Privacy in Data Collection, Management, and Use

Karen Sollins

CFP Meeting, October 27, 2015

The Dilemma

- The value of data
  - All about positives
  - Aggregation and fusion
  - Data Lakes
- The personal cost: privacy
  - What are the risks/costs
  - To whom
  - Who has rights
  - Who has responsibilities
  - What is feasible: now and in the future

Privacy at CSAIL

- CFP: Privacy and Security WG
- BigData Privacy Working Group
- The same set of issues
  - Large amounts of data
  - About human subjects and the results of their activities in cyberspace
  - All the same questions about the tradeoffs in
    - Beneficial and useful opportunities to use the data (and all the tools to do that)
    - Individual's right to privacy (and tools to do that)

Considering real examples

- MOOCs and other online educational systems
- Use of social media information for research
- Sensor and mobile device tracking data for public and individual health
- Privacy on aggregated datasets
- Privacy and user consent: challenges and privacy concerns
- Consumer privacy and marketing
- Genomics and Health

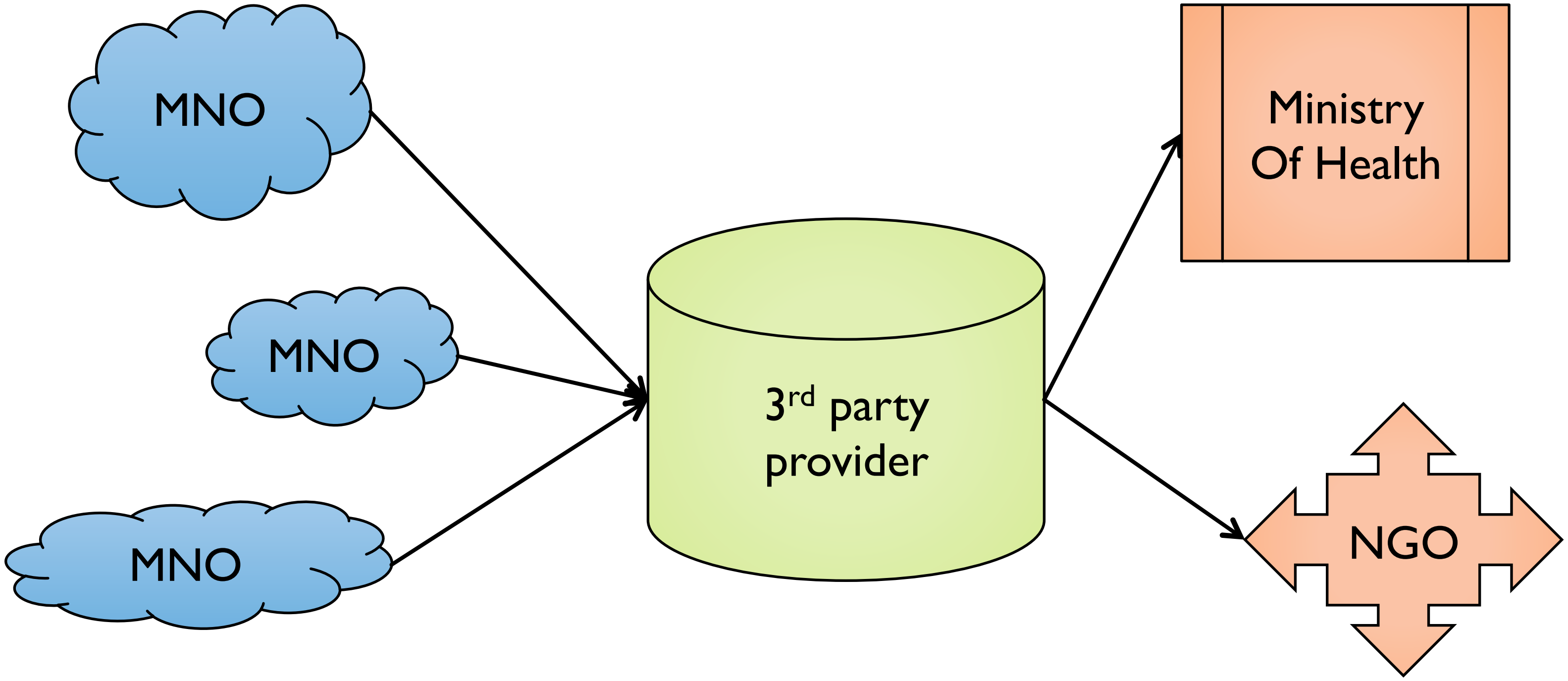
One example: sensor and mobile data for public and individual health

- The challenge
  - NGOs and Ministries of Health (governmental): concerns about health and mitigation of infectious diseases
  - Data: Population (or individual) mobility and infectious diseases
  - Two alternative objectives
    - Scenario 1: Understand and quantify spread of specific diseases
    - Scenario 2: Micro-target individuals for individualized responses
- The data: mobile phone metadata
  - Locations
  - Distances traveled
  - Duration and frequency of travel
  - Recharging patterns (e.g. reflects socio-economic status)
  - Texting patterns (frequency and timeliness of responses, etc.)

Privacy considerations in systems design: Scenario 1

- Participants (besides subjects)
  - MNO (Mobile Network Operator): a number of these
  - 3rd party data provider
  - Ministry of Health or other "customer"
- System modularity options
  - Design 1
    - Each MNO anonymizes and coarsens data
    - Data provider aggregates all records
      - All the problems of conformance of data from different sources
      - Ability to reverse anonymization techniques
  - Design 2
    - Each MNO provides only aggregate or summary data: must provide conformance for data provider
    - Data provider aggregates the summaries

System design



What is difficult in this?

- Regulatory environment
  - Two (French civil code tradition carried into EU and English common law) qualitatively different regulatory regimes
  - Human mobility not reflective of those boundaries
- Data utility: tradeoff of work and privacy against flexibility, extensibility and utility of data
- MNOs: Not in the business of social or health data analysis, or providing data to NGOs and ministries of health (e.g. storage, curation of data, etc.)

Observations: issues

- Scale: amount of data growing exponentially
- Diversity of stakeholders with new interests and objectives
- Integration across previously unmerged datasets
- Secondary subjects: others not normally included in "privacy policies" increasing affected and targeted
- Emergent information will require emergent privacy policies

Observations: stakeholders

- Data subject(s): primary and secondary
- Decision makers
- Data collectors
- Data curators
- Data analysts
- Data platform providers (maybe “stewards”)
- Policy enforcers
- Auditors

What are the options?

- What technologies are available?
- To what are they applicable?
  - What are their strengths and weaknesses
  - What are their underlying assumptions about the data, the policies, and the whole process of definition and application of policies?
- Where are they in their evolution from idea to practice?
- How can we frame their utility?
  - Which problems does each one solve?
  - How might they "fit together"?
  - And, what is missing from the larger picture?

Technologies for privacy in Big Data and Communications

- Two broad categories of "data use"
  - Aggregate results
  - Individual insights
  - Lead to significantly different technology opportunities

- Points in data cycle where privacy technologies can be applied
  - Data collection
  - Data access
  - Data processing (incl. fusion) and analytical methods
  - Data compliance and audit
  - Data destruction

Privacy Provision Challenge: consider technologies, their utilities, their scope of applicability and where they are in their evolution from idea to practice

Data Collection

Application of privacy policies at point and time of data collection

- Approaches
  - Online notice and consent
  - Informed consent
  - Setting of personal attributes
  - Human subjects/ethics review boards
  - Inference: application of machine learning to individuals' behaviors

- Challenges
  - Too complex for average user
  - Too disruptive
  - Users feel they have "no choice"
  - Doesn't capture future uses
  - Doesn't clarify who (which stakeholders) have responsibility and opportunity to define policies
  - Doesn't include secondary subjects effectively

Data Access Controls

Application of particular privacy policy to specific data resource

- Approaches
  - Data use agreements (in some parts of the world, defined by law, in others by policy statement)
    - Examples: data tagging, DRM for personal data
  - Authentication/authorization protocols (both software and hardware): OAuth & access control
  - Encryption (and related key management): Functional encryption

- Challenges
  - Different levels and models for different subjects
  - Key management (including revocation problem)
  - Tagging may be too simplistic
  - Lack of "extensibility"
  - Lack of "evolvability"
  - Functional encryption both restrictive and computational intensive

Data processing and analytical methods

Privacy preserving analysis including anonymization

- Approaches
  - Data access
    - Remove PII and other personally identifying data
    - Statistical anonymization (e.g. k-anonymity)
  - Individual queries
    - Personal Data Stores
    - Secure multi-party computation
    - Functional and homomorphic encryption
  - Statistical approaches
    - Differential privacy and algorithms
    - Synthetic data sets

- Challenges (just a few examples)
  - Personal Data Stores: only as trustworthy as the underlying system, no control once data has "left" the store
  - Functional and homomorphic encryption: limited to small set of possible operations, computationally intense
  - Differential privacy
    - Static data
    - Data set must be large enough to "hide" individuals
    - Restriction on queries: cannot ask too many queries
    - Defining and understanding ε

Compliance and monitoring

Tracking and enforcing use policies

- Approaches
  - Accountable systems
    - Logging metadata
    - Focus on recourse
  - Formalizing legal constraints and enforcement
    - Example: Microsoft on Bing
      - Legalease: formal representation
      - Grok: strong type enforcement

- Challenges
  - Accountable systems
    - Scalability
    - Generality vs. specificity
    - Sources of trust
    - Privacy of logging data
  - Microsoft approach
    - Specific to Map/Reduce type interactions
    - Static (compile time)

Data Destruction

Eliminating access to data

- Approaches (besides just deletion)
  - Elimination: Garfinkel's proposal: on schedule probabilistically lose one bit at a time of encryption key
  - Overload: Forgetting functions
    - Specification of when to forget which data
    - Achieving it through aggregation and/or sampling plus deletion
  - Machine unlearning: retain learning set and summaries, and subtract what is to be forgotten from summaries

- Challenges
  - Figuring out what should be deleted
  - Figuring out when to delete
  - Trusting that the deletion happens

Cross-cutting: metrics for privacy

- Observations
  - Privacy is not binary
  - Privacy is dynamic
  - Privacy is context sensitive (Nissenbaum)
- Metrics
  - K-anonymity
  - L-diversity
  - ε differential privacy
  - Privacy-approximation ratio of function *f*: quantifying the amount of privacy afforded to participants providing sensitive information to the distributed function *f* (Feigenbaum)
  - Information theoretic approach to modeling disclosure risk measures (Bezzi)

What's missing? People & policies

- Only at data collection points?
  - Policies for data curation
  - Policies for data management
  - Policies for data fusion
  - Policies for data use: what can be asked and what can be done with the answers
- How to balance
  - Humans' ability to understand and make choices about risks/benefits
  - Legal responsibilities
  - Societal expectations and norms

What is missing? Trust

- Have talked about this in this context before
- Issues
  - Must understand the risks
  - Must understand the cost of those risks
  - Must understand the value of the trust that those risks will not occur
- Who are we trusting to behave in what ways in which contexts, and what is the cost or recourse if they do not?
- Trust frameworks: significant work on trust frameworks for identity management, but limited domain

Challenges to the collection, management and use of large amounts of data

- Notion of privacy evolving over time
  - Not binary
  - May change with time
  - Hence notions of "harms", "risks", and "costs" may change
- The whole data life-cycle is important
- Trust is critically important, but becomes increasingly complex as data management, fusion and use evolves
- Negative social implications
  - Increased and more subtle opportunities for discrimination
  - Freedom of speech
  - Reduction in possibilities for anonymity (not the same as privacy)

Thank you

- Questions?
- Contact me: Karen Sollins
    - sollins@csail.mit.edu
    - +1 617 253 6006