Web Tracking

ADVANCED BROADBAND
COMMUNICATIONS CENTER (CCABA)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA (UPC)

Ismael Castell Uroz

Email: icastell@ac.upc.edu

Departament d'Arquitectura de Computadors

Table of contents

Table of contents

Defining Web Tracking

- What is Web Tracking?
  - Cookies

Defining Web Tracking

- What is Web Tracking?
  - Cookies
  - Personalize the web experience

Defining Web Tracking

- What is Web Tracking?
  - Cookies
  - Personalize the web experience
  - Is that all?

Defining Web Tracking

- Cambridge Analytica
  - Company focused on data mining and analysis for electoral purposes

Defining Web Tracking

- Cambridge Analytica
  - Company focused on data mining and analysis for electoral purposes
  - Took part in Donald Trump's presidential campaign as well as for Leave.EU

Defining Web Tracking

- Cambridge Analytica
  - Company focused on data mining and analysis for electoral purposes
  - Took part in Donald Trump's presidential campaign as well as for Leave.EU

  - Acquired and used personal data from millions of Facebook users without consent

Defining Web Tracking

Iglesias reclama a Sánchez una "negociación integral" y lanza la consulta a sus bases

Podemos consultará desde este viernes al jueves que viene a sus bases qué prefieren: "Llegar a un acuerdo íntegro para un Gobierno de coalición" o "un Gobierno diseñado únicamente por el PSOE"

650

JOSÉ MARCOS

Madrid - 12 JUL 2019 - 12:40 CEST

Pedro Sánchez y Pablo Iglesias, en su última reunión en el Congreso. En vídeo, declaraciones de la portavoz de Podemos Noelia Vera. ULY MARTÍN | ATLAS

Pablo Iglesias insiste en una negociación "integral" de cargos en el Gobierno, programa y presupuestos. El secretario general de Podemos rechaza la última propuesta de Pedro Sánchez, que este jueves se abrió a la presencia de ministros de su socio preferente de perfil técnico pero no político. Una condición que

Defining Web Tracking

- Third Party Trackers

Defining Web Tracking

- Third Party Trackers

Defining Web Tracking

- Third Party Trackers

Defining Web Tracking

- Google is present in more than 50% of the websites
  - Google Analytics
- Between 50%~70% of the total Internet has tracker systems

Table of contents

Purposes and implications

1) User-Oriented Search

- Personalization of the search based on user's interests
- Also called *Filter Bubble*
- An algorithm selectively guesses what information the user would like to see
- In theory this might be interesting
  - *The user receives results as per his/her likings*
- In practice it results in *user isolation*
  - *Does not receive points of view, topics or events that are not in line with the previous activity*

Purposes and implications

2) Online Advertising

- Facilitate marketing and increase sales profit
- Behavioral tracking, audience segmentation and targeting
- Gmail scans sent, received and stored e-mails for trends to be used in targeted advertising

- AdStack permits sending marketing e-mail whose advertisement is selected in real time

Purposes and implications

3) Web Analytics and Usability Tests

- Usually only used inside own website (not third party tracker)
- Improve the website
- Not a threat for the user
- Record and playback cursor movements, positions and timing
- Some “malicious” usages
  - *Pass them to third party trackers*
  - *Scan the selected text before copying to clipboard*

Purposes and implications

4) Assessing Financial Credibility

- Lenddo: Facebook friends are late with their loan payments

- Kreditech: Use Facebook, eBay and Amazon accounts, and the location to decide the credibility

- Kabbage: Looks PayPal, eBay and other payment accounts. Creditworthiness improvement if linking Facebook and Twitter

Purposes and implications

5) Price Discrimination

- Price depends on:
  - *Geographical location*
  - *Affluence of the user*
  - *The referrer*
- Interest rates of the credit cards depending on ZIP code and date of birth
- Hotel offers differ for Mac and PC
  - *Mac users get more expensive hotels ads*

- Car rental price depends on user identity
  - *Profile from work was cheaper than profile from home*
  - *Both of them were taken at the same time*

Purposes and implications

6) Determining Insurance Coverage

- Lifestyle, interests, habits and hobbies

- Infer by:
  - *Product warranties*
  - *Consumer surveys*
  - *Magazine subscriptions,*
  - *Credit card spending*
  - *Social networks*
  - *Online articles read*

Purposes and implications

7) Impact on the job Market

- Background check prior employment

- Often data is of bad quality, outdated or confusing
  - *Same name for different person*

Purposes and implications

8) Government Surveillance

- Between January and June 2014, the U.S. government made 12.539 requests for 21.576 person's information from Google including search history
- Google complied with 84%
- Use of cookies to distinguish between flows from different users in the same connection
- Cookie unicity can be used to track user change of location during time
- Edward Snowden - NSA

Purposes and implications

9) Identity Theft

- Data revealed by people on the Internet when combined are in many cases enough to predict social security number (broadly used as ID in the USA)

- LinkedIn and Facebook users are more likely to become victims of a fraud
  - *They share the date of birth, school name, and other relevant data*

Table of contents

Tracking Mechanisms

- Classified by type
  - Session-only
  - Storage-based
  - Cache-based
  - Fingerprinting
  - Others

Tracking Mechanisms

- Session-only
  - Non-persistent tracking method
  - Session identifiers stored in hidden fields
    - *Use GET or POST in the URL fields to pass identifier*
  - Explicit web-form authentication

  - window.name DOM property
    - *Represents all web elements into a tree structure*
    - *window.name can store until 2MB of data as a single string*
    - *Resistant to page reloads*
    - *Accessible by other domains*
    - *Stringify JSON structure*

Tracking Mechanisms

- Storage-based
  - Persistent tracking mechanism
  - HTTP cookies
    - *Session cookies*
    - *Persistent cookies*
    - *Lightweight, fast and completely transparent*
    - *The probability of being a tracker cookie increase with the size and the expiration period of the cookie*

  - Flash cookies and Java JNLP PersistenceService
  - Flash LocalConnection Object
  - Silverlight Isolated Storage
  - HTML5 Global, Local, and Session Storage
  - Web SQL Database and HTML5 IndexedDB
  - Internet Explorer userData Storage

Tracking Mechanisms

- Cache-based
  - Use non-explicit storage
    - *Exploit possibilities to identify browsers and determine the previously visited websites*
  - Web Cache
    - *Prior 2010 it could be predicted using link colors*
    - *Embed identifiers in cached documents*
    - *Loading performance tests*
  - DNS Cache
    - *Use JavaScript to indirectly case a DNS lookup and measure its time*
  - Operational Caches
    - *HTTP Redirect, HTTP Authentication, etc…*

Tracking Mechanisms

- Fingerprinting
  - The most complex of all of them
  - Uses a combination of technologies to create a profile of the user using identifiers for things like:
    - *Device used*
    - *Network parameters*
    - *OS*
    - *Browser version*
    - *Browser instance,*
    - *More…*
  - Do not uses cookies
  - Completely transparent for the user

Tracking Mechanisms

- Fingerprinting

Network and location	IP address, user's country, city, neighborhood
Device	Device id, Ip address, operative system, screen resolution, timezone, list of system fonts, web browser, information about hardware (mouse, keyboard, accelerometer, multitouch capability, microphone, camera), TCP timestamps
OS	OS instance id, OS versión, OS architecture, system language, user-specific language, local timezone, local date and time, list of system fonts, color depth, screen dimensions, audio capabilities, camera, microphone, hard disk, printing support, computer name, Inter Explorer id, Windows Digital Product Id, Installed drivers, more.
Browser version	Detailed browser versión
Browser instance (canvas)	Browser instance id
Browser instance (browsing history)	Browser instance id, browsing history
Browser instance (other)	Browser instance id, supported image and media files formats, preferred and accepted languages, insatlled plugins, language, dimensions, flash versión, screen resolution, color depth, timezone, system fonts, IP address, accepted HTTP headers, cookies enabled, supercookies limitations

Tracking Mechanisms

- Others
  - Headers attached to outgoing HTTP requests
  - Using telephone metadata
    - Using call logs it can be determined*:*
      - *the health (including mental)*
      - *the religion*
      - *addictions*

  - Timing attacks
  - Unconscious collaboration of the user
    - *Fake captcha*
  - Clickjacking
  - Evercookies
    - *Auto-recoverable cookies*

Table of contents

Identification of the tracked user

- Associate the digital identity with the real identity
  - 87% of the U.S. population can be unambiguously identified based on 3 attributes:
    - Date of birth
    - Gender
    - ZIP code

Identification of the tracked user

- Associate the digital identity with the real identity
  - Methods:
    - Legitimate first-party services tracking third parties
      - Google and Facebook
    - Leaking information to third parties
      - Setting the e-mail in the *Referrer* of HTTP headers
      - Set sensitive information in the *Request-URI* header
      - More
    - Selling information to third parties
    - Using web hacks
    - Intended deanonymization
      - Use matching data to deanonymize
      - Link multiple accounts with the same username
        - Over 70% of usernames contained the first or the last name
        - Around 30% of the usernames were a concatenation of the first and last names

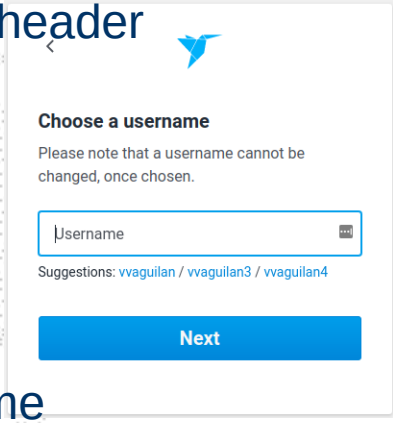
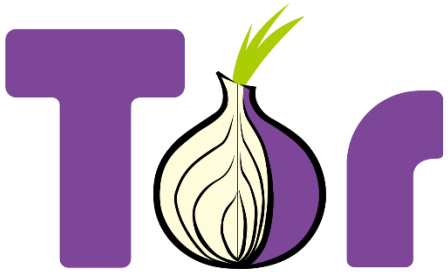


Table of contents

Methods to improve privacy

- Block advertisement services
  - Content-filtering plugins
    - *uBlock Origin*
    - *AdBlock Plus*
    - *Ghostery*
    - *Privacy Badger*
    - *…*



- Hiding the IP address
  - Proxy servers
  - Tor
  - VPN's

Methods to improve privacy

- Modify data sent over the network
  - Privoxy
    - *Modify HTTP headers and message content on the fly*
    - *Block ads, banners, popups,…*
    - *Filter cookies*
- Opt-Out cookies
  - Cookies designed to reject other cookies and advertisement
  - Limited effect
    - *Only work for one domain*
    - *Also expire*
    - *Deleted when the cookies are cleaned*

Methods to improve privacy

- Do Not Track
- Use privacy-focused search engines
  - DuckDuckGo, Startpage, etc…
- Private Browsing Mode
- Clearing the browser cache and history
- Execution blocking
  - NoScript, Flashblock
- E-mail aliases

Table of contents

Research results

- Main problem to fight web tracking:
  - Find new web tracking methods is very hard!
    - *High grade of expertise*
    - *Massive environment*
  - Most current tools only search for already known tracking

Research results

- Content-bloquers:
  - Pros:
    - *Fast; They only look at the URL*
    - *Easy to implement as a browser plugin*
    - *Robust against minification/obfuscation*
  - Cons:
    - *Based on manually-curated pattern lists*
    - *Slow adaptation to new tracking methods*

Research results

- Alternatives:
  - Current methods proposed by the research community:
    - *Mostly based on applying ML algorithms over the website code*
    - *Usually complex and hard to implement inside plugins*
    - *Only detect already known patterns*

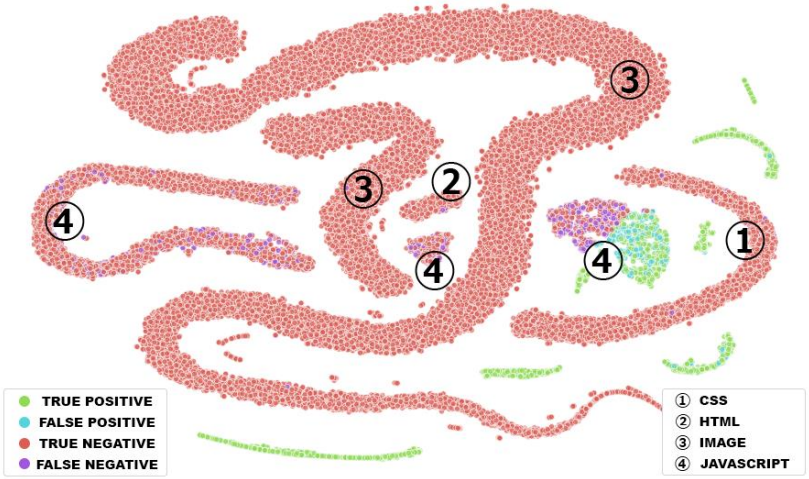
Research results

- Our goals:
  1. Improve blocking tools (content-blockers)
  2. Detect new web tracking code

Research results

1) Improve blocking tools:

- Use ML only over the URL to improve content-blockers:
  - *Deep Neural Network (DNN)*
  - *Automatically study the URL characteristics*
- Results:
  - *Classification accuracy 97%*
  - *Can potentially generalize to new tracking methods*



Research results

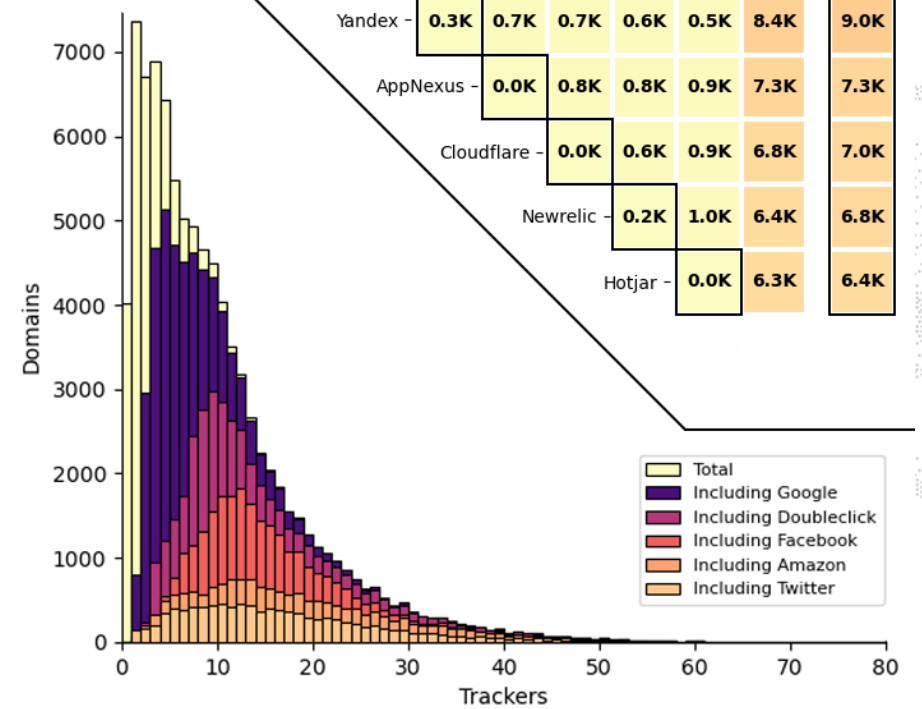
2) Detect new web tracking code:

- Look for code with high probability of being tracking code
- How?
  - *Computing code signatures to find:*
    - *Code present in many different and non-related webs*
    - *Code similar to already known tracking code*

- Results:
  - Scanned top 100K most popular websites
  - Detected a total of 73.641 web trackers (30.000 of them not previously known)
  - 95% of websites include at least one tracker!

Research results

	Google	Doubleclick	Facebook	Amazon	Twitter	Yandex	AppNexus	Cloudflare	Newrelic	Hotjar	Others	Total
Google	1.6K	42.7K	28.1K	15.9K	9.3K	7.3K	7.1K	6.5K	6.0K	6.4K	66.6K	70.5K
Doubleclick		0.0K	20.9K	11.8K	6.8K	4.5K	6.5K	4.3K	4.1K	4.6K	41.5K	43.0K
Facebook			0.0K	9.2K	6.7K	2.9K	3.8K	3.3K	3.3K	4.8K	28.4K	29.0K
Amazon				0.1K	2.7K	1.3K	3.7K	1.9K	2.0K	2.4K	16.1K	16.6K
Twitter					0.2K	0.8K	1.4K	1.1K	1.0K	1.2K	9.6K	10.0K
Yandex						0.3K	0.7K	0.7K	0.6K	0.5K	8.4K	9.0K
AppNexus							0.0K	0.8K	0.8K	0.9K	7.3K	7.3K
Cloudflare								0.0K	0.6K	0.9K	6.8K	7.0K
Newrelic									0.2K	1.0K	6.4K	6.8K
Hotjar										0.0K	6.3K	6.4K



Web Tracking

ADVANCED BROADBAND
COMMUNICATIONS CENTER (CCABA)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA (UPC)

Ismael Castell Uroz

Email: icastell@ac.upc.edu

Departament d'Arquitectura de Computadors

References

T. Bujlow, V. Carela-Español, J. Solé-Pareta, P. Barlet-Ros, “A survey on Web Tracking: Mechanisms, Implications, and Defenses,” Proceedings of the IEEE, Vol. 205, Issue 8, 2014.

J. Mikians, L. Gyarmati, V. erramilli, and N. Laoutaris, “Detecting price and search discrimination on the Internet,” in Proceedings of 11th ACM Workshop Hot Topics on Networking, pages 79-84, Seattle, WA, USA, October 2012

F. Roesner, T. kohno, and D. Wethereall, “Detecting and defending against third-party tracking on the web,” in Proceedings 9th USENIX Conference on Networked Systems Design and implementation, Pages 12-12, San Jose, CA, April 2012.

G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, “The Web never forgets: Persistent tracking mechanisms in the wild,” in Proceedings of 2014 ACM SIGSAC Conference on Cimputer and Communications Security, Pages 674-689, Scottsdale, Arizona, USA, November 2014