# Privacy leakage vs. Protection measures: the growing disconnect

Balachander Krishnamurthy (AT&T Labs–Research)
Konstantin Naryshkin (Worcester Polytechnic Institute)
Craig E. Wills (Worcester Polytechnic Institute)

# Presenter

Balachander Krishnamurthy

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

## Ongoing work since 2005

- Cat and Mouse: Content delivery tradeoffs in Web access (WWW'06)
- Generating a privacy footprint (IMC'06)
- Privacy Loss and impact of Privacy Protection (SOUPS'07)
- Characterizing privacy in OSNs (WOSN'08)
- Privacy Diffusion on the Web: A Longitudinal Perspective (WWW'09)
- On the leakage of PII via OSNs (WOSN'09) *Moral: People read WSJ*
- Privacy issues on Twitter (ICA'10)
- PII leakage in *mobile* OSNs (WOSN'10)

Situation has steadily worsened with each paper....

## This study: Popular *Non-OSN* sites with user accounts

Why?

- Millions of users create accounts with varying degrees of personal information
- Hitherto unexamined
- Beyond traditional leakage, potential to examine *linkage*
- Examine various protection measures' effectiveness in combating leakage/linkage

Focus on *direct* leakage of bits of private information to third party aggregators

## What is *leakage*?

Depends on the viewpoint!

- User: Personal information shared with any site other than first party
- First party:

  Contracted work outsourced to third party (e.g. analytics)
  Tracking by third party for marketing/demographic information

A first party site can legitimately state that a third party working on its behalf obtaining user data is *not* leakage. Risks can be lowered by explicit statements in privacy policies covering all data shared in any transaction.

But not *all* first parties have unambiguous, clear statements and thus from user's POV the concern of leakage remains.

# Categories and choice of sites for study

- 10 popular + 1 sensitive category + OSNs = 12 categories (120 sites)
- Site must have a minimum of 100K registered users (most have millions)
- Select among 17 Alexa (sub)categories and end up with:
  Arts, Employment, Video Game News, Photosharing, News, Travel, Shopping, Relationships, Generations and Age Groups, and Sports.
- Sensitive category: Health
- OSN category: beyond popularity, mainly for contrast

# Data gathering

- Use Fiddler Web proxy to capture HTTP requests/responses
- Ignore SSL (tiny fraction)
- Steps: create account, confirm verification message, create profile
- In many sites we can use existing accounts in Facebook, Google, or Twitter but we created accounts directly whenever possible

## Interacting with sites

- Actions available only to *registered* users, and some common actions available to all users
- Share content with 'friends' via email or OSN links when feasible
- Create a set of sensitive strings from user's profile (name, email, address). Include search keywords sent to Health and Travel sites
- Stored HTTP request/response and POST data and then look for any of these strings transmitted to 3d-parties
- Eliminate false positives by hand (e.g., 90210 being part of long number string)

Obviously we only find *lower bound* of leakage, only for set of actions done, and only what is not encrypted/modified

# Results: executive summary

- Worse than our (jaded) expectations
- Search strings to healthcare sites and travel details going to aggregators was a tad surprising as was extent of leakage (9 of 10 popular sites in these categories)
- Potential of linkage is high through cookies and other linkable elements
- Existing techniques are largely ineffective

Recall our caveat about what is "leakage"

Next: selected results in more detail on per-action class basis

# Action: Account creation/confirmation

User's email address leaked via a Sports category website via Referer header to a doubleclick.net server.

GET http://ad.doubleclick.net/adj/...
Referer: http://submit.SPORTS.com/...?email=**jdoe@email.com**
Cookie: id=35c192bcfe0000b1...

We will come back to this example when talking about linkage via email address later.

# Action: Account login/navigation

Some sites store private information in site-specific *first-party* cookies

Email, Name, Zipcode leakage Via 1st-party cookies to hidden 3d-Party

GET http://metrics.AGEGROUPS.site/b/ss/..global/...
Host: metrics.AGEGROUPS.site
Referer: http://www.AGEGROUPS.site
Cookie: ...e=**jdoe@email.com**&f=**John**&l=**Doe**&...&p=**12201**...

metrics.AGEGROUPS.site is AGEGROUPS.122.2o7.net i.e., Omniture (2nd largest aggregator, now part of Adobe). Hidden DNS issue is not addressed by *any* protection measures

domain cookie assumes metrics.AGEGROUPS.site is same as AGEGROUPS.site and is sent to Omniture.

# Action: entering content

User's input is often sent as parameters in a GET request instead of POST. Fetching embedded 3d-party objects in page result in sending user output in Referer header of subsequent request.

Example showing age, zip code and gender of a user on PHOTOSHARE site being leaked to specificlick.net

GET http://afe.specificclick.net/?l=7654&sz=200x250...
Referer: http://a.PHOTO.com/.../age=**30**/zip=**12201**/gender=**M**/...

# Action: Searching for Sensitive Terms

Users have higher expectation of privacy here

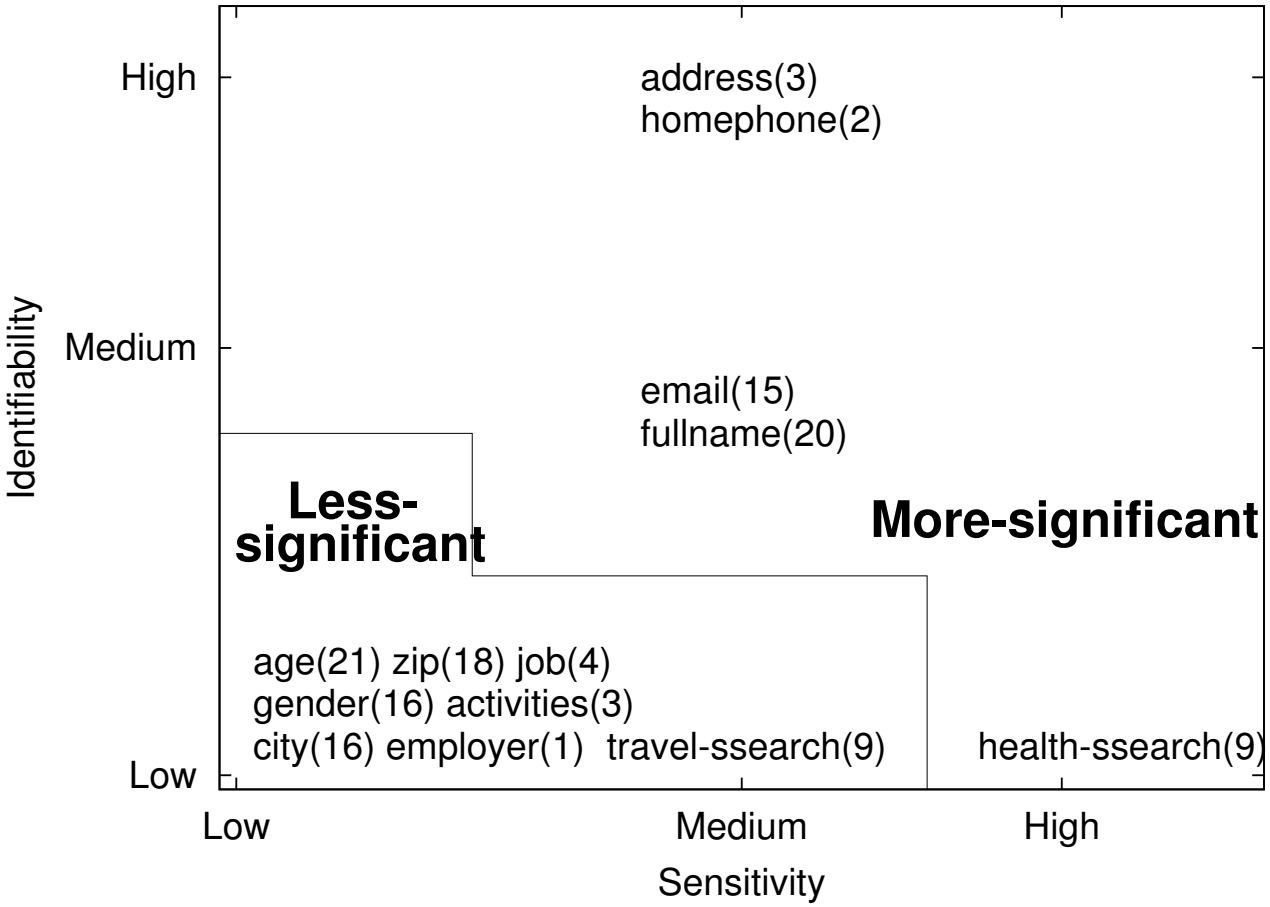
Search in Health site: sensitive string to quantserve via Referer header

GET http://pixel.quantserve.com/pixel;r=1423312787...
Referer: http://search.HEALTH.com/search.jsp?q=**pancreatic+cancer**

Search in Travel site: travel itinerary origin/destination/dates sent to doubleclick

GET http://pix04.revsci.net/...TRAVELSITE
Referer: http://fls.doubleclick.net/...u11=Economy/Coach;u3=**BOS**;
u4=**20110415**| **20110417**;u1=Flight;u2=**MCO**...

# Sensitivity and identifiability of leaked bits




Numbers in paren are number of sites leaking that bit of data.
Search strings in Health and Travel are shared in 9 (out of 10) cases.

# Leakage across categories

| Category | Sites w/ Direct Leakage | Action | | | | |
|---|---|---|---|---|---|---|
| | | Create Account | Account Login/ Navig. | View/ Edit Profile | Input Content | Sens. Search |
| Health | 9 | 0 | 1 | 0 | 0 | 9 |
| Travel | 9 | 0 | 1 | 0 | 0 | 9 |
| Employment | 8 | 0 | 2 | 2 | 7 | 0 |
| OSN | 7 | 0 | 3 | 5 | 0 | 0 |
| Arts | 7 | 0 | 3 | 4 | 1 | 0 |
| Relationships | 7 | 0 | 3 | 2 | 2 | 0 |
| News | 5 | 0 | 5 | 0 | 0 | 0 |
| PhotoShare | 4 | 3 | 3 | 0 | 1 | 0 |
| Sports | 4 | 1 | 2 | 0 | 1 | 0 |
| Shopping | 3 | 0 | 2 | 0 | 2 | 0 |
| AgeGroups | 2 | 0 | 1 | 1 | 0 | 0 |
| VideoGames | 2 | 0 | 1 | 1 | 0 | 0 |
| Tot. Sites/Cat. | 67/12 | 4/2 | 27/12 | 15/6 | 14/6 | 18/2 |

67 (56%) of 120 sites directly "leak" private info to at least one 3rd-party.

# What is linkage?

Aggregators receive data from multiple sites (sources). Similar or identical identifying information of users is often present across sites.

Aggregators are in a position to *link* data to get a broader picture about users.

Linking can be done via Globally unique IDs and even in the absence of cookies (via browser fingerprinting)

Are they linking? We don't know.

We report. You decide.

# Linking via Globally Unique IDs

48% of sites leaked userid to a 3d-party

*Potential* linkage of records through email address

GET http://ad.doubleclick.net/activity;...
Referer: http://f.nexac.com/...http://www.EMPLOYMENT.com/...
na_fn=**John**&na_ln=**Doe**&na_zc=**12201**&
na_cy=**Albany**&na_st=NY&na_a1=**24 Main St.**&
na_em=**jdoe@email.com**...
Cookie: id=22a348d29e12001d...

Same email address was present in our first example of leakage from a Sports site.

Email leaked by 13% of sites – many users use same email across sites.

# Linking in the absence of cookies

Cookies are just one vector of linkage. Other vectors include client IP addresses and browser fingerprints.

One fingerprint is the set of plugins a user has installed on their browser. Omniture gathers this in the Request-URI in several browsers (Firefox, Chrome).

A *good* use of the browser fingerprint is for the first party to know what versions of plugins users have installed.

A possibly *risky* use is the ability that third parties now have to link users across sites even in the absence of cookies, as the specific set of plugins and versions is likely to be unique.

## Summary of existing protection measures

**block**: Blocking requests to targeted 3d-parties (AdBlock, IE9, block-hidden)

**NoCook** and **Nojs**: Refusing all or 3d-party cookies, blocking JavaScript

**Referer**: filtering (modifying/removing) protocol headers

**Anon**: Anonymizing user's IP address (via a proxy, Tor)

**Opt-out**: Opting out of tracking by using NAI opt-out cookie

**DNT header**: Request not to be tracked

**FTC Dec '10 report**: input to legislation, advocated Privacy By Design (PBD)
PBD: embed privacy early, make default private, enable access to user data

# How effective are these techniques/proposals?

| | | Protection Measure | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Leakage/Linkage Scenario | block | block-hidden | nocook | no3rdcook | nojs | referer | anon | opt-out |
| a) | User visit | ✓ | | | | | | | |
| | Hidden third party | | ✓ | | | | | | |
| | 3rd-party tracking linkage | ✓ | | ✓ | ✓ | | | | ✓ |
| | 1st-party tracking linkage | | | ✓ | | ✓ | | | |
| b) | Leakage via Referer | ✓ | | | | | ✓ | | |
| | Leakage via cookies | | ✓ | ✓ | | | | | |
| | Leakage via JS | ✓ | | | | ✓ | | | |
| c) | Linkage via IP addr | ✓ | | | | | | ✓ | |
| | Linkage via Flash cookies | ✓ | | | | | | | |
| | Linkage via fingerprint | ✓ | | | | ✓ | | | |
| | Linkage via GUID | ✓ | | | | | ✓ | | |
| | Linkage w/ Other Sources | | | | | | | | |

a) Expected, b) Known, c) Potential

# Shortcomings

Most of the protection measures are ineffective (including current proposals)

**block** seems to be a user's best friend.

Turning of JavaScript has negative consequences (not always..)

DNT: Enforcement, fines, and reverse engineering technology that can analytically demonstrate users are not being tracked are possible way forward.

Economic acquisitions are accelerating

"Right to be forgotten" is needed

Hidden parties should be brought to the front.

Privacy policies have to be less opaque.

3d-parties can be more demonstrative of their non-linkage guarantees.

# Conclusion

Vectors of leakage abound in non-OSN sites.

Notion of sensitive and identifiable bits of information.

Additional concerns of possible linkage.

User's point of view should carry more weight.

First party sites can contribute more.