Privacy Seminar - Obfuscation

Presenters:
Dor Alter (s1027021)
Michel de Boer (s1011542)
Mustafa Ekici (s4549449)
Bart Hofman (s1018982)
Paolo Scattolin (s1023775)

Date: April 22, 2022

Table of Contents

Introduction

Tracking and online monitoring

Obfuscation methods

Is obfuscation justified?

Why is obfuscation necessary?

Would it work?

Conclusions

Literal definition

To throw into shadow
or
To make obscure

Practical definition

Introduction

Tracking and online monitoring

Obfuscation methods

Is obfuscation justified?

Why is obfuscation necessary?

Would it work?

Conclusions

Cookies

- Thought up in 1994
- Originally used to make a site stateful
- A small bit of data saved in the browser

Figure 1: Lou Montulli

Third-party cookies

- User for de-anonymising traffic
- Link browsing behaviour to a real person
- Mainly used by advertisers to **track people across the web**

Example case: coolblue.nl

- bing.com
- linkedin.com
- pinterest.com
- clarity.ms
- ...

Third-party cookies

- User for de-anonymising traffic
- Link browsing behaviour to a real person
- Also used by governments to **track people across the web (and world)**

Third-party cookies

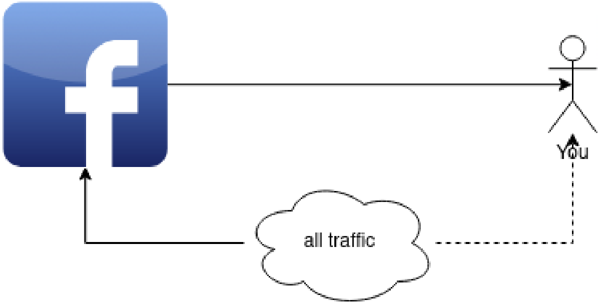


Figure 2: Cookie piggybacking [1]

Cookies that give you away

Can we use third party cookies to associate traffic to real people?

Attacker model

- "No SSL stripping allowed" - 2014 (only 30% of the web was https)
  - Allow our attacker model to mitm
- assuming attacker has access to major web traffic junctions (think of an ISP)
- The attack is fully passive
- IP addresses of the target can change frequently

Cookies that give you away

- Long lasting cookies
- non changing cookies
- User specific cookies
- cookies we can use to create a user "profile"

Cookies that give you away

Good way to find these cookies?

run two browser simulations at the same time and check what cookies are unique

Matching traffic to users

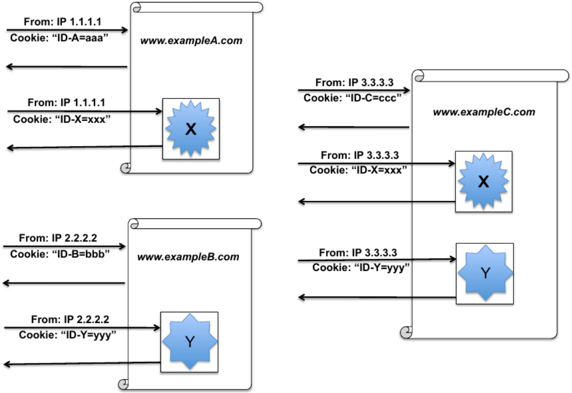


Figure 3: Associating traffic [1]

Blockers

Figure 4: Ublock origin

Tracking pixels

- Simple transparent image object
- No actual use ... besides tracking
- Harder to block (but not impossible)
- quick visit to bol.com: (?) tracking pixels

Tracking pixels

- Simple transparent image object
- No actual use ... besides tracking
- Harder to block (but not impossible)
- quick visit to bol.com: 2 tracking pixels
- product data sent?

Tracking pixels

- Simple transparent image object
- No actual use ... besides tracking
- Harder to block (but not impossible)
- quick visit to bol.com: 2 tracking pixels
- product data sent: 121 parameters sent....

FLoC

- Anonymisation by grouping
- Everyone hates it because no one got asked

Hardware tracking

- DrawnApart [2]
- Browser render times
- Nearly impossible to block and cross browser
- Sadly shows great promise

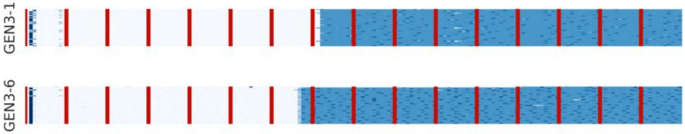


Figure 5: Raw traces from 2 gpus [2]

Introduction

Tracking and online monitoring

Obfuscation methods

Is obfuscation justified?

Why is obfuscation necessary?

Would it work?

Conclusions

Defining the scope

- Many different methods, each designed for specific purposes
- We are going to look at 2 examples of techniques
- CacheCloak and TrackMeNot

CacheCloak

- False echoes and imitations to achieve obfuscation
- Privacy vs functionality trade-off
- k-anonymity

CacheCloak: Goals

- Best features of all existing solutions to the trade off
- Relevant data can still be extracted by the user
- Not temporary time-buying strategy

CacheCloak: Solution

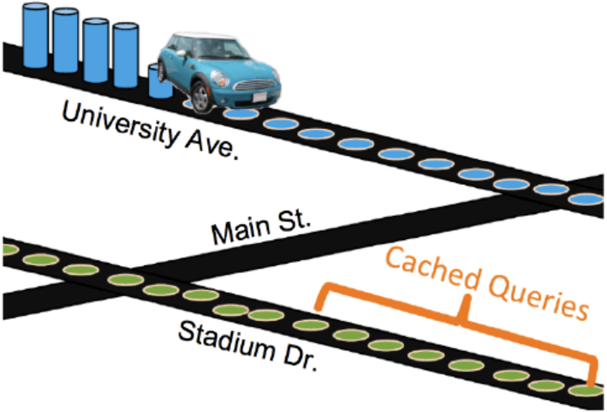
- It generates mobility predictions from historical data ..
- .. and submits intersecting predicted paths.
- Each new predicted path is made to intersect with other users' paths.
- Users retrieve cached query responses for successive new locations from the trusted server.
- New prediction only when no cached response is available for their current locations.

CacheCloak: Solution

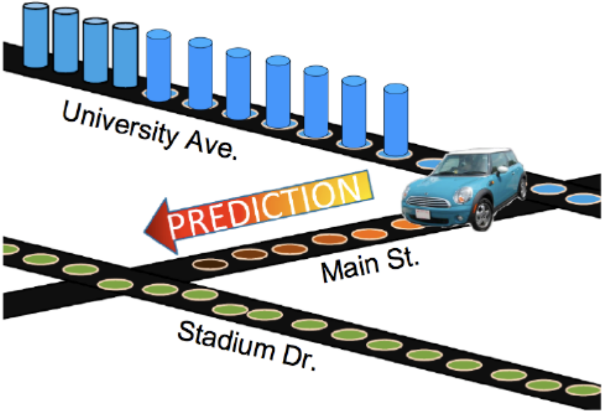
1 Cache Hit
  - Cachecloak has cached data on this location, return cached location

2 Cache Miss
  - Cachecloak doesn't have cached data on this location
  - Calculate predicted path with that location a known cached location
  - query LSB for all locations in that path

CacheCloak: Example



CacheCloak: Example



CacheCloak: Distribution

- Trust needed in CacheCloak server
- Can be solved by doing the caching locallly
- Using the CacheCloak server as "lookup proxy"

CacheCloak: Conclusion

- No spatial degrading
- Minimal time delay
- The more users, the more privacy

TrackMeNot

- Hiding the user in the crowd (that you create)
- Benefit of personalization against the risk of sensitive information disclosure
- Best of both words?

TrackMeNot: The Problem

- Provide personalized searches in a privacy friendly way
- It is challenging!
- A lot of data that can (and will) be misused

TrackMeNot: The Problem

- Many queries are too ambiguous
- Reformulation techniques are used
- Mostly based on previous search history
- Search logs!

Previous attempts

- By search engines themselves
- Tor
- Private information retrieval techniques (PIT)
- Obfuscation of search queries

TrackMeNot: Goals

- Extension for Chrome and Firefox, available for free
- Simplicity by design
- Achieve ‘query indistinguishability’
- Prevent adversaries from using side channels and fingerprints

TrackMeNot: Goals Limitations

- Only aiming to hide the user queries and specific user interests
- Some info can still leak
- Destruction vs devaluation

TrackMeNot: Adversary Model

- TMN queries vs real user ones
- Topic based analysis
- Timing based analysis
- Frequency based analysis

TrackMeNot: Query indistinguishability

- Topic-exposed Query Indistinguishability (TEQI)
- Queries are generated across the same broad topics as the original topics
- Topic-obfuscated Query Indistinguishability (TOQI)
- Query indistinguishability plus a larger set of topics

TEQI: Against frequency analysis

- Frequency of queries across topics
- Frequency of keywords within topics
- Relative popularity of query n-grams
- TMN maintains an obfuscation profile with similar relative frequencies

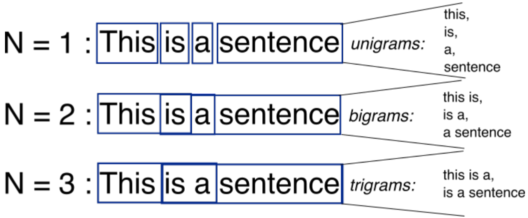


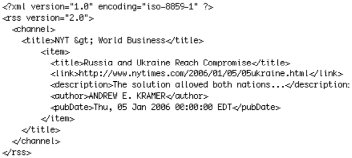
Figure 6: N-grams example

TEQI: Against timing analysis

- TMN uses user's history to be resilient to timing analysis
- TMN maintains a weekly and daily profiles of timing information of user requests
- Exact replication does expose TMN to timing analysis
- Randomness is introduced
- TMN also tries to keep track of active periods

TOQI: Against topic analysis

- Hiding the user topics into a larger set
- Infer the user topic interests as well as construct a list of obfuscation topics
- Selection from the universe of topics U
- The user can then chose among the topics selected
- RSS feeds are retrieved from RSS seach engines or Twitter
- Keywords are extracted (mainly titles and capital letters)



```
<?xml version="1.0" encoding="iso-8859-1" ?>
<rss version="2.0">
  <channel>
    <title>NYT &gt; World Business</title>
        <item>
          <title>Russia and Ukraine Reach Compromise</title>
          <link>http://www.nytimes.com/2006/01/05/05ukraine.html</link>
          <description>The solution allowed both nations...</description
          <author>ANDREW E. KRAMER</author>
          <pubDate>Thu, 05 Jan 2006 00:00:00 EDT</pubDate>
        </item>
    </title>
  </channel>
</rss>
```

Figure 7: RSS example

TOQI: topic analysis limitations

- The obfuscation topic and the actual topic have fundamentally different semantics
- TMN needs to choose at least one obfuscation topic to hide a real user topic
- Only against a simplistic query analysis model

TrackMeNot: Side Channel Attacks

- Query scheduling
- HTTP Header
- External element downloading
- Favicon
- Active Content Handling
- Query suggestions
- Click Stream

Side Channel Attacks: Query scheduling

- No queries if the browser is closed
- Query Bursts
- Search patterns

Side Channel Attacks: HTTP Header

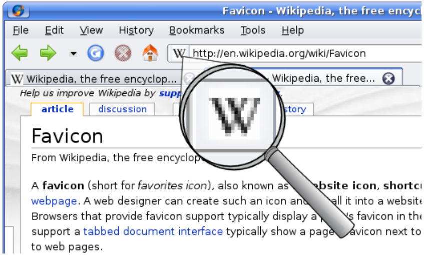
- Miscomputed Cookie and User Agent headers can flag TMN queries
- TMN uses Regular Expression to catch the search URL
- The referrer is never set according to the websites that a user visits

Side Channel Attacks: External element downloading

- TMN loads the search result page in a collapsed browser element
- Same fingerprints as the user's queries
- Search engines cannot filter artificial queries

Side Channel Attacks: Favicon

- Is displayed along the title of the page
- Is downloaded by TMN when the page is rendered
- If present in the cache it is not downloaded


Favicon - Wikipedia, the free encyc
File
Edit
View
History
Bookmarks
Tools
Help
http://en.wikipedia.org/wiki/Favicon
W Wikipedia, the free encyclop...
- Wikipedia, the free...
article
discussion
story
W

Help us improve Wikipedia by supp

Favicon

From Wikipedia, the free encyclop

A **favicon** (short for *favorites icon*), also known as **bsite icon, shortc** webpage. A web designer can create such an icon an ll it into a website Browsers that provide favicon support typically display a s favicon in the support a tabbed document interface typically show a page avicon next to to web pages.

Figure 8: Favicon example

Side Channel Attacks: Active Content Handling

- Active content is supported on the search result page
- Malicious JavaScript?
- Personalized depending on the search engine

Side Channel Attacks: Query suggestions

- TMN mimics interactions with search engine interfaces
- TMN simulates every DOM event that is monitored on the search engine web page
- When TMN simulates keystrokes in the search box, query suggestions are requested to the search engine

Side Channel Attacks: Click Stream

- TMN sometimes follows search result links
- but never actually downloads content from clicked URLs
- Clicks on sponsored links are prevented

Introduction

Tracking and online monitoring

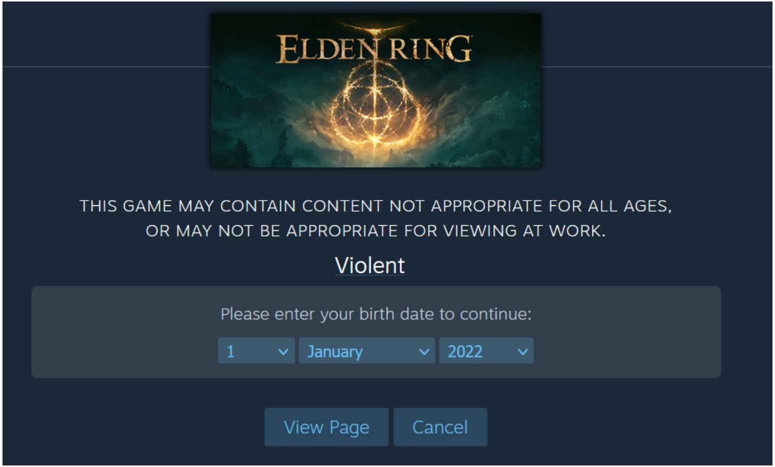
Obfuscation methods

Is obfuscation justified?

Why is obfuscation necessary?

Would it work?

Conclusions



THIS GAME MAY CONTAIN CONTENT NOT APPROPRIATE FOR ALL AGES, OR MAY NOT BE APPROPRIATE FOR VIEWING AT WORK.

Violent

Please enter your birth date to continue:

1 January 2022

Figure 9: Steam enter date of birth

Dishonesty of obfuscation

- How is the data processed?
- What data is collected?
- GDPR Art. 12 and 13
- Facebook

Figure 10: Image from Wikipedia

Philosophy: Immanuel Kant

- Merely mean to achieve goals
- Lying always wrong
- Acts morally good independent on outcome
- Responsible for consequence of lie
- Nuance

Kant, Immanuel, On a supposed right to lie because of philanthropic concerns.

Philosophy: John Stuart Mill

- Maximize happiness
- Protect privacy for many

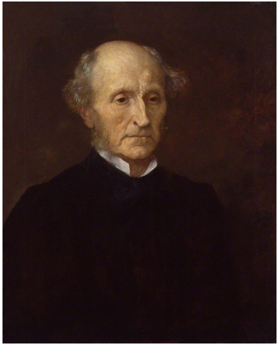


Figure 11: Image from Wikipedia

Free riding

- Benefits
- Without giving up privacy

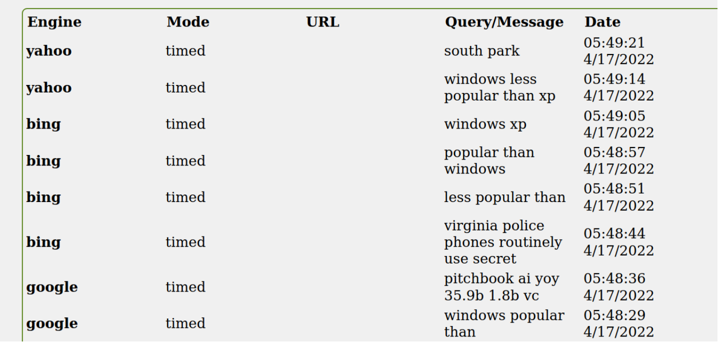
Waste and pollution

- Unnecessary usage
- Pollution

Privacy Enhancing Technologies

- TrackMeNot
- DuckDuckGo
- TOR
- AdNauseam

TrackMeNot



Engine	Mode	URL	Query/Message	Date
yahoo	timed		south park	05:49:21 4/17/2022
yahoo	timed		windows less popular than xp	05:49:14 4/17/2022
bing	timed		windows xp	05:49:05 4/17/2022
bing	timed		popular than windows	05:48:57 4/17/2022
bing	timed		less popular than	05:48:51 4/17/2022
bing	timed		virginia police phones routinely use secret	05:48:44 4/17/2022
google	timed		pitchbook ai yoy 35.9b 1.8b vc	05:48:36 4/17/2022
google	timed		windows popular than	05:48:29 4/17/2022

Figure 12: TrackMeNot

TrackMeNot

- Honesty?
- Free riding?
- Waste?
- ?

Figure 13: TrackMeNot

TOR

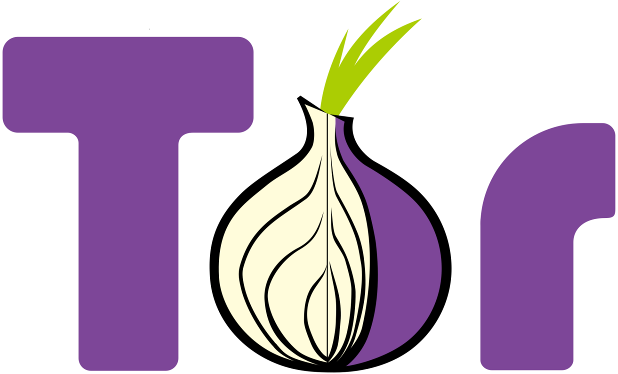


Figure 14: TOR

AdNauseam

- Built on top of uBlock origin
- Clicks all ads
- Discontent with ads, a protest

Figure 15: AdNauseam

Summary: justification of obfuscation

- Dishonest
- Free riding
- Waste and pollution

Summary: justification of obfuscation

- Dishonest
- Free riding
- Waste and pollution

However... Why should we use obfuscation?

Introduction

Tracking and online monitoring

Obfuscation methods

Is obfuscation justified?

Why is obfuscation necessary?

Would it work?

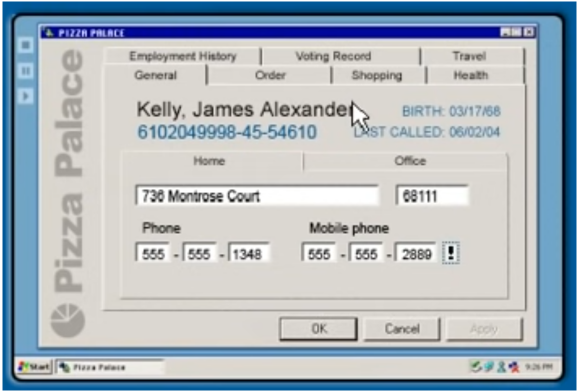
Conclusions

"I've got nothing to hide."

"Arguing that you don’t care about the right to privacy because you have nothing to hide is no different than saying you don’t care about free speech because you have nothing to say."

Edward Snowden

A world with no privacy



Kelly, James Alexander

6102049998-45-54610

BIRTH: 03/17/68

LAST CALLED: 06/02/04

736 Montrose Court

68111

Phone 555 - 555 - 1348

Mobile phone 555 - 555 - 2889 !

https://www.youtube.com/watch?v=RNJl9EEcsoE

Information asymmetry

Information asymmetry

- Data collection about user (by default)
- User no choice regarding collection
- Hence, asymmetrical relationship

Opting out

Opting out

- Just opt out

Opting out

- Just opt out
- ... right?

Opting out

- Just opt out
- ... right?
- Easier said than done

Responsibility

- User
- Businesses
- Government

Responsibility

- ~~User~~
- Businesses
- Government

Businesses

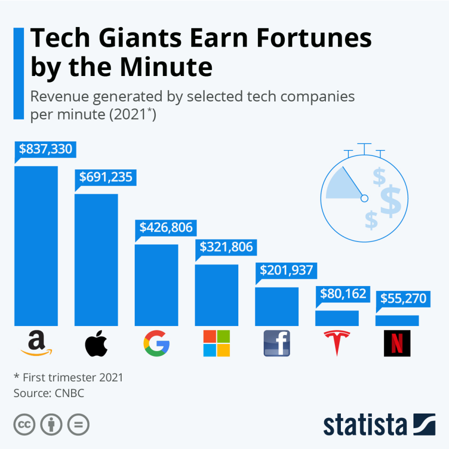


Figure 16: Earnings of Big Tech companies in first trimester of 2021

Businesses

- Gathering, bundling and selling individual data (DoubleClick, Acxiom)
- Customer data to improve its operations (Amazon, Wal-Mart)
- User-data-driven advertising revenue (Google)

Businesses

- Gathering, bundling and selling individual data (DoubleClick, Acxiom)
- Customer data to improve its operations (Amazon, Wal-Mart)
- User-data-driven advertising revenue (Google)

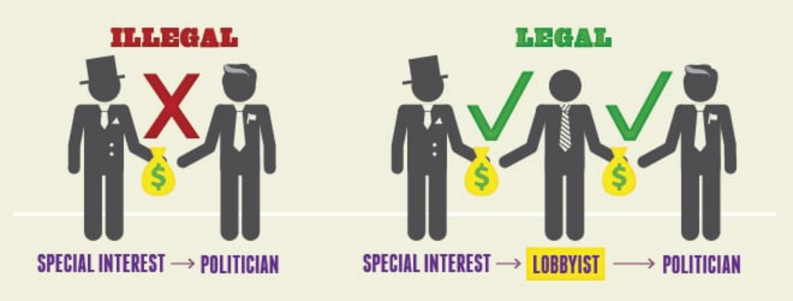
It is not in the company's interest to support restraints on access to this information.

Government

- Balancing interests
- Protecting values and principles
- Defining and enforcing data-collection and data-management practices for businesses (?)

Government - problems

- Legislation operates relatively slow
- Opposition slows it down even further (corporations, other institutes, sometimes the government itself!)


ILLEGAL
SPECIAL INTEREST → POLITICIAN
LEGAL
SPECIAL INTEREST → LOBBYIST → POLITICIAN

Privacy is served by ...

- Improved opting out systems
- Disclosure limits imposed by organizational best practices
- Constraints of law and regulation

Vulnerability remains

- Obfuscation, a troublemaker strategy
- An additional layer of cover

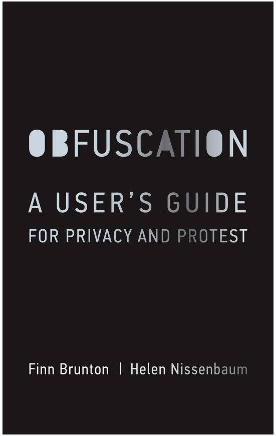


Figure 17: The book on obfuscation by Finn Brunton and Helen Nissenbaum

"[Obfuscation] can be used for data disobedience under difficult circumstances and as a digital weapon for the informationally weak." [3]

Introduction

Tracking and online monitoring

Obfuscation methods

Is obfuscation justified?

Why is obfuscation necessary?

Would it work?

Conclusions

Obfuscation in Nature

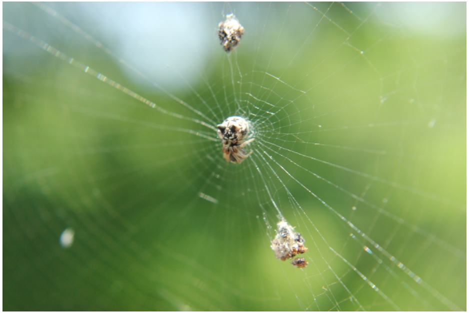


Figure 18: Image of the webs and decays of the spider taken from [4]

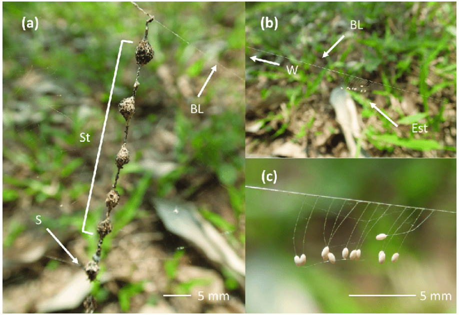


Figure 19: Image of the decays of the spider taken from [5]

Hiding from the internet

- Living like a cage man?
- WhatsApp vs Signal and Telegram

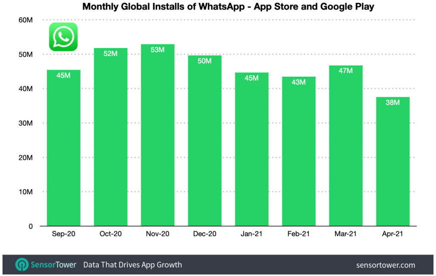


Figure 20: Installation of WhatsApp during 20210-2021 taken from [6]

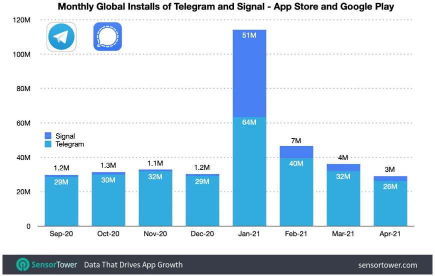


Figure 21: Installation of Telegram and Signal during 20210-2021 taken from [6]

To buy time

- We do not need all the time just enough time
- The safe is as good as the security around it

Figure 22: Buying time [7]

Figure 23: Safe as the system around it [8]

To provide cover

- Hide your identity in a pool of others
- If there is enough noise in the data, the data is useless
- Most often it results in the first goal of buying time

Figure 24: Cover your identity taken from [9]

For deniability

- Hide the link between the user and action
- Linkability is a key concept in privacy
- Tor

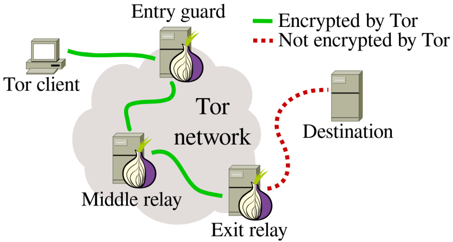


Figure 25: Tor taken from [10]

To hide individual's identity in a database

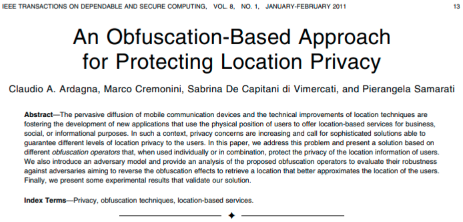
- Privacy-friendly storage of data
- Tracker's action
- In reality we have laws to enforce this

To express protest

- Unlike others this is an aggressive action
- Intentionally inserting wrong information into the collected data.
- Wrong data or conclusions make by the tracker.

Yes but it depends ...

- What is the goal?
- What are the capabilities of the tracker/attacker?
- Probably enough for most real life cases.



IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 8, NO. 1, JANUARY-FEBRUARY 2011 13

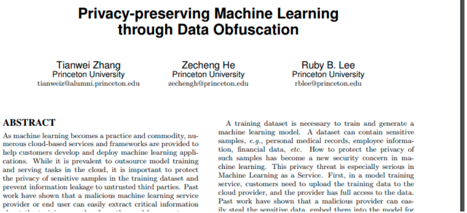
An Obfuscation-Based Approach for Protecting Location Privacy

Claudio A. Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, and Pierangela Samarati

Abstract—The pervasive diffusion of mobile communication devices and the technical improvements of location techniques are fostering the development of new applications that use the physical position of users to offer location-based services for business, social, or informational purposes. In such a context, privacy concerns are increasing and call for sophisticated solutions able to guarantee different levels of location privacy to the users. In this paper, we address this problem and present a solution based on different *obfuscation operators* that, when used individually or in combination, protect the privacy of the location information of users. We also introduce an adversary model and provide an analysis of the proposed obfuscation operators to evaluate their robustness against adversaries aiming to reverse the obfuscation effects to retrieve a location that better approximates the location of the users. Finally, we present some experimental results that validate our solution.

Index Terms—Privacy, obfuscation techniques, location-based services.

Figure 26: Obfuscation for locations [11]



Privacy-preserving Machine Learning through Data Obfuscation

Tianwei Zhang
Princeton University
tianweiz@alumni.princeton.edu

Zecheng He
Princeton University
zechengh@princeton.edu

Ruby B. Lee
Princeton University
rblee@princeton.edu

ABSTRACT

As machine learning becomes a practice and commodity, numerous cloud-based services and frameworks are provided to help customers develop and deploy machine learning applications. While it is prevalent to outsource model training and serving tasks in the cloud, it is important to protect the privacy of sensitive samples in the training dataset and prevent information leakage to untrusted third parties. Past work have shown that a malicious machine learning service provider or end user can easily extract critical information

A training dataset is necessary to train and generate a machine learning model. A dataset can contain sensitive samples, *e.g.*, personal medical records, employee information, financial data, *etc.* How to protect the privacy of such samples has become a new security concern in machine learning. This privacy threat is especially serious in Machine Learning as a Service. First, in a model training service, customers need to upload the training data to the cloud provider, and the provider has full access to the data. Past work have shown that a malicious provider can easily steal the sensitive data, embed them into the model for

Figure 27: Obfuscation for machine learning [12]

Small use case - slang

- The trackers can understand as much as we let them to
- Encryption is key but it is hard in real life
- Parent who do not understand the slang and abbreviation used by their kids
- Event related words that "you need to be there to understand"
- These words are easy to use and might be hard to decode

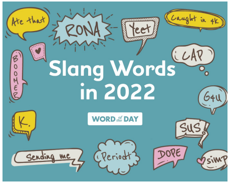


Figure 28: Example of slang from 2022 [13]

Figure 29: Old man getting 'good soup' [14]

Conclusions

- Advanced tracker wants your data.
- Obfuscation methods to fight back the trackers.
- Ethics about obfuscation.
- Why do we need it?
- Would it work?
- Relatively new filed with a lot to research and discover.

Thank you for your attention.

Questions ?

References I

- Dillon Reisman, Steven Englehardt, Christian Eubank, Peter Zimmerman, and Arvind Narayanan.
  Cookies that give you away: Evaluating the surveillance implications of web tracking.
  https://www.cs.princeton.edu/~arvindn/publications/cookie-surveillance.pdf.

- Durey A. Laor T., Mehanna N.
  Drawnapart: A device identification technique based on remote gpu fingerprinting.
  2022.

- Finn Brunton and Helen Nissenbaum.
  Obfuscation: A user's guide for privacy and protest.
  Mit Press, 2015.

References II

T. Thorell.
Primo saggio sui ragni birmani.
Annali del Museo Civico di Storia Naturale di Genova 25, pages 5–417, 1887.

Timothy Hawes.
Egg sacs of the orb-weaving genus cyclosa (araneae: Araneidae) targeted by ovipositing lacewings.
Neuroptera: Chrysopidae, pages 83–89, 12 2018.

A. Ahmed.
Signal and telegram have witnessed a rise of 1200% in usage before the implementation of the controversial whatsapp privacy policy.
Digital information world, 2022.

References III

- Nastudio Studio.
  Buying time.hand holding money and hand holding clock or time. exchanging time for money. time is money.
  dreamstime, 2022.
  https://www.dreamstime.com/buying-time-hand-holding-money-clock-exchanging-image179827160.

- M600maxx.
  Two guards protect a safe.
  dreamstime, 2022.
  https://www.dreamstime.com/stock-photography-two-guards-protect-safe-image23863282.

References IV

S. Cox.
Anonymous, hacktivism and the rise of the cyber protester.
BBC, 2012.
https://www.bbc.com/news/technology-20446048.

D. Goodin.
Scientists detect "spoiled onions" trying to sabotage tor privacy network.
Ars Technica, 2014.
https://arstechnica.com/information-technology/2014/01/scientists-detect-spoiled-onions-trying-to-sabotage-tor-privacy-network/.

Claudio A. Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, and Pierangela Samarati.
An obfuscation-based approach for protecting location privacy.
IEEE Transactions on Dependable and Secure Computing, 8(1):13–27, 2011.

References V

Tianwei Zhang, Zecheng He, and Ruby B. Lee.
Privacy-preserving machine learning through data obfuscation, 2018.

Word of the day.
Slang word you will need in 2022.
medium, 2022.
https://medium.com/@wordofthedayapp/slang-word-you-will-need-in-2022-7692ea7d0ce7.

M. Rocky.
i-hear-they-make-good-soup.
memegenerator, 2022.
https://memegenerator.net/instance/57648314/rocky-mickey-i-hear-they-make-good-soup.