Big Data and Privacy

Josep Domingo-Ferrer

Universitat Rovira i Virgili, Tarragona, Catalonia

Helsinki, Oct. 5-7, 2015

Personal data protection principles in the EU law

Personal data, or more precisely personally identifiable information (PII) mean any information related to an identified or identifiable natural person.
Principles applicable to PII (European Data Protection Directive, Art. 29 DP Working party, proposed GDPR):

- **Lawfulness** (consent obtained or processing needed for: a contract or legal obligation or the subject's vital interests or a public interest or legitimate processor's interests compatible with the subject's rights)
- **Consent** (simple, specific, informed and explicit)
- **Purpose limitation** (legitimate and specified before collection)

Personal data protection principles in the EU law (II)

- **Necessity and data minimization** (collect only what is need and keep only as long as needed)
- **Transparency and openness** (subjects need to get info about collection and processing in a way they understand)
- **Individual rights** (to access, rectify, erase/be forgotten)
- **Information security** (collected data protected against unauthorized access and processing, manipulation, loss, destruction, etc.)
- **Accountability** (ability to demonstrate compliance with principles)
- **Data protection by design and by default** (privacy built-in from the start rather than added later)

Personal big data conflict with principles

- Big data result from collecting and linking data from several sources, often in a continuous way
- Unless personal data are anonymized, potential conflicts with the above principles:
  - **Purpose limitation**. Big data often used secondarily for purposes not even known at collection time.
  - **Consent**. If purpose is not clear, consent cannot be obtained.
  - **Lawfulness**. Without purpose limitation and consent, lawfulness is dubious.
  - **Necessity and data minimization**. Big data result precisely from accumulating data for potential use.
  - **Individual rights**. Individuals do not even know which data are stored on them.
  - **Accountability**. Compliance does not hold and hence cannot be demonstrated.

Pros and challenges of anonymization

- If personal big data are anonymized in such a way they are no longer identified/identifiable to specific persons, they become big data *tout court*.
- Challenges:
  - Too little anonymization (*e.g.* mere de-identification by just suppressing direct identifiers) may not be enough to ensure non-identifiability (*e.g.* AOL scandal, Netflix, etc.).
  - **Too much anonymization may prevent linking data coming from several sources and corresponding to the same/similar individuals.**

Desiderata for big data anonymization

- *Linkability*. Linking data on the same individuals coming from several sources should remain feasible to some extent on anonymized data.
- *Composability*. The privacy guarantees given by a privacy model for several separate data sets should hold to some extent when the data sets are merged.
- *Computational cost*. SDC methods used to reach a certain privacy model should be scalable to large data volumes.

Desiderata for big data anonymization (II)

How well k-anonymity and ε-differential privacy satisfy the above desiderata is examined in:

Jordi Soria-Comas and Josep Domingo-Ferrer,
"Big data privacy: challenges to privacy principles and models",
Data Science Engineering (to appear)

Recommendation: tunable and verifiable anonymization

- Privacy-first anonymization (based on enforcing a privacy model, like k-anonymity, t-closeness or ε-differential privacy) often leads to poor data utility/linkability.
- Utility-first anonymization (iteratively changing parameters until empirical disclosure risk is low enough, as usual in official statistics) is slow and lacks formal privacy guarantees.
- Verifiable anonymization (based on the permutation model) allows *exactly tuning anonymization to achieve the desired linkability while offering formal privacy guarantees to the data administrator and the subjects*.

On verifiable anonymization

Josep Domingo-Ferrer and Krishnamurty Muralidhar,
"New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users",
Technical Report, Jan. 17, 2015.
`http://arxiv.org/abs/1501.04186`

Outstanding challenges

- If data released on the same "anonymous" individuals grow over time, anonymization is unfeasible.
- If more and more "anonymized" data is linkable to the same "anonymous" individual, in the end that individual will no longer be anonymous.
  $\Longrightarrow$ Newly released anonymized data should not be linkable to previously released anonymized data.