A Calculus of Tracking: Theory and Practice

Giorgio Di Tizio, Fabio Massacci

UNIVERSITÀ DI TRENTO

VU VRIJE UNIVERSITEIT AMSTERDAM

Cyber Security for Europe

Analysis of Web Tracking

- As Web Tracking is a ubiquitous activity on the Internet, a variety of tracker-blocking tools has been proposed
- The de-facto approach to evaluate the efficacy of these tools or to determine policy compliance is by mean of large-scale crawling
  - Results are often contradictory and lack transparency
  - Do the users need a Top X million analysis if they only visit few well-known domains?
- Manual inspection is simply impractical

A framework for Web Tracking

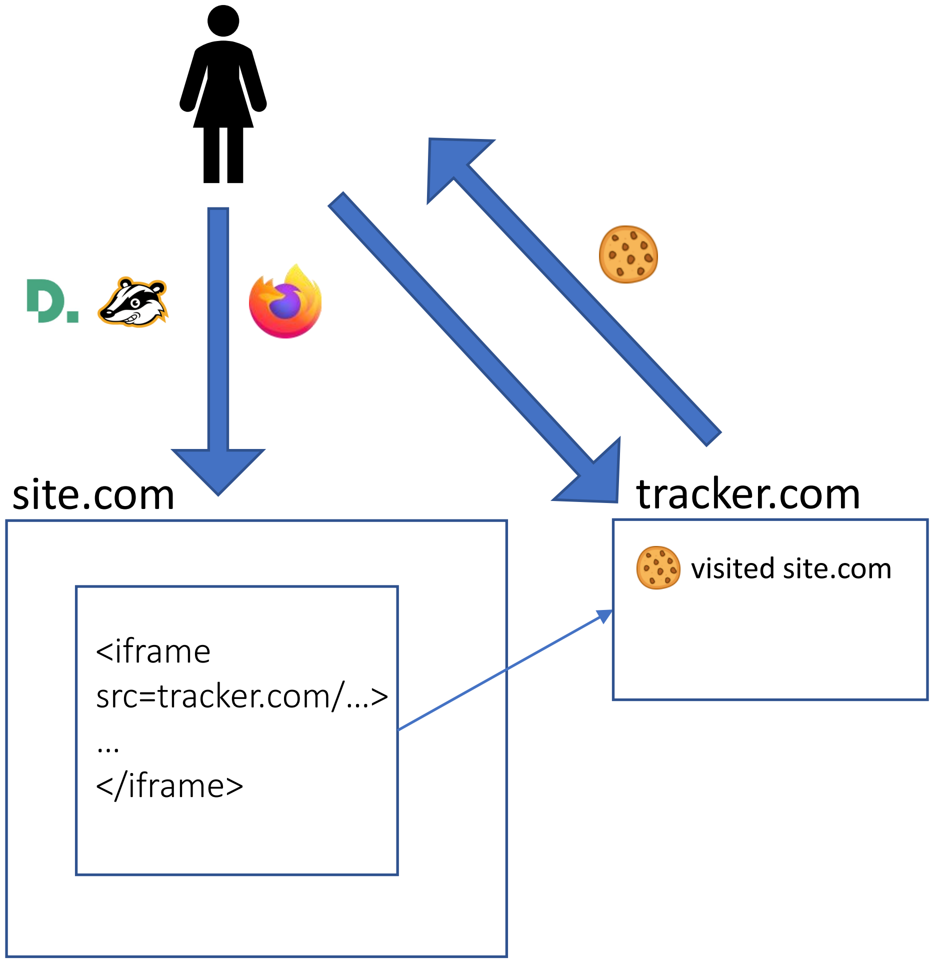
- Technical Contributions:
  - a framework for independent verification of tracking practices
  - based on tracking techniques and data exchange from the client perspective
  - formal rules based on IFOL
  - automated extraction of rules from snapshots of the Internet (OpenWPM)
  - extension to probability

- Demonstrated Applications:
  - Compare trade-off of tracker-blocking extensions
  - Determine potential need for compliance with COPPA

A Formal Model for Web Tracking

- Tracking is decomposed as a sequence of pre- and post-conditions observable as network interactions between websites and the user visiting them.
- Tracker-blocking extensions are modeled as pre-conditions that disable tracking techniques:
  - Block cookies
  - Block connections

Example: Modeling 3rd-*party Tracking*



$$\frac{\mathit{Visit}(\text{site.com}) \quad \dfrac{\dfrac{\mathit{IncludeContent}(\text{site.com,tracker.com})}{\mathit{Link}(\text{site.com,tracker.com})} \quad \sim\mathit{Block\_request}(\text{tracker.com})}{\mathit{Access}(\text{site.com,tracker.com})} \quad \sim\text{Block\_cookie}(\text{tracker.com})}{\mathit{Knows}(\text{tracker.com,site.com})}$$

General Rules for Modeling Web Tracking

Inclusion rule:

$$\frac{\mathit{IncludeContent(w,w')}}{\mathit{Link(w,w')}}$$

Description

If a website *w* includes content from a website *w'*, there is a link that allows an exchange of information

Access rule:

$$\frac{\mathit{Link(w,w')} \qquad \mathit{\sim Block\_request(w')}}{\mathit{Access(w,w')}}$$

Description

If a website *w* has a link with a website *w'* that is not blocked by any tracker-blocking tool, then the user access *w'* from *w*

3rd-party tracking rule

$$\frac{\mathit{Visit(w)} \qquad \mathit{Access(w,w')} \qquad \mathit{\sim Block\_cookie(w')}}{\mathit{Knows(w',w)}}$$

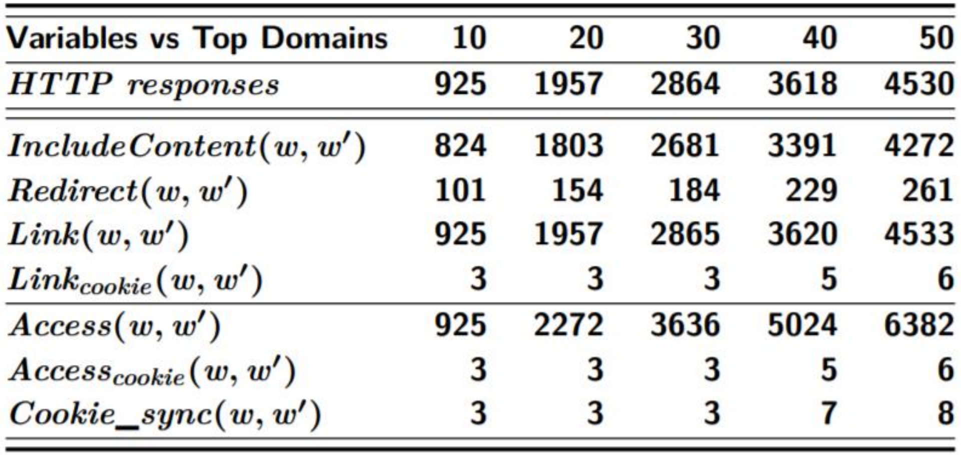
Description

If a user visits a website *w* that forces to access a website *w'* not blocked by any tool, then *w'* knows that the user visited *w*

From Theory to Practice: Predicates instantiation

- The framework automatically instantiates ground predicates from OpenWPM databases
- The remaining predicates are derived by applying the rules in the model

# of instantiated ground predicates for the Top Alexa



Variables vs Top Domains	10	20	30	40	50
$HTTP\ responses$	925	1957	2864	3618	4530
$IncludeContent(w, w')$	824	1803	2681	3391	4272
$Redirect(w, w')$	101	154	184	229	261
$Link(w, w')$	925	1957	2865	3620	4533
$Link_{cookie}(w, w')$	3	3	3	5	6
$Access(w, w')$	925	2272	3636	5024	6382
$Access_{cookie}(w, w')$	3	3	3	5	6
$Cookie\_sync(w, w')$	3	3	3	7	8

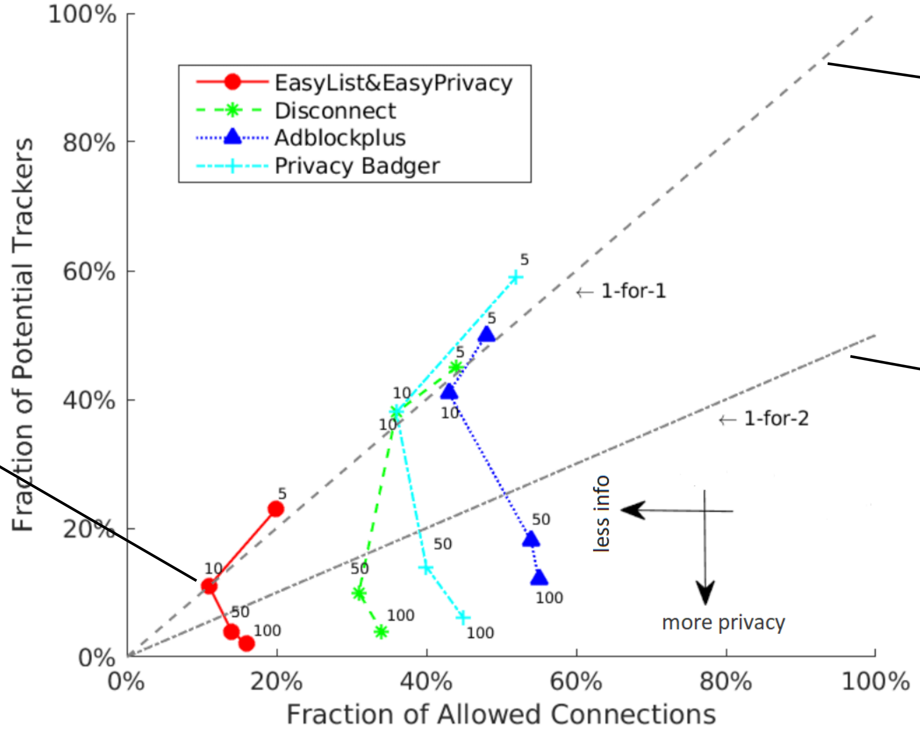
Analysis of Mitigations for Individual Cases

- The *Knows* and *Access* predicates are used to compare the trade-off of different tracker-blocking tools:
  - # unique *Knows* → measures the potential trackers a user can encounter while browsing the Web
  - # unique *Access* → measures the connections established and thus site breakage

- A tool *A* is better than *B* if the # of unique *Access* predicates is greater or equal than B, while the # of unique *Knows* predicate is smaller.

Analysis of Mitigations for Individual Cases

- Different tracker/ad-blocking tools have different trade-offs when visiting 5 to 100 Top Alexa websites. *EasyList&EasyPrivacy* blocks most trackers at the cost of fewer connections



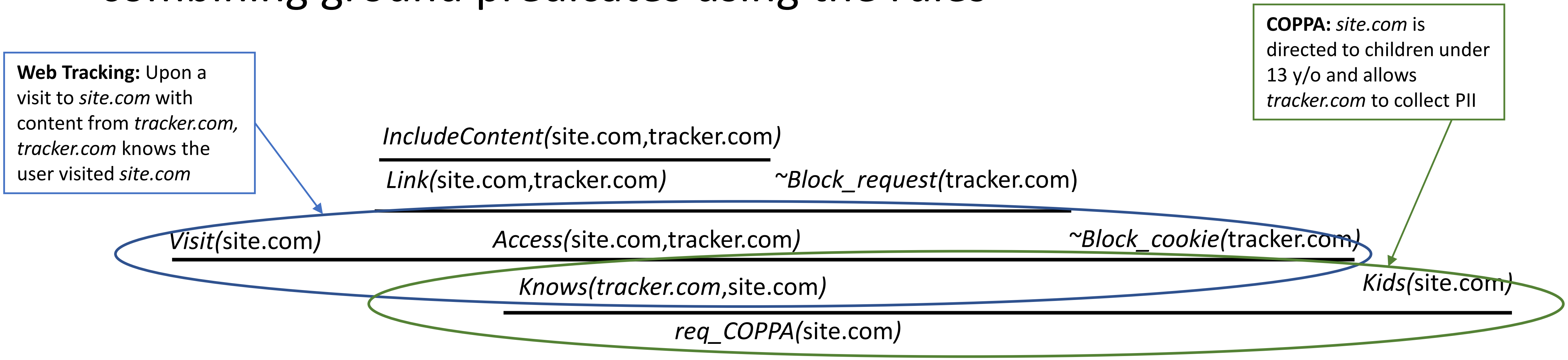
By visiting the Top 10 Alexa with *EasyList&EasyPrivacy* one observes ~10% of the potential trackers and ~10% of the connections compared to the case without any tool

By allowing 50% connections, 50% trackers is allowed

By allowing 50% connections, 25% trackers is allowed

Do you have to respect COPPA?

- The rules are encoded as axioms for the *Slakje* intuitionistic prover
- Given a conjecture: is *req_COPPA(site.com)?*
- The framework produces a proof (if exists) of the conjecture by combining ground predicates using the rules



Web Tracking: Upon a visit to *site.com* with content from *tracker.com, tracker.com* knows the user visited *site.com*

COPPA: *site.com* is directed to children under 13 y/o and allows *tracker.com* to collect PII

*IncludeContent(*site.com,tracker.com*)*

*Link(*site.com,tracker.com*)* *~Block_request(*tracker.com)

*Visit(*site.com*)* *Access(*site.com,tracker.com*)* *~Block_cookie(*tracker.com*)*

*Knows(tracker.com,*site.com*)* *Kids(*site.com*)*

*req_COPPA(*site.com*)*

A Calculus of Tracking: Theory and Practice

- We presented a framework for the analysis of web tracking that fills the gap between large-scale and manual inspection by providing an explanation in the form of a proof
- The framework can be used to:
  - Directly take data from OpenWPM
  - Compare tracker-blocking tools
  - Determine potential need for compliance with COPPA
- Future directions can:
  - Extend the model with more tracking techniques and mitigations e.g. browser fingerprinting