

CS47300: Web Information Search and Management

Search Ethics: Data Privacy

Prof. Chris Clifton

19 October 2020



Ethics Issues for Web Search

What's the Problem?

- Privacy
 - Query
 - Pages clicked
 - Profiles
- Inappropriate search results
 - Children
 - “Picking” what you want people to see
 - Racial/Gender/Ethnic/... bias

What is Privacy?

- “The right to be let alone” - *Warren & Brandeis, 4 Harvard L.R. 193 (Dec. 15, 1890)*
 - My information protected so it doesn’t adversely affect me in the future
- Control over data
 - My information used only in ways I approve
- Issues:
 - Disclosure / sharing
 - Approved use
 - Recourse

3

Data Privacy: The Goal

- Protect the Individual
 - “Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.” – Charter of Fundamental Rights of the European Union
- Challenges: What do we mean by
 - “concerning” an individual
 - Protection
 - Consent
 - Access / rectified



4

“Obvious” answers

- Concerning an individual
 - Has your name/address/other identifying information
- Protection
 - Only used/accessed in expected, intended, authorized ways
- Consent
 - You know and agree to what is done with the data
- Access/Rectify
 - You can see the data and correct errors

5

Consent?



[The Guardian](#)

Maev Kennedy

Thu 11 Jun 2009 07.17 EDT

American family's web photo ends up as Czech advertisement

Smiths from Missouri only heard about it when a friend travelling in Prague saw them on a grocery store poster

6

Concerning an Individual: IC 24-4.9-2-10

Sec. 10. "Personal information" means:

- (1) a Social Security number that is not encrypted or redacted; or
- (2) an individual's first and last names, or first initial and last name, and one (1) or more of the following data elements that are not encrypted or redacted:
 - (A) A driver's license number.
 - (B) A state identification card number.
 - (C) A credit card number.
 - (D) A financial account number or debit card number in combination with a security code, password, or access code that would permit access to the person's account.

9

Concerning an Individual: IC 24-4.8-1-10

- **IC 24-4.8-1-10 "Personally identifying information"**
- Sec. 10. "Personally identifying information" means the following information that refers to a person who is an owner or operator of a computer:
 - (1) Identifying information (as defined in [IC 35-43-5-1](#)).
 - (2) An electronic mail address.
 - (3) Any of the following information in a form that personally identifies an owner or operator of a computer:
 - (A) An account balance.
 - (B) An overdraft history.
 - (C) A payment history.

10

Identifying Information: IC 35-43-5-1(j)

"Identifying information" means information that identifies a person, including a person's:

- (1) name, address, date of birth, place of employment, employer identification number, mother's maiden name, Social Security number, or any identification number issued by a governmental entity;
- (2) unique biometric data, including the person's fingerprint, voice print, or retina or iris image;
- (3) unique electronic identification number, address, or routing code;
- (4) telecommunication identifying information; or
- (5) telecommunication access device, including a card, a plate, a code, a telephone number, an account number, a personal identification number, an electronic serial number, a mobile identification number, or another telecommunications service or device or means of account access that may be used to:
 - (A) obtain money, goods, services, or any other thing of value; or
 - (B) initiate a transfer of funds.

11

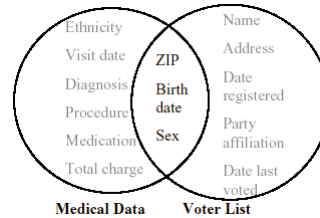
The AOL Awakening

- In Aug 2006, AOL released its customers web searches for research studies
- 20 Million unique queries of 650K unique users
- <user-id> AOL fired its CTO over this issue;
- NY Times Two researchers were forced out
individual from the queries
 - Queries included “60 single men” “landscapers in Lilburn, Ga”
 - Many more queries contained enough information to uniquely identify the person
- *And it keeps going (Netflix, NYC Taxi, ...)*

14

Re-identifying “anonymous” data (Sweeney '01)

- 37 US states mandate collection of information
- Dr. Sweeney purchased the voter registration list for Cambridge Massachusetts
 - 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three



- Solution: k-anonymity
 - Any combination of values appears at least k times
- Developed systems that guarantee k-anonymity
 - Minimize distortion of results

15

Redaction: [IC 24-4.9-2-11](#)

(a) Data are redacted for purposes of this article if the data have been altered or truncated so that not more than the last four (4) digits of:

- (1) a driver's license number;
- (2) a state identification number; or
- (3) an account number;

is accessible as part of personal information.

(b) For purposes of this article, personal information is "redacted" if the personal information has been altered or truncated so that not more than five (5) digits of a Social Security number are accessible as part of personal information.

16

Right to be Forgotten

- EC 95/46 Article 12(b)
as appropriate the rectification, erasure or blocking of data the processing of which does not comply with the provisions of this Directive, in particular because of the incomplete or inaccurate nature of the data;
 - Applied to removal of even correct information from search engines
- [GDPR Article 17](#) – much more clearly spelled out

18

Anonymity: The Goal

- Prevent Disclosure of Personal Information
 - GDPR: ‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly
 - Qatar Law 13 of 2016: Personal Data: Data belonging to an Individual with specified or reasonably specifiable identity whether through such Personal Data or through combining the same with any other data
 - *But still use the data where appropriate!*
- Problem: It can’t be done!
 - “Perfect” privacy requires zero utility (e.g., the data must be encrypted.)
 - As soon as we can use the data (e.g., decrypt), it is at risk

20

Why Perfect Privacy is Impossible

(Dwork, McSherry, Nissim, and Smith '06)

- Background Knowledge
 - Adversary may already know a lot
 - Whatever we provide (even de-identified or anonymized data) may add to that knowledge
- It may just take that “last bit of knowledge” to give the adversary the ability to violate privacy
 - *We can formally prove 1 bit may be too much*


21

What We Can Do

- Encryption
 - Reduce risk to minimal levels when data not in use
- Anonymization
 - Produce usable data that is hard to link to individuals
- Noise addition
 - Usable data where any link to individuals (or information we surmise about individuals) is guaranteed to be uncertain/suspect

22

What We Need: Legal Incentives

- “Notice and Consent” framework discourages application of technological advances
 - We can’t guarantee your privacy, so please allow us to use your data in unsafe ways
 - U.S.: [Enforcement action against Snapchat](#) for promising to protect privacy and not doing a good enough job 
 - Companies get away with not even trying, as long as they tell you so
- Can legal frameworks acknowledge that privacy is at risk?
 - Require efforts to manage, not eliminate, that risk

35

Fair Information Practices

1. Notice/Awareness
2. Choice/Consent
3. Access/Participation
4. Integrity/Security
5. Enforcement/Redress
 - Self-Regulation
 - Private Remedies
 - Government Enforcement

<http://www.ftc.gov/reports/privacy3/fairinfo.shtm>

40