Privacy and Big Data: Challenges and Promise

Graham Cormode
Goals for privacy in companies:
– Enable appropriate use of data while protecting customers
– Keep chairman and CTO off front page of WSJ

Security is binary: allow access to data iff you have the key
– Encryption is robust, reliable and widely deployed

Privacy comes in many shades: reveal some information, disallow unintended uses
– Hard to control what may be inferred
– Possible to combine with other data sources to breach privacy
– Privacy technology is still maturing
What is Private?

- Almost any information that can be linked to individual
- Telecoms are privy to much private customer information:
  - **Personally Identifiable Information (PII):** SSN, DOB, address
  - **Financial data (SPI):** bill amount, payment schedule, bank details
  - **Phone activity (CPNI):** called numbers, durations, times
  - **Internet activity:** visited sites, search queries, entered data
  - **Video activity:** live TV, on-demand (Video Privacy Protection Act)
  - **Location activity:** where and when
Aspects of Privacy

♦ First-person privacy: Who can see what about me?
  – Example: Who can see my holiday photos on a social network?
  – Failure: “Sacked for complaining about boss on Facebook!”
  – Controls: User sets up rules/groups for other (authenticated) users

♦ Second-person privacy: Who can share your data with others?
  – Example: Does a search engine share your queries with advertisers?
  – Failure: MySpace leaks user ids to 3rd party advertisers
  – Controls: Policy, regulations, scrutiny, “Do Not Track”

♦ Third-person (plural) privacy: Can you be found in the crowd?
  – Example: Can trace someone’s movements in a mobility dataset?
  – Failure: AOL releases search logs that allow users to be identified
  – Controls: Access controls and anonymization technology
Dimensions to consider

- How much **privacy** do we need?
- How much **utility** do we want from the anonymized data?
- How will data be accessed: as data feed, as data set, via API?

Who will use the data?

1. Permanent employees
   Temporary employees (students, contractors)
2. Outside contractors
   Data purchasers
3. General Public
Privacy Tools and Algorithms

- Many efforts from k-anonymity to differential privacy
- Same questions arise for every proposal:
  - What privacy guarantee is made (if any)?
  - How robust to attack, background knowledge?
  - What is the format of the output, how useful/usable?
- And some bigger questions:
  - How to reach widely accepted privacy standards?
  - General purpose tools for privacy transformation?
  - Align with other changes: legal, social, political
  - No more privacy catastrophes?