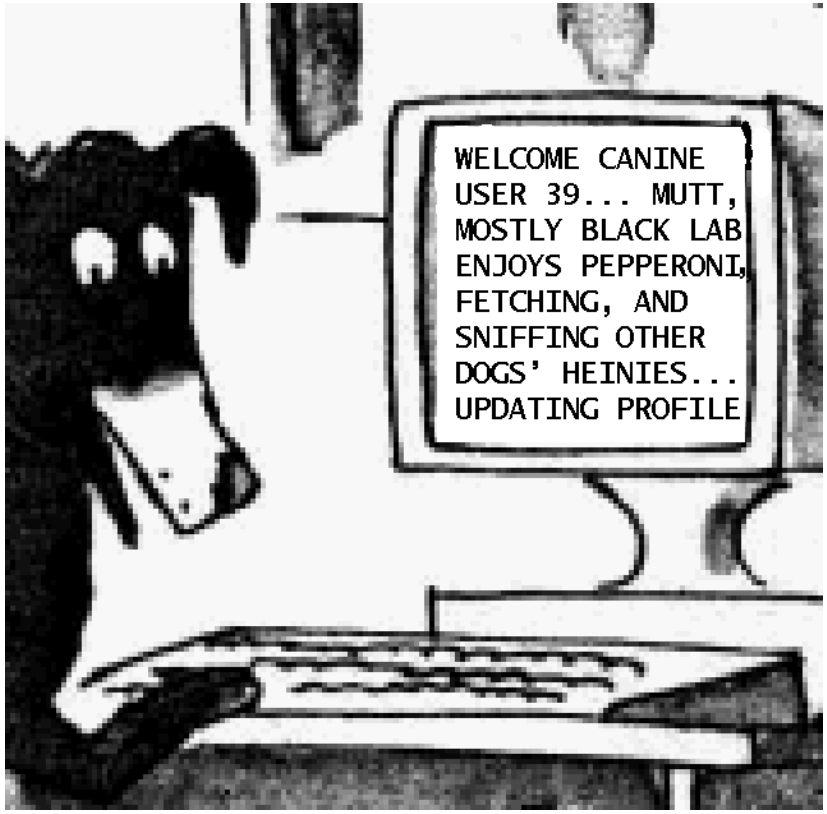
Internet Privacy: Then and Now

"On the Internet, nobody knows you're a dog."



Right cartoon obtained from `http://www.unc.edu/depts/jomc/academics/dri/idog.html`.
Left cartoon by Peter Steiner has been reproduced from page 61 of July 5, 1993 issue of The New Yorker.

WPI

Understanding What is Happening Regarding Internet Privacy

Craig E. Wills
Worcester Polytechnic Institute

Presented at the SENCER Spring Regional Meeting
Southern Connecticut State University
April 5, 2014

WPI

Privacy is a Big Topic

Potential concern for private information loss to many entities:

- Commercial/Business
- Governments/Defense & Security Agencies
- Neighbors
- ...

Focus of this talk is on individual control of private information to commercial entities on the Internet.

More specifically on the control of information to third parties.

Roadmap for Presentation

- Longitudinal privacy footprint
- Understanding what is done with information
- Leakage, Linkage and Lifetime

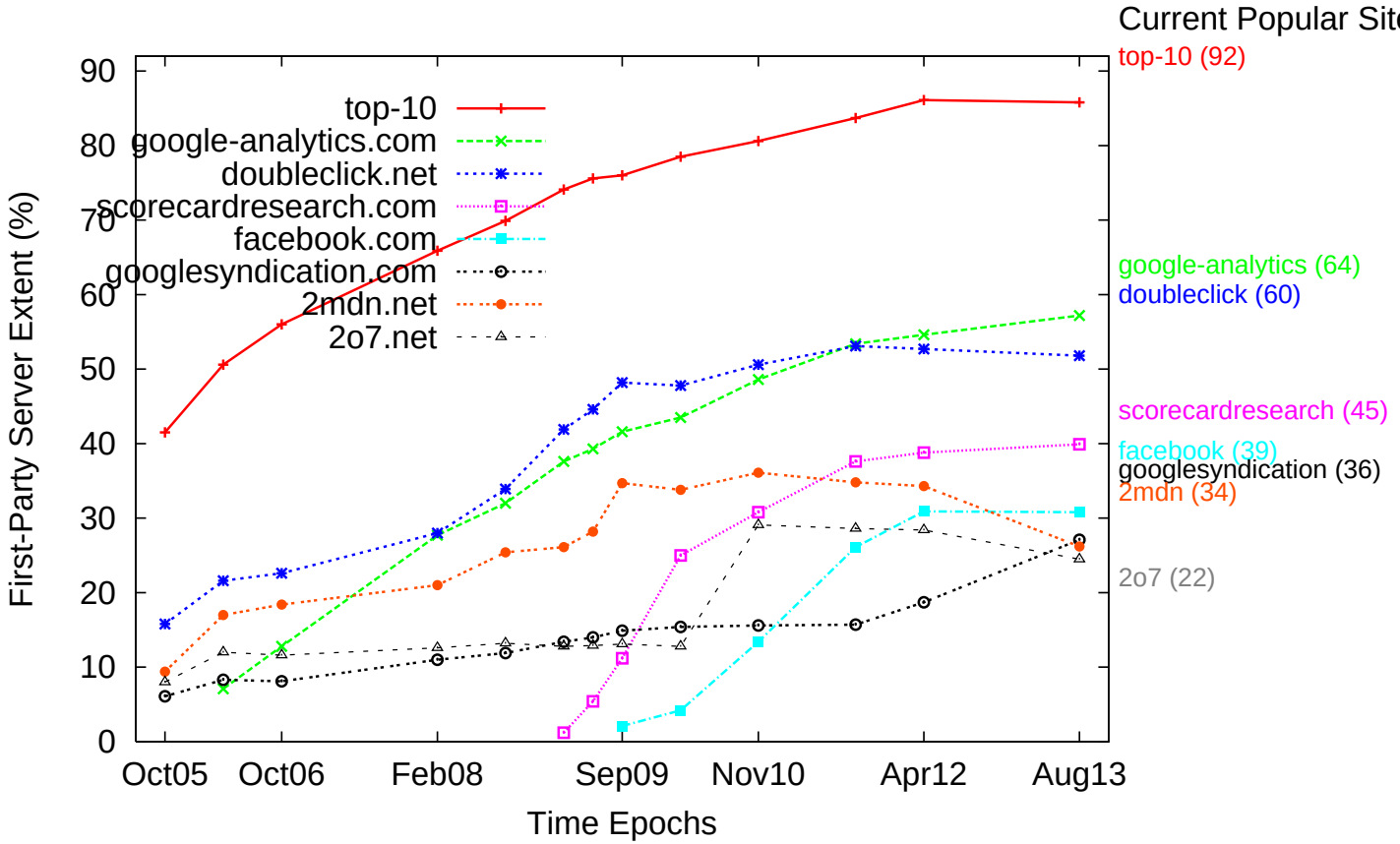
Longitudinal Privacy Footprint

Aggregation of user browsing behavior by third parties (e.g. doubleclick.net) across *unrelated* first-party sites (e.g. cnn.com, hulu.com, espn.com) typically in the presence of third-party tracking cookies.

Originally built a list of 1100+ popular Web sites across 12 categories of 100 sites each based on Alexa rankings in 2005. Have periodically accessed the home pages of these sites and recorded the set of downloaded objects since 2005—most recently August 2013.

Similarly built a list of popular Web sites in 2012 based on Alexa rankings at that time. Retrieved this set of popular pages as well.

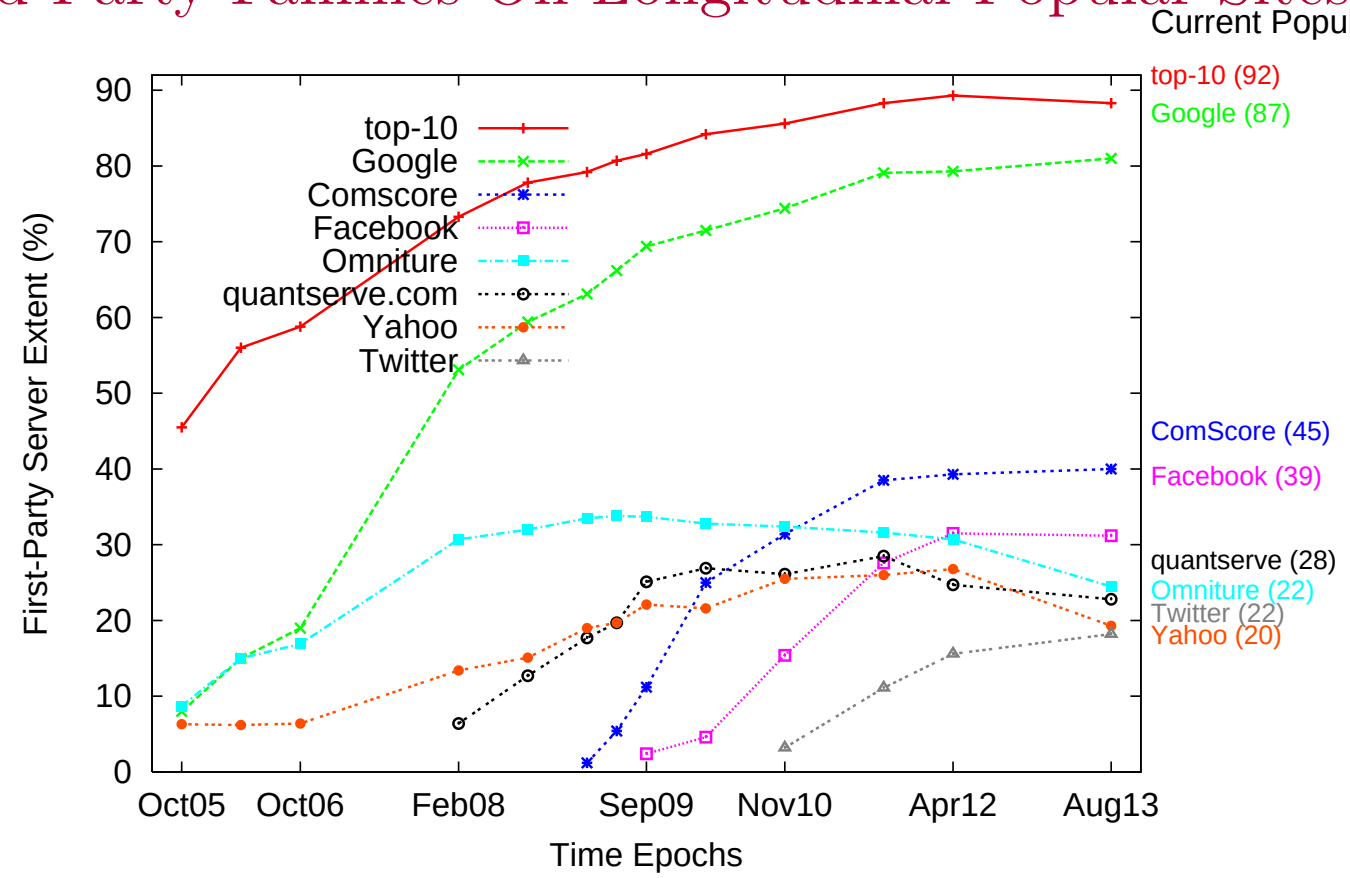
Top Third-Party Domains On Longitudinal Popular Sites



Current (2012) popular sites extent results even larger.

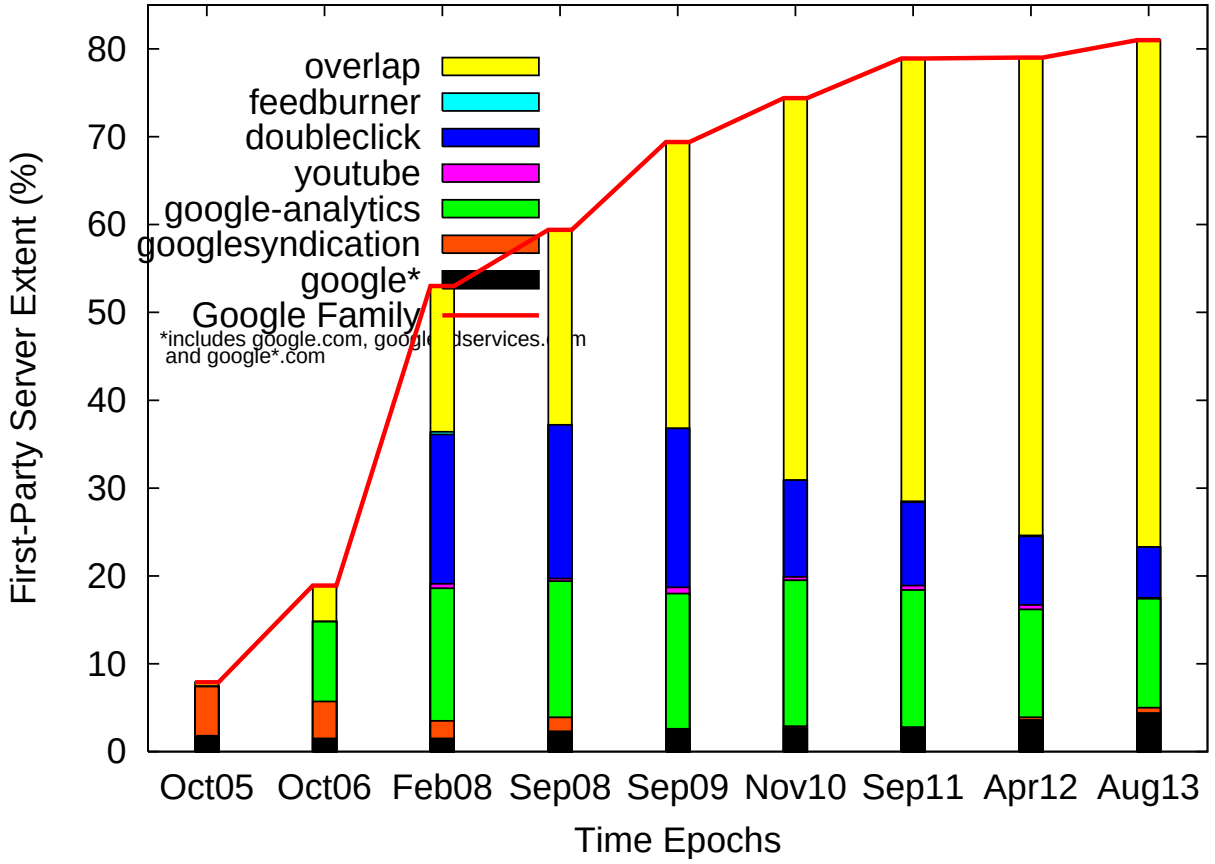
Over time we see continuing as well as new players.

Top Third-Party Families On Longitudinal Popular Sites



Facebook and Comscore have developed a prominent third-party presence.

Anatomy of Google Family Growth



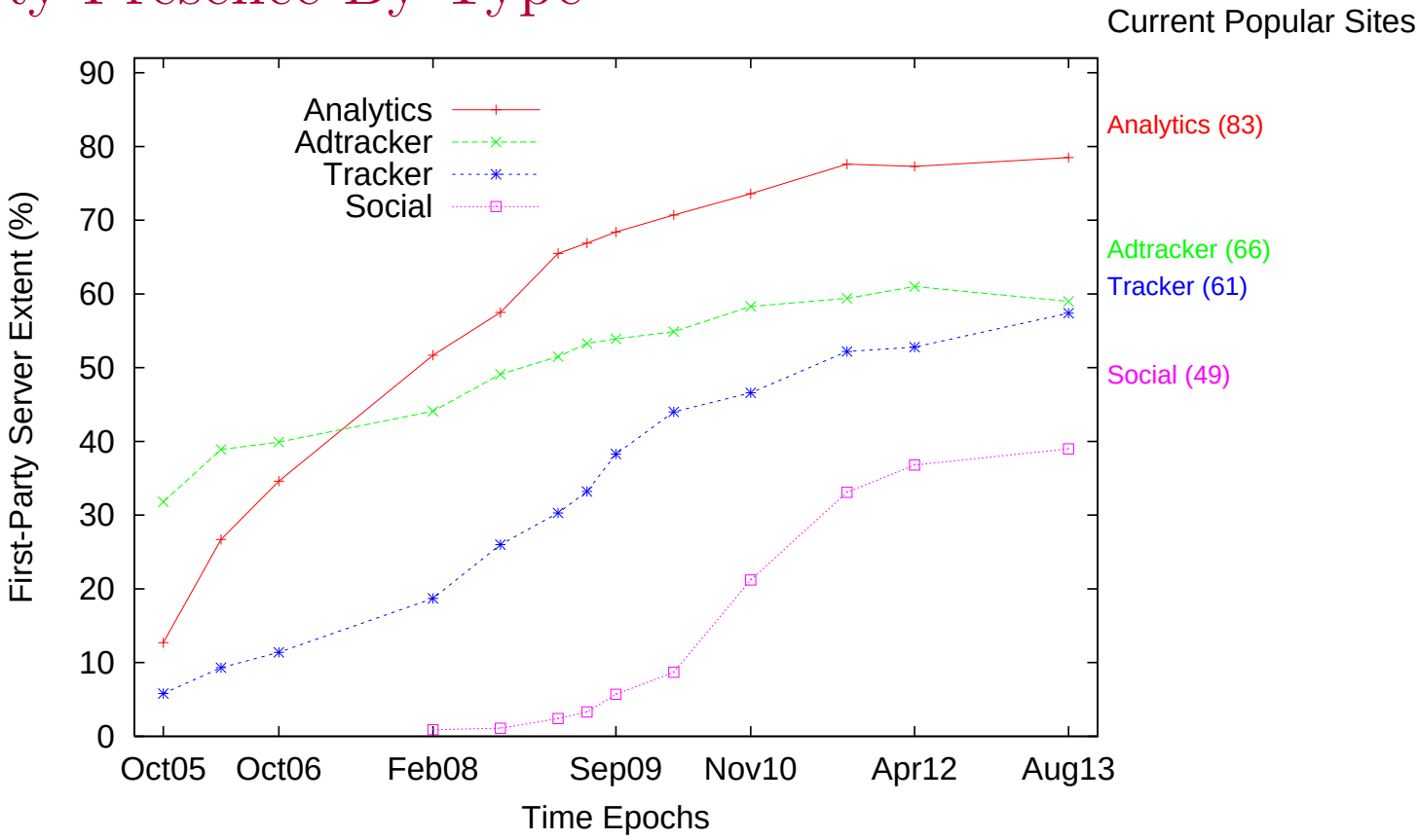
Acquisitions fueled early growth. Increasing presence of multiple Google entities on Web sites.

What Are These Third Parties Doing?

Classify third parties into four types—based on behavior as well as categorization defined by Ghostery, Evidon and Privacy Choice

1. Analytics (e.g. google-analytics, omniture, imrworldwide): provide data aggregation for first-party sites.
2. AdTracker (e.g. doubleclick, googleadservices, atdmt, yieldmanager): serve ads and track user activity across third-party sites.
3. Tracker (e.g. scorecardresearch, quantserve, revsci, bluekai): do not directly serve ads, but track and aggregate user activity
4. Social (e.g. facebook, googleplus, twitter): icons/links to connect user activity with social media sites.

Third-Party Presence By Type



Presence of third-party ads initially the most prevalent, but shows least-dramatic growth.

Biggest growth in the presence of third-party analytics.

Significant growth of Tracker and Social categories.

Roadmap for Presentation

- Longitudinal privacy footprint
- Understanding what is done with information
- Leakage, Linkage and Lifetime

Understanding What is Done with Information

Much work showing third-party advertisers are in a position to observe behavior and *infer* user characteristics across a broad range of first-party Web sites.

Other work has shown advertisers also in a position to obtain *known* information—social networking and many other sites where users reveal information about themselves.

Research question of this work: Understand what "they" (the advertisers) actually do with this information available to them.

Examine ad networks providing "Ad Preference Managers" and Facebook, which is in a position to display ads based on user-provided information.

Contributions

- not only examine how advertisers use *behavioral* information in serving ads, but take a more comprehensive approach to see if and how this information is combined with *location* and *personal* characteristics of a user,

- introduce a variety of sensitive topics,

- examine more than just text ads,

- examine behavior of Ad Preference Managers in terms of inferred demographics and interests.

Google Ads Preferences Manager

Found approximately ten ad networks with an "Ad Preference Manager."

Information Received by Third Parties

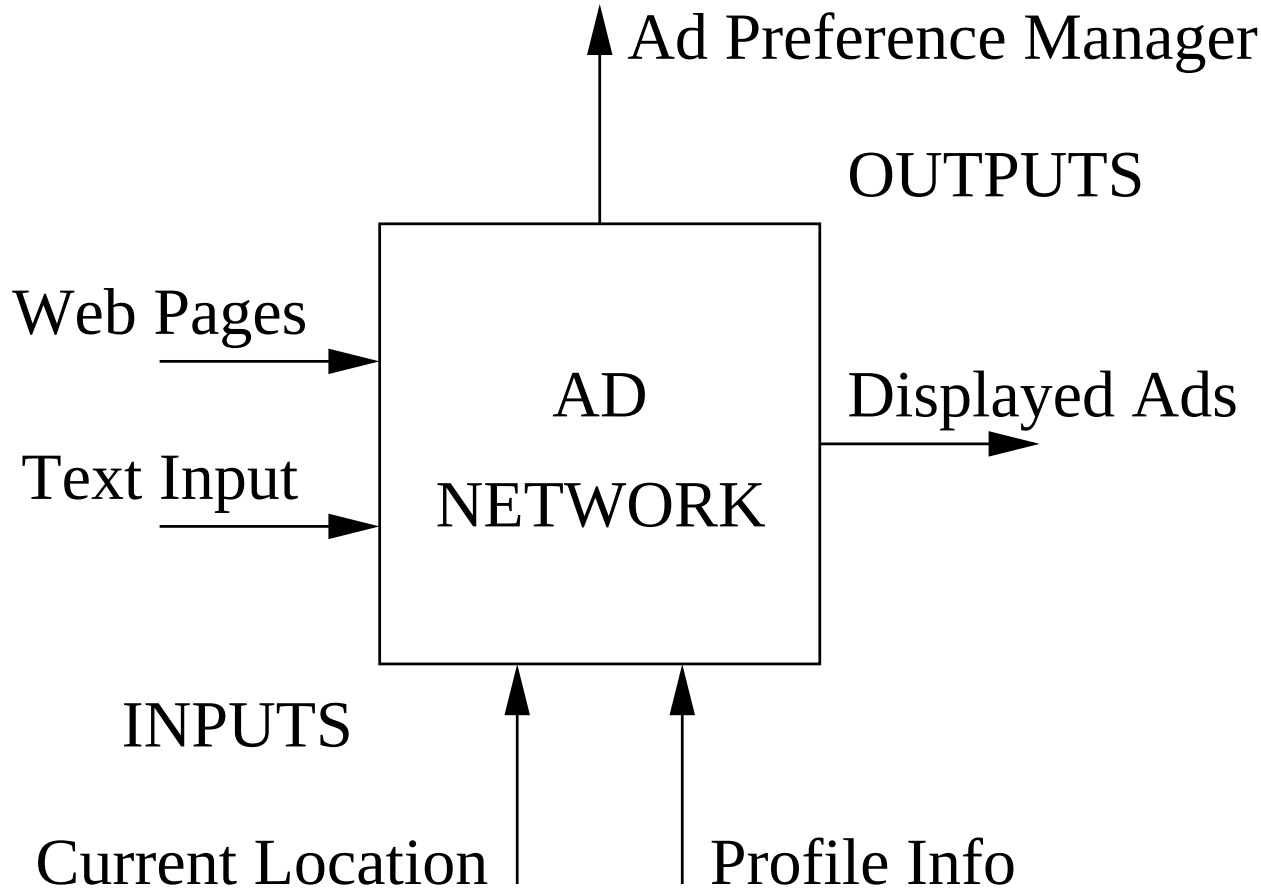
User browsing behavior and profile information is transmitted to third-party advertisers via first-party sites.

Example: Leakage of Private Information from Pandora Profile to DoubleClick (Google)

GET http://ad.doubleclick.net/pfadx/pand.default/...;
artist=S1421673;genre=love;ag=32;gnd=1;zip=11201
Host: ad.doubleclick.net
Referer: http://www.pandora.com/
Cookie: id=223d4200013312||t=1292486411|et=730|cs=p999khn4

Methodology

Control inputs to an ad network then examine output of ad network for evidence that these inputs are used.



Initial Results

Initially studied four larger ad networks employing an APM—AOL, BlueKai (actually a data exchange), Google and Yahoo!

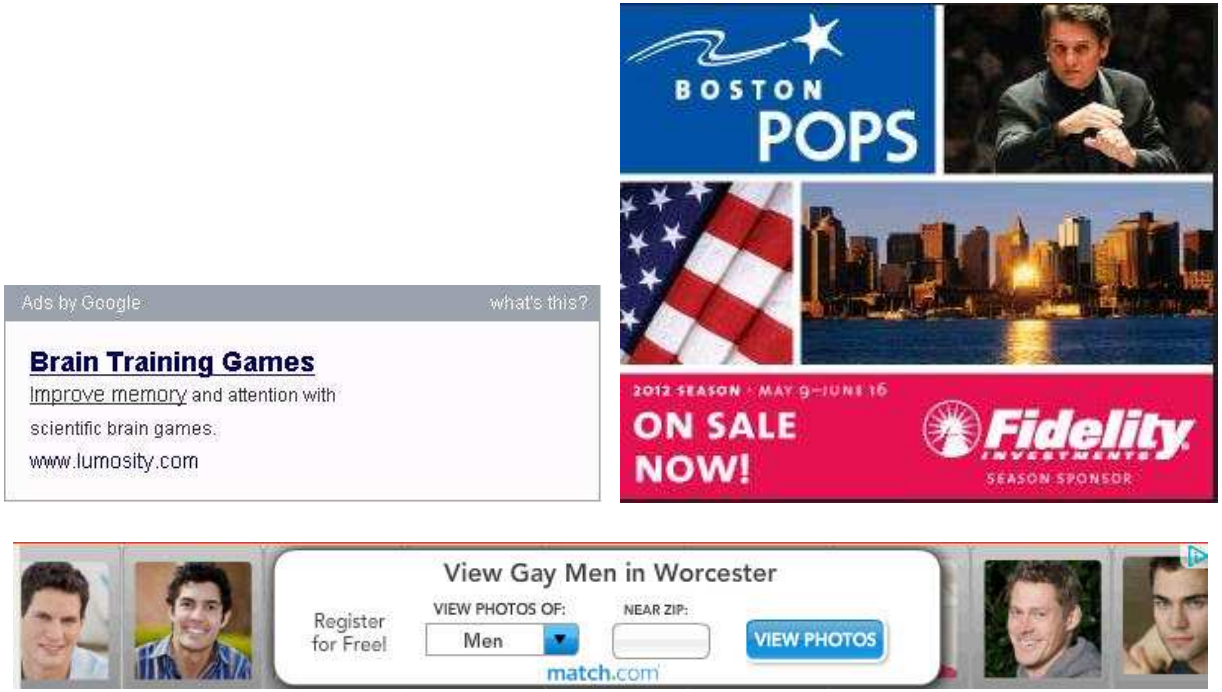
- Unable to characterize types of ads served based on information known by BlueKai;
- Generic, contextual, location and behavioral ads (consistent with APM contents) for other three networks; and
- For Google, also observed ads for topics not in APM and ads for sensitive topics.

Focused study of Google—55 daily sessions across 6 experiments, where different interests were induced or not induced in each experiment starting with a clean browser.

Sites Used for Google Ad Network

Web Site	Category
bloomberg.com	Financial News
accuweather.com	News/Weather
tripadvisor.com	Travel Planning
yelp.com	User Reviews
ford.com	Automotive
toyota.com	Automotive
gaylife.about.com	Gay Life
thenewgay.net	Gay/Lesbian
linkedin.com	Professional Networking
pandora.com	Radio
medhelp.org	Health/Support
menshealth.com	Men/Health
metrolyrics.com	Music/Lyrics
tmz.com	Entertainment
cbsnews.com	News
cnn.com	News
huffingtonpost.com	News
nytimes.com	News
macmillandictionary.com	Dictionary
thefreedictionary.com	Dictionary

Contextual and Location-Based Ads



1. Contextual Ad on `nytimes.com` Science Page (numerous)
2. Location-Based Ad on `pandora.com` (100% of sessions)
3. Contextual and Location-Based Ad on `gaylife.about.com`

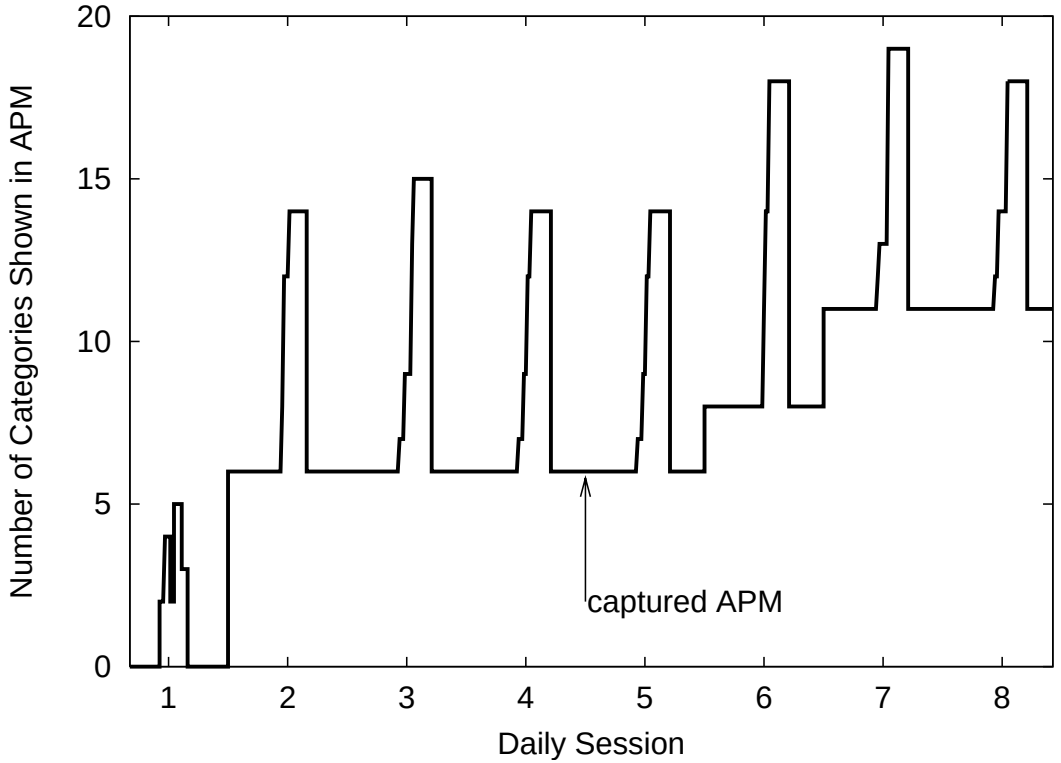
Profile-Based Ads on `pandora.com` using Profile Location (New York) as well as Profile Age (32) and Current Location (Worcester)

Subsequent use of profile information? Found 59% of sessions contained at least one ad for "nyc" or "new york" (not on LinkedIn or Pandora), although no control for comparison.

Induced Behavioral Interests

Induced Interest	How Induced?	Match Keyword(s)
cars	Ford, Toyota sites	ford, toyota, cars, autos, mazda, honda, jeep
dogs	search term	dog, k-9, pets, veterinarian, puppies
golf	search term	golf
investment	Bloomberg site	finance, invest, stock, market, trusts
miami	location selection	miami, south beach
tennis	search term	tennis, racquet

Evolution of APM Over Time



Appears to be a two-stage process for mapping browsing behavior to the set of categories:

1. Short-term based on input text and content of pages—such as `nytimes.com` and `bloomberg.com` (not `cnn.com` or `cbsnew.com`).
2. Long-term that persist between daily sessions.

Behavioral Ads for Interests Golf and Dogs (Each Shown in Respective APM) on `accuweather.com` and `macmillandictionary.com`

Ads by Google

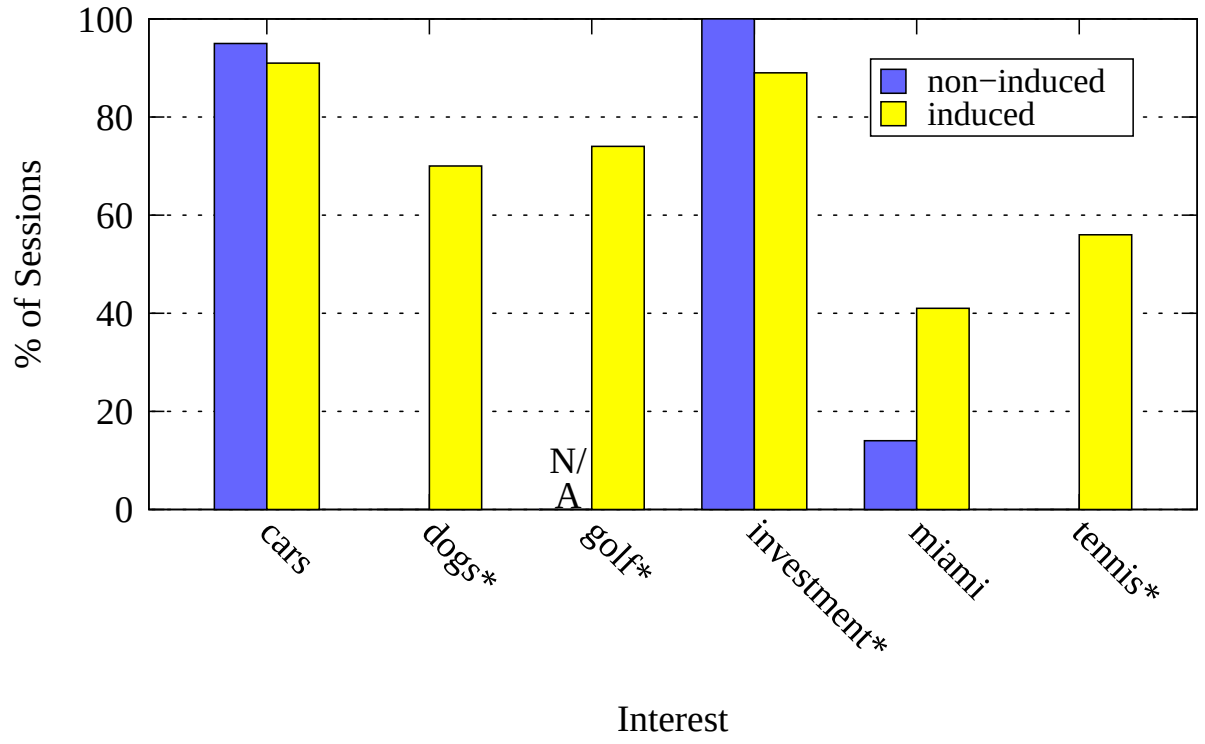
Top 10 Swing Killers
These Faults Make It Impossible To
www.RevolutionGolf.com

Callaway Golf
Buy Top Rated Callaway Golf Clubs.
www.Golfsmith.com/Callaway

Pine Needles Golf Academy
Receive top quailty instruction
www.pineneedles-midpines.com

Behavioral Ad for Miami Interest (Not Shown in APM) on `pandora.com`

Percentage of Sessions Displaying Non-Contextual Ad Matching Induced Interest



(* indicates interest was shown in APM when induced)

Behavior matching displayed ads is often, but not always shown in APM.

Induced Sensitive Interests

Induced Interest	How Induced?	Match Keyword(s)
bankruptcy	search term	bankrupt, chapter 7, debt, tax relief, foreclosure
depression	health search term	depression
diabetes	health search term	diabetes
gay/lesbian	gaylife, thenewgay sites	lgbt, lesbian, gay, mat_boy
pregnancy	health search term	pregnant, ob/gyn, infant, baby, birth
skin cancer	health search term	skin cancer, melanoma

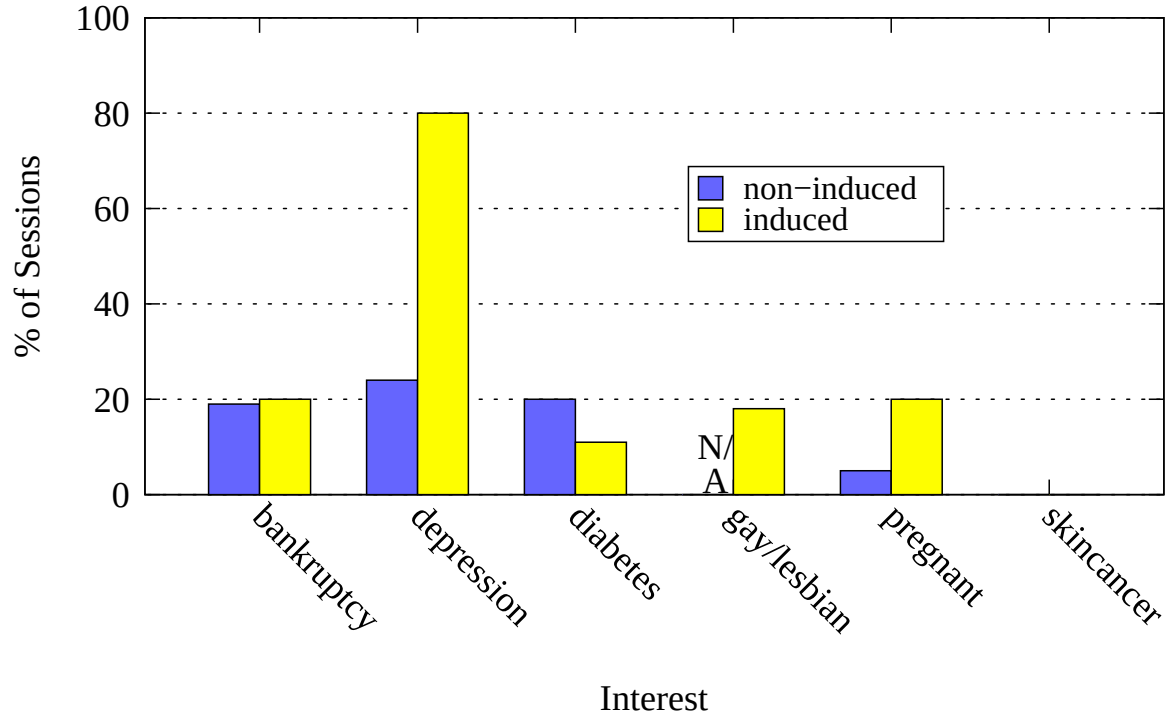
Non-Contextual Ads for Sensitive Induced Interest

1. Depression on `nytimes.com`

2. Bankruptcy on `accuweather.com` Page for Miami Weather

3. Pregnancy on `pandora.com`

Percentage of Sessions Displaying Non-Contextual Ad Matching Induced Sensitive Interest



(* indicates interest was shown in APM when induced)

Non-contextual ads for sensitive topics are being shown—could be viewed by users as behavioral ads.

Facebook Methodology

Facebook has a growing third-party presence and serves ads on its own site. How are ads for interests handled?

Four experiments, each with 15-20 daily sessions:

1. interest not induced;

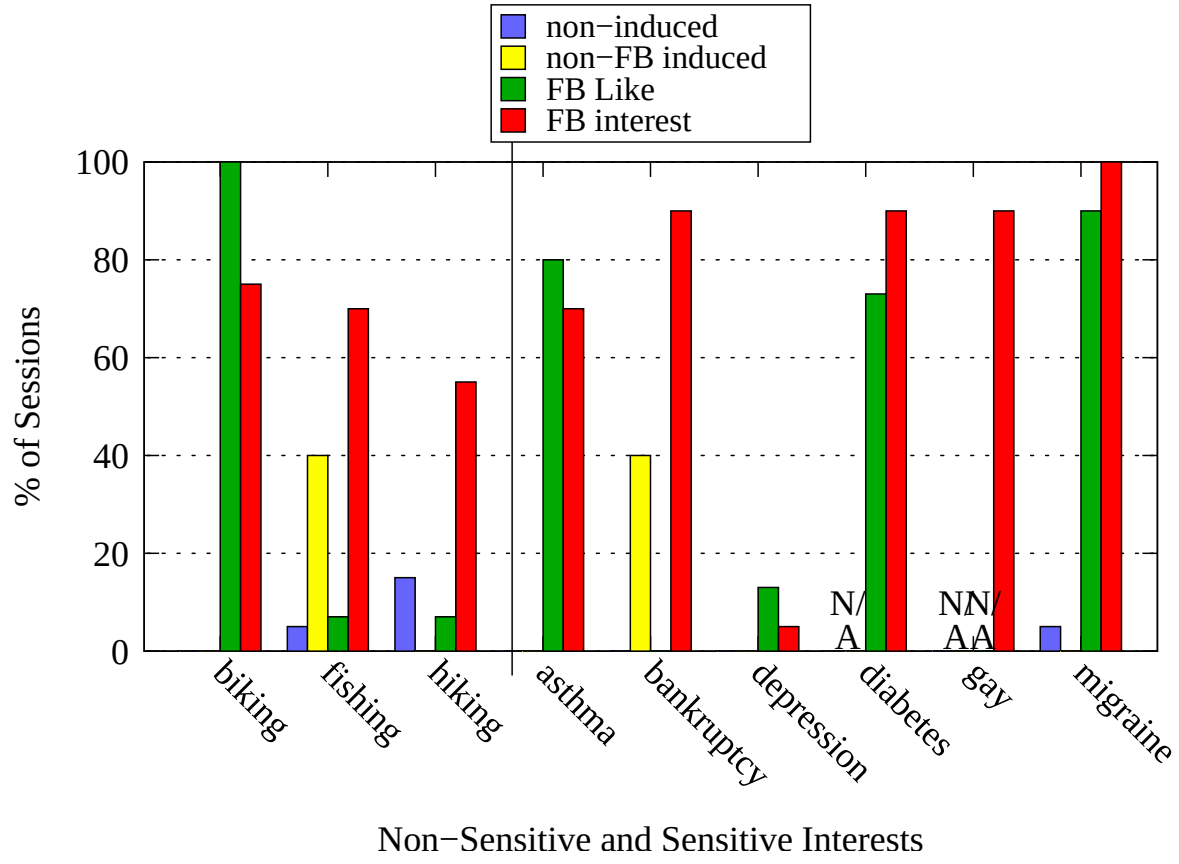
2. interest induced by visiting a site containing Facebook Like button, but button not clicked on;

3. interest induced by visiting a site containing Facebook Like button and clicking on button; and

4. no sites visited, but interest induced as a Facebook interest.

Facebook Ad Based on a User's Interest

Percentage of Sessions Displaying Facebook Ad Matching Induced Interest



No clear evidence of Facebook ads shown based on non-Facebook behavior.

Facebook Ads Targeting Diabetes and Sexual Orientation

"Ad text may not assert or imply, directly or indirectly, within the ad content or by targeting, a user's personal characteristics within the following [sensitive] categories: ..."

Ad text *is* targeting personal characteristics.

Facebook Ads Asserting Age, Sexual Orientation and Diabetes

Ads assert personal characteristics by asking if a user has condition or encouraging user to join *others* with this characteristic—these ads violate ad text guidelines as confirmed by Facebook.

We found ads asserting a sensitive characteristic in *each* of the experimental sessions where the interest was induced as a Facebook interest, primarily through ads for diabetes, migraines and sexual orientation.

Summary

- Google ad network generally shows evidence of "expected" behavior. Some "unexpected" evidence of ads for interests not displayed in Ad Preferences Manager as well as evidence of non-contextual ads for sensitive topics—could be construed by users as behavioral.
- No clear evidence of Facebook ads shown based on non-Facebook behavior. Facebook targeting of ads for sensitive interests matches expectations, but appears to not match stated policy. Ads with admitted inappropriate language for sensitive topics were found.
- Have developed a methodology for ongoing monitoring of what third-party advertisers are doing with the information they received. We are in the process of more fully automating it.
- Ongoing methodology is important to verify stated practices of advertisers as well as help consumers understand what is being done with their information.

Roadmap for Presentation

- Longitudinal privacy footprint
- Understanding what is done with information
- Leakage, Linkage and Lifetime

Leakage, Linkage and Lifetime

Unwanted dissemination of private information to a third party requires the presence of three conditions:

1. leakage of information,
2. linkage of information from different sources, and
3. a non-trivial lifetime for the information.

Elimination of any of these three conditions prevents potential harmful privacy loss from occurring.

So what can be done?

Elimination of Data Lifetime

What if lifetime of data can be controlled?

- "Ephemeral messages are incredibly freeing and make people communicate more authentically and freely with their friends." —Sep'13 Communications of the ACM

- Snapchat as an example where pictures last 3-10 seconds.

- Timed revocation (or expiration) of private data using keys from trusted server.

Problems: local caching/snapshots, is data really ephemeral?, trust in a shared entity

Elimination of Linkage

Many vectors:

- Heavy use of tracking cookies by most sites.
- Browser fingerprinting—2013 paper showing it is being used.
- IP Addresses
- Globally unique identifers: email addresses, usernames, social network identifiers (enabling linkage across devices)
- AdID as a replacement for third-party cookies, basically a browser identifier (report of Google activity, Sep'13)

Must control all vectors, not just one.

Leakage—Is It Still Happening?

Pandora (Leaks Info to Doubleclick)

GET http://ad.doubleclick.net/adj/pand.default/...;**artist=S1421673**;
genre=love;**ag=33**;**gnd=1**;**zip=01609**
Host: ad.doubleclick.net
Referer: http://www.pandora.com/radioAdEmbed.html?cb=...
Cookie: id=223d4200013312||t=1292486411|et=730|cs=p999khn4

AARP (Leaks Email, First Name and Zip to Omniture)

GET http://metrics.aarp.org/b/ss/aarpglobal/...
Host: metrics.aarp.org
Referer: http://www.aarp.org/
Cookie: ...e=**jdoe@email.com**&f=**John**&...&p=**01609**...

Current Leakage of Sensitive Info, Which Can Be Linked

Expedia (Leaks Itinerary to Doubleclick and Bluekai)

(and Casalemedia and Scorecardresearch and Adtechus and Adnxs and Pointroll and ...) (but is this just "unintentional" leakage?)

GET http://tags.bluekai.com/...u2=**SFO**&phint=u3=**BOS**&
phint=u4=**20131016**—**20131018**...
Host: tags.bluekai.com
Referer: http://2588797.fls.doubleclick.net/...u2=**SFO**&phint=u3=**BOS**&
phint=u4=**20131016**—**20131018**...
Cookie: bkc=KJh5NkkQQaWDOabc4aIg7Ej+...

WebMD (Leaks Health Search Term to Scorecardresearch)

GET http://b.scorecardresearch.com/...www.webmd.com/
search/...?query=**pancreatic+cancer**
Host: b.scorecardresearch.com
Referer: http://www.webmd.com/search/...?query=**pancreatic+cancer**
Cookie: UID=25a44ab3-208.27.224.13-1333356933;;...

Elimination of Leakage

First-party sites are often in the best position to prevent leakage of information.

Users can block content, but existing tools are one-size-fits-all (or used as such) where users do not understand what is being blocked.

Pagefair reports (Aug'13) an average ad blocking rate of 23% and growing. PageFair has software to measure ad blocking and help sites educate visitors.

What Can a User Do?

1. Tension between users wanting to protect information and aggregators wanting to collect it—more information leads to better ad targeting.
2. Ads pay for content, this is a way for users to see more relevant ads.
3. First-party sites are often in the best position to prevent leakage of information.
4. Users can make it harder for aggregators by refusing their cookies and blocking their content—can be done via browser settings and extensions, but may impact page display/function.
5. Need to look at semantic solutions where users can better understand and control what happens with their information.
6. Users should minimize information given to sites—cannot leak what they do not know!

Internet Privacy: Then and Now

"On the Internet, nobody knows you're a dog."