

Query Log Analysis

Social and Technological Challenges

WWW2007 Workshop
May 8, 2007

John Morris
Center for Democracy & Technology

Slides available at
<http://www.cdt.org/privacy/20070508querylogs.pdf>



Hard Issues

- Competing valid goals/concerns
 - Protect Privacy
 - Improve Search - Study Search
- Some essential elements of privacy:
 - Data minimization
 - Notice
 - Informed Consent
 - Ability to refuse consent
- Three-legged triangle of competing interests
 - Search companies, users, academics
- CDT is at the starting point of its analysis
 - Raising questions rather than providing answers



Reasons to Retain Query Logs

- Improve search results
 - Specific results for specific users
 - General results for all users
- Improve marketing to user or users
- Provide info to third parties for marketing or other commercial purposes
- Provide info to third parties for academic research -- or preserve history
- Comply with law ?
 - Not now, but possible in future
- Others ??



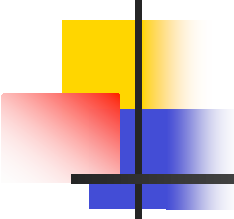
Threats Raised by Query Log Retention

- Disclosure of “PII”
 - Accidental disclosure
 - Malicious disclosure - rogue employee/researcher
- Compelled disclosure to third parties
 - Subpoena in civil case -- divorce, anything
- Disclosure to government
 - Compelled disclosure pursuant to lawful court order
 - Unauthorized/warrantless disclosure
- Marketing based on query logs
 - Ad targeting/delivery by search engine
 - Marketing by third parties
- Others ??



Personally Identifiable Information (PII) in Searches

- Single queries can disclose PII
 - U.S. Social Security numbers
 - Credit card numbers
 - Risky even if “unlinked” with identity
- Multiple linked queries can disclose PII - even without identification info
 - Health info + location/identity signals
 - Personal info (sexual orientation) ...
 - Criminal activity (medical marijuana) ...
 - Civil rights (free expression in repressive regime) ...
- Risk much higher if queries linkable to specific individuals



Closest Analog in Offline World: Library Records

- Libraries are direct precursors to search engines
- Broad societal consensus to protect library records (at least in United States)
 - In U.S., all 50 states have laws prohibiting the disclosure of library records
- Similarly, forced disclosures of bookstore records have been blocked

Arkansas Law:

Section 13-2-701 et seq.

- “Library records which contain names or other personally identifying details ... shall be confidential and shall not be disclosed except as permitted by this subchapter.”
- Main exceptions:
 - With informed written consent of patron
 - Pursuant to lawful court order
- Penalties: a misdemeanor - punishable by a fine of not more than two hundred dollars (\$200) or thirty (30) days in jail, or both



Arkansas Law (continued)

- "Confidential library records" means documents or **information in any format retained** in a library that identify a patron as having requested, used, or obtained specific materials, **including, but not limited to**, circulation of library books, materials, **computer database searches**, interlibrary loan transactions, **reference queries**, patent searches, requests for photocopies of library materials, title reserve requests, or the use of audiovisual materials, films, or records.



Possible “States” of Query Log Entries

- Linked to individual
 - To IP address
 - By cookie
 - By express registration
- Pseudonymized
 - Key back to individual retained
 - Key back to individual discarded
 - Key and sensitive data discarded
- Anonymized
 - With queries linked - not possible?
 - Links discarded
 - Links discarded with sensitive data discarded
- Other states ??



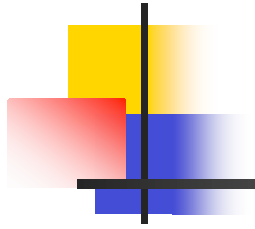
Positive ideas

- Intriguing technical proposals
 - Eytan Adar
 - Li Xiong and Eugene Agichtein
- Institutional Review Boards
 - Critical element of IRB system and the “Common Rule” is on affirmative, informed consent
 - May not be sufficient for companies - different evaluations by IRBs may not be adequate to guard against liability
 - May provide useful model nevertheless



Goals and next steps

- Key goals and aspirations
 - Provide notice and obtain consent
 - Effectively anonymize information
 - Discard or mask sensitive info
- “Best Practices” working group ?
 - Aimed initially at company actions
 - Possibly also aimed at researchers
 - Companies, researchers and users must all participate
 - Example: <http://www.antispywarecoalition.org/>



John Morris
Center for Democracy & Technology
jmorris@cdt.org

Slides available at
<http://www.cdt.org/privacy/20070508querylogs.pdf>

