# Privacy in Online Social Networks: An Oxymoron?

Balachander Krishnamurthy

AT&T Labs–Research

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

# 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Sadly, this cartoon is out of date.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.

## What does privacy mean in a tell-all OSN world?

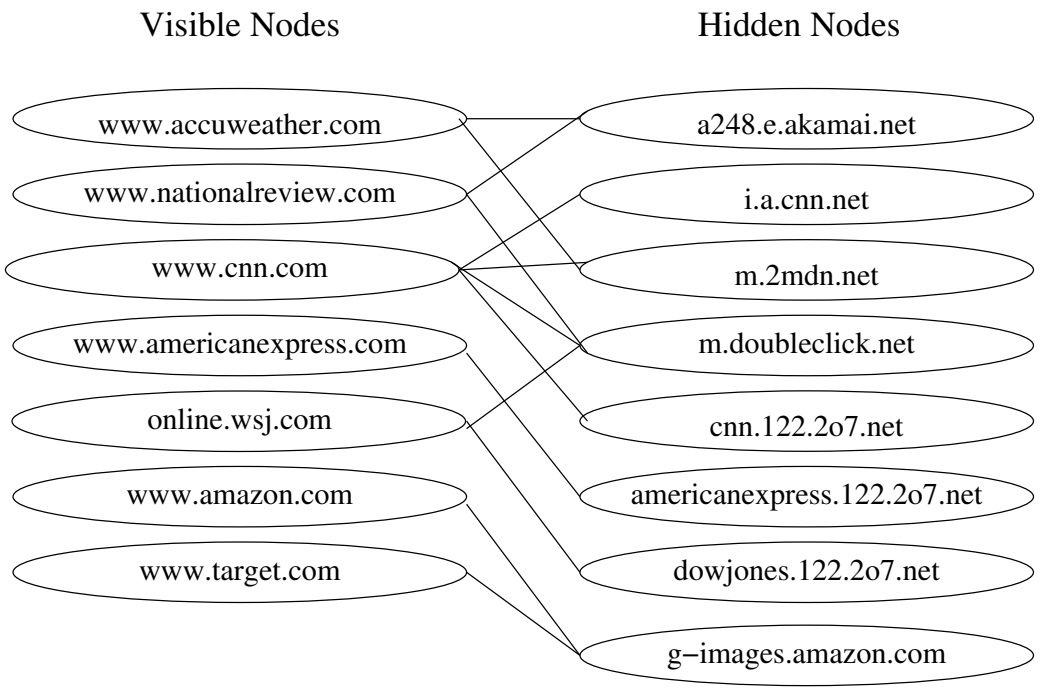
- Still depends on the information disseminated, ability to combine external data, what data collectors *might* do with it
- Good to know *what* information is being diffused, *who* is tracking it, and *how*
- Chrome: URL completion leaks *any* URLs to Google *by default*
- Google toolbar on *by default* on every Dell sold
- Specific Media (175M individual profiles)

Goal is to allow standard network activity while preserving desired privacy

# 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)

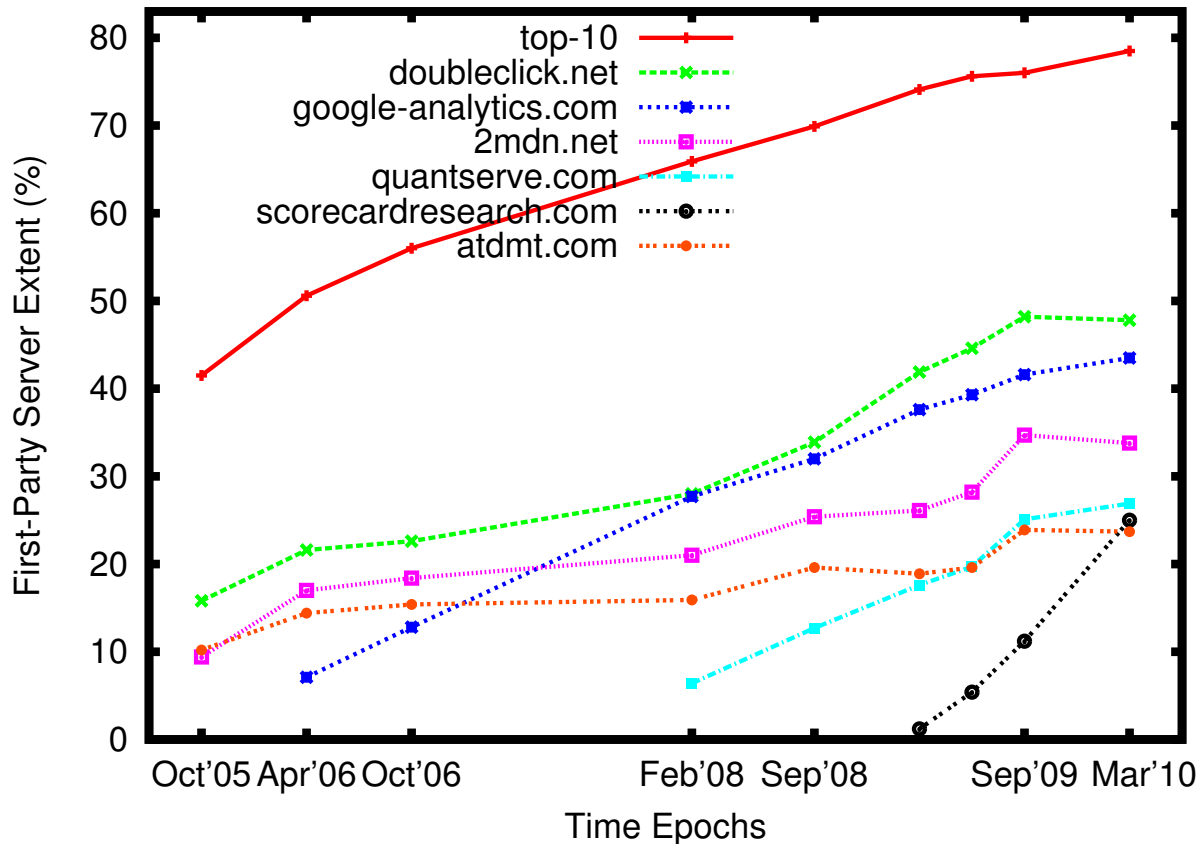
3. CDNs: Serve images, rarely JavaScript. e.g., akamai.net, yimg.com

Privacy leaks to all of them.

# Privacy footprint: longitudinal study (WWW 2009)

- Privacy footprint: measure of dissemination of user-related information across *unrelated* sites. Shows the number and diversity of 3d-party sites visited as a result of a user visiting first party sites.
- Examined the penetration of the top 3d-party *domains* that aggregate information about user's movements on the Web
- Examined the role of economic acquisitions of aggregator companies that buy others and increase their tracking ability as *families*
- A few domains/families can track across most popular Web sites.

# Top 3d-party domains over time

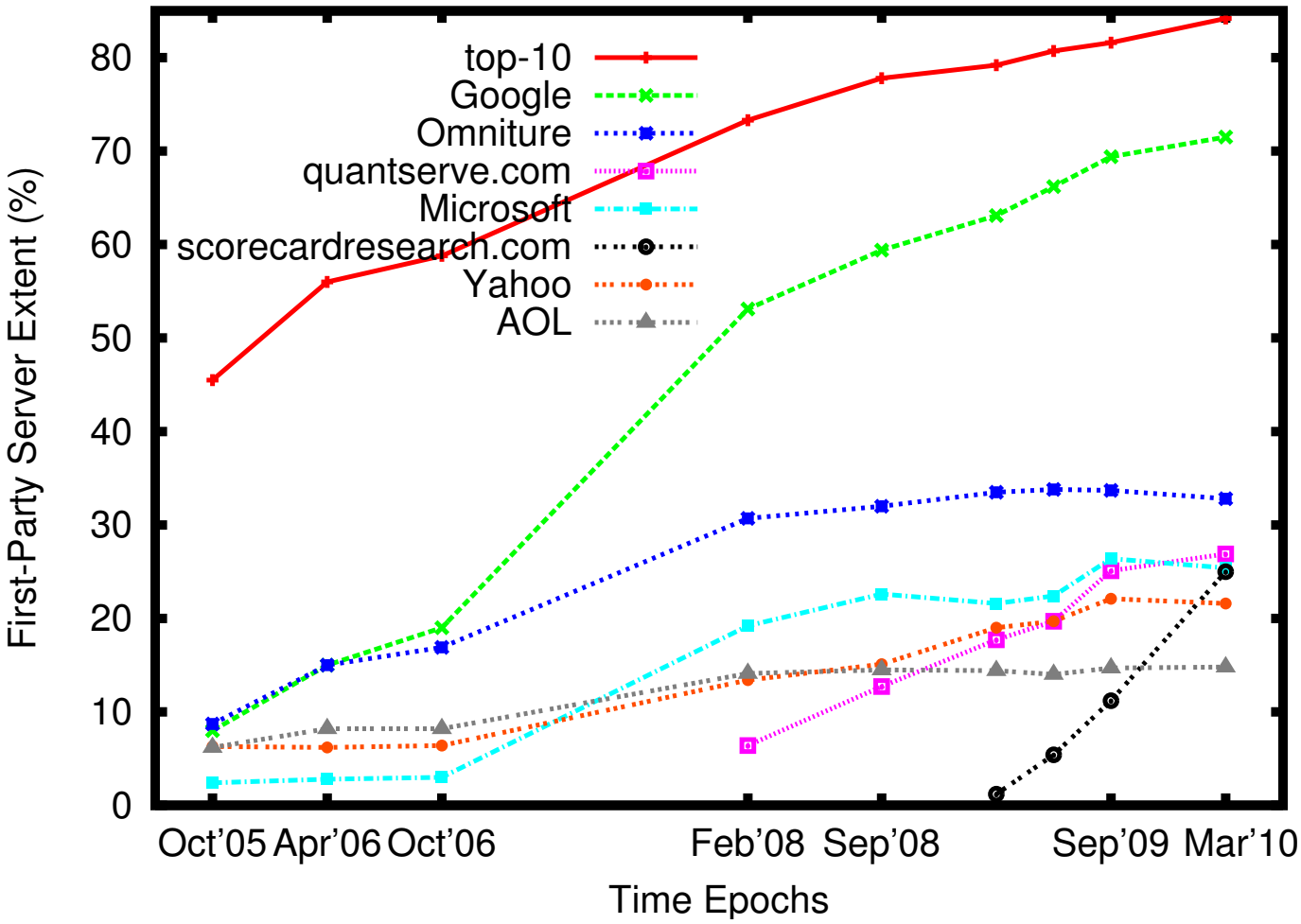



Combined impact of the top-10 domains: up from 40% to over 70%.

# Situation grimmer in the face of acquisitions

| Family | Acquired | Date | Reach % |
|---|---|---|---|
| AOL | advertising.com | Jun'04 | |
| | tacoda.net, adsonar.com | Jul'07/Dec'07 | 14 |
| Doubleclick | falkag.net | Mar'06 | |
| Google | youtube.com ($1.65B) | Oct'06 | |
| | doubleclick.net ($3.1B) | Mar'07 | |
| | feedburner.com | Jun'07 | |
| | admobs.com ($750M) | Nov '09 | 70 |
| Microsoft | aquantive.com (atdmt.com, $6B) | May'07 | 25 |
| 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 | |
| | yieldmanager.com, adrevolver.com | Apr'07/Oct'07 | 18 |
| Adobe | Omniture ($1.8B) | Sept 15 2009 | 28 |

# Top-10 Family Growth




Top-10 families cross 80% in Sep'09

## Web sites vs. OSNs

- The amount of personal information on traditional Websites is relatively low
- The voluminous amounts of private information OSNs ask for and users provide (often above and beyond what they are asked) is of particular concern
- User's expectation of limits of information spread often vary from reality.
- This necessitates examination of what OSNs ask for and examining the default availability of user's information to others in the OSN
- User's OSN information being available outside the OSN is even more problematic
- Finally, if the actors who may receive such information are the same as in the popular Websites case, linkage is of obvious concern.

# 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.

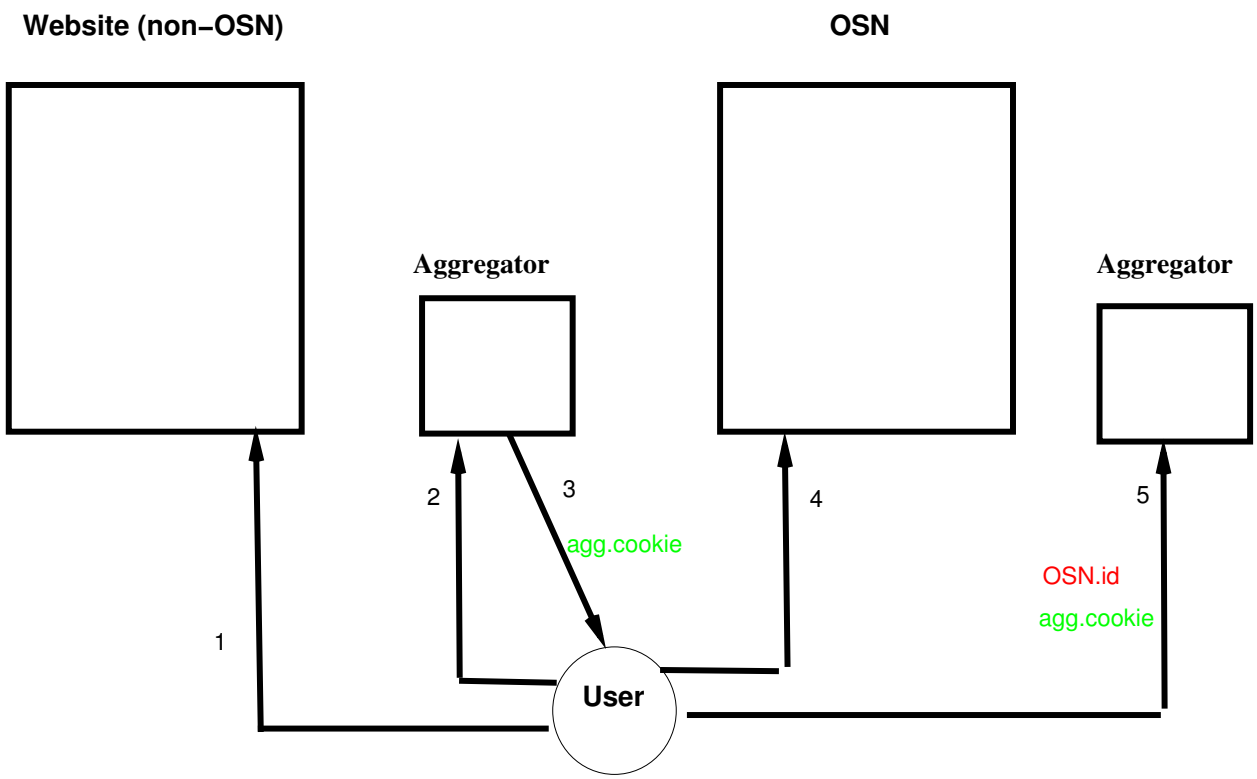

 

# Source of leakage

- 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

## Recent examples of other uses of OSN information

Several legal cases settled with information from Facebook/MySpace:

- 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.

Cases are common because US Congress mandated changes to the federal rules of civil procedure, expanding the acceptance of electronically stored information as evidence in 2006. http://www.uscourts.gov/rules/Reports/ST09-2006.pdf

# National Advertising Initiative: NAI

NAI: a cooperative of online marketing and analytics companies let concerned users to opt-out of targeted ads via third-party cookies.

- Users can visit http://networkadvertising.org/managing/opt_out.asp
- Initiative by third-party aggregators
- Opt-out of targeted ads by any or all of the 38 NAI members.
- Opting out causes an "opt-out" cookie for the given member to be set

Drawbacks: removing all cookies will remove opt-out (TACO extension makes this persistent).

More importantly this does not address tracking at all!

# Evolution of privacy

- Awareness: users need to know what private information is being leaked, to whom, and how.
- Control: users should be presented with the opportunity to control the degree and extent to which private information is being shared.
- Negotiation: Next logical step – is there a way to bridge the gap between legitimate need for (say, demographic data) and what the users are willing to share?

# Unknown author's cartoon