Written Evidence submitted by Yves-Alexandre de Montjoye and Andrea Gadotti (Computational Privacy Group, Imperial College London) (DDA0016)

Yves-Alexandre de Montjoye
Associate Professor, Department of Computing and Data Science Institute, Imperial College London
Head, Computational Privacy Group
Director, Algorithmic Society Lab

Andrea Gadotti
Doctoral researcher, Department of Computing (Computational Privacy Group), Imperial College London

The Computational Privacy Group at Imperial College studies the privacy risks arising from large-scale datasets. We develop attack models to test mechanisms and design solutions to collect and use data safely.

Summary:

This submission focuses on paragraph 123 on anonymisation in the UK Government's consultation *Data: A new direction* and on the role of modern privacy engineering technologies for the safe use and sharing of data for research.

Key points include:

- Paragraph 123 of the Government's consultation proposes that data may be considered anonymous whenever it is not identifiable by the controller who processes it.
- This might be interpreted to suggest that even if data were to be highly identifiable in the hands of someone other than the controller, such data would fall outside the scope of the GDPR.
- Such an interpretation would be likely to substantially lower the technical standards for anonymisation.
- Lowering the technical standards on anonymisation would be both unnecessary and counterproductive. Modern privacy engineering can be used to collect, share, and analyse data while providing strong privacy guarantees. Lower standards on anonymisation would weaken the incentives to their development and adoption and create significant risks to the privacy of individuals.

Anonymisation standards and privacy engineering for data sharing

1. Sharing data for research is of the utmost importance. The GDPR already contains several provisions (e.g. Article 89) that make it easier for controllers to share data for research purposes. These provisions apply to the sharing of personal data — i.e. data that relates to individuals who can be identified (directly or indirectly). Alternatively, the data can be shared in anonymised form. In this case, the data can be shared without any restrictions as the GDPR does not apply to data that has been rendered anonymous. In particular, it can be legally transferred to any country outside the UK. The rationale is that, if the data is anonymous, then individuals are given strong guarantees that the data cannot be feasibly linked back to them or negatively impact them in any way. To ensure that anonymous data in fact achieves such guarantees and doesn't become a data protection loophole, the GDPR sets a high standard on what constitutes anonymous data — for which all reasonably likely means of identification available to any person should be taken into account (Recital 26).

2. At a technical level, anonymisation has been historically performed through de-identification techniques. These modify the dataset (containing individual-level record) to obtain an altered version of the dataset which preserves the statistical properties of the data while trying to prevent re-identification. However, these techniques are usually not effective to safely anonymise modern types of data — often called big data — where many data points are available for each individual. For example, our recent work published in Nature Communications shows that only 15 attributes (such as age, postcode, number of kids, etc) are enough to uniquely identify 99.98% of the individuals in the population[1]. To allow anyone to explore their own risk of re-identification, we have released an interactive tool[2]. Overall, the literature shows that traditional de-identification techniques do not, most of the time, provide an acceptable privacy/utility trade-off for most modern data.

3. Due to these limitations, some have argued that policy-makers should weaken the technical standards required by anonymisation. We are under the impression that some proposals contained in the Government's consultation (Data: A new direction) document might be interpreted as going in this direction. For example, paragraph 123 states that *"If the data controller has no way of obtaining the means of identification, the data is not identifiable in the hands of the controller (and so is anonymous in that particular controller's hands even though it may be identifiable in another controller's hands)"*. While we agree that the risk of re-identification depends also on the means available to the recipient, this formulation might lead highly identifiable data (such as pseudonymous data) to be readily considered anonymous — and hence outside the scope of the GDPR. If this interpretation were to become correct, it would pose significant risks to the privacy of UK citizens. In particular, anonymous data can be legally transferred abroad or freely sold.

4. Importantly, we believe that weakening the standards for anonymisation is counterproductive and unnecessary. Modern privacy engineering can be used to collect, share, and analyse data while providing strong privacy guarantees. Some of these technologies are particularly promising and can be combined to achieve very high privacy standards:

    - Query-based systems, which allow controllers to store datasets safely behind their servers without sharing the raw data with researchers, but allowing them to send queries on the data and receive only answers aggregated over many users. These systems typically support strong security measures (e.g. authentication, activity logging, and others) that can reduce even further the chances of privacy violations. Several solutions already exist, such as OPAL[3] — developed by our group at Imperial, together with MIT, Orange, and other partners — and OpenSafely by the University of Oxford.

---

[1] Rocher, L., Hendrickx, J.M., de Montjoye, Y.-A., 2019. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications 10, 1–9. https://doi.org/10.1038/s41467-019-10933-3

[2] Available at https://cpg.doc.ic.ac.uk/observatory/

[3] Oehmichen, A., Jain, S., Gadotti, A., de Montjoye, Y.-A., 2019. OPAL: High performance platform for large-scale privacy-preserving location data analytics, in: 2019 IEEE International Conference on Big Data (Big Data). Presented at the 2019 IEEE International Conference on Big Data (Big Data), pp. 1332–1342. https://doi.org/10.1109/BigData47090.2019.9006389

- Differential privacy, a framework of mechanisms to analyse data which provide strong mathematical guarantees of privacy. These can be particularly useful to share aggregate data broadly for transparency and reporting purposes[4].
- Synthetic data, which can be mostly useful to share broadly datasets which "look like" the original one in order for researchers to run tests and simulate analysis (which can then be run on the real data in a more controlled environment). Importantly, while synthetic data can help test hypotheses, these have then always to be validated against the real data.

5. Beyond technical standards, we believe that transparency and sensible governance are key to achieving trust. Review of projects by an ethics committee, open and accessible documentation on which projects are carried out on which data, and active engagement with communities who provide their data and can be affected by decisions made based on the data (a principle at the heart of OPAL)— are all best practices that should be encouraged.

6. Making data easier to share for research is crucial, but this does not require to weaken the standards on anonymisation. Modern techniques can provide an excellent privacy/utility trade off, and UK research institutions and companies are already at the forefront in the development of such techniques.

7. In summary, we recommend that:
   - The anonymisation standards required by the GDPR should be preserved.
   - The Government and authorities such as the ICO should discourage the use of weak anonymisation techniques and instead encourage the use of modern privacy engineering technologies. This includes providing guidance on how anonymisation techniques can and cannot be employed to meet the intent and standards of the law.

January 2022

[4] The practical guarantees of differential privacy mechanisms should be however evaluated carefully based on the context, see for example our recent article: Houssiau, F., Rocher, L., de Montjoye, Y.-A., 2022. On the difficulty of achieving Differential Privacy in practice: user-level guarantees in aggregate location data. Nat Commun 13, 29. https://doi.org/10.1038/s41467-021-27566-0