Anonymous Big Data Workshop

Executive Summary

Vienna University of Economics and Business

Vienna, 20/01/2020

ANITA Anonymous big data

FFG Forschung wirkt.

Bundesministerium Verkehr, Innovation und Technologie

Program "ICT of the Future" – an initiative of the Federal Ministry for Transport, Innovation, and Technology

Workshop format

The *goal* of the workshop was to explore the topic of Synthetic Data from multiple perspectives and to create a collaborative dialogue around the following questions:

1. Opportunity: Which type(s) of privacy-sensitive data assets are of interest for (market) research?
2. Utility: What are requirements with regard to accuracy and representativeness for synthetic data?
3. Legal: Which legal frameworks are to be considered for synthetic data generation?
4. Trust: What is required to establish trust in synthetic data or other forms of privacy preservation (e.g., data minimization), in terms of accuracy and privacy?
5. Communication: How are data synthetization and other forms of privacy preservation perceived by the general public?
6. Ethics: Are there other ethical questions, aside from privacy, with respect to synthetic data?

23 experts with data science, marketing, legal, privacy, ethics and philosophy backgrounds participated in the workshop.

The workshop was conducted in the form of a carousel brainstorming, where the participants were divided into 6 small groups. They rotated through 6 stations (each of the above mentioned questions constituted a separate station) and collaboratively brainstormed responses to the question at each station.

Discussion results

Opportunity. There are four large groups of potentially privacy sensitive data that could be of interest for (market) research: (i) behavioral tracking data, (ii) demographic and socio-economic data, (iii) attitudinal/preferential data, and (iv) sensor data.

Utility. Synthetic data have to be as close to the original data as possible. This contradicts the requirement of privacy. As a result, there has to be a trade-off between privacy and utility. This trade-off will highly depend on the prediction task.

Legal. GDPR and industry-specific legislations are the main legal frameworks to consider for synthetic data generation. There could also be a need for ethical guidelines. Certifications, standards and external auditing procedures could be beneficial for the synthetic data generators.

Trust. Standard metrics, quality controls, trust frameworks, hands-on trainings, information about benefits for data protection, additional ethical requirements could be introduced to gain trust among the data suppliers, data users, and society in general.

Communication. The perception of the synthetic data by the general public clustered into the following topics: (i) motivation to be interested in the synthetic data, (ii) lack of understanding of the methodology itself and its quality metrics, (iii) necessity to disclose the methods, (iv) ways to communicate the synthetic data topic effectively, and (v) general trust issues.

Ethics. The following groups of ethical questions with respect to synthetic data were identified: (i) synthetic data creation, (ii) synthetic data usage, (iii) data ownership, (iv) information disclosure, and (v) fundamental ethical questions.