Anti-tracking Capability with Real World Web Sites

<table>
<thead>
<tr>
<th>Tracking Host</th>
<th>Prevalence (# Domains)</th>
<th>Tracking Token(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b.scorecardresearch.com</td>
<td>133</td>
<td>UIDR</td>
</tr>
<tr>
<td>ad.doubleclick.net</td>
<td>117</td>
<td>id, __gads</td>
</tr>
<tr>
<td>ib.adnxs.com</td>
<td>75</td>
<td>anj</td>
</tr>
<tr>
<td>p.twitter.com</td>
<td>70</td>
<td>__utma</td>
</tr>
<tr>
<td>cm.g.doubleclick.net</td>
<td>56</td>
<td>id</td>
</tr>
<tr>
<td>ad.yieldmanager.com</td>
<td>52</td>
<td>bx</td>
</tr>
<tr>
<td>bs.serving-sys.com</td>
<td>40</td>
<td>A4</td>
</tr>
<tr>
<td>cdn.api.twitter.com</td>
<td>40</td>
<td>__utmz</td>
</tr>
<tr>
<td>secure-us.imrworldwide.com</td>
<td>38</td>
<td>IMRID</td>
</tr>
<tr>
<td>adfarm.mediasplex.com</td>
<td>31</td>
<td>svid</td>
</tr>
</tbody>
</table>

**Top 10 Tracking Hosts**
Performance

<table>
<thead>
<tr>
<th>Latency Overhead Source</th>
<th>Cost (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Construction</td>
<td>322.36</td>
</tr>
<tr>
<td>Extra IPC</td>
<td>349.06</td>
</tr>
<tr>
<td>Render/JS Engine Instrumentation</td>
<td>139.21</td>
</tr>
</tbody>
</table>

Overall Overhead: ~3% - ~20%

(1). Address Bar Navigation without Principal Avg. Overhead 8.29%
(2). Address Bar Navigation with Principal Avg. Overhead 3.36%
(3). Cross Site Navigation Avg. Overhead 19.43%
(4). Within-site Navigation Avg. Overhead 4.70%
## Memory/Disk Overhead

### Memory Overhead on 12 Web Pages (~25MB/Principal)

<table>
<thead>
<tr>
<th>Memory</th>
<th>Chromium</th>
<th>TrackingFree</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Principal</td>
<td>477.1(MB)</td>
<td>505(MB)</td>
<td>27.9(MB)</td>
</tr>
<tr>
<td>4 Principals</td>
<td>623.6(MB)</td>
<td>702.8(MB)</td>
<td>79.2(MB)</td>
</tr>
<tr>
<td>12 Principals</td>
<td>434.6(MB)</td>
<td>642.5(MB)</td>
<td>297.9(MB)</td>
</tr>
</tbody>
</table>

### Disk Overhead on 12 Web Pages (~0.6MB/Principal)

<table>
<thead>
<tr>
<th>Memory</th>
<th>Chromium</th>
<th>TrackingFree</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Principal</td>
<td>21.3(MB)</td>
<td>21.8(MB)</td>
<td>0.5(MB)</td>
</tr>
<tr>
<td>4 Principals</td>
<td>22.5(MB)</td>
<td>25.9MB</td>
<td>3.4(MB)</td>
</tr>
<tr>
<td>12 Principals</td>
<td>23.7(MB)</td>
<td>29.4(MB)</td>
<td>5.7(MB)</td>
</tr>
</tbody>
</table>
Compatibility

- Manually tested TrackingFree’s compatibility on Alexa Top 50 websites
- Compatibility on first-party websites
  - Results: 50/50
- Compatibility on third-party services
  - Cross-site online payments (1/1)
  - Cross-site content sharing (31/31)
  - Single sign-on (35/36)
- Overall results: 67/68
Case study: Logging Yahoo using Facebook Account

Client-side

Server-side

1. Click login using FB
2. Data(Y) carried on URL
3. Data(Y)
4. Data(F)
5. User login FB
6. Data(F) carried on URL
7. Data(F)
8. Data(F)
Summary

We designed and implemented TrackingFree browser that completely protect users from third-party web tracking by isolating resources in different principals.

We theoretically and experimentally proved TrackingFree’s anti-tracking capability.

TrackingFree incurs affordable overhead and compatibility cost.