Data privacy and big data

Lorrie Faith Cranor

November 12, 2015

8-533 / 8-733 / 19-608 / 95-818:
Privacy Policy, Law, and Technology

Carnegie Mellon University
CyLab

institute for SOFTWARE RESEARCH

Engineering & Public Policy

Today’s agenda

- Quiz
- Data privacy
- Big data

Data privacy through de-identification

- De-identification: Process of removing the association between a set of identifying data and the data subject
  - Sometimes it prevents re-identification, sometimes it does not
  - Auxilary datasets may allow for re-identification through linkage attacks
  - Data Usage Agreements can prohibit re-identification
- Reduces privacy risks, while preserving some utility of the data
- Some US laws provide exceptions for de-identified data: e.g. FERPA, HIPAA

Simson L. Garfinkel. De-Identification of Personal Information. NISTIR 8053. October 2015. http://dx.doi.org/10.6028/NIST.IR.8053

De-identification of direct identifiers

- Remove direct identifiers
  - Remove completely
  - Replace with categories, e.g. PERSON NAME or ANYTOWN, USA
  - Replace with random strings
- Pseudonymization
  - Replace direct identifiers with pseudonyms
  - Allows linking across records
  - Often can be reversed

De-identification of quasi-identifiers

- identifiers that by themselves do not identify a specific individual but can be aggregated and “linked” with other information to identify data subjects
- Approaches
  - Suppression – remove quasi-identifier
  - Generalization – describe more generally as group or range
  - Perturbation – systematic adjustment, e.g. randomly add between -10 and 10
  - Swapping (between records)
  - Sub-sampling – release only part of the sample

De-identification of Protected Health Information (PHI) under HIPAA

- Expert determination – an expert examines data, determines appropriate way to de-identify to make risk of re-identification “very small”, and documents this
- Safe Harbor - remove 18 specific types of data for “the individual or relatives, employers, or household members of the individual”
  - Geographic divisions smaller than a state, dates other than year, telephone, fax, email, SSN, medical record numbers, health plan numbers, account numbers, license numbers, vehicle IDs, device IDs, URLs, IP addresses, biometrics, photos of faces, any other unique identifying number or code

K-anonymity (Sweeney 2002)

- A data set is k-anonymous if for all records there are at least k records with matching quasi-identifiers

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Example data set

Zip code	Car Make	Gender	Income
15218	Hyundai	M	High
15218	BMW	M	Low
15218	BMW	F	Medium
15216	Kia	F	Low
15213	Ford	F	Low
15213	Toyota	M	Medium
15213	Toyota	M	High
15213	Honda	M	Low

Suppress last digit of ZIP

Zip code	Car Make	Gender	Income
1521*	Hyundai	M	High
1521*	BMW	M	Low
1521*	BMW	F	Medium
1521*	Kia	F	Low
1521*	Ford	F	Low
1521*	Toyota	M	Medium
1521*	Toyota	M	High
1521*	Honda	M	Low

Generalize car country

Zip code	Car Make	Gender	Income
15218	Korean	M	High
15218	German	M	Low
15218	German	F	Medium
15216	Korean	F	Low
15213	American	F	Low
15213	Japanese	M	Medium
15213	Japanese	M	High
15213	Japanese	M	Low

Suppress and/or generalize multiple elements

Zip code	Car Make	Gender	Income
1521*	Hyundai/Toyota/ Honda	M	High
1521*	BMW/Kia/Ford	*	Low
1521*	BMW/Kia/Ford	*	Medium
1521*	BMW/Kia/Ford	*	Low
1521*	BMW/Kia/Ford	*	Low
1521*	Hyundai/Toyota/ Honda	M	Medium
1521*	Hyundai/Toyota/ Honda	M	High
1521*	Hyundai/Toyota/ Honda	M	Low

De-identification scenario

- Happiness survey…

Benefits of big data

- Scientific American “How Big Data Can Transform Society for the Better’ Oct 13
- Understanding the spread of Malaria in Kenya through mobile phone usage patterns (Wesolowski, *Science* 2012)
- Better public transportation through GPS tracking
- Better public health through search queries
- Fraud detection
- Recommendations

Concerns about big data

- Incremental Effect
- Automated Decision-Making
- Predictive Analysis
- Lack of Access and Exclusion
- Analytics
- Chilling Effect

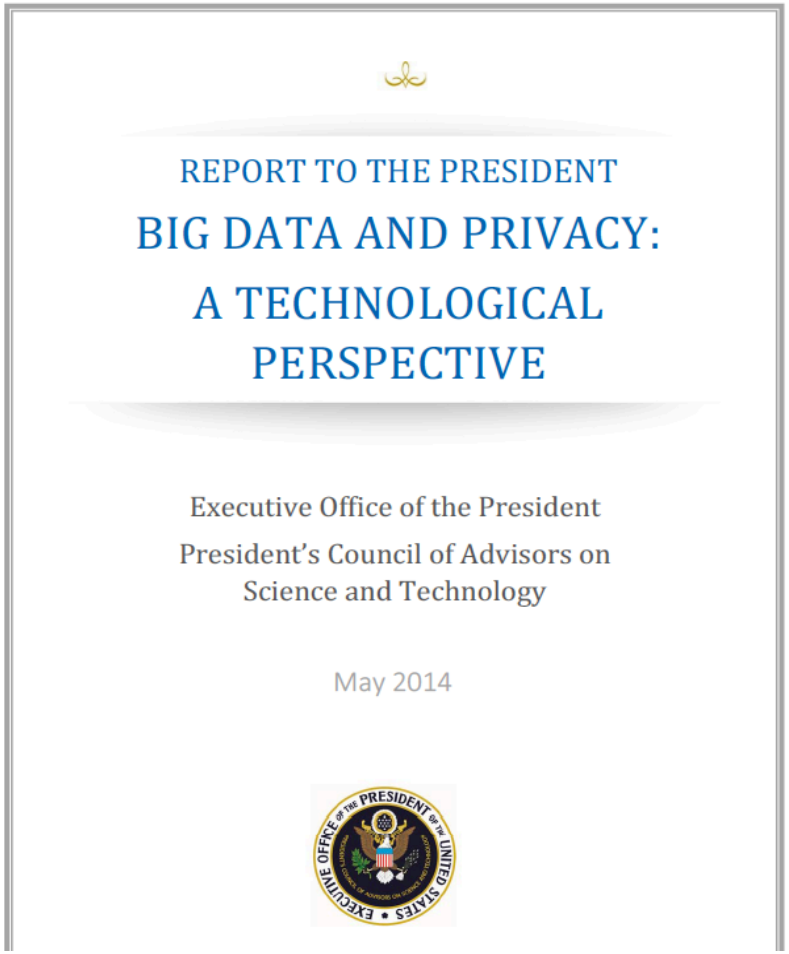
Omer Tene and Jules Polonetsky,
Big Data for All: Privacy and User Control in the Age of Analytics, 11 Nw. J. Tech. & Intell. Prop. 239 (2013).

Big data and privacy protection

- Is big data compatible with privacy protection?
  - Data minimization
  - Consent
  - Deletion
  - Encryption

Solutions to the concerns?

PCAST report on big data



REPORT TO THE PRESIDENT

BIG DATA AND PRIVACY:

A TECHNOLOGICAL PERSPECTIVE

Executive Office of the President

President's Council of Advisors on Science and Technology

May 2014

What's new about big data

- The quantity and variety of data that are available to be processed.
- The scale of analysis, inferences, and conclusions
- Data fusion: "when data from different sources are brought into contact and new facts emerge"

PCAST Policy Recommendations

1. Focus more on use of data than collection and analysis
2. Policy should be on intended outcomes, not technology solutions
3. Strengthen U.S. research in privacy-related technologies
4. Encourage increased education and training opportunities concerning privacy protection
5. US should take the lead through standards and procurement practices

Carnegie Mellon University
CyLab

institute for SOFTWARE RESEARCH

Engineering & Public Policy