Privacy leakage on the Internet

Balachander Krishnamurthy

AT&T Labs–Research

`http://www.research.att.com/~bala/papers`

Joint work with Craig E. Wills, `http://www.cs.wpi.edu/~cew`

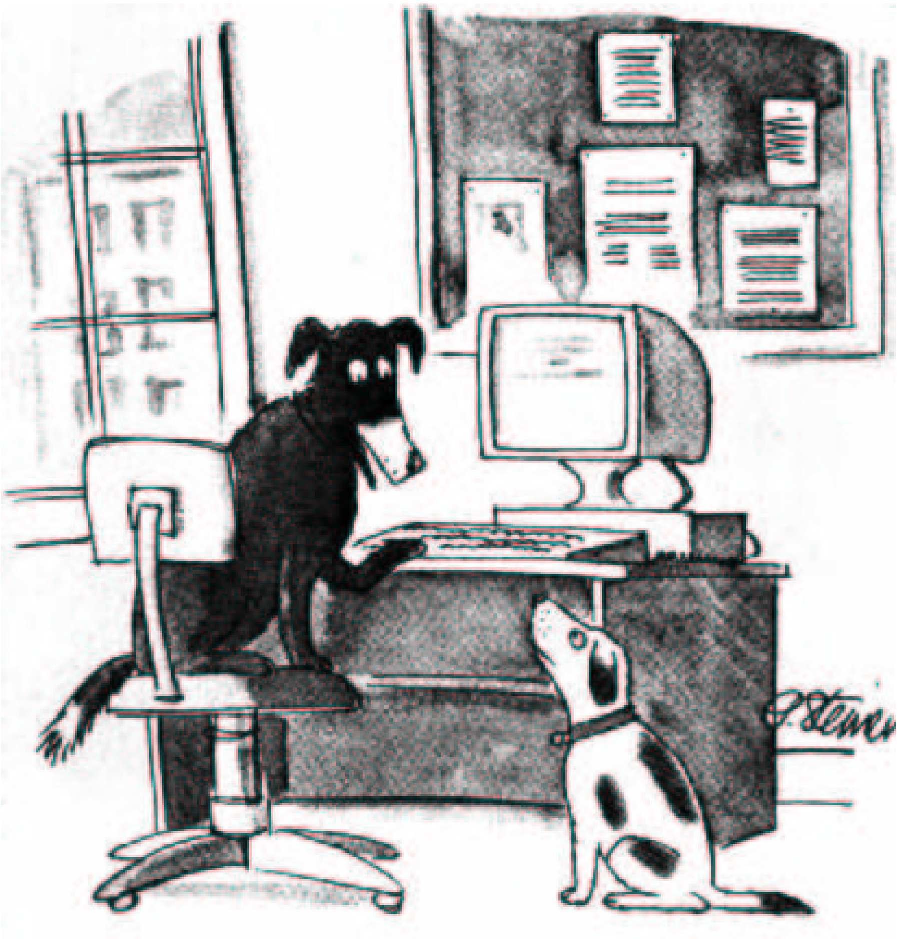
Talk outline

1. Privacy footprint: a longitudinal study report

2. Personally identifiable information leakage in Online Social Networks

3. Some IETF mumbling

July 5 1993, New Yorker, Peter Steiner's cartoon



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

Sadly, this cartoon is out of date.

Internet and Web Privacy

- Security is about keeping *unwanted* traffic from entering our network
- Privacy is about keeping *wanted* information from leaving our network
  Privacy is thus the dual of security
- Privacy can be examined at user-, organizational-, ISP-level
- Higher awareness due to e-commerce, new demographics (e.g., children) identity theft, and Online Social Networks.

Should we care about privacy?

- Depends on the information disseminated, ability to combine external data, what data collectors *might* do with it

- We need to know *what* information is being diffused, *who* is tracking it, and *how*

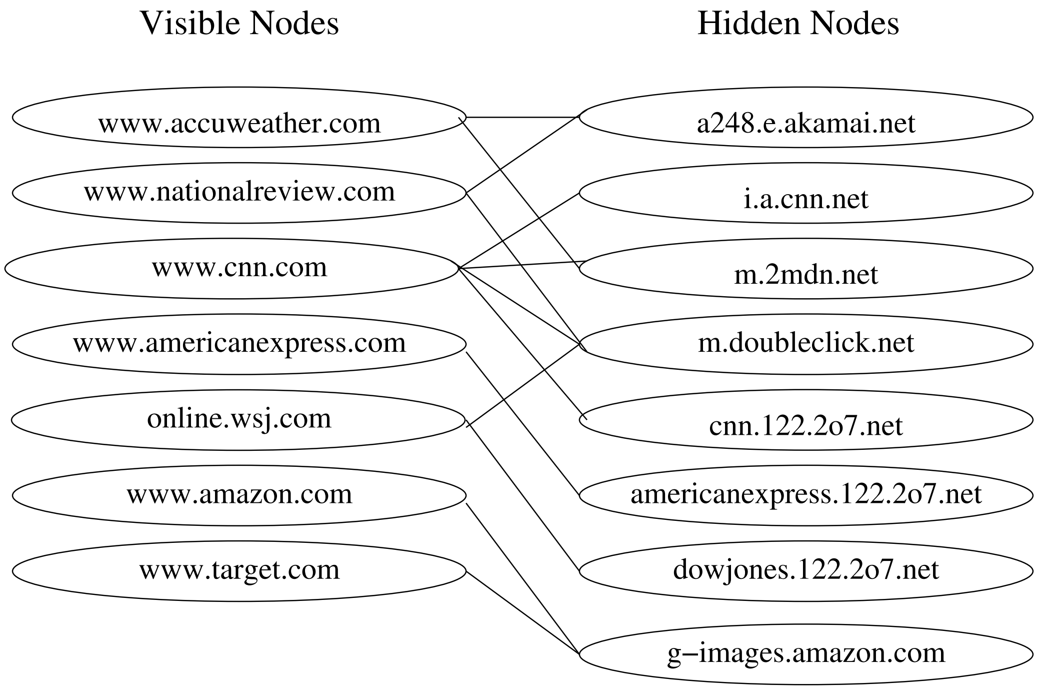
Goal is to allow standard network activity while preserving desired privacy

Privacy footprint

- Various daily interactions on the Web (commerce, email, search...):
- Sites use many techniques to track users (1x1 pixel Web bugs, tracking cookies, JavaScript)
- Aggregators track across sites (`dclk, googlesyndication, tacoda`)
- Privacy footprint: measure of dissemination of user-related information across *unrelated* sites

First-party vs. Third-Party nodes

Connections between first-party visible (servers explicitly visited) and hidden third-party (visited as by-product) nodes



Third parties

1. Ad Networks: First-party sites (publishers) arrange with ad networks to place ads on their pages via images or javascript code.
   E.g., Google's Adsense (googlesyndication.com, doubleclick.net),
   AOL (advertising.com, tacoda.net), Yahoo!(yieldmanager.net)

2. Analytics companies: measure traffic, characterize users by downloading a JavaScript file and send back information in a URL.
   E.g., google-analytics.com (urchin.js), 2o7.net (Omniture),
   atdmt.com (Microsoft/aquantive), quantserve.com (Quantcast)

3. CDNs: Serve images, rarely JavaScript. e.g., akamai.net, yimg.com

Privacy leaks to all of them.

Mechanics of our data collection

- Visible nodes: Popular 1200 Web sites in dozen Alexa categories

- Extracted hidden nodes corresponding to each visible node via a Firefox extension that fetches objects and records request/response

- Tests of popular Web sites in 68 countries and 19 languages.

- Examined cookies, JavaScript, identifying URLs (those with ? $=$ &)

- Narrowed examination to *consumer* and *fiduciary* sites: subset of sites that raise more privacy concerns.

- Study carried out nine times over a five year period:
  Oct '05, April/Oct '06, Feb/Sep '08, March/June/Sept '09, March '10

Node association

Two visible nodes are *associated* if accessing them results in accessing the same hidden node.

Association can be due to several reasons:

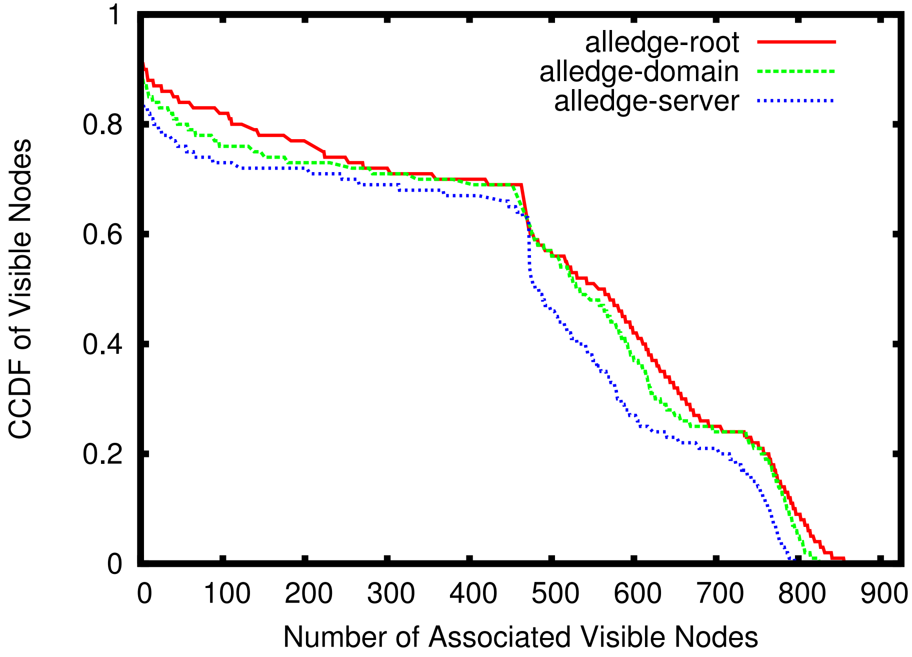
1. server: Identical server name (`www.google-analytics.com`)
2. domain: Aggregated by merging hidden nodes with same 2nd-level domain names. E.g. `cnn.112.2o7.net` and `dowjones.112.2o7.net`
3. adns: Aggregated by merging hidden nodes that share the same ADNS (authoritative DNS server). e.g. `doubleclick.net` and `ebayobjects.com` have the same ADNS. (try dig ... NS)

Cleaning up domain association

- DNS for third-party servers may be provided by sites like ultradns.net
- CDNs are increasingly used to serve content for *third* party servers (e.g., JavaScript or images with cookies)
- We check ADNS of 3d-party and 1st-party servers—if they differ and the ADNS server is *not* that of a known CDN or DNS service, we use the 3d-party server as the domain
- e.g. pixel.quantserve.com's ADNS is akamai, so *root* domain is quantserve.com, but w88.go.com's *root* domain is omniture.com (based on its ADNS).
- Root domain: identifies the root cause of the origin for each server

Association: Common hidden node between two visible nodes

CCDF of number of other visible nodes associated with each visible node



X-axis: Single visible node's *maximal* association: (www.vonage.com)
Server: 813 (75%), Domain: 850 (78%), ADNS: 885 (81%) of 1086 nodes.

Y-axis: Degree of association: 87% server, 91% domain, 94% ADNS
75% of all visible nodes are associated with over 100 visible nodes

Cumulative count of unique associated visible nodes

Some visible nodes are associated via more than one hidden node.

E.g., (www.cnn.com, online.wsj.com) with (doubleclick.net, 2o7.net) domains

Top-10 associated ADNS nodes connected to 78.5% of visible nodes
doubleclick.net, google-analytics.com, 2mdn.net, quantserve.com, scorecardresearch.com, atdmt.com, omniture.com, googlesyndication.com, yieldmanager.com,2o7.net

Merging holding companies: Google, Omniture, MSFT, Yahoo, etc.

OK to focus on these.

Hidden Nodes in 68 countries (older data)

Hidden nodes appearing in at least 20% of Per-Country Top-10 Lists

Hidden Node	Number of Appearances in Country Top-10 Hidden Node List (%)
google-analytics.com	61 (90%)
yahoo.com	58 (85%)
yimg.com	47 (69%)
googlesyndication.com	44 (65%)
doubleclick.net	39 (57%)
2o7.net	31 (46%)
atdmt.com	24 (35%)
2mdn.net	22 (32%)
statcounter.com	15 (22%)
imrworldwide.com	14 (21%)
adbrite.com	14 (21%)

Google is thus present in 90% of countries' top-10 lists.

Hidden Nodes in 19 languages Top-100 Lists (older data)

French, Italian, Portugese, Spanish, English, German, Dutch, Greek,

Danish, Norwegian, Finnish, Swedish,

Arabic, Turkish, Czech, Russian,

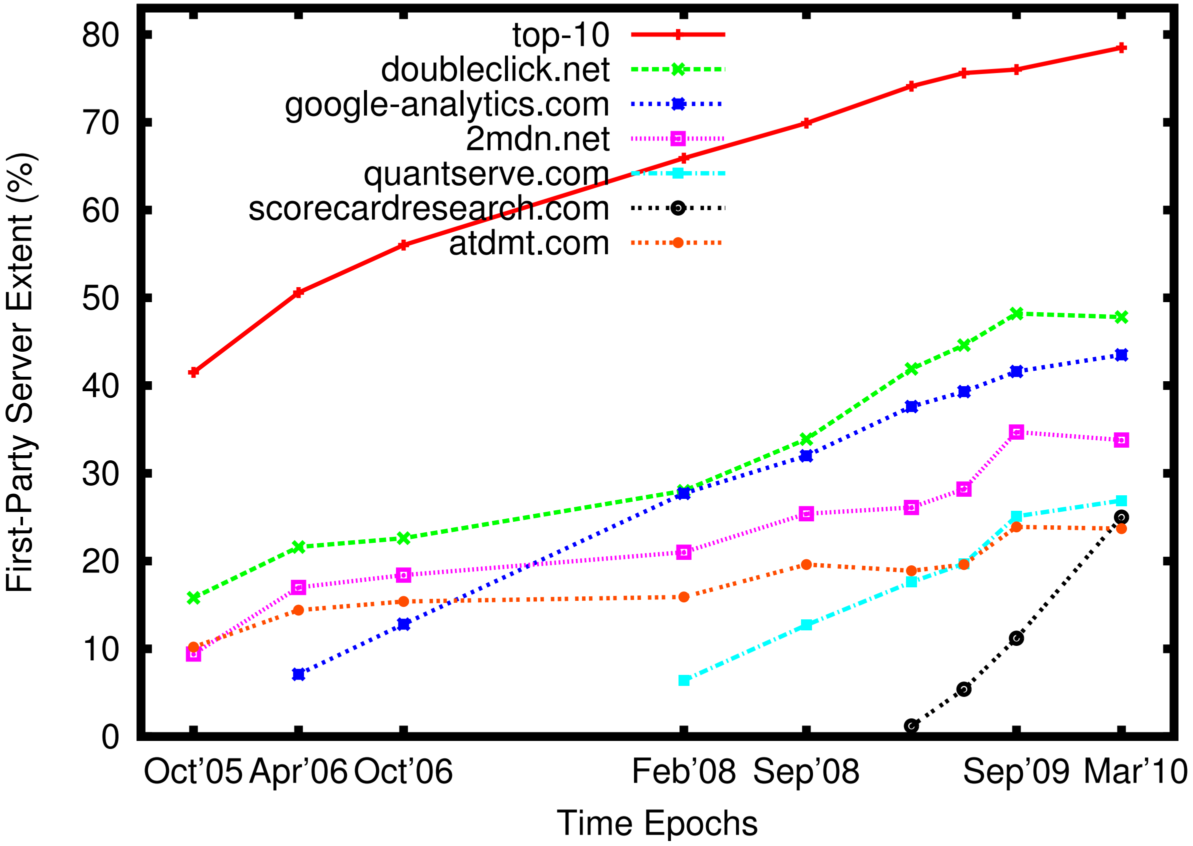
Korean, Japanese, Chinese.

Weighted average of three footprint metrics: visible nodes association range from 76% to 92%.

Privacy footprint: longitudinal study

- Footprint shows the number and diversity of 3d-party sites visited as a result of a user visiting first party sites.
- We examine the penetration of the top 3d-party domains that aggregate information about user's movements on the Web
- Multiple 3d-parties may track users on a given first-party site and so this is examined as well
- Finally, we examine the role of economic acquisitions of aggregator companies that buy others and increase their tracking ability

Top 3d-party domains over time



Combined impact of the top-10 domains: up from 40% to nearly 80%.

Manner of tracking

Initially just 3d-party cookies, but now through 1st-party cookies and JavaScript. We examined traces of requested objects, cookies and JavaScript downloaded.

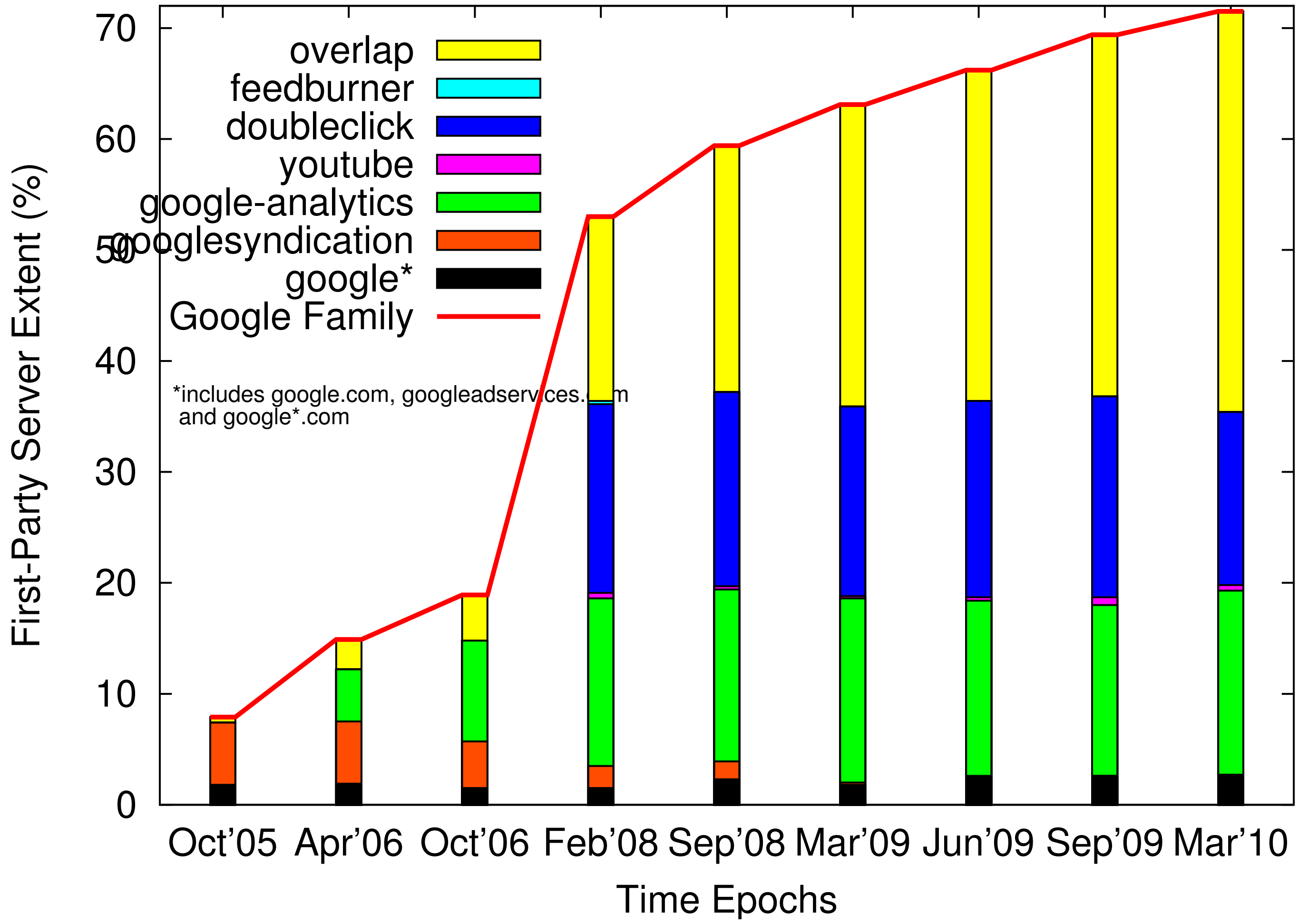
Four categories of 3d-party domains:

1. Only set 3d-party cookies, no JS (dclk, atdmt, 2o7.net)
2. Use JS with state saved in 1st-party cookies (google-analytics: urchin.js examines 1st-party cookies, forces retrieval via an identifying URL to send information to 3d-party server)
3. Both 3d-party cookies and JS to set 1st-party cookies (quantserve)
4. 3d-party cookies and JS not used to set 1st-party cookies but instead serve ad URLs with tracking information (adbrite, adbureau)

Situation grimmer in the face of acquisitions

Family	Acquired	Date
AOL	advertising.com tacoda.net, adsonar.com	Jun'04 Jul'07/Dec'07
Doubleclick	falkag.net	Mar'06
Google	youtube.com ($1.65B) doubleclick.net ($3.1B) feedburner.com,admobs.com ($750M)	Oct'06 Mar'07 Jun'07/Nov '09
Microsoft	aquantive.com (atdmt.com, $6B)	May'07
Omniture	offermatica.com visual sciences (hitbox.com, $0.4B)	Sep'07 Oct'07
Valueclick	mediaplex.com fastclick.net	Oct'01 Sep'05
Yahoo	overture.com ($1.6B) yieldmanager.com, adrevolver.com	Dec'03 Apr'07/Oct'07
Adobe	Omniture ($1.8B)	Sept '09

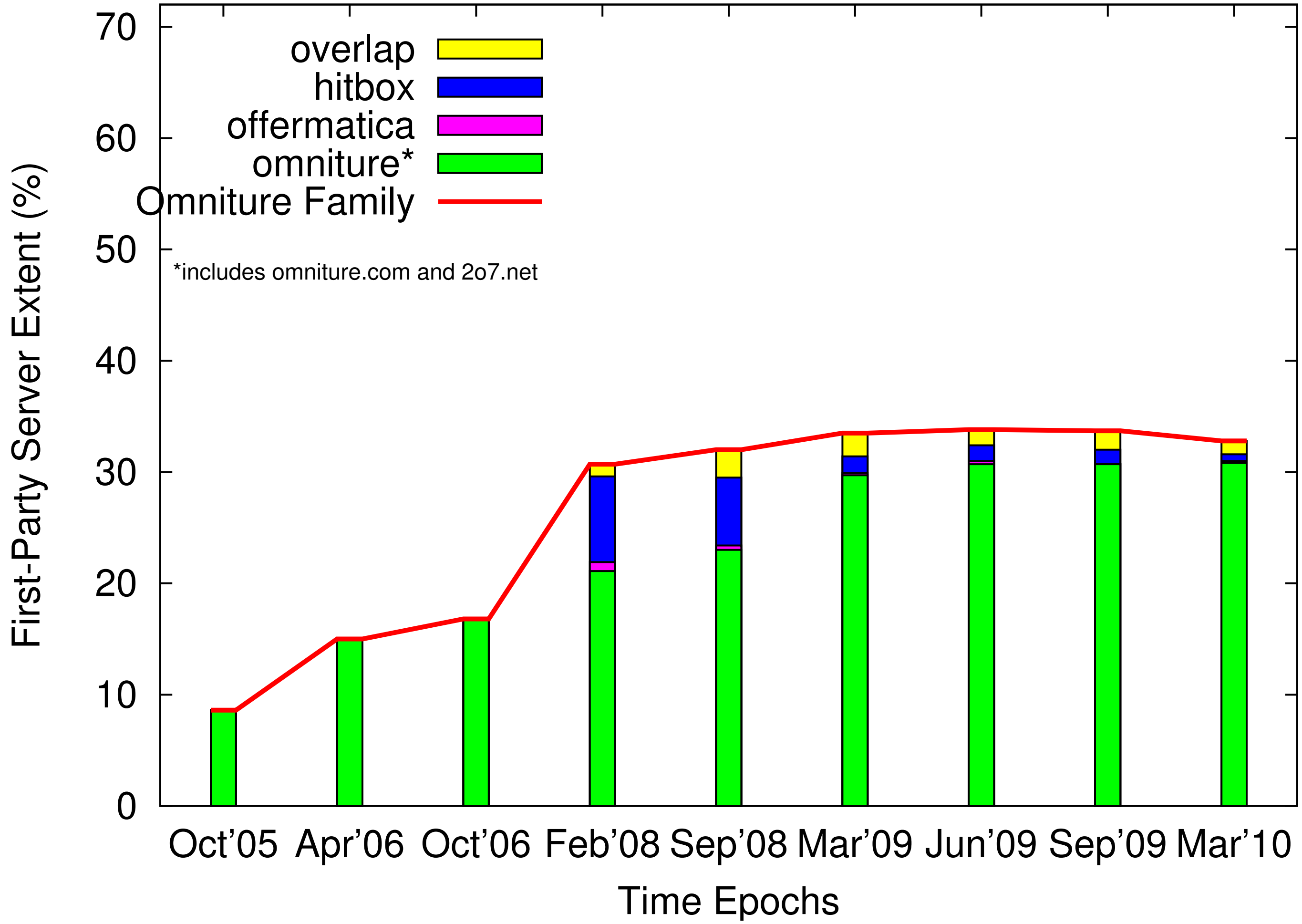
Family 1: Growth of Google Family



includes google.com, googleadservices.com and google.com

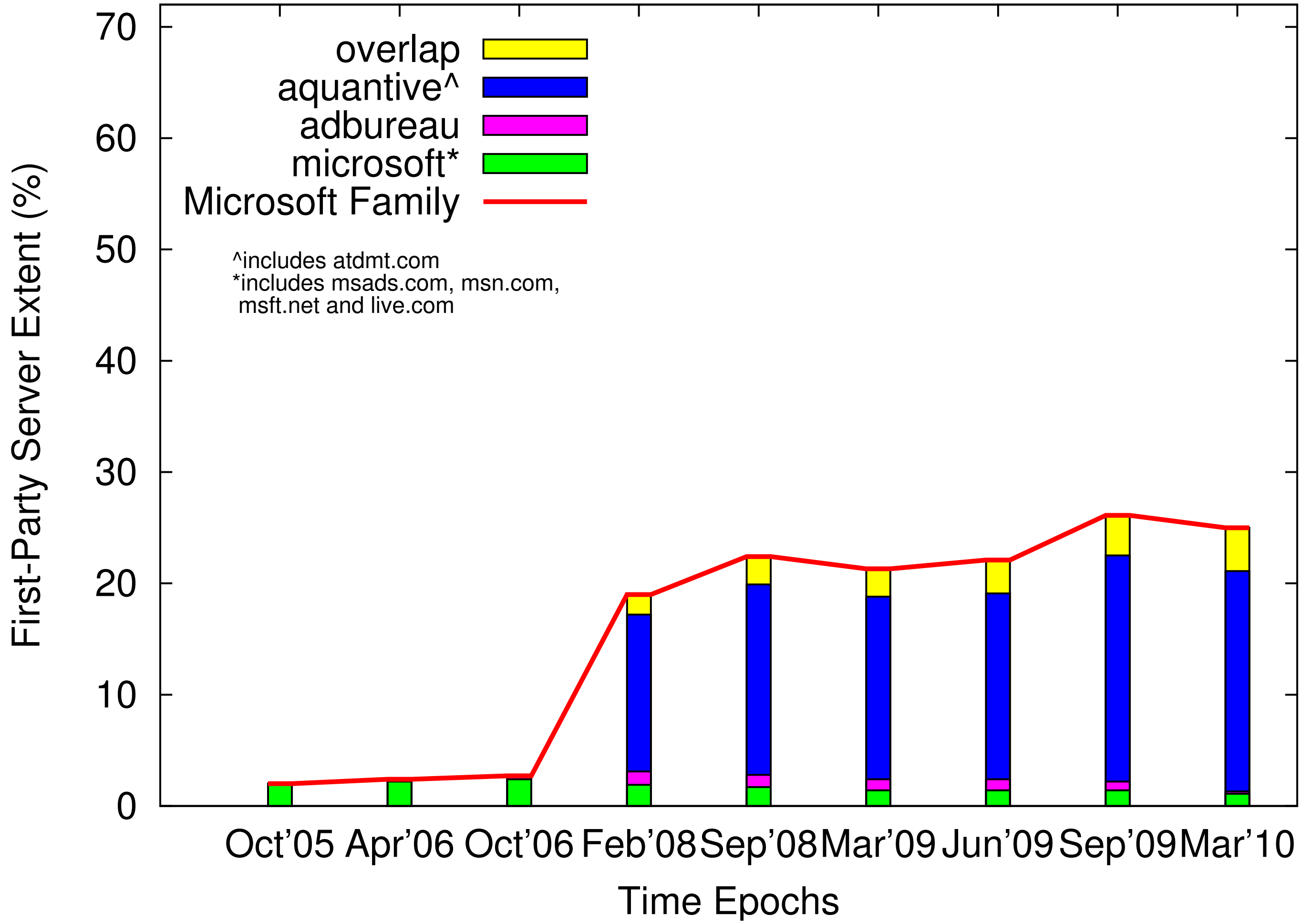
Sep'09 Google family reach: over 70%—highest among all third parties by far.

Family 2: Growth of the Adobe (ne Omniture) Family



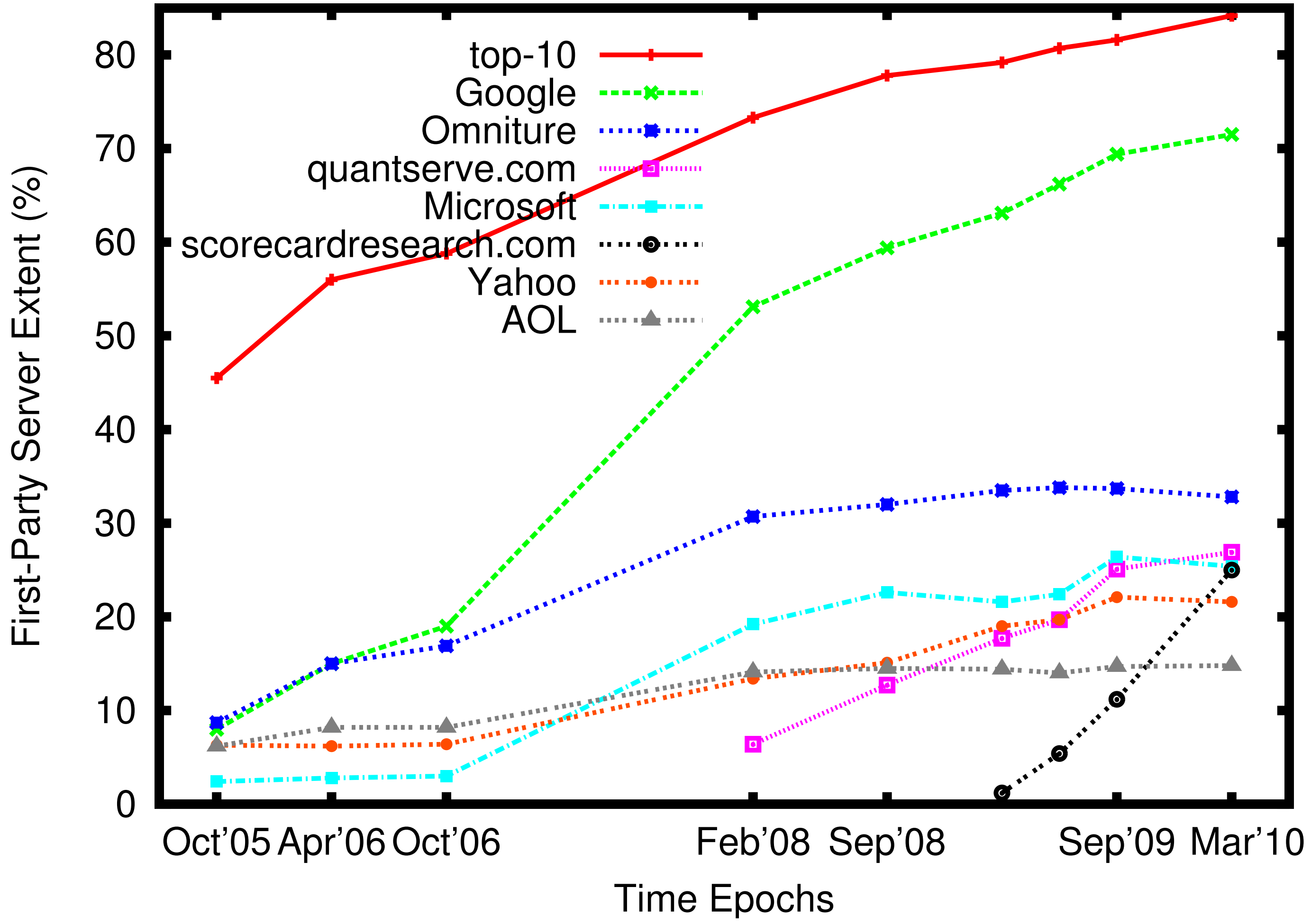
Primarily 2o7.net domain and then acquisitions–reach of over 30%
Adobe acquired Omniture in September '09

Family 3: Reach of the Microsoft Family



Reach of over 25% in Sep'09, growth from buying Aquantive (atdmt.com).
Other families: Yahoo: 22%, AOL: 14% in Sep'09

Top-10 Family Growth



Extent of top-10 families cross 84% in Mar'10

Depth of Tracking has also increased

Users are being tracked by two or more third-party entities.

- In Oct'05, 24% of 1200 popular Web sites contained more than one of the top-10 3d-party domains.
- In Sep'08 this figure had risen to 52% (34% with more than two 3d-parties).
- It is not enough just to block a single tracking entity.

Consumer sites

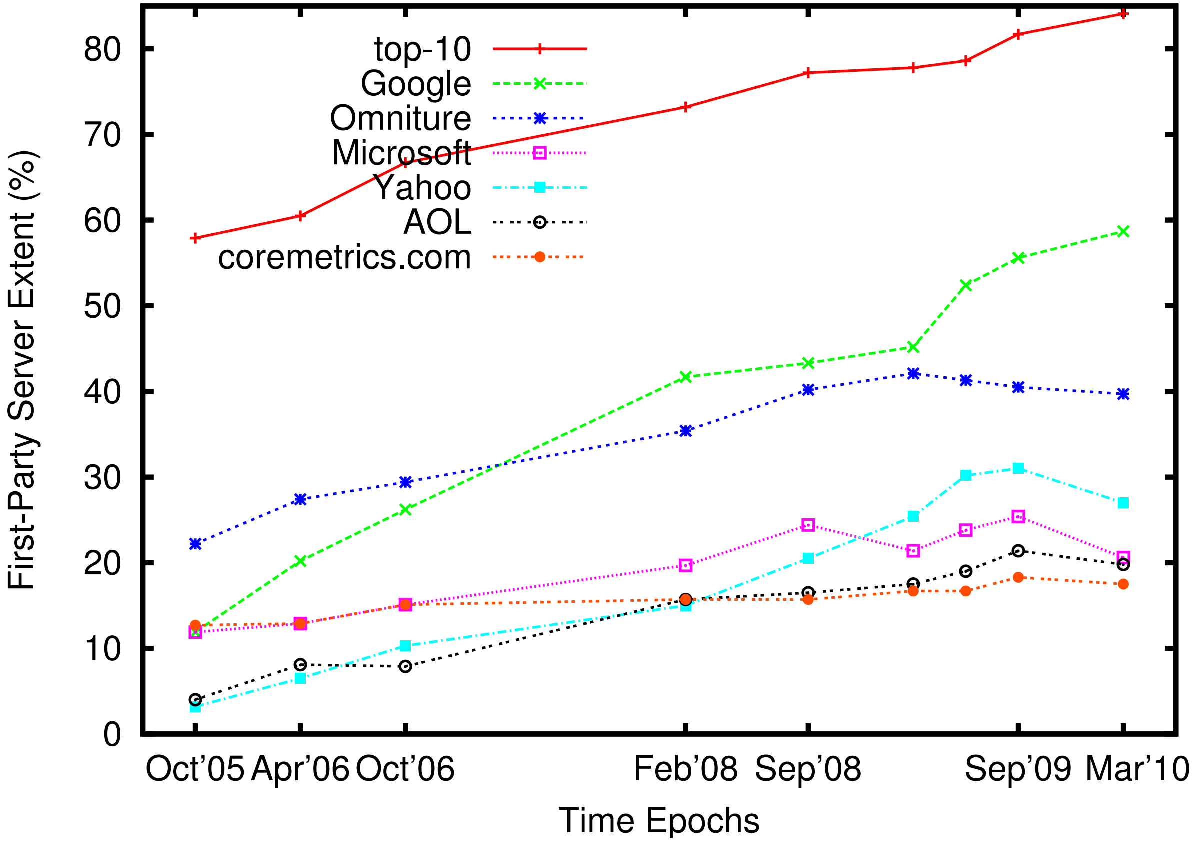
Examined 127 consumer sites' longitudinal privacy leakage.

E.g., apple, blockbuster, buy, ebay, expedia, gap, hilton, ikea, kayak, netflix...

Steadily increasing node associations:
Oct '05: 58%, Oct '06: 66%, Feb '08: 74%, Sep '09: 80, Mar'10: 84.1%

Top aggregator families:
google, omniture, yahoo, microsoft, aol, coremetrics, quantserve

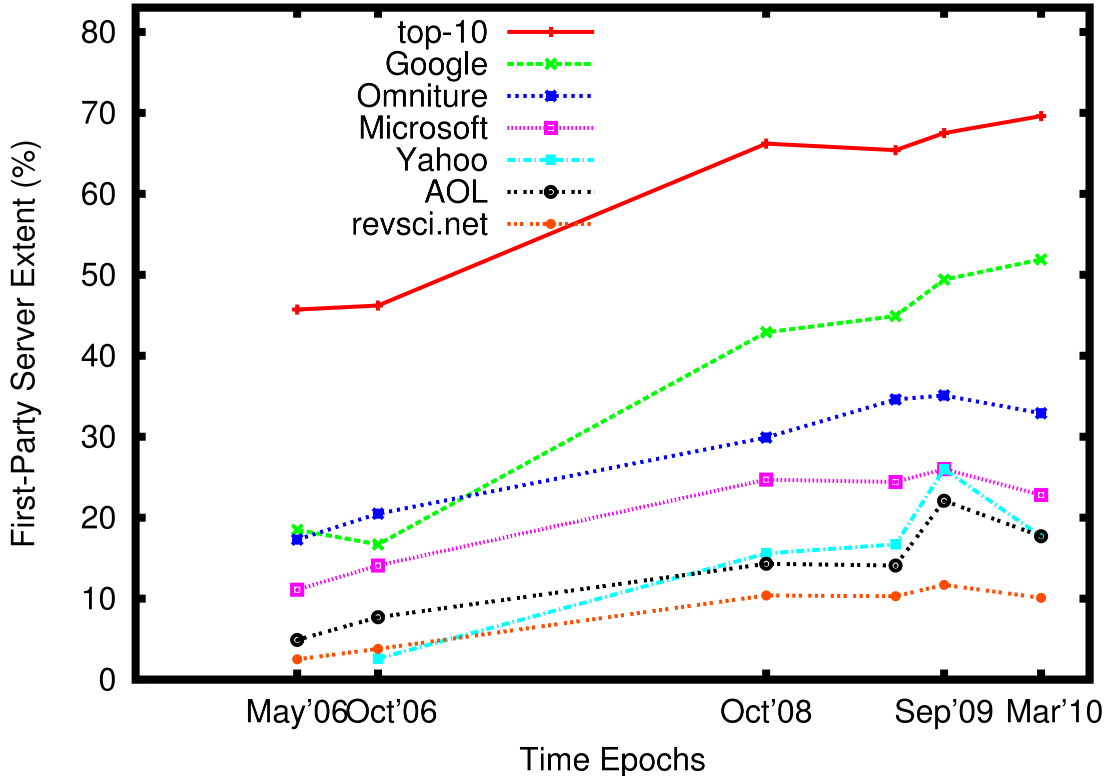
Top-10 3d-party families in Consumer sites over time



Top-10 domains account for over 84%. Google family is largest since '08.

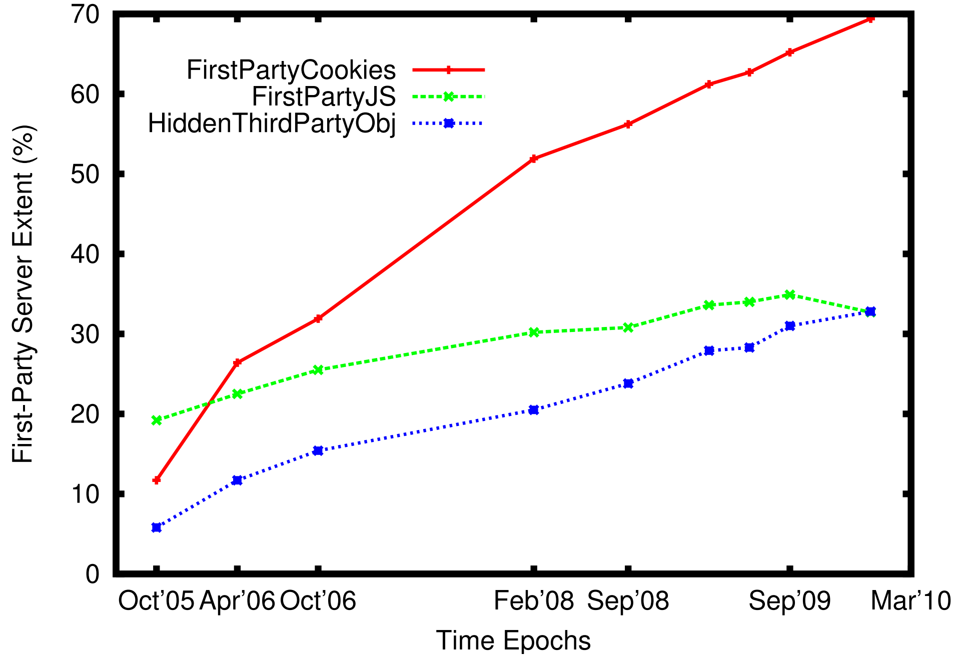
Top-10 3d-Party families in Fiduciary sites over time

81 sites in 9 categories:
credit financial insurance medical mortgage shopping subscription travel utility



Top-10 families account for over 70%.

Growth of Hidden Third-Party Content



3d-party aggregators are using *1st-party* cookies to track users via 3d-party JavaScript - 70%. Can't reject all 1st-party cookies..

3d-party JavaScript served by 1st-party server: cannot auto block - over 30%.

20% have 3d-party objects "hidden" in seemingly 1st-party servers (Omniture's JS on abc.go.com: ident URL for w88.go.com, ADNS shows it is in Omniture)

Privacy protection measures

Currently one can disable cookies and JavaScript execution, filter ads, and block images.

Not directly available today, but possible to filter *all* third-party objects, remove JavaScript content, filter requests with *identifying* URLs or objects from top aggregation servers.

Page rendering may break Best techniques: *no3obj*, *no3js*, *noaggregators*
Worst technique: *noimg* (See our SOUPS07 paper for details)

Proxies can also help in obscuring users to some extent.

https://panopticlick.eff.org rates browser configuratiions

Some recent external developments: IE

- IE 8.0's InPrivate Filtering (nee InPrivate Browsing), lets users block downloading of content that appear on a specific number of websites

- Such a "line of sight" blocking targets specific .js files being downloaded when user visits different sites.

- On average 41% of 3rd-party domains accessed are in the top-10 domain set and half of these set cookies.

Privacy policies/leakage/default profiling

- cuil.com's simple privacy policy
- Chrome: URL completion (Suggest) leaks *any* URLs to Google *by default*
- Switch to Iron?
  `http://www.srware.net/en/software_srware_iron_chrome_vs_iron.php`
  (Client-ID, Timestamp, Suggest, Error Reporting, RLZ-Tracking, Updater, URL-Tracker)
- Google toolbar on *by default* on every Dell sold
  `http://news.cnet.com/Dell-embraces-Google/2100-1032_3-6077051.html`
- Specific Media (175M individual profiles)
  `http://www.washingtonpost.com/wp-dyn/content/article/`
  `2008/08/29/AR2008082903178_2.html`

Part 2: Leakage of PII in OSNs

Lots of ‘vague’ talk about user and privacy loss until now.

Aggregators: We only know IP address, no PII about user is ever recorded.

Executive Excerpt from June 2008 article by Saul Hansell, NYT

> ”Google is quick to point out that some of these systems are not connected to each other. And most of the information it gets is not what is generally considered to be personally identifiable, like a name or e-mail address.”

http://bits.blogs.nytimes.com/2008/06/26/
google-tests-using-your-search-data-to-tailor-ads-to-you

Well, they certainly have the opportunity to do so...

Personally Identifiable Information

OMB memorandum ”Safeguarding Against and Responding to the Breach of Personally Identifiable Information”

http://www.whitehouse.gov/omb/memoranda/fy2007/m07-16.pdf

- Information which can be used to distinguish or trace an individual's identity. e.g., name, social security number, biometric records.
- Alone or when combined with other personal or identifying information
- Linked or linkable to a specific individual. e.g. date and place of birth, mother's maiden name, ...

Longer list of what constitutes PII

1. Name (full name, maiden name, mother's maiden name)

2. Personal ID number (e.g., SSN), address (street/email), telephone numbers

3. Personal characteristics (photo of face, X-ray, fingerprint, biometric image: retina scan, voice signature, facial geometry)

4. Asset information (IP or MAC address, persistent static ID that consistently links to a particular person or a small, well-defined group

5. Information identifying personally owned property (vehicle registration/VIN)

6. Linked/linkable information to any of the above: date/place of birth, race, religion, activities, or employment/medical/education/financial information

Well-known result in linking pieces of PII: *most Americans (87%) can be uniquely identified from a birth date, zip code, and gender* (Sweeney)

Pieces of PII in OSNs

Users are specifically asked for these as part of their OSN profile

1. Name (first and last)
2. Location (city and zip code), address (street/email)
3. Telephone numbers
4. Photos (both personal and collections)
5. Linkable: gender, birthday, age, birth year, schools, employer, friends, activities

Not all profile elements are filled in by users; entries may be false. We did not parse contents of OSN users' pages.

12 OSNs studied: Bebo, Digg, Facebook, Friendster, Hi5, Imeem, LinkedIn, LiveJournal, MySpace, Orkut, Twitter and Xanga.

Degree of availability of PII (to OSN users) in 12 OSNs

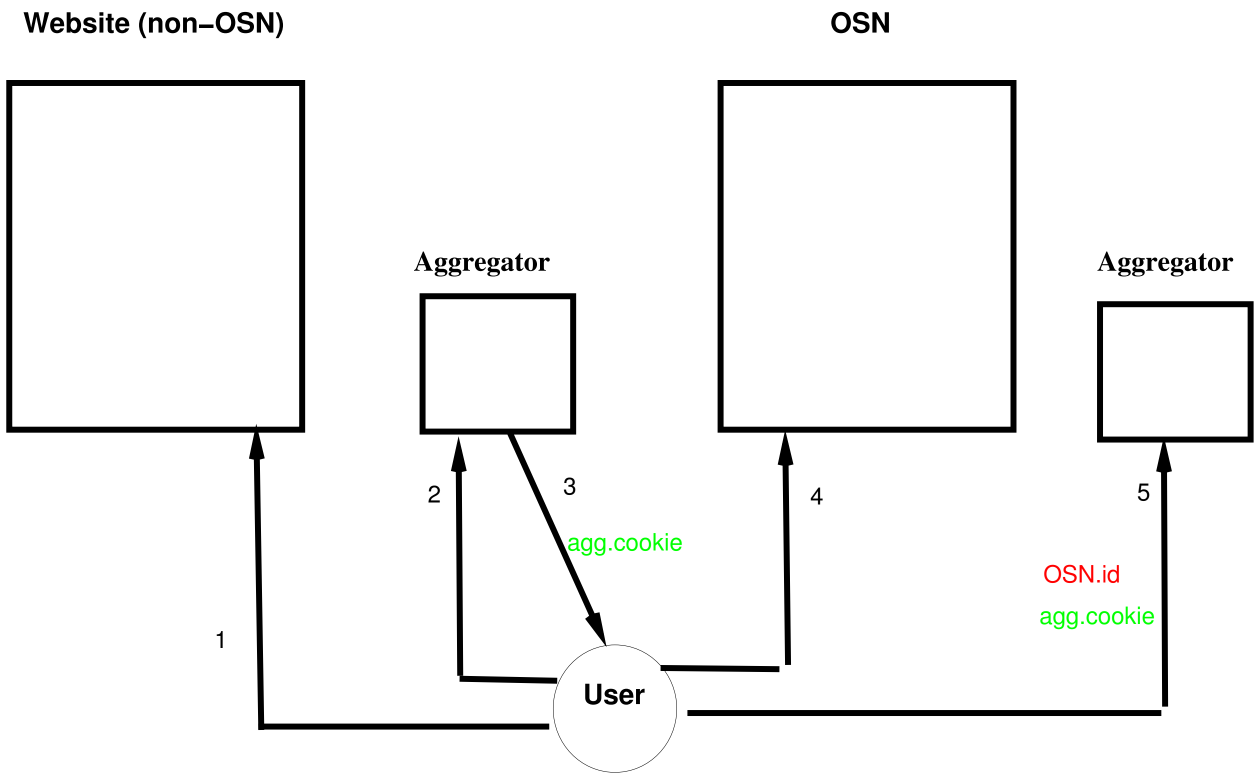
Piece of PII	Always Available	Available by default	Unavailable by default	Always Unavailable
Personal Photo	9	2	1	0
Location	5	7	0	0
Gender	4	6	0	2
Name	5	6	1	0
Friends	1	10	1	0
Activities	2	8	0	2
Photo Set	0	9	0	3
Age/Birth Year	2	5	4	1
Schools	0	8	1	3
Employer	0	6	1	5
Birthday	0	4	7	1
Zip Code	0	0	10	2
Email Address	0	0	12	0
Phone Number	0	0	6	6
Street Address	0	0	4	8

Entries are counts of OSNs; columns go from bad to good wrt privacy concerns.

Source of leakage

- OSNs assign unique IDs for their users that may be displayed as part of URL when user navigates around the OSN
- If the ID stays *within* the OSN, it is not a problem
- However, ID is ‘leaked’ to multiple outsiders, including 3d-party aggregators
- The ID, in conjunction with the aggregator’s tracking cookie leads to the actual privacy leakage
- The *same* tracking cookie is sent to the aggregator when the user visits other sites that trigger connections to the aggregator

Simple illustration



Aggregator knows who went to (or may go to) non-OSN sites as well

Typical sequence of actions to trigger leakage

- Purely *internal* actions within an OSN – e.g., user clicks on a list of friends.
- Action that results in an ad being downloaded from an aggregator site
- Clicking on an ad

Different actions result in OSN ID leakage in different ways.

Technical manners of leakage

At least four broad categories of leakage

- OSN identifier (pointer to PII) via HTTP headers
- OSN identifier through external applications
- Specific pieces of PII
- Linkages across OSNs and non-OSNs

Category 1: OSN ID leakage via HTTP headers

1. via Referer (sic) header (9 of 12 OSNs), problem noted in RFC 1945, May ’96

```
GET /link/click?lid=43000000170958623 HTTP/1.1
Host: clickserve.dartsearch.net
Referer: http://www.facebook.com/profile.php?id=123456789&ref=name
```

2. via Request-URI (5 of 12 OSNs)

```
GET /utm.gif?...&utmp=utmhn=twitter.com&utmp=/profile/jdoe...
Host: www.google-analytics.com
Referer: http://twitter.com/jdoe
```

3. via Cookie (2 of 12 OSNs)

```
GET ...g=http://digg.com/users/jdoe...
Host: z.digg.com
Referer: http://digg.com/users/jdoe
Cookie: s_sq=...http://digg.com/users/jdoe...
```

Users can potentially block 1 and 3, but not 2 easily

Category 2: OSN ID leakage via external applications

OSNs warn users that their information will be given to external applications. These in turn use ads and can hand out user's ID to aggregators. The direct source of leakage here are external applications that run on non-OSN servers.

1. Via Referer Header (MySpace external application "iLike")

```
GET /TLC/...
Host: view.atdmt.com
Referer: http://delb.opt.fimserve.com/adopt/..&puid=123456789&..
Cookie: AA002=123-456/789;...//
```

OSN ID leakage via external applications (contd.)

2. Via Request-URI (Facebook external application “iLike”)

```
GET /...&utmhn=www.ilike.com&utmr=http://fb.ilike.com/facebook/
    auto_playlist_search?name=Springsteen&..fb_sig_user=123456789&..
Host: www.google-analytics.com
Referer: http://www.ilike.com/player?app=fb&url=http://
www.ilike.com/player/..._artistname/q=Springsteen
```

3. Via Request-URI and Cookie (Facebook external application: Kickmania!)

```
GET /track/?...&fb_sig_time=1236041837.35&fb_sig_user=123456789&..
Host: adtracker.socialmedia.com
Referer: http://apps.facebook.com/kick_ass/...
Cookie: fbuserid=123456789;...=blog.socialmedia.com..cookname=anon; cookid=594...074;
```

Category 3: Direct leakage of specific pieces of PII

1. Age and gender via Request-URI

```
GET /show?gender=M&age=29&country=US&language=en...
Host: ads.sixapart.com
Referer: http://jdoe.livejournal.com/profile
```

2. Age, gender, zipcode and email via Request-URI and Cookie

```
GET /st?ad_type=iframe&age=29&gender=M&e=&zip=11301&...
Host: ad.hi5.com
Referer: http://www.hi5.com/friend/profile/displaySameProfile.do?userid=123456789
Cookie: LoginInfo=M_A—US_0_11;Userid=123456789;Email=jdoe@email.com
```

The hi5 example is in clear contravention of their own privacy policy `http://www.hi5.com/friend/displayPrivacy.do` as of October 1, 2009

Category 4: Linkages across OSNs and non-OSNs

- A user on two different OSNs may leak ID and thus be linked across OSNs
- A user moving through list of friends may leak friends' OSN id and thus aggregator could know some of the friends.
- A user visiting an external non-OSN Web site could have their action linked with their OSN PII (see example below)

Example of third-party cookie for non-OSN server:

```
GET /pagead/ads?client=ca-primedia-premium_js&...
Host: googleads.g.doubleclick.net
Referer: http://pregnancy.about.com
Cookie: id=2015bdfb9ec—t=1234359834—et=730—cs=7aepmsks
```

The same Cookie is sent to doubleclick.net when the user is on a OSN and the OSN ID is leaked.

What can aggregators do with PII

- Tracking cookie from any other site is trivially linkable with OSN user
- Visits to non-OSN websites in the *past* and *future* can be linked with the information
- Searches are identifiable potentially with a person assuming OSN ID is not falsified

Note that aggregators *may* have contractual agreements not to exploit data that they may have access to as a result of actions by users on OSNs.

Protection against leakage

- Users: block Referer header and third-party Cookie headers, filter for all OSNs URI's with appropriate ID syntax (latter is problematic)
- Aggregators: could ignore PII related information...
- OSNs: strip ID from all headers, internally remap IDs

What about mobile OSNs?

Ongoing study

- Studied 20 popular mOSNs
- Similar problems with additional issues: location, presence, connections to traditional OSNs
- Device unique ID leaking, privacy settings on mOSNs and connected traditional OSNs (FB, Twitter) are often different

Recent examples of other uses of OSN information

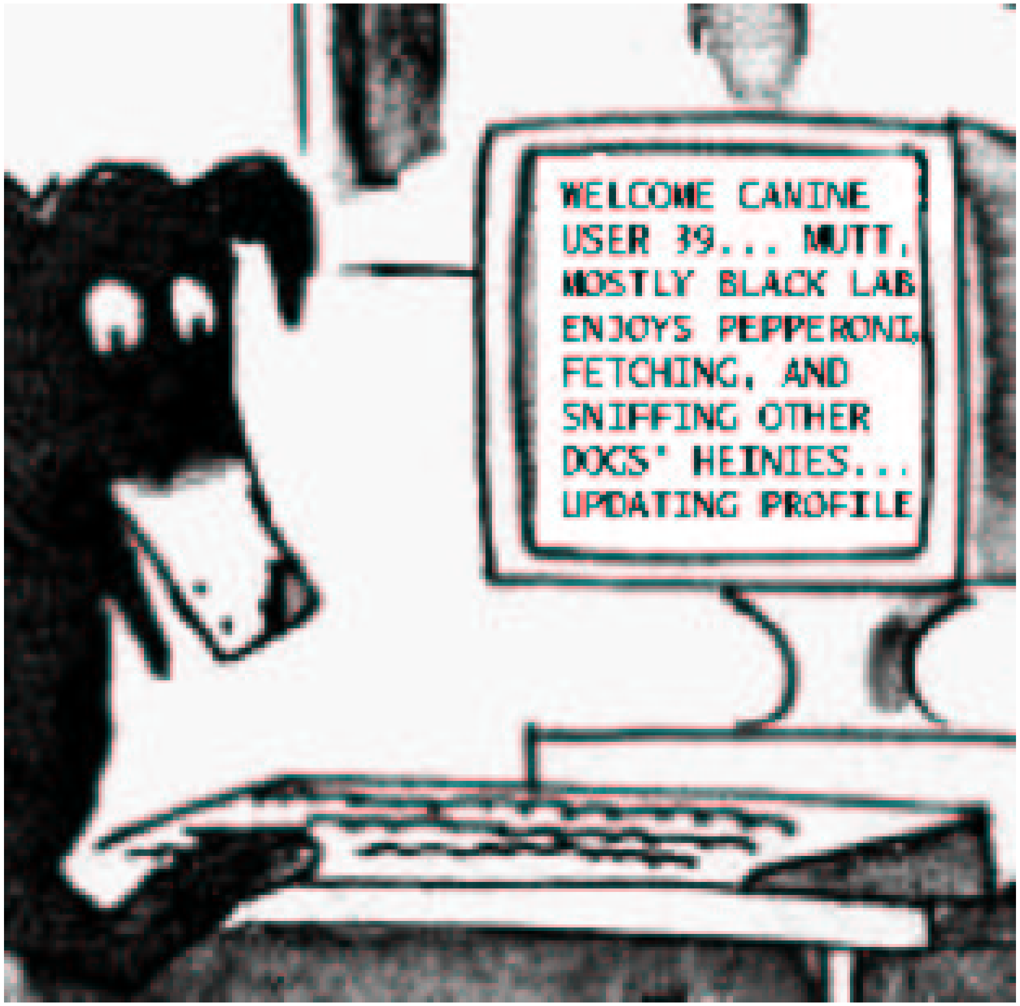
Several legal cases settled with information from Facebook:

- Arrest in Pennsylvania: Jonathan Parker, 19, of Fort Loudoun, Pa on 8/28/09, while committing a daytime burglary checked his FB status on the computer of the house he broke into and left himself logged in.

- Arrest in Mexico: 10/14/09 Maxi Sopo, a 26-year-old criminal (bank fraud in Seattle) hiding in Cancun updated his Facebook status to say "having a good time" and also made the "elementary error" of friending a former justice department official - faces 30 years - without access to FB.

- Exoneration in Harlem: Rodney Bradford 11:49 a.m. 10/17/09: "wherer my i hop" from a computer at an apt at 71 W 118th St. Arrested/exonerated when FB confirmed status update time.

Unknown author's cartoon



IETF and privacy

- Privacy is not an entirely new concept to IETF
- Multiple efforts related to privacy; e.g., Geopriv, identity management
- Broadly three efforts in IM: OAUTH, OpenID, SAML – all dealing with managing and protecting a user's identity.
- RFC 4505 SASL Simple Authentication and Security Layer - - an application framework to generalize authentication.
- Related recent drafts discuss how identity providers may log information about relying parties (issue of tracking visits, potential collusion between different RPs).
- Possibility of a privacy consideration section in relevant IDs/RFCs? (sure, says mnot, for httpbis, if i write it..)

Some recent IETF discussions re privacy

Two topics:

- Blue sheet retention – retaining list of attendees at IETF meetings that might result in subpoenas (realization that IETF has no privacy policy). Should old blue sheets be destroyed? Scott Bradner refers to list of attendee requirement from RFC lore.
- Controversy over use of RFID tag reader as an experiment in addition to blue sheets.

Usual issues: opt-in vs. opt-out (most recently seen in Buzz)

Ongoing discussion - no final conlusions reached yet

Thanks to Allison Mankin and Lucy Lynch for their valuable pointers.

Thanks to Hannes Tschofenig and unnamed colleague for comments on slides.

Summary

- An overview of privacy leakage
- Specific look at leakage on popular OSNs
- Strong concern about concentration of significant amounts of information in a few hands
- Things not necessarily getting better in newer technologies
- IETF might want to start a discussion on the privacy issue