1st Training School
COST Action IC1206
Limassol, Cyprus, October 7-11, 2015

Legal Aspects of Personal Data De-Identification

Dr. Oleksandr (Alex) Pastukhov
Senior Lecturer
Dept. of Information Policy & Governance

L-UNIVERSITÀ TA' MALTA
UNIVERSITY OF MALTA

Sorting out the Terminology

- Revocable de-identification: severing a data set from the identity of the data contributor (= data subject)
- Irrevocable de-identification (= anonymization): de-identifying a data set irrevocably meant to prevent any future re-identification even by the anonymizer(s)
  - EU: rendering data "anonymous in such a way that the data subject is no longer identifiable" (DPD Recital 26)
  - U.S. (text data context): "technology that converts clear text data into a nonhuman readable and irreversible form, including but not limited to preimage resistant hashes (e.g., one-way hashes) and encryption techniques in which the decryption key has been discarded" (DOJ 2006 recommendations)
- The opposites: re-identification and de-anonymization, correspondingly

Personal Data

DPD Art. 2

(a) 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

Regulation Art. 4

(1) 'data subject' means an identified natural person or a natural person who can be identified, directly or indirectly, **by means reasonably likely to be used by the controller or by any other natural or legal person**, in particular by reference to an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, **genetic**, mental, economic, cultural or social identity of that person;

(2) 'personal data' means any information relating to a data subject;

Relevant Types of Personal Data

- CCTV footage
  - What if there are no "**means reasonably likely to be used by the controller or by any other natural or legal person**"? "As the purpose of video surveillance is, however, to identify the persons to be seen in the video images in all cases where such identification is deemed necessary by the controller, the whole application as such has to be considered as processing data about identifiable persons, even if some persons recorded are not identifiable in practice." (WP29, Opinion 4/2007)
- Biometrical data
  - Draft Regulation: "any data relating to the physical, physiological or behavioural characteristics of an individual which allow their unique identification, such as facial images, or dactyloscopic data" (Art. 4(11))
  - Both part of the "information relating to an identified or identifiable natural person" and the identifiers that act as a link between the information and the person
  - Can't be discarded and replaced
- Sensitive data
  - Health, religion, racial or ethnic origin, etc.

Irreversibility of De-Identification

- Legal consequences:
  - EU: “the principles of protection shall not apply to data rendered anonymous” (DPD Recital 26)
  - U.S. (e-gov context): “the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation” afterwards (DOJ 2006 recommendations)
- Criteria for establishing irreversibility:
  - EU: none
  - U.S. “Standards for Privacy of Individually Identifiable Health Information”, a regulation under the Health Insurance Portability and Accountability Act (HIPAA), a.k.a. the ‘HIPAA Privacy Rule’:
    - Expert Determination
    - Safe Harbor

Expert Determination

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable, must:

- determine that the risk is very small that the information could be used by the intended recipient, alone or in combination with other reasonably available information, to identify an individual; and
- document the methods and results of the analysis leading to such determination (HIPAA s. 164.514(b)(1))

Safe Harbor

The following 18 identifiers of the individual or of relatives, employers, or household members of the individual must be removed and the covered entity must have no actual knowledge that the information could be used alone or in combination with other information to identify the individual:

(A) Names;
(B) Street address, city, county, precinct, zip code, and their equivalent geocodes, except for geographic units with population under 20,000 people;
(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death;
(D) Telephone numbers;
(E) Fax numbers;
(F) Electronic mail addresses;
(G) Social security numbers;
(H) Medical record numbers;
(I) Health plan beneficiary numbers;
(J) Account numbers;
(K) Certificate/license numbers;
(L) Vehicle identifiers and serial numbers, including license plate numbers;
(M) Device identifiers and serial numbers;
(N) Web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) Biometric identifiers, including finger and voice prints;
(Q) Full face photographic images and any comparable images; and
(R) Any other unique identifying number, characteristic, or code.

(HIPAA s. 164.514(b)(2)).

Degrees of Irreversibility

- Biometric data are extremely hard to de-identify (esp. DNA sequences)
- In the age of 'big data', it's more appropriate to speak of degrees of irreversibility
- It can be useful to preserve some identifiers that can be re-linked to the data subject by a trusted party in certain situations (e.g. epidemic)
- HIPAA provides for a 'limited data set', i.e. individually identifying health information with all the 18 identifiers, except town or city, State, zip code and date of birth, removed (s. 164.514(e)(2)) . A covered entity may use or disclose a limited data set only
  - for the purposes of research, public health, or health care operations (s.164.514(e)(3)); and
  - if the covered entity enters into a data use agreement that meets the requirements of s. 164.514(e)(4)

Past Failures

- Narayanan & Shmatikov: “Any information that distinguishes one person from another can be used for re-identifying data”
- They identified Netflix Prize Dataset (competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings only) users by matching the data sets with film ratings on IMDb (2007)
- MIT & UCLouvain identified 95% of 1.5 mln cell phone users in an unnamed small European country by analyzing their telecom data over 15 months and found that just 4 points of reference, with fairly low spatial and temporal resolution, were enough

Lessons for the Future

- De-identification at the point of collection, i.e. camera
- Privacy by design and by default: camera either ‘smart’ or ‘dumbed down’
- Original data are processed onboard and discarded
- When only detection of human presence is needed, motion or heat detectors or images useless for identification purposes mandated
- Laws containing rules on de-identification similar to those of the HIPAA Privacy Rule
- Industry self-regulation: codes of conduct (encouraged by DPD, but dropped in the draft Regulation)
- Operational safeguards (internal policies & procedures), awareness rising, training, education and incident response planning

LexConverge

Law ▸ Across Disciplines ▸ Across Technologies ▸ Across Cultures

Thank you for your attention!

oleksandr.pastukhov@um.edu.mt
opastukhov@mappingtheinternet.eu

www.smartsurveillance.eu www.respectproject.eu
www.mappingtheinternet.eu

L-UNIVERSITÀ TA' MALTA
UNIVERSITY OF MALTA