Private Communications
Issues in the Social Web era

CS315 – Web Search and Data Mining

The AOL search data release

By the way, …

… search companies log your searches …

17556639 how to kill your wife
17556639 how to kill your wife
17556639 how to kill your wife
17556639 how to kill a wife
17556639 poop
17556639 dead people
17556639 pictures of dead people
17556639 killed people
17556639 dead pictures
17556639 dead pictures
17556639 murder photo
17556639 steak and cheese
17556639 photo of death
17556639 death
17556639 dead people photos
17556639 photo of dead people
17556639 www.murderpeople.com
17556639 decapitated photos
17556639 decapitated photos

Privacy concerns

- Data is often collected silently
  - Web allows large quantities of data to be collected inexpensively and unobtrusively
- Data from multiple sources may be merged
  - Non-identifiable information can become identifiable when merged
- Data collected for business purposes may be used in civil and criminal proceedings
- Users given no meaningful choice
  - Few sites offer alternatives
Cookies 101

- Cookies can be useful
  - Used like a staple to attach multiple parts of a form together
  - Used to identify you when you return to a web site so you don’t have to remember a password
  - Used to help web sites understand how people use them
- Cookies can do unexpected things
  - Used to profile users and track their activities, especially across web sites

How cookies work - the basics

- A cookie stores a small string of characters
- A web site asks your browser to “set” a cookie
- Whenever you return to that site your browser sends the cookie back automatically

Browser Chatter

- Browsers chatter about
  - IP address, domain name, organization,
  - Referring page
  - Platform: O/S, browser
  - What information is requested
    - URLs and search terms
  - Cookies

- To anyone who might be listening
  - End servers
  - System administrators
  - Internet Service Providers
  - Other third parties
    - Advertising networks
  - Anyone who might subpoena log files later

Cookies

Cookies can be useful

- Used like a staple to attach multiple parts of a form together
- Used to identify you when you return to a web site so you don’t have to remember a password
- Used to help web sites understand how people use them

Cookies can do unexpected things

- Used to profile users and track their activities, especially across web sites
How cookies work – advanced

- Cookies are only sent back to the “site” that set them – but this may be any host in domain
  - Sites setting cookies indicate path, domain, and expiration for cookies

- Cookies can store user info or a database key that is used to look up user info

Cookie terminology

- **Cookie Replay** – sending a cookie back to a site
- **Session cookie** – cookie replayed only during current browsing session
- **Persistent cookie** – cookie replayed until expiration date
- **First-party cookie** – cookie associated with the site the user requested
- **Third-party cookie** – cookie associated with an image, ad, frame, or other content from a site with a different domain name that is embedded in the site the user requested
  - Browser interprets third-party cookie based on domain name, even if both domains are owned by the same company

“Web bugs”

- Invisible “images” (1-by-1 pixels, transparent) embedded in web pages and cause referrer info and cookies to be transferred
  - Also called web beacons, clear gifs, tracker gifs, etc.

- Work just like banner ads from ad networks, but you can’t see them unless you look at the page source

- Also embedded in HTML formatted email messages, MS Word documents, etc.

Ad networks

- Ad company can get your name & address from CD order and link them to your search
  - (This is NOT how Google Ads work)
What They Know

Personal data:
- Email address
- Full name
- Mailing address (street, city, state, and Zip code)
- Phone number

Transactional data:
- Details of plane trips
- Search phrases used at search engines
- Health conditions
- Phone number

"It was not necessary for me to click on the banner ads for information to be sent to DoubleClick servers."
- Richard M. Smith

Spyware

- Spyware: Software that employs a user’s Internet connection, without their knowledge or explicit permission, to collect info
  - Most products use pseudonymous, but unique ID
- Over 800 known freeware and shareware products contain Spyware, for example:
  - Beeline Search Utility
  - GoZilla Download Manager
  - Comet Cursor
- Often difficult to uninstall!
- Anti-Spyware Sites:
  - http://grc.com/oo/spyware.htm
  - http://www.adcop.org/smallfish
  - http://www.spychecker.com
  - http://cexx.org/adware.htm

Offline data goes online...

- The Stop and Shop grocery store began posting purchase information for customers who had frequent shopper cards
- The Cranor family’s 25 most frequent grocery purchases (sorted by nutritional value)