Internet monitoring and web tracking

Lorrie Faith Cranor

October 3, 2013

8-533 / 8-733 / 19-608 / 95-818:
Privacy Policy, Law, and Technology

Carnegie Mellon University
CyLab

institute for SOFTWARE RESEARCH

Engineering & Public Policy

Video

- http://cironline.org/reports/easily-obtained-subpoenas-turn-your-personal-information-against-you-5104

How online tracking works

Browser Chatter

- Browsers chatter about
  - IP address, domain name, organization,
  - Referring page
  - Platform: O/S, browser
  - What information is requested
    - URLs and search terms
  - Cookies
- To anyone who might be listening
  - End servers
  - System administrators
  - Internet Service Providers
  - Other third parties
    - Advertising networks
  - Anyone who might subpoena log files later

Typical HTTP request with cookie

- GET /retail/searchresults.asp?qu=beer HTTP/1.0
- Referer: http://www.us.buy.com/default.asp
- User-Agent: Mozilla/4.75 [en] (X11; U; NetBSD 1.5_ALPHA i386)
- Host: www.us.buy.com
- Accept: image/gif, image/jpeg, image/pjpeg, */*
- Accept-Language: en
- Cookie: buycountry=us; dcLocName=Basket; dcCatID=6773; dcLocID=6773; dcAd=buybasket; loc=; parentLocName=Basket; parentLoc=6773; ShopperManager%2F=ShopperManager%2F=66FUQULL0QBT8MMTVSC5MMNKBJFWDVH7; Store=107; Category=0

Referer log problems

- GET methods result in values in URL
- These URLs are sent in the referer header to next host
- Example: http://www.merchant.com/cgi_bin/order?name=Tom+Jones&address=here+there&credit+card=234876923234&PIN=1234&->index.html
- Access log example: http://www.sdr.info/logs/access_log
- Click from this page to see the referer too: http://cups.cs.cmu.edu/courses/pplt-fa13/referer.html

Cookies

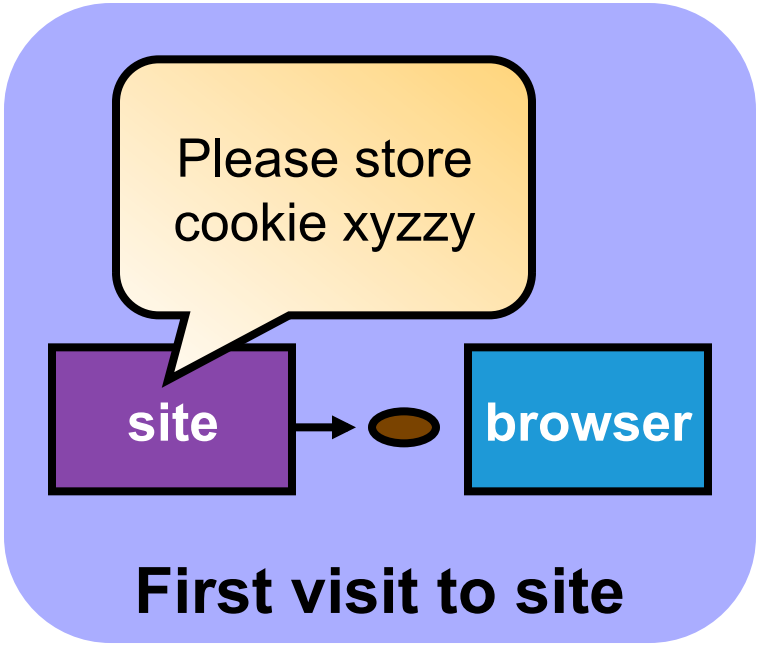
- What are cookies?
- What are people concerned about cookies?
- What useful purposes do cookies serve?

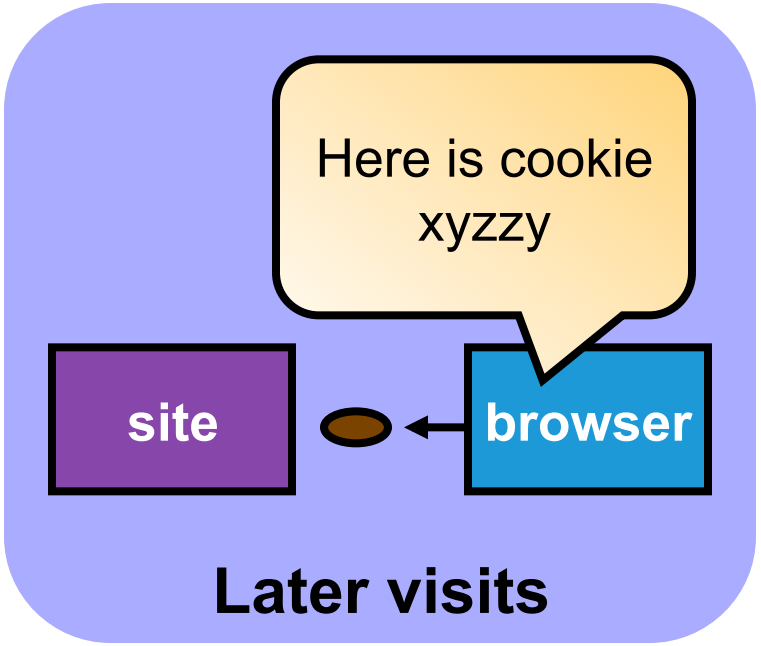
Cookies 101

- Cookies can be useful
  - Used like a staple to attach multiple parts of a form together
  - Used to identify you when you return to a web site so you don't have to remember a password
  - Used to help web sites understand how people use them
- Cookies can do unexpected things
  - Used to profile users and track their activities, especially across web sites

How cookies work – the basics

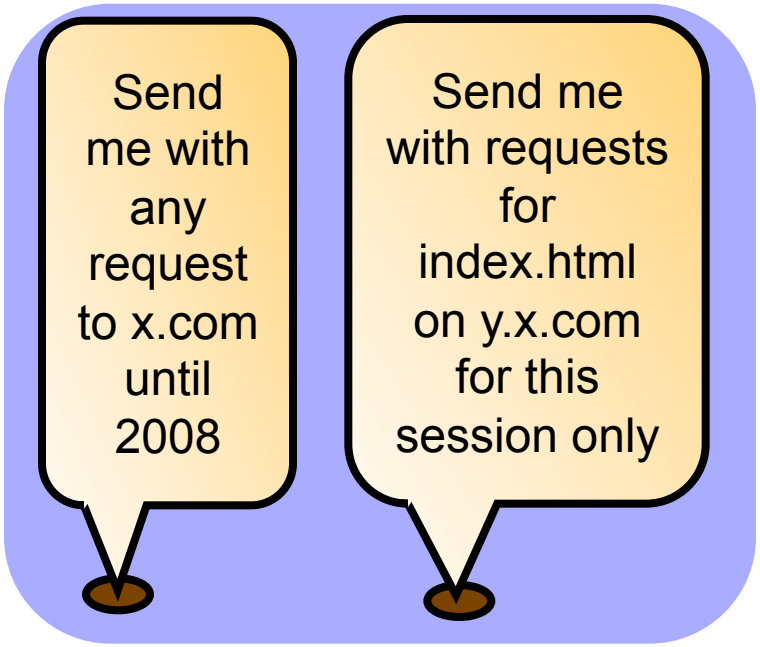
- A cookie stores a small string of characters
- A web site asks your browser to “set” a cookie
- Whenever you return to that site your browser sends the cookie back automatically

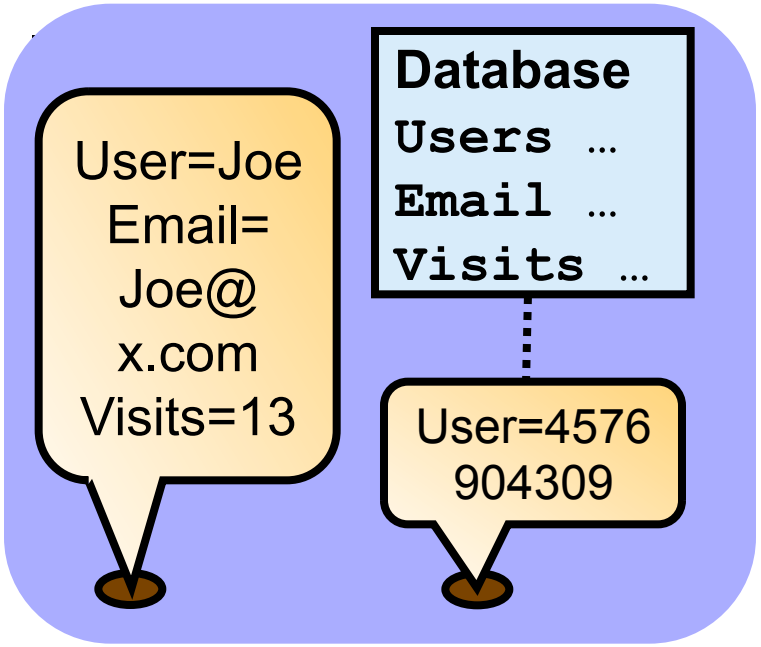




How cookies work – advanced

- Cookies are only sent back to the “site” that set them, but this may be any host in domain
- Cookies can store user info or a database key that is used to look up user info





Cookie terminology

- Cookie replay
  - sending a cookie back to a site
- Session cookie
  - cookie replayed only during current browsing session
- Persistent cookie
  - cookie replayed until expiration date
- First-party cookie
  - cookie associated with the site the user requested
- Third-party cookie
  - cookie associated with an image, ad, frame, or other content from a site with a different domain name that is embedded in the site the user requested
  - Browser interprets third-party cookie based on domain name, even if both domains are owned by the same company

Web bugs

- Invisible “images” (1-by-1 pixels, transparent) embedded in web pages and cause referer info and cookies to be transferred
- Also called web beacons, clear gifs, tracker gifs,etc.
- Work just like banner ads from ad networks, but you can’t see them unless you look at the code behind a web page
- Also embedded in HTML formatted email messages, MS Word documents, etc.

How data can be linked

- Every time the same cookie is replayed to a site, site may add information to the record associated with that cookie
  - Number of times you visit a link, time, date
  - What page you visit
  - What page you visited last
  - Information you type into a web form
- If multiple cookies are replayed together, they are usually logged together, linking their data
  - Narrow scoped cookie might get logged with broad scoped cookie

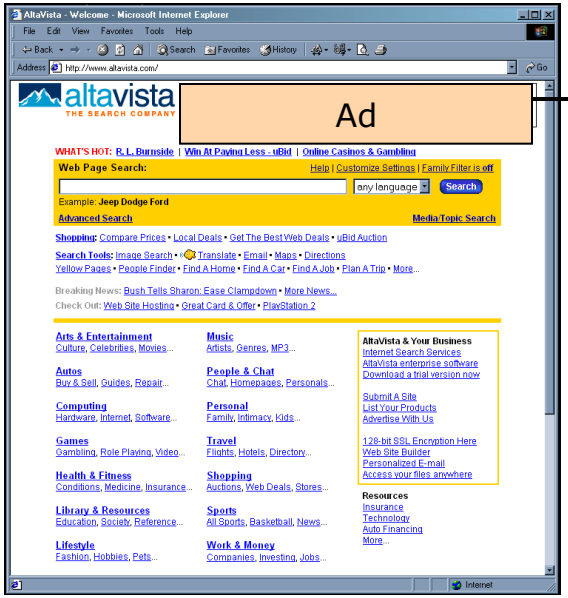
Ad networks

search for medical information

buy CD

set cookie

replay cookie



Search Service

Ad company can get your name and address from CD order and link them to your search

CD Store

What ad networks may know…

- Personal data:
  - Email address
  - Full name
  - Mailing address (street, city, state, and Zip code)
  - Phone number
- Transactional data:
  - Details of plane trips
  - Search phrases used at search engines
  - Health conditions

"It was not necessary for me to click on the banner ads for information to be sent to DoubleClick servers."

– Richard M. Smith

Online and offline merging

- In November 1999, DoubleClick purchased Abacus Direct, a company possessing detailed consumer profiles on more than 90% of US households
- In mid-February 2000 DoubleClick announced plans to merge “anonymous” online data with personal information obtained from offline databases
- By March 2000 the plans were put on hold
  - Stock dropped from $125 (12/99) to $80 (03/00)

Network Advertising Initiative

- NAI formed in 2000 and published NAI principles, guided by the FTC
  - No use of sensitive PII for OBA
  - Opt-in to merge PII with previously collected non-PII
  - Robust notice and choice for future merging of PII with non-PII
  - Robust notice and choice for merging offline and online PII
  - Websites that have third-party OBA will provide notice and choice
- Updated in 2008

Behavioral targeting

- In 2007/2008, more concerns raised about “behavioral” targeting as a new round of companies started deploying systems to target ads based on previous online behavior
- FTC privacy roundtables in 2009/2010 raised more questions about this practice
  - What is the distinction between behavioral and contextual advertising?
  - How do you implement effective notice and choice?
    - Where should notice be provided?
    - Opt-in? Opt-out? When? Where?
  - Do opt-out cookies work?
  - Do we need a “do not track” list?

Tracking without cookies

- Browser fingerprinting
  - What are the components of a browser fingerprint?
  - https://panopticlick.eff.org
- How else can users be tracked?

Tracking email

- What mechanisms can be used to track email?
- What can be learned through email tracking?

Can you control Behavioral Advertising?

Measuring the effectiveness of privacy tools for limiting behavioral advertising

Rebecca Balebako, Pedro G. Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor

Carnegie Mellon University
CyLab

Objective of this work

- Measure behavioral advertising based on web history (build on Guha, et. al 2010)
- Develop method to measure any reduction in behavioral advertising with privacy tools

Tools Tested

- Block third party content
  - Abine TACO
  - Ghostery
  - Block third party cookies
- Opt-out
  - Digital Advertising Agency (DAA)
  - Network Advertising Initiative (NAI)
- Do Not Track headers

Method

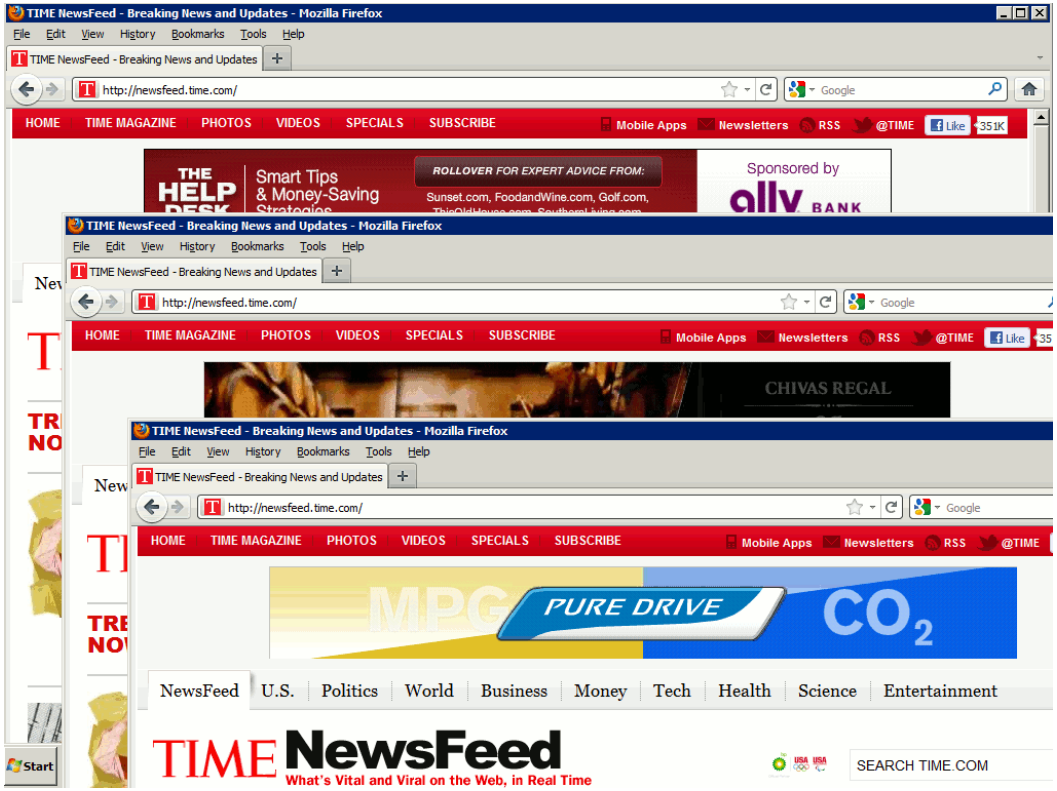
1. Automatically run scenarios that could induce behavioral advertising with training and testing
2. Measure ad turnover
3. Confirm behavioral advertising exists
4. Run scenarios with privacy tools
5. Compare tools

Scenarios - Training

- Training: visit 10-20 pages (~7 unique domains) on a topic
- Topics:
  - European Travel
  - Digital Camera
  - Bicycling
  - Wedding planning
  - Pregnancy
  - Blank (no training)

Scenarios - Testing

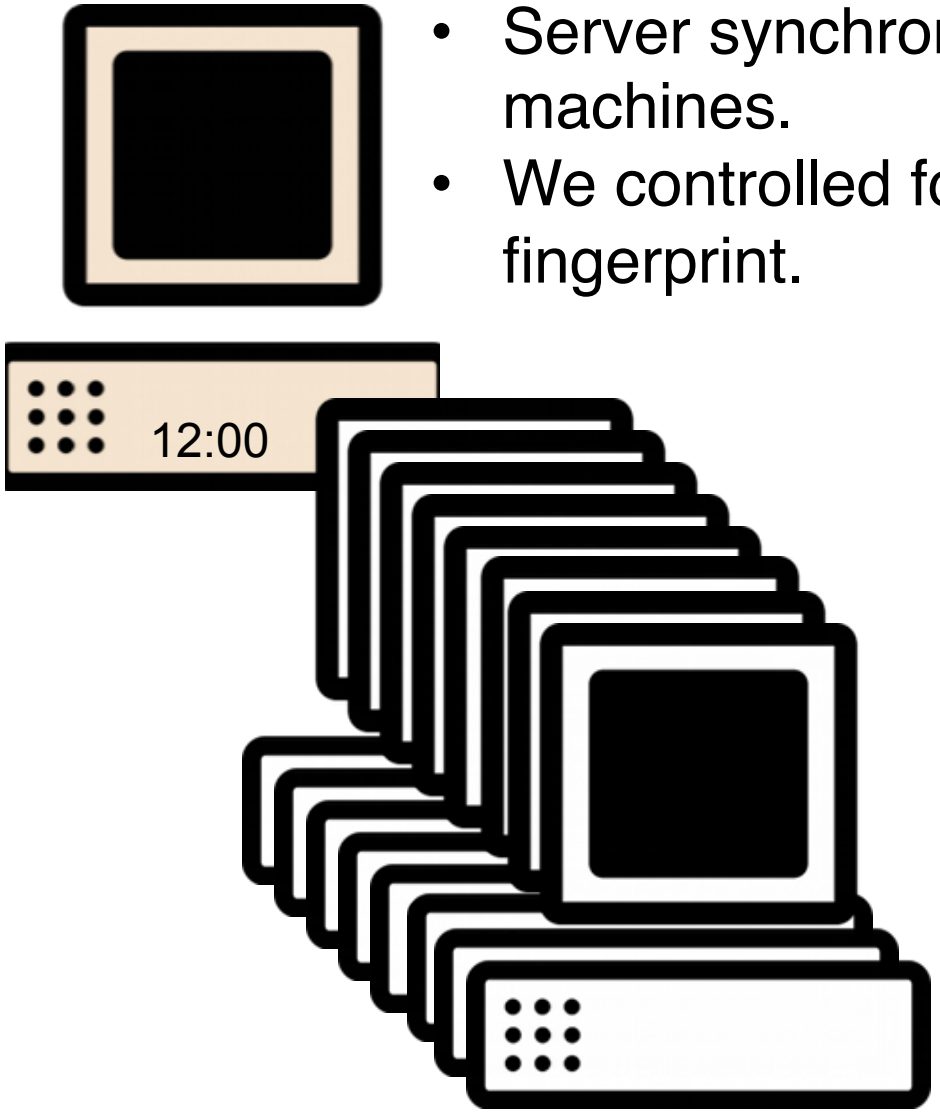
- Test: Unrelated sites with little context
  - New York Times
  - LA Times
  - Chicago Tribune
  - HowStuffWorks
  - CNN
- 7 hits
- Save the text ads



Two different automated tests

goal	control	synchronization
measure OBA	no training	all topics run simultaneously
test tools	no tool	all tools run simultaneously for each topic

Automated Testing



- Server synchronizes identical virtual machines.
- We controlled for time, IP, & browser fingerprint.

1. Control
2. Control2
3. Abine Taco
4. Ghostery
5. DAA
6. NAI
7. Firefox 3rd Party Cookies
8. Firefox DNT

Analysis: Cosine Similarity

- Cosine similarity used to compare frequency vectors of words or URLs
- A and B are frequency vectors of elements in A ∪ B
- Cosine similarity defined as

$$\frac{\overline{A} \bullet \overline{B}}{\left\|\overline{A}\right\| \left\|\overline{B}\right\|}, \text{ where } \quad \overline{A} = \left[w_{A,e} \right]$$

- Weight of element *e* in A is the frequency it appeared
- *e* is either word or URL

Anatomy of an Ad

- Display URL: www.GoAheadTours.com
- Stemmed Words: tour beauti itali $2,199 9-dai tour across itali includ air hotel more

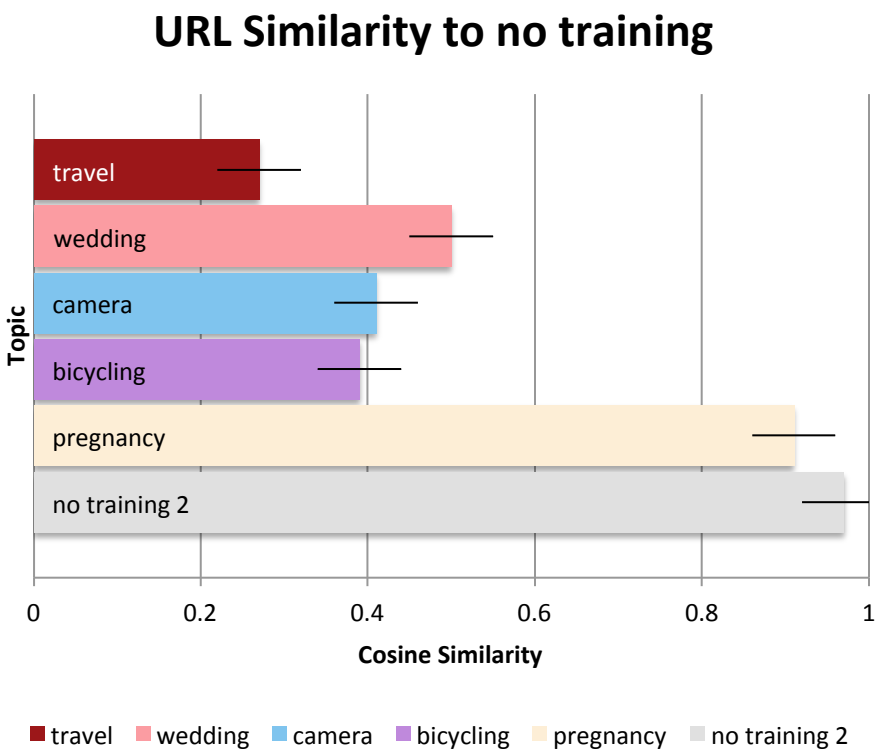
Comparing Ads

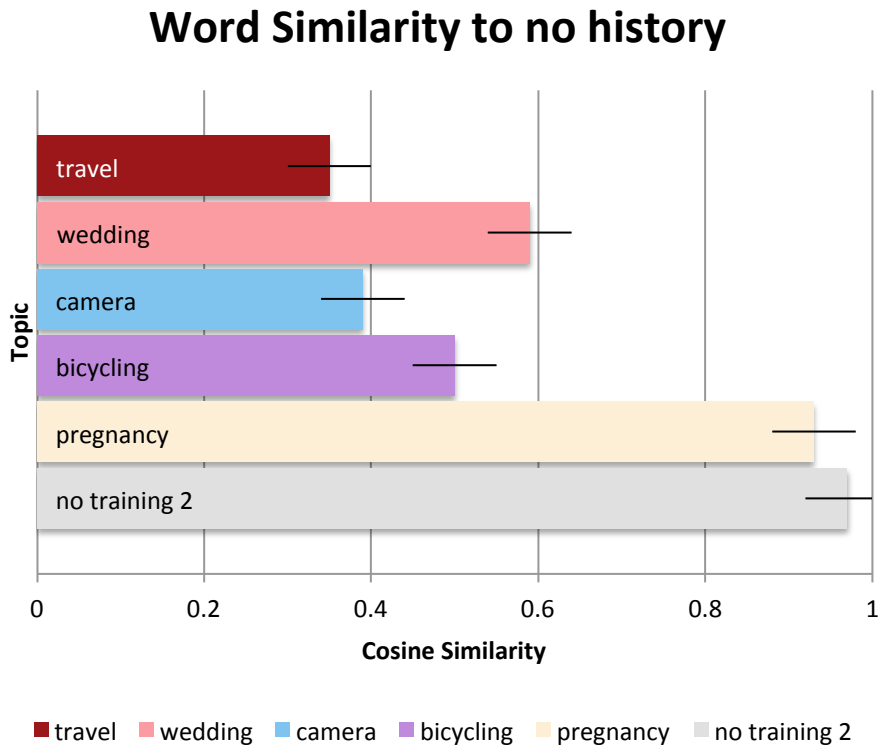
- Compare Ads:
  - Use the display URL to determine if ads are unique
  - Use the stemmed words in the title and the description to determine contextual differences between sets of ads

Ad Turnover

- Similarity between “notraining” and “notraining2”
  - Test 1: .97 for word frequency and .97 for URL frequency
  - Test 2: .97 for word frequency and .95 for URL frequency
  - Therefore a conservative .9 = same set

OBA found in 4 topics

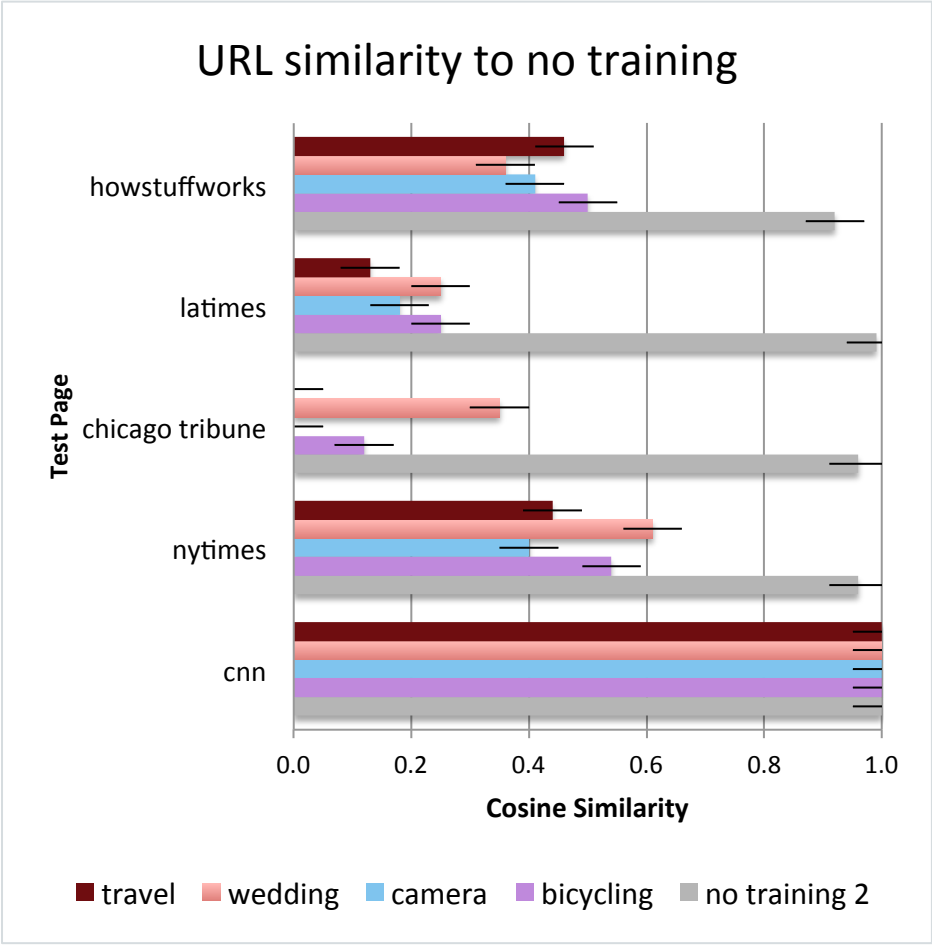


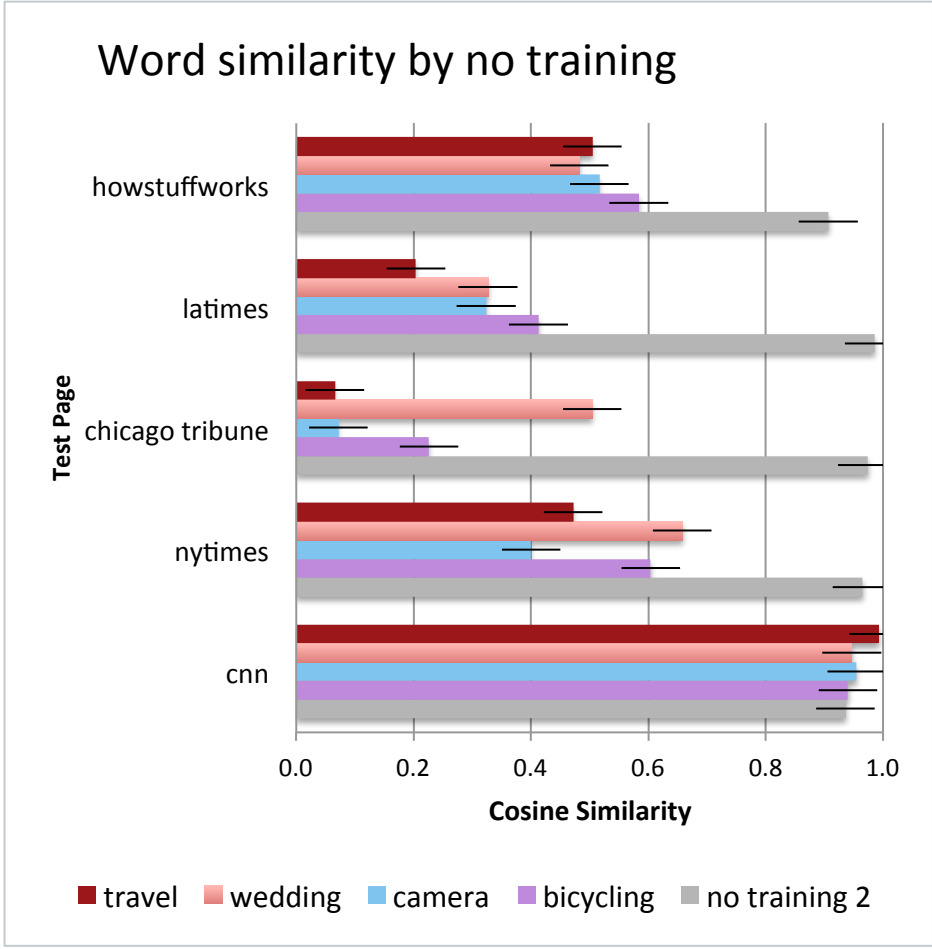


OBA demonstrated by frequent words

Topic	5 Most Frequent Words
travel	on, eurail, pass, sapson, to
wedding	free, for, wed, label, your
camera	camera, free, sale, ship, for
bicycle	bike, mountain, and, you, for
pregnancy	depress, for, symptom, free, have
no training	depress, for, symptom, a, now
no training 2	depress, for, symptom, now, new

OBA found on 4 test pages





Tool Effectiveness

- Similarity between tool and no tool
- Similarity should be less: ads are different because tool stops behavioral advertising

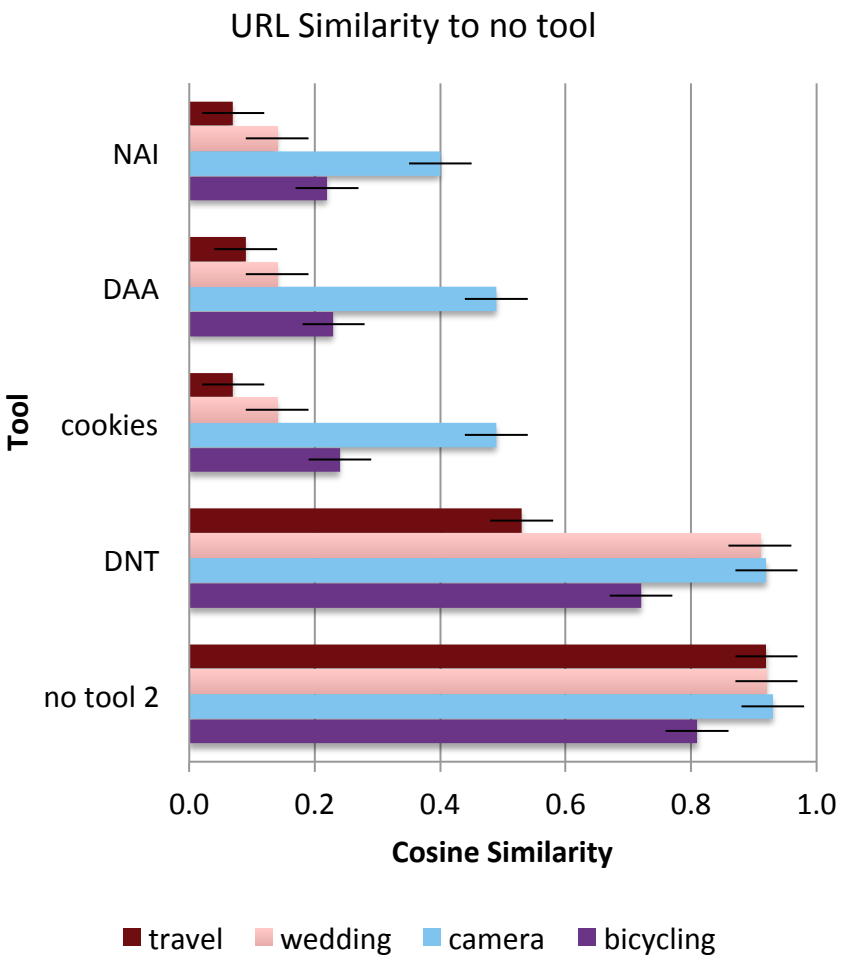
- All ads are “Ads by Google”

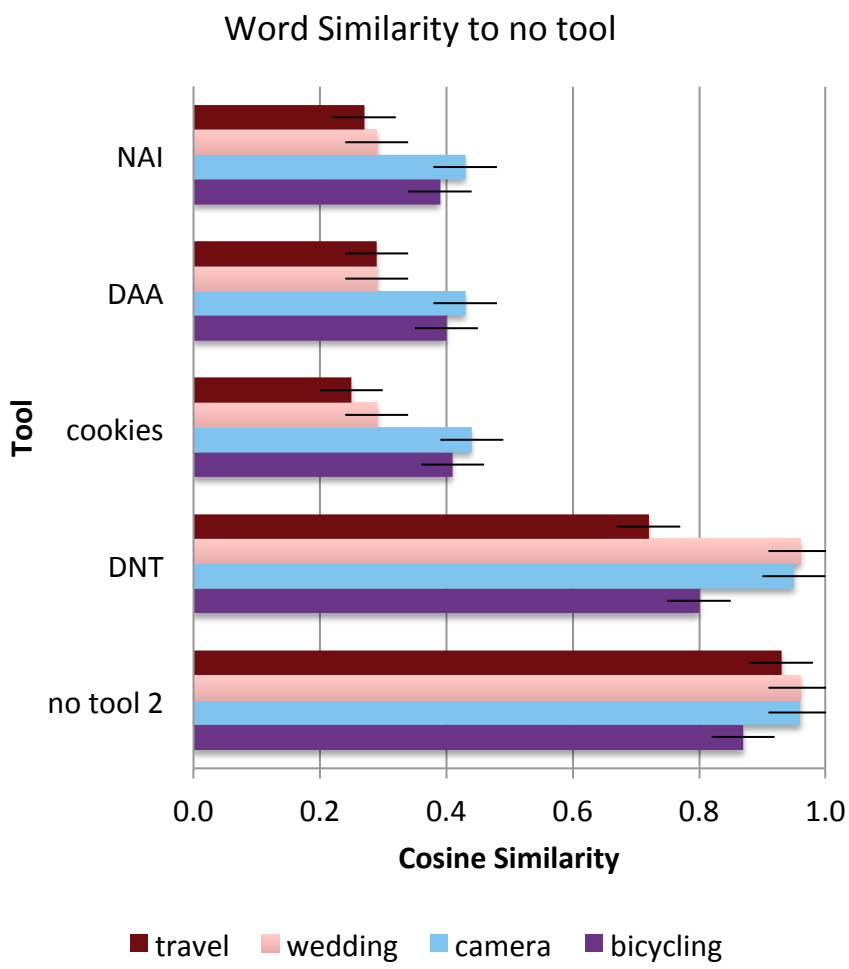
Blockers Blocked Ads

- Ads by Google completely eliminated
  - Abine Taco
  - Ghostery
- Do not block all ads

Tool Effectiveness

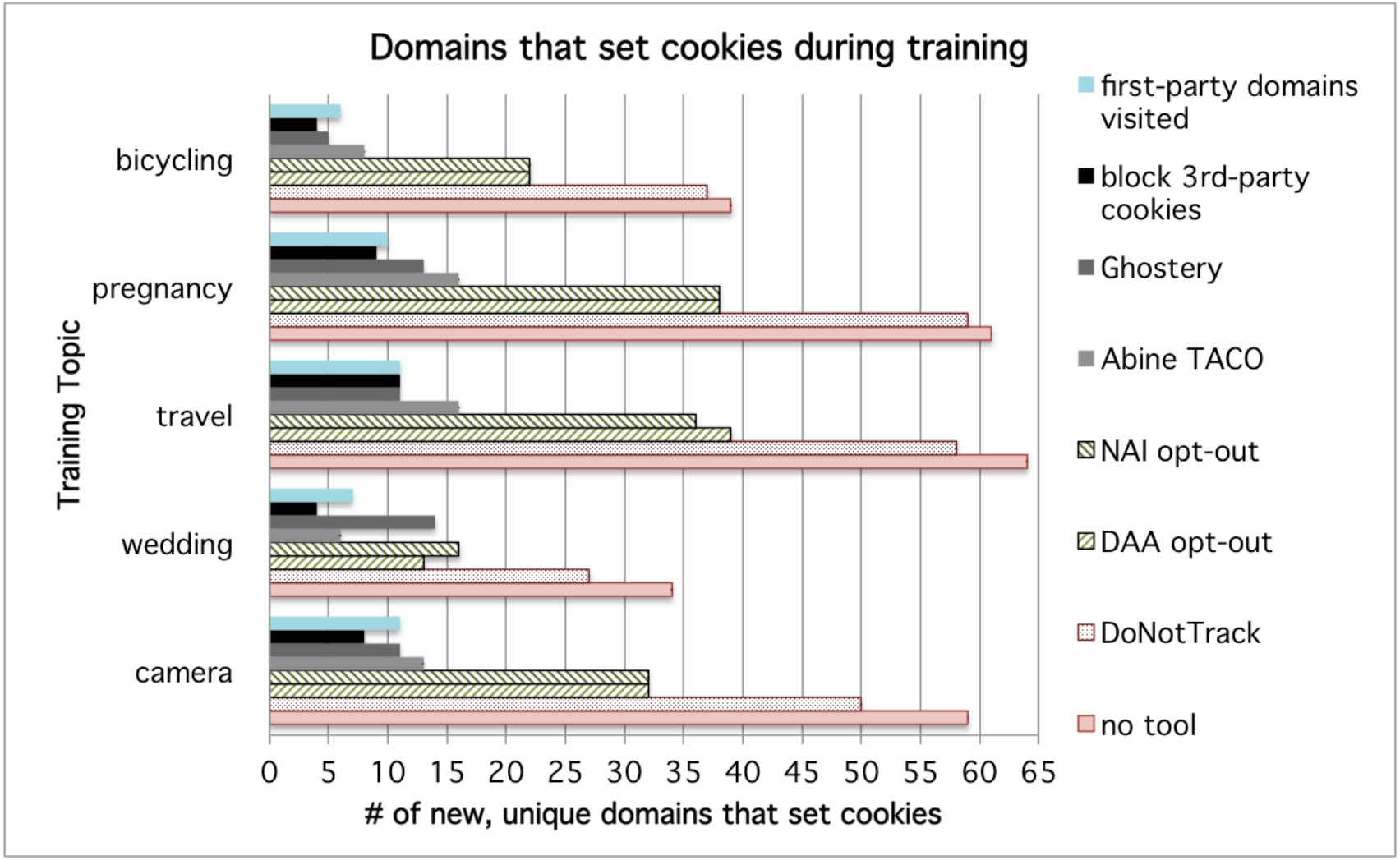
DNT not effective





Cookies

DNT and opt-out not very effective



Carnegie Mellon University
CyLab

isr institute for SOFTWARE RESEARCH

Engineering & Public Policy