Privacy leakage on the Internet

Balachander Krishnamurthy

AT&T Labs–Research

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

Talk outline

1. Privacy footprint
2. Personally identifiable information (PII) leakage in Online Social Networks
3. PII leakage in $mobile$ OSNs
4. Current status and where do we go from here

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

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

Sadly, this cartoon became out of date even before the advent of OSNs.

Privacy

- The right of privacy is a fundamental right in the Universal Declaration of Human Rights
- Some disagree: libertarians, too-late-to-do-anything crowd who repeatedly announce its death (McNeely, O'Reilly, Schmidt, Mark Z)
- Many (often uninformed) enable (over?)sharing – DNA (23andme), credit card purchases (blippy), everything else (OSNs)
- Kundera: police destroy it in Communist countries, journalists threaten it in democratic countries, ...people themselves lose their taste for private life and their sense of it

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

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)

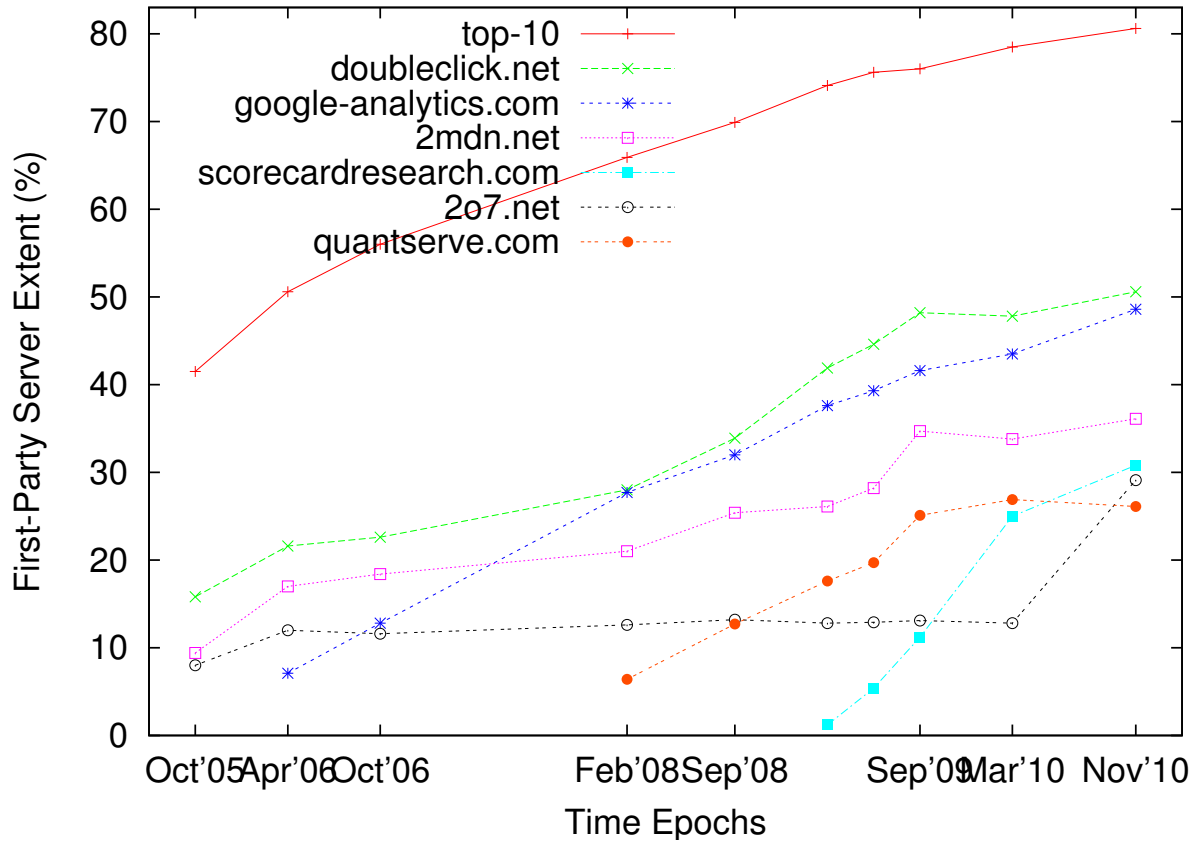
3. CDNs: Serve images, rarely JavaScript. e.g., akamai.net, yimg.com

Privacy leaks to all of them.

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

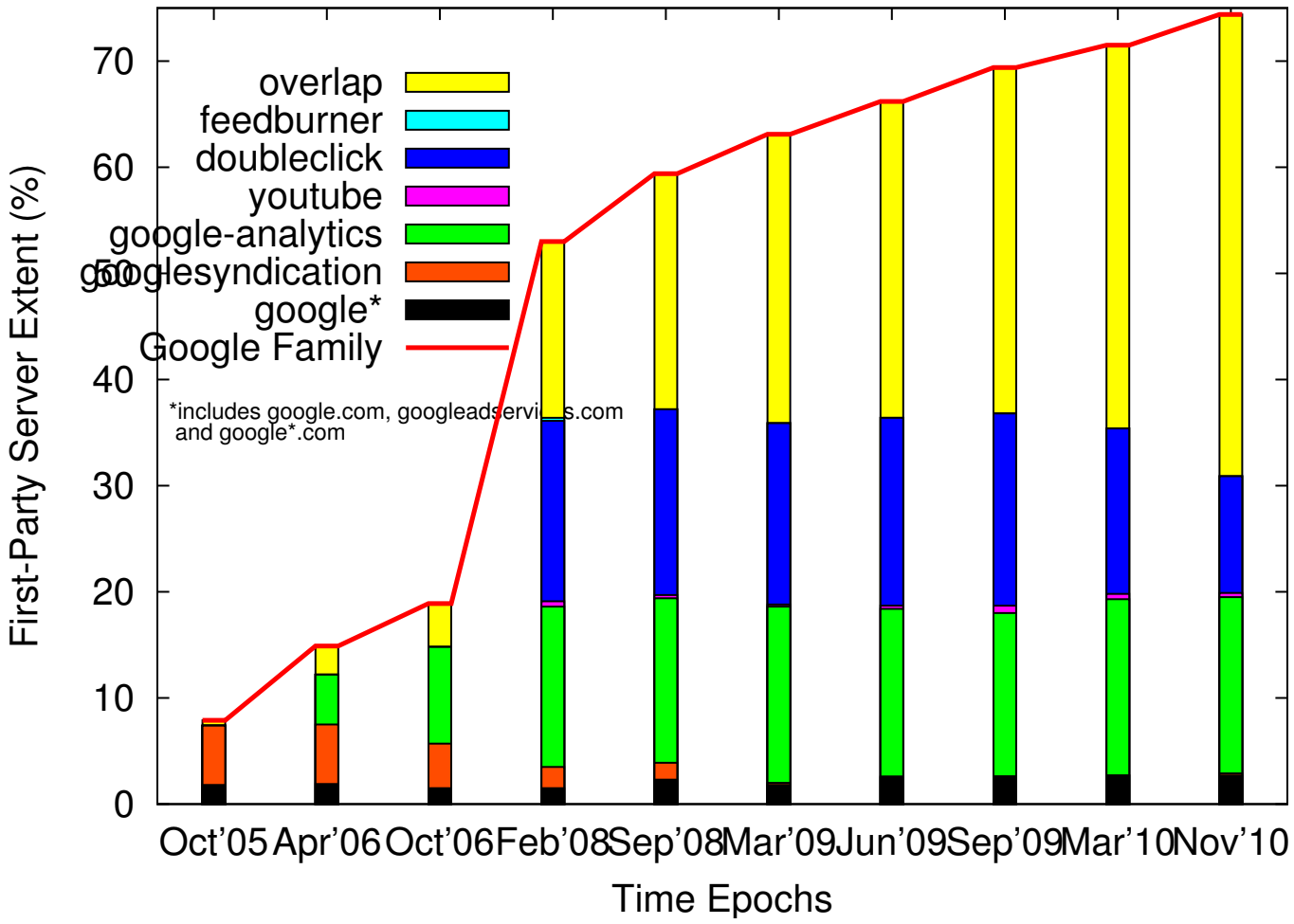


Combined impact of the top-10 domains: up from 40% to nearly 80%.

Situation grimmer in the face of acquisitions

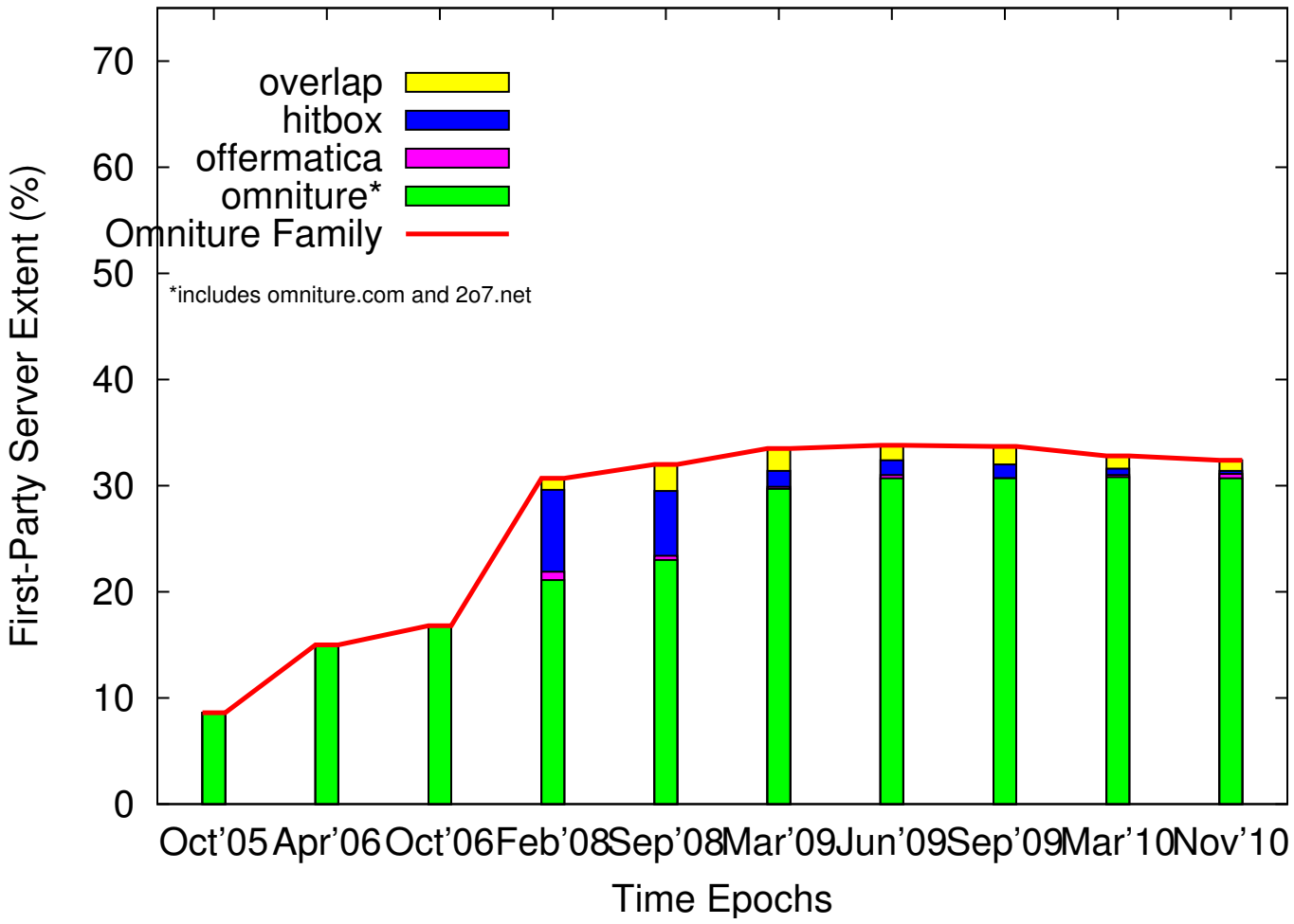
Family	Acquired	Date
AOL	advertising.com	Jun'04
	tacoda.net, adsonar.com	Jul'07/Dec'07
Doubleclick	falkag.net	Mar'06
Google	youtube.com ($1.65B)	Oct'06
	doubleclick.net ($3.1B)	Mar'07
	feedburner.com,admob.com ($750M)	Jun'07/Nov '09
Microsoft	aquantive.com (atdmt.com, $6B)	May'07
Omniture	offermatica.com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

Growth of Google Family



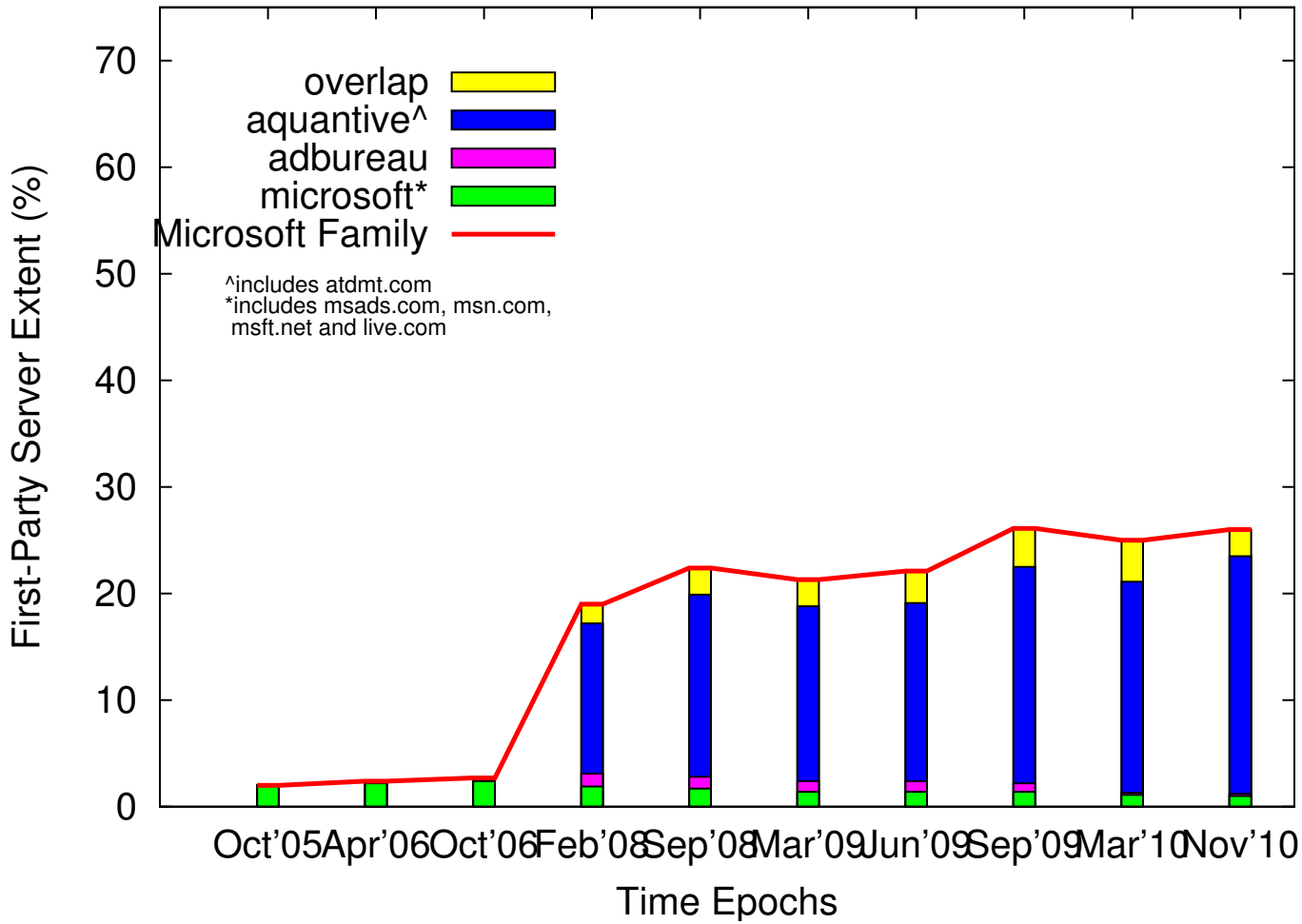
Nov'10 Google family reach: nearly 75%—highest among all third parties by far.

Growth of Adobe/Omniture Family

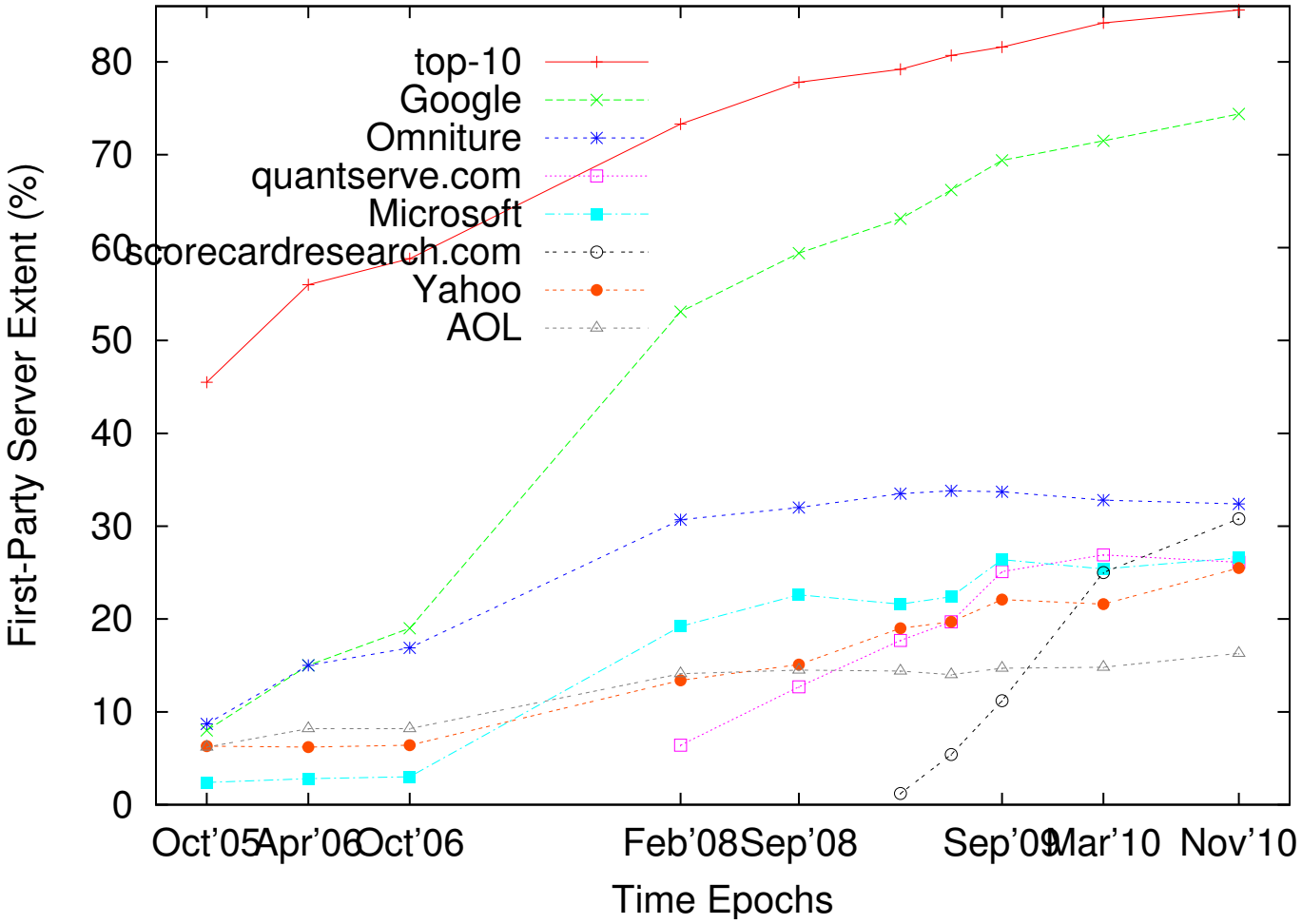


Adobe also sets Flash cookies (hard to delete, respawn)

Growth of Microsoft Family



Top-10 Family Growth



Extent of top-10 families cross 85% in Nov'10

Leakage of PII in OSNs

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... 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

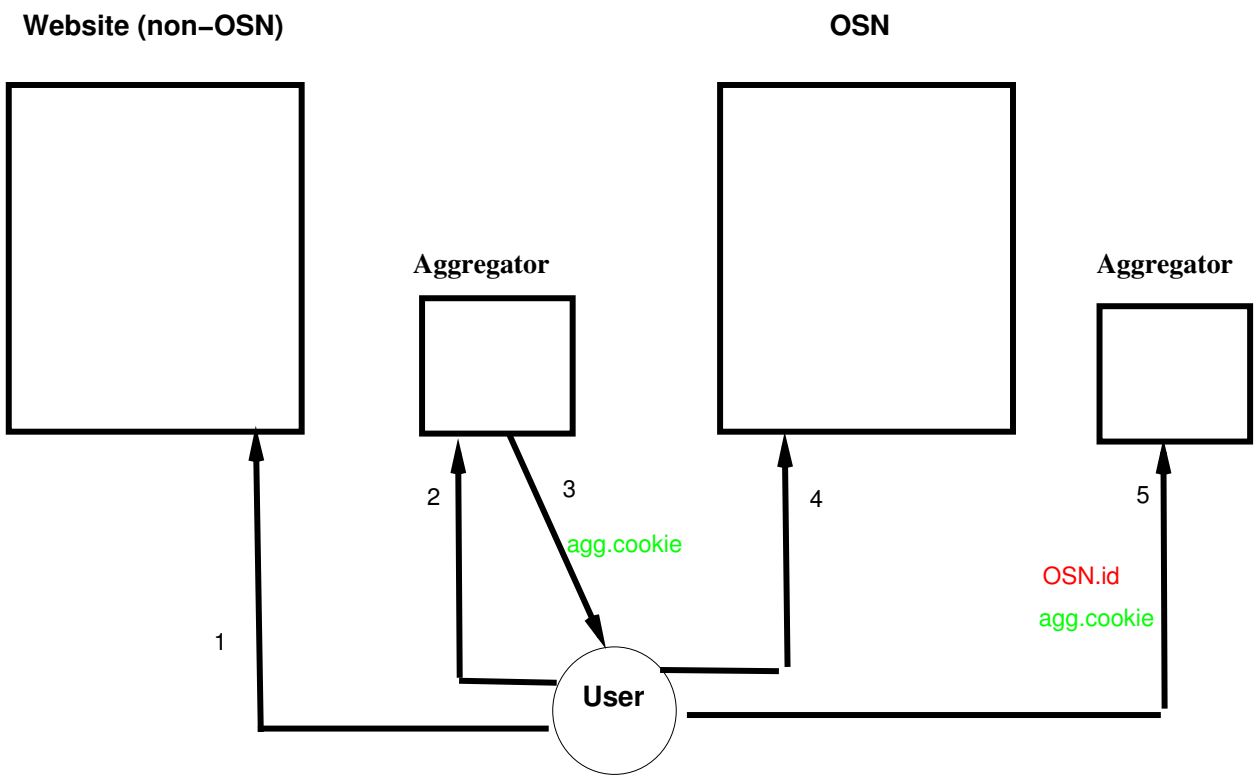
Piece of PII	Always Available	Available by default	Unavailable by default	Always Unavailable
Personal Photo	9	2	1	0
Location	5	7	0	0
Gender	4	6	0	2
Name	5	6	1	0
Friends	1	10	1	0
Activities	2	8	0	2
Photo Set	0	9	0	3
Age/Birth Year	2	5	4	1
Schools	0	8	1	3
Employer	0	6	1	5
Birthday	0	4	7	1
Zip Code	0	0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

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.

What about mobile OSNs?

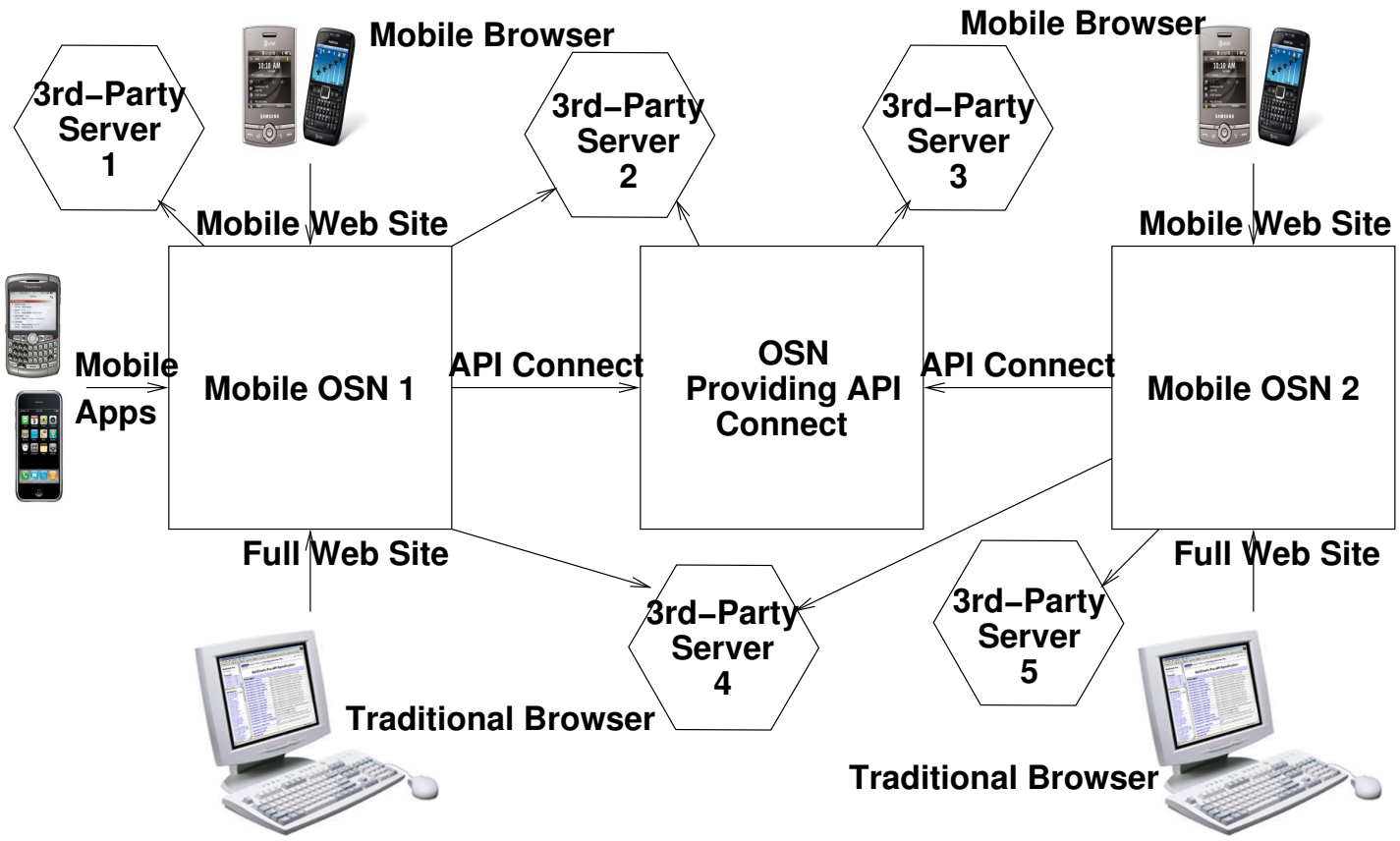
Studied 20 popular mOSNs

- Bebo, Facebook, Hi5, Linkedin, Livejournal, MySpace and Twitter (traditional OSNs)
- Flickr, Yelp (special-purpose social networks),
- Brightkite, Buzzd, Dopplr, Foursquare, Gowalla, Gypsii, Loopt, Radar, Urbanspoon, Wattpad and Whrrl (11 new mOSNs)

Already more than 1 in 4 accesses to Facebook are from mobile devices.

Mobile access to Web sites designed specifically for mobile devices and through the development of "apps" that are specific to a site and to a mobile platform.

Interfaces and Interconnections for mOSNs



Interfaces include full Web site, mobile Web site and mobile app. Connections with OSNs and third-parties.

New concerns in mobile OSNs

Many mOSNs have a "check-in" mechanism—both establishes a user's *presence* on the mOSN and the user's current *location*.

Mobile devices typically have a unique device identifier, which is often used as verification for installing approved apps on a user's mobile device.

If this unique identifier is leaked to a third-party via an application and can be associated with a user's identity, this becomes a privacy problem.

Linkage between mOSNs and traditional OSNs: information shared with a mOSN connected to a traditional OSN is also shared with that OSN

Study methodology

Examined each available interface for 19 of the mOSNs (all but Hi5) using the Apple iPhone application as the app interface.

Used a Web proxy to capture HTTP traffic for full and mobile Web sites.

Almost all applications also used HTTP for communication, which was capturable with proxy. Network sniffer revealed no leakage for non-HTTP traffic.

Recorded multiple sessions for each mOSN interface while performing actions appropriate for the given interface.

Availability of PII Pieces to Users in 13 new mOSNs

These 13 mOSNs request and make available less information about each user in comparison to OSNs previously studied.

Each mOSN allows the sharing of information to be controlled by a user via the full Web site interface of the mOSN.

Only a minority of these mOSNs provide any privacy settings via the mobile and mobile application interfaces.

Location and Presence

Seven traditional OSNs: 5 provide means to post public comments (presence), but only one allow a user to establish a current location.

Of the 13 others:

Information	Level of Availability			
	Always Available	Available by default	Unavailable by default	Always Unavailable
Check-In Location	3	3	2	5
Comments	4	7	1	1

Location available by default to all mOSN users for roughly half of the mOSNs with comments available by default for most mOSNs.

Interconnectedness of mOSNs

Number of mOSNs with Connections to Twitter, Facebook and Flickr (out of 12 excluding seven traditional plus Flickr):

- Twitter—10
- Facebook—8
- Flickr—2

Actions in mOSNs passed through to these OSNs, but not privacy controls.

A user's location may now be posted to all on Facebook or Twitter.

Example Leakage of OSN Identifier

```
GET /e0?rt=1&amp;...
Host: p.admob.com
Referer: http://buzzd.com/m/buzz/.../id/OSN-ID
Cookie: uuid=ef07qb76f47b19173389f27a9ae1d391
```

Via Referer Field of Buzzd Mobile Web Site

Example of Direct PII Leakage

```
GET /ad_source.php?d[gender]=m...
Host: r.admob.com
X-Admob-Isu: IPHONE-UDID
Cookie: uuid=ef07qb76f47b19173389f27a9ae1d391
```

Direct PII Leakage to a Third-Party Via Request-URI of Radar App (since deceased).

Note that admob is owned by Google now.

Example Leakage of User Location

```
POST http://beacon.pinchmedia.com/
Host: beacon.pinchmedia.com
User-Agent: buzzd/2.2.0 CFNetwork/459
   Darwin/10.0.0d3
beacons="did":"IPHONE-UDID",..
   "lat" : "20.00", "lon" : "-70.00"
```

Location Leakage to a Third-Party Via POST from Buzzd App

Observe that the location is shared with a map service by the application interface of eight mOSNs, the mobile Web site of four mOSNs and the full Web site of one mOSN.

Type of Leakage: Explicit Vs. Implicit

Do *not* know intention when leakage occurs, but can classify leakage into two types:

1. *Explicit* if leak via Request-URI or POST. Difficult to prevent unless done so on a per-server basis.

2. *Implicit* if leak via Referer or Cookie header. Possible for user to prevent.

Explicit leakage for 9 of 26 instances of OSN identifier leakage.

All leakage instances of PII pieces and location are explicit leakage.

Example Leakage of Unique Device Identifier to Third-Parties

```
GET /?i=xxxxxxxx-xxxx-...&u=IPHONE-UDID
Host: ads.mobclix.com
User-Agent: Wattpad/1.6.1 CFNetwork/459
   Darwin/10.0.0d3
```

Observed such explicit leakage for six mOSNs.

Some type of private information is leaked to a third-party via all 20 of the mOSNS in our study.

Observations

Leakage problems found earlier in traditional OSNs continue to be a problem with new mOSNs.

New leakages found—location, device identifiers.

Multi-dimensional privacy protection problem for the user: including duration of privacy settings, transitive closure of information arising from connections with traditional OSNs, what information is shared with different OSNs and third parties.

Needs continued monitoring as sites evolve.

Also need to extend app study to a broader set of devices.

Where do we go from here?

No quick resolution but solution will necessarily involve

1. Publicity/raising awareness (e.g., this talk, various papers, interviews)
2. Role of various legislative bodies (FTC, Canadian Privacy Commissioner)
3. Alternative technical solutions (e.g., Private OSNs)
4. Other research in privacy space (increasing steadily)
5. Smart use of economic analysis to bring aggregators into a modus vivendi

My thoughts covering these are outlined in an October 2010 CCR paper
"I know what you will do next summer"
`http://www.research.att.com/~bala/papers/ccr10-priv.pdf`

1. Publicity

- Do people really care about privacy? Yes.
- Strict opt-in is more expensive for newOSNs: Google's Buzz(fail) forced opt-in led to leakage, massive negative publicity, reverse position in 72 hours and settle multi-million dollar class-action lawsuit
- PII leakage paper was largely ignored until WSJ interview (May '10)
- Over 10K media links within a few hours
- Op-ed response by Mark Z in Washington Post 3 days later claiming addressing of all privacy issues...

On the positive side, the Internet Architecture Board, in charge of all Internet protocols, has started a program on privacy after a plenary talk and ran a 2-day workshop in December 2010 with 60 attendees representing diverse viewpoints.

2. Government/legal

- Internet spans too many countries for large-scale impact
- Individual governments can pressure some companies
- Economic concern may be what drives companies to be compliant
- EU laws are generally more stringent
- The U.S. Federal Trade Commission's Do Not Track proposal is pending in front of US Congress
- FTC is the only U.S. federal agency with both consumer protection and competition jurisdiction in broad sectors of the economy; pursues vigorous and effective law enforcement.

FTC December 2010 report

- Privacy policies are too complex and incomprehensible
- Asymmetricity with respect to low cost of invisible data collection and potential harm to consumers
- Post-F2F discussion: 3 FTC round tables held
- Executive summary of report summarizes the problems (page ii)

1. Browsing online: advertisers might collect/share information
2. On OSN sites 3d-party applications have access to your information
3. Location-enabled smartphone applications leak location

1-1 match of results from:

Privacy Diffusion on the Web: A Longitudinal Perspective (WWW'09)

On the Leakage of PII via Online Social Networks (WOSN'09)

Privacy Leakage in Mobile Online Social Networks (WOSN'10)

FTC proposed framework for businesses and policymakers

122 page report advocates:

- Adoption of Canadian Privacy Commissioner's "Privacy by design" model:
  1. Embedding privacy at design stage, pro-actively
  2. Sets default to be private
  3. Points out false choice of security or privacy
  4. Urges transparency (who has user's information and what they plan to do with it) and respect for users
- Providing streamlined, clear privacy choices to consumers to enable informed consent (e.g, Do Not Track approach)
- Access to all user-related sensitive data stored in aggregators

Open for comment; DNT may be technically hard and enforcement harder still

3. Are private OSNs the future?

- pip.io — a social OS with layers of relationship

  'friend' may not be able to see your wall
  rooms: subsets of friends get topic-specific feeds (collaborative narcissism)

- pidder – strong encryption to protect controllable, permanently deletable private data shared with deliberately chosen friends using pseudonyms with reputation and multiple strong passwords

- others include Thefridge (containers for different sets of friends, use that to filter data) and tool providers like reclaimprivacy.org (open src tool to scan a user's privacy settings)

- Worrying signs: togetherville.com (parents must be FB friends, for now..)

I doubt it: economic problem has to be solved first

4. Role of other researchers

This is more a call to arms!

- Lots of ongoing work in leakage detection beyond our work
- Interesting work in examining role of JavaScript/leakage in OSDI and CCS
- Possible areas of contribution
  1. Where should protection be provided?
  2. Architectural solutions
  3. Identity management
  4. Anonymization
  5. Security
  6. Usability

Privacy protection location and architectural issues

Where to provide privacy?

- Browser (via extensions/plugins, currently most popular)
- Intermediaries (Gateways/proxies)

Decentralized OSNs?

- Storage in distributed nodes or cloud
- Lockr (CoNext), Vis-a-vis (WOSN), Safebook
- Incentives for enough nodes to be always available low
- No sustainable economic model yet

Identity management

- OAUTH and OpenID efforts already well underway - might be harnessed
- PsuedoID proposal
- Several newer proposals and some aimed at privacy in the offing

Anonymization and security

Anonymization

- Multiple research efforts: variants on k-anonymity (VLDB'09)
- Differential privacy being tailored to OSNs and privacy (data anonymization in SIGCOMM'10)

Security: much more potential to re-use

- Cryptographic approaches
- Breaking queries into sub-queries (Private Information Retrieval, CACM)
- Tor and friends

Usability

May be key reason why not remotely enough users bother with changing default privacy settings

- Simplest technique of Firefox extension still has very few takers
- Sophisticated techniques get fewer takers if ever
- Opt-in always has fewer adopters (now you see why Buzz-fail was tried as opt-out)
- Concern about loss of fidelity of page when using privacy protection – possibly unfounded

5. Can we arrive at a modus vivendi with aggregators?

- Interesting work in the space of economics of privacy: Alessandro Acquisti
- Goal: customer convenience, economic gain for supplier
- Separation of on-line and off-line identities would be useful
- Question: how to arrive at appropriate trade-off - maybe make linkage of off- and on-line identities much harder
- A nice formulation that at least sounds good (Acquisti)

  1. Economics will decide what to protect and what to share
  2. Law will send appropriate signal to the free market
  3. Technology will implement whatever is chosen

Current Status

- Multiple vectors of violations breaking out regularly (now with cross-OSN potential) while large aggregators have grown steadily.
- Awareness: growing but rather slowly. 1M joining Facebook daily with no indication of being any more privacy-aware than the existing 500M.
- Risk of users being inured about periodic press announcements.
- No real damage/impact on large violators
- Government agencies are yet to demonstrate impact (but Dec'10 FTC report may change minds)
- Tangible technical gains are minimal
- Economic impact due to public opinion, press, and government combined: some hope

Summary

- Privacy leakage is widespread
- Am I optimistic? Well, cup is half-full..
- Impossible to convince masses
- Governments should play a stronger role
- Brand risk may be the strong motivating factor for compliance

Unknown author's cartoon

Another updated cartoon

Acknowledgment

Much of the work described here is joint work with Craig Wills, WPI.

Thanks to numerous colleagues who commented on my SIGCOMM CCR paper
"I Know what you will do next summer"
http://www.research.att.com/ bala/papera/ccr-priv10.pdf