Machine Learning: A Promising Direction for Web Tracking Countermeasures

Jason Bau, Jonathan Mayer, Hristo Paskov and John C. Mitchell
Stanford University

Stanford Computer Security Lab

Motivation

- Consumers want control over third-party online tracking*
- Regulatory agencies (US, Canada, EU) want to empower consumer preference
- Do Not Track

* Detailed definitions of “third party” and “tracking” are hotly contested. For purposes of this presentation, we mean simply unaffiliated websites and the collection of a user’s browsing history.

Motivation

All internet users were posed the following choice regarding targeted advertising:

68% say… I'm NOT OKAY with targeted advertising because I don't like having my online behavior tracked and analyzed

28% say… I'm OKAY with targeted advertising because it means I see advertisements and get information about things I'm really interested in

Source: http://pewinternet.org/~/media//Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf

Do Not Track

- Central technology discussed for standardization
- HTTP header (`DNT: 1`) sent by browser
- Voluntary observation by industry sites receiving header
- Stalled at W3C standardization
  - Limitations enforced when enabled
  - Defaults

Do Not Track

"It will be dead in a couple of weeks You don't have to worry about that." – Tracking Industry CEO

http://www.mediapost.com/publications/article/201052/evidon-w3cs-effort-to-forge-do-not-track-agreeme.html#ixzz2UAy68HOz

Do Not Track Framework Doc Stirs Controversy Ahead of W3C Meeting

Document was not a DAA proposal, W3C co-chair explains By Katy Bachman

April 30, 2013, 1:12 PM EDT | Technology

Bloomberg BNA

Login to Your Bloomberg BN

LEGAL & BUSINESS | TAX & ACCOUNTING | ENVIRONMENT, HEALTH & SAFE

Home » News » Legal & Business News » FTC's Ramirez Says W3C Efforts Are Last Hope for Do Not Track Tool

Latest News

Legal & Business

FTC's Ramirez Says W3C Efforts Are Last Hope for Do Not Track Tool

Monday, April 22, 2013

from Privacy & Data Security Law Resource Center™

ELECTRONIC FRONTIER FOUNDATION
DEFENDING YOUR RIGHTS

HOME | ABOUT | OUR WORK

Digital Civil Rights in Europe

Home » EDRi-gram newsletter - Number 11.9, 8 May 2013

ENDitorial: Last Call For The W3C Do Not Track Process

8 May, 2013 » Privacy

This article is also available in:
Deutsch: ENDitorial: Letzte Chance für die Do-Not-Track-Initiative des W3C

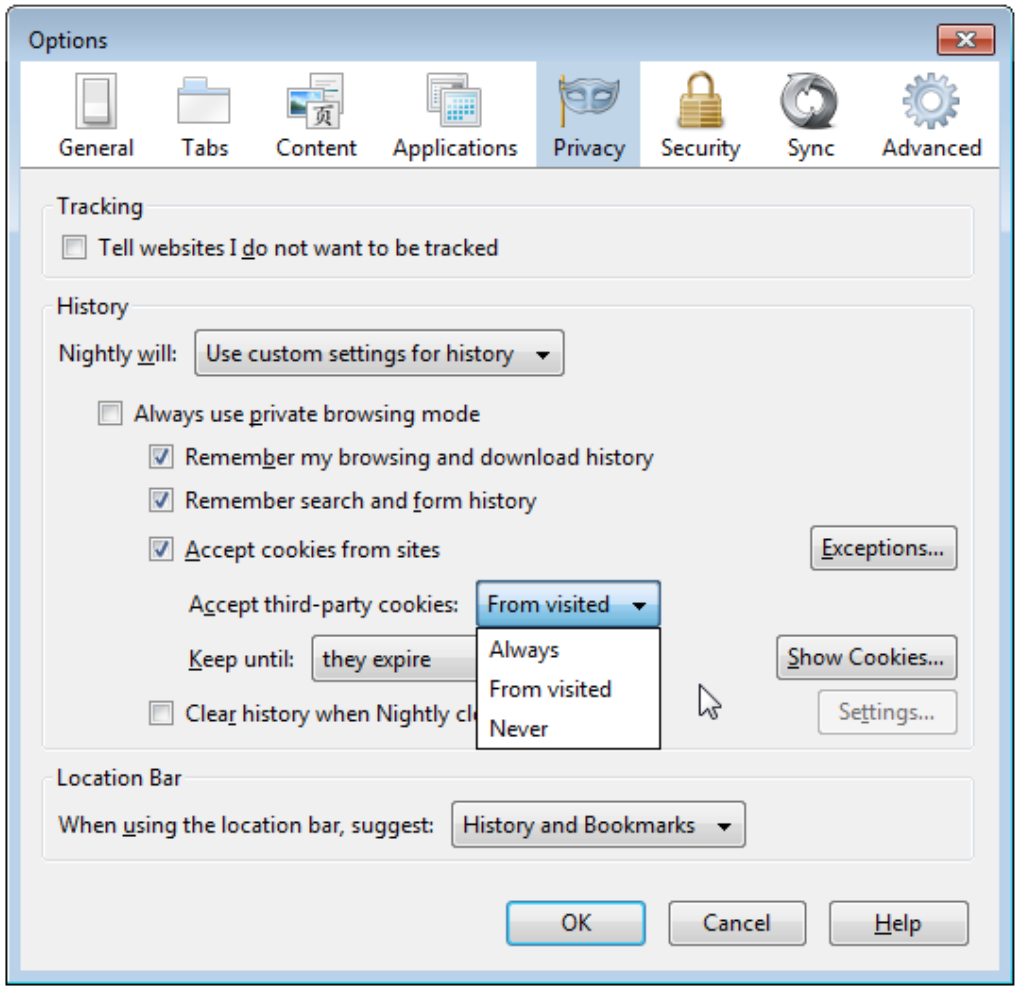
MAY 14, 2013 | BY DAN AUERBACH

Do Not Track: Are Weak Protections Worse Than None At All?

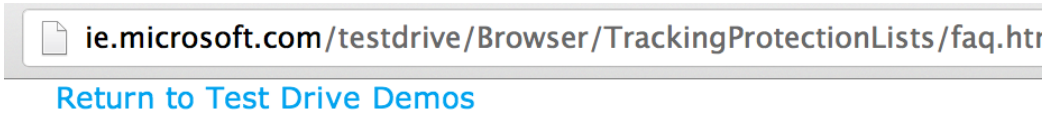
Renewed Interest in Technical Solns

Examples:

Firefox new third party cookie policy



IE Tracking Protection Lists



Technical Solution Considerations

- Usability (in-browser)
- Collateral impact (false positive rate)
- Distance Human expert judgment
  - Singling out individual or groups of entities
  - Maintainbility
- Objective standards and confidence measures
  - Possibly tied into different grades of countermeasure (e.g. blocking cookies vs blocking HTTP)

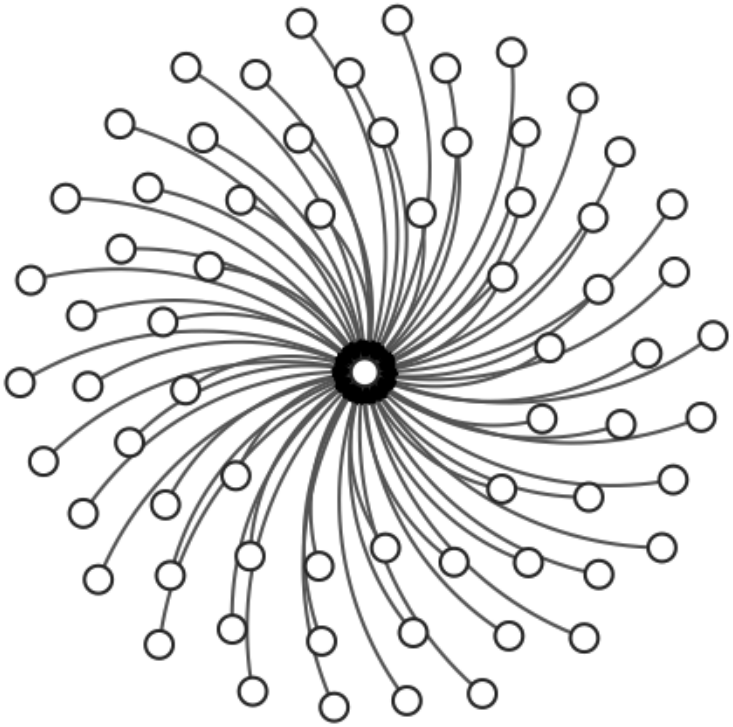
Technical Solution Considerations

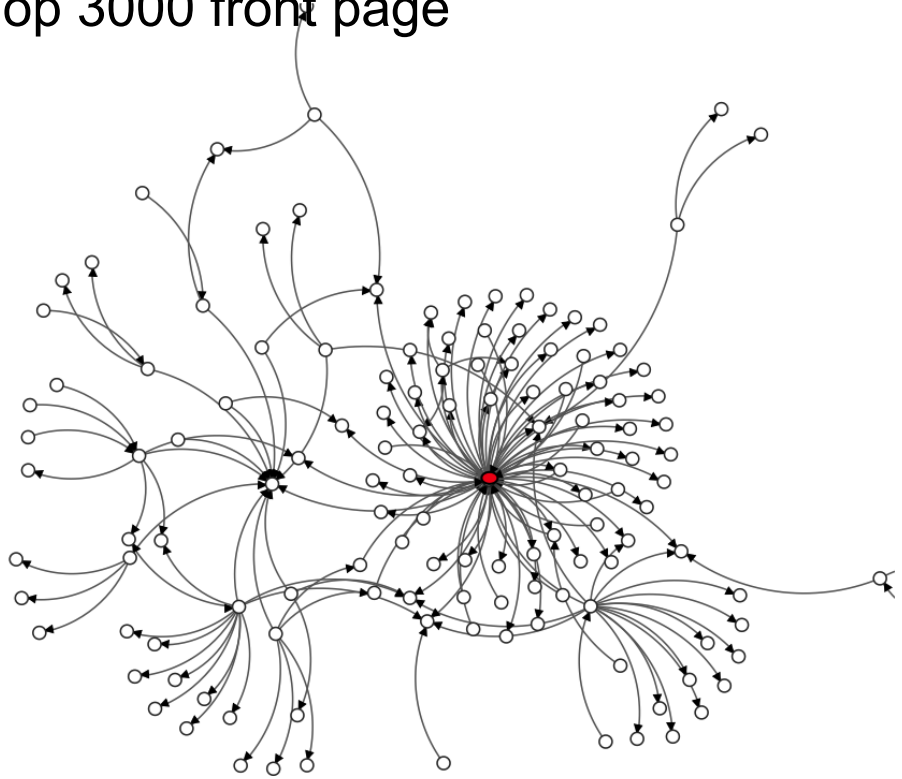
- Usability (in-browser)
- Collateral impact (false positive rate)
- Distance Human
  - Singling
  - Maintai
- Objective
  - Possibly tied into different grades of countermeasure (e.g. blocking cookies vs blocking HTTP)

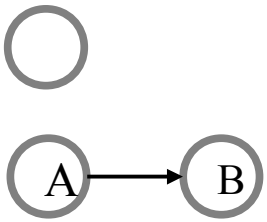
Machine Learning?

Telling Apart Non-Trackers vs Trackers

Data from Alexa Top 3000 front page







domains (PS+1)

<script> from A loads <script> from B into DO

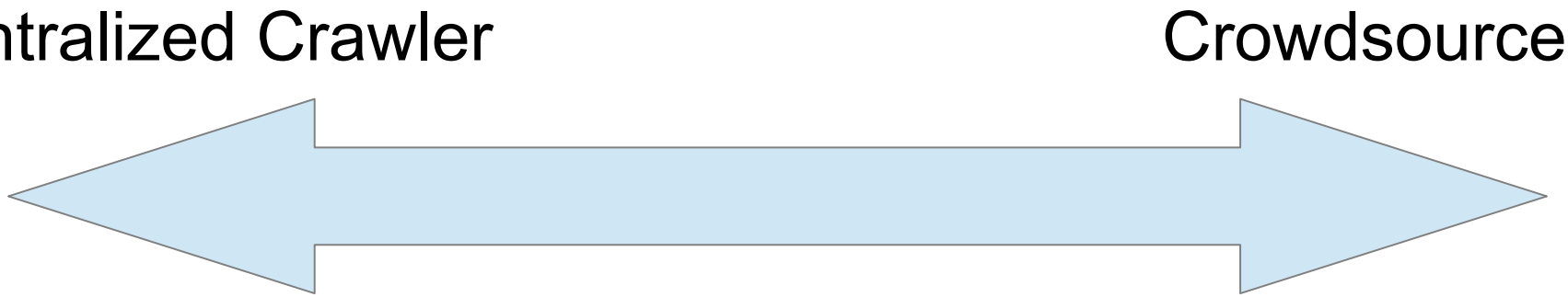
Note: simple prevalence won't do here

2 Categories of Data to Collect

- Relationship between entities (domains) in page DOMs
  - “Caused to load” tree statistics
  - imgs, iframes, scripts, redirects, objects
  - Communications for tracking

- Properties of loaded content (HTTP header)
  - Type
  - Size (1px)
  - Cache params
  - Set-Cookie
  - HTTP/browser features for tracking

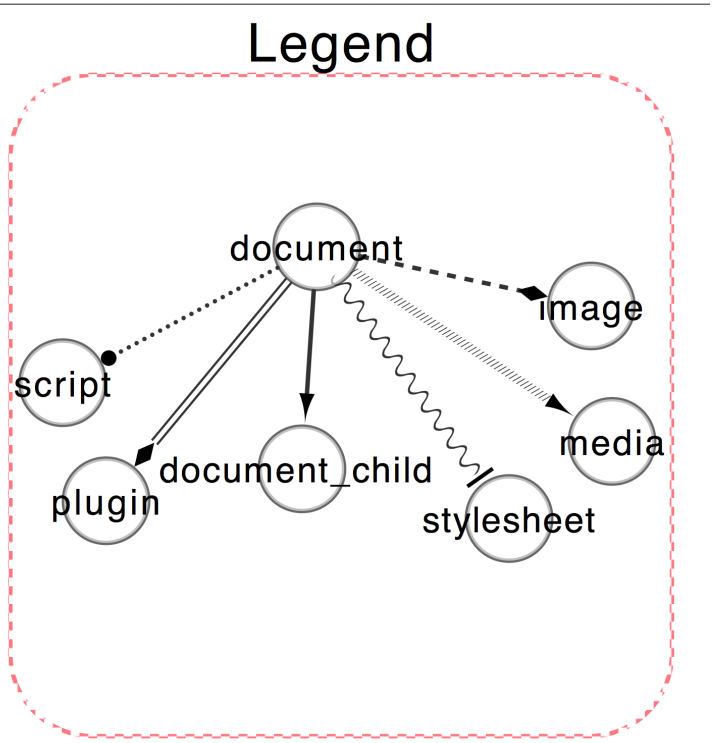
Possible Data Collection Architectures



- Both can use instrumented browser for fidelity

Our Preliminary Experiment

- Crawler (4th Party)
  - Quantcast US Top 32K – 5 random links from landing
  - Collect DOM-like hierarchy
    - Tree rooted at visited page
    - Interior nodes: documents
    - Leaf nodes:
      - Script
      - Image
      - Stylesheet
      - Media
      - Plugin



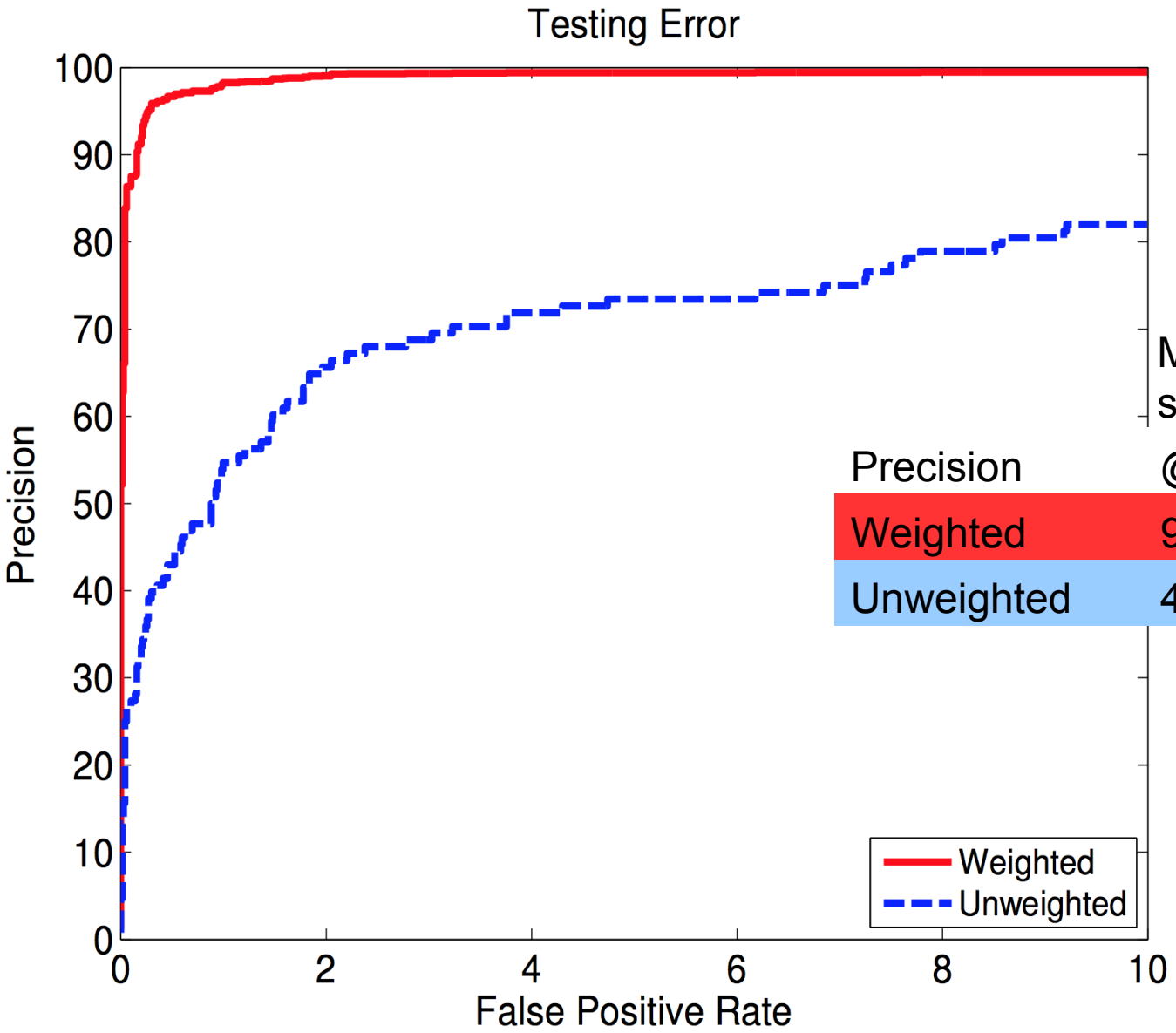
ML Features and Training

- For each domain:
  - Min / Max / Median statistics based on trees appeared in
    - Depth
    - Occurrences
    - Degree
    - Siblings
    - Children
    - Unique parents
    - Etc

- Training Labels from popular blocklist, hand curated to remove 1st party domains and add missing 3rd party domains

- Elastic Net trained on 20% of the data, 80% used for testing

Results



Median of results on 10 randomly selected training/test sets

Precision	@0.5% FPR	@1% FPR
Weighted	96.7%	98%
Unweighted	43%	54%

Weighting each tracker by its prevalence in crawl data.

Tracker changes to evade detection

- Regulatory precedent against actions judged as evasion

- Changing tracking domain names
  - Loses historical data (already-installed cookies)
  - Changes required for their business partners, clients, etc
  - No change to classification algorithms

- New browser features for tracking
  - ETAGs, other supercookies, etc
  - Browser-based data collection will notice
  - Adapt classification algorithm

- "1st party" stand-in for 3rd party tracking
  - Simple CNAMEs can be detected in DNS
  - Server-side proxying to 3rd party possible, but too drastic?

Improvements to Prelim Work

- Better unweighted precision
  - Incorporation of HTTP header features
  - More advanced ML algorithms
- Objectivity
  - Relate features to “fundamentally objectionable” tracking
- Future:
  - Identifier extraction
  - Script provenance graph
  - DNS info
  - Decentralization

Conclusions from prototype

- Machine learning is promising direction for browser controls over third-party tracking reflecting user preference
- Good precision (getting better) at low false positive rates
- Can collect data + classify in days (or less w/infrastructure)
  - Adaptable to changes in tracking landscape
  - Maintainable
- Expert judgement bootstraps, but ultimate criteria can have
  - Understandable objective features
  - Confidence measures

Thanks!

jbau@stanford.edu

Motivation

Table 1: What Americans Want "Do Not Track " to Do

If a 'do not track' option were available to you when browsing the internet, which of the following things would you most want it to do? Should do not track... (READ AND RANDOMIZE)(N=1203)	
Prevent websites from collecting information about you	60%
Block websites from showing you advertisements	20%
Prevent websites from tailoring advertisements based upon the websites you have previously visited	14%
Don't know/refused	6%

Source: Hoofnagle, Urban and Li (2012)