TRUST
Team for Research in Ubiquitous Secure Technology

Privacy and Search Engines

Chris Jay Hoofnagle
Samuelson Clinic
Berkeley Ctr. for Law and Tech.

Carnegie Mellon · Cornell University · MILLS COLLEGE · San José State UNIVERSITY
SMITH COLLEGE · STANFORD UNIVERSITY · Berkeley · VANDERBILT UNIVERSITY

Defining Privacy

TRUST
Team for Research in Ubiquitous Secure Technology

The desire by each of us for physical space where we can be free of interruption, intrusion, embarrassment, or accountability and the attempt to control the time and manner of disclosures of personal information about ourselves.

--Robert Ellis Smith, *Ben Franklin's Web Site*

Fair Information Practices (FIPs)

TRUST
Team for Research in Ubiquitous Secure Technology

- Assumes that entities that collect personally identifiable information assume certain risks and responsibilities.
- FIPs are a consensus standard for addressing those risks and responsibilities.
- Developed by the Health, Education, and Welfare committee on Automated Personal Data Systems in 1973.

OECD Privacy Guidelines (1980)

TRUST
Team for Research in Ubiquitous Secure Technology

- Collection limitation
- Data quality
- Purpose specification
- Use Limitation
- Security safeguards
- Openness
- Individual participation
- Accountability

APEC Privacy Principles (1994)

- Focus shifts from human rights to protecting individuals from harm.
- Loophole for publicly available information--
  - personal information...made available to the public, or is legally obtained and accessed from:
    - government records that are available to the public;
    - journalistic reports; or
    - information required by law to be made available to the public.

Search Engines Break the Mold

TRUST
Team for Research in Ubiquitous Secure Technology

- Not collecting personal data, per se. But, search engines mediate access to content. Therefore, they are a central point of privacy vulnerability
  - Search strings
    - Access
    - Retention
  - Personalization/Customization
  - Idea of personally identifiable information may be a dated concept
    - Metadata, data about others may identify you too

Anonymous?

TRUST
Team for Research in Ubiquitous Secure Technology

- Demographic data can occur infrequently, allowing “reidentification”
  - ...87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides...In general, few characteristics are needed to uniquely identify a person.
    - Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population* (using 1990 census data)

Search Strings

- AOL releases 20m queries based on 600k users to help researchers
  - Were trying to make routine access more efficient
- Users are uniquely enumerated
- Some easy to identify
  - Users vanity searched name, SSN
- Many others identifiable based on searches unrelated to PII
  - Thelma Arnold, Lilburn, Ga

Google's Search Policy

SEARCH PRIVACY PRACTICES Companies ordered according to share of U.S. searches.	How long after search data has been collected will it be removed?			How will search data be removed?		
	IP address	**Cookie ID**	**Query**	**IP address**	**Cookie ID**	**Query**
Google *Policies will be in place by December 2007, applied retroactively.*	18 months	18 months	Indefinite	Deletes last octet of address.	Deletes partial or complete ID (specifics TBA).	Does not remove.
Yahoo!	13	13	Indefinite	Deletes last	Deletes	Applies

Source: Search Privacy Practices: A Work In Progress, CDT Report -- August 2007

SEARCH PRIVACY PRACTICES Companies ordered according to share of U.S. searches.	How long after search data has been collected will it be removed?			How will search data be removed?		
	IP address	Cookie ID	Query	IP address	Cookie ID	Query
Google *Policies will be in place by December 2007, applied retroactively.*	18 months	18 months	Indefinite	Deletes last octet of address.	Deletes partial or complete ID (specifics TBA).	Does not remove.
Yahoo! *Policies will be in place by July 2008. Currently reviewing how to apply policies to historical data.*	13 months	13 months	Indefinite. Some queries will be removed automatically by personal information filter after 13 months.	Deletes last octet(s) of address.	Deletes some portion of ID (specifics TBA).	Applies personal information filter to remove names, SSNs, etc.
Microsoft *Policies will be in place by July 2008, applied retroactively.*	18 months	18 months	Indefinite	Deletes complete address.	Deletes complete ID.	Does not remove.
Ask.com *Policies will be in place in 2007. Currently reviewing how to apply policies to historical data.* For users who opt out of having Ask retain their search data (via AskEraser):	Few hours	Few hours	Few hours	Deletes complete address.	Deletes complete ID.	Deletes complete query.
For all other users:	18 months	18 months	Indefinite	Deletes complete address or last octet(s) (specifics TBA).	Deletes complete ID.	Does not remove.
AOL *Policies will be in place in 2007, applied retroactively.*	13 months	13 months	13 months	Deletes complete address.	Deletes complete ID.	Retains only aggregate statistics about search query frequency.

Personalization/Customization

TRUST
Team for Research in Ubiquitous Secure Technology

- Tracking is present, even to sites with “sensitive” topics
- Goal is to present ads across multiple platforms (desktop, laptop, xbox)

Looking Ahead

TRUST
Team for Research in Ubiquitous Secure Technology

- Search will move from presenting documents to presenting distilled information.
  - What will this mean for privacy?
  - How can students of information science design for privacy?