The technical fix for international data transfers — a word of caution

Sophie Stalla-Bourdillon, Senior Privacy Counsel and Legal Engineer, and Alfred Rossi, Research Scientist, at Immuta, analyse the EDPB's recommendations on pseudonymisation and split or multi-party processing to bolster the operation of SCCs to achieve lawful international data transfers

In November 2020, the European Data Protection Board ('EDPB') issued draft recommendations ('the Recommendations', copy at: www.pdpjournals.com/docs/888116) in an attempt to shed some light on the implications of the *Schrems II* (C-311/18) decision and explain how Standard Contractual Clauses ('SCCs') and other 'appropriate safeguards' under Article 46 of the General Data Protection Regulation ('GDPR') could help legitimise international data transfers.

The Recommendations confirm the main message stemming from the Court of Justice of the European Union's ('CJEU') judgment: that contractual and organisational measures are not enough to prevent access to data by intelligence services. The question that remains is whether supplementary measures — and in particular technical measures — could ever usefully complement SCCs when the third country's overall legal framework does not provide EU citizens with enforceable rights and effective legal remedies against unlawful foreign intelligence surveillance.

The draft Recommendations attempt to detail and assess the potential of various technical safeguards. However, not many appear promising. This article explains how two of these technical safeguards in particular — pseudonymisation and split or multi-party processing — do not easily legitimise computation and data analysis at the data importer's end.

1. Pseudonymisation

The Recommendations offer 'use case studies' (pages 21 to 27 of the document) and use case study 2 covers the transfer of pseudonymised data to a third country for analysis. The Recommendations are slightly confusing here, however. This is because they seem to refer to the standard for anonymisation rather than pseudonymisation.

It is important to understand these two key concepts properly. De-identifying personal data requires classifying data into two groups: direct identifiers and indirect identifiers. Identifiers are personal attributes that can be used to help identify an individual. Identifiers that are unique to a single individual, such as social security numbers, passport numbers, and taxpayer identification numbers are known as 'direct identifiers'. The remaining kinds of identifiers are known as 'indirect identifiers', and generally consist of personal attributes that are not unique to a specific individual on their own. Examples of indirect identifiers include height, ethnicity, hair color, and more. Indirect identifiers can be used in combination to single out an individual's records.

Pseudonymising personal data is often thought as a means to transform personal data in such a way that the individual is not directly identifiable. This has been confirmed in prior guidance issued by the Article 29 Working Party (which has been replaced by the EDPB), the French Supervisory Authority (the CNIL), German Supervisory Authorities, and the UK Supervisory Authority.

The most recent guidance from the European National Security Agency ('ENISA') adds that best practice mandates the adoption of a risk-based approach to be in a position to optimise the tradeoff between security and utility, and states that:

"When considering the application of pseudonymisation to real-world scenarios, this trade-off should be analysed carefully, so as to optimise utility for the intended purposes while keeping the protection of the pseudonym holders (data subjects) as strong as possible."

Importantly, even if the data are pseudonymised, it is possible that by combining pseudonymised data with information that is publicly available or attainable, the individual to whom the data pertain remains (indirectly) identifiable. This is implicit in GDPR Article 4(5), which only considers the additional information that is kept separately, and is subject to technical and organisational measures to determine whether it can be attributed to an identified or identifiable natural person.

However, the EDPB states in its use case 2 that in order to make pseudonymisation an effective supplementary measure: "the controller

[must] establish by means of a thorough analysis of the data in question taking into account any information that the public authorities of the recipient country may possess that the pseudonymised personal data cannot be attributed to an identified or identifiable natural person even if cross-referenced with such information."

In other words, to make pseudonymisation an effective supplementary measure, it is necessary to look beyond the additional information held exclusively by the data exporter, and consider publicly available or attainable information to which an attacker has or could have access to, which is not necessarily required under Article 4(5) of the GDPR. As a result, in order to assess whether pseudonymisation is sufficient, it is necessary to determine whether the individual is indirectly identifiable as well. The end goal is thus now moving towards that of anonymisation.

As a reminder, data have successfully undergone anonymisation if 'the data subject is not or no longer identifiable' both directly and indirectly, as per Recital 26 of the GDPR, which should be read together with Article 4(1). To determine whether the individual does remain (indirectly) identifiable, best practice suggests implementing a 'formal attack' model.

One model that could make sense for international data transfers to the US, where the concern is surveillance by US intelligence services, is the database cross match model (DBXM, Elliot & Dale, 1999), which posits an attacker who aims to enrich their database by linking it to a target data set. This model is related to the common attack models used in privacy risk assessments, such as the prosecutor and journalist attack models (PAM and JAM, respectively, defined in (Marsh et al., 1991). Attack models are often used when anonymising data for a public release or for clinical research purposes.

"Pseudonymisation under EDPB's use case 2 would thus require the implementation of sophisticated 'anonymisation techniques.' Although these techniques can have serious utility implications, the best solutions available on the market are able to prioritise the most useful data attributes for the analysis and optimise utility for a given set of queries."

The European Medicines Agency acknowledged this by distinguishing between the approach followed by the Article 29 Working Party in its 2014 Opinion, and a risk-based approach that is seemingly not precluded by that Opinion. Note that the 'motivated intruder test' set forth by the ICO is less restrictive than DBXM in that it assumes the attacker, called a motivated intruder, is not a specialist.

More specifically, under DBXM, an attacker wants to associate records within their database with individual records in a target data set. We may assume the attacker's database is quite rich, and has already incorporated all publicly available information or, if we would like to be conservative, publicly attainable information. The overall re-identification risk is a conditional probability consisting of the product of two factors: the probability of identification given that an attempt is occurring, and the probability of an attempt. The former factor, which is amenable to theoretical and objective analysis, can be viewed as the re-identification risk given the use of privacy enhancing data transformation techniques, (i.e., the data risk). The latter factor (i.e., the context risk) remains situationally dependent, and is more subjective since it primarily depends upon situational, environmental, and/or incentive-based factors that often cannot be reliably quantified. This explains why the EDPB rejects the context risk, as it is hard to work with such a factor when the attacker, such as an intelligence service, is not well understood. The EDPB expressly states that "if you still wish to proceed with the transfer, you should look into other relevant and objective factors, and not rely on subjective factors such as the likelihood of public authorities' access to your data in a manner not in line with EU standards."

Still, it is possible to argue that from analysis of the data risk alone, the overall re-identification risk is remote. This is because the overall re-identification risk cannot exceed the data risk, and thus the overall risk is mitigated if the data risk is sufficiently controlled.

A few data transformation techniques can potentially be used to achieve a 'remote' data risk. The most obvious ones are k-anonymisation and differential privacy. While k-anonymisation perturbs indirect identifiers to make it possible for individuals to hide in groups of a k number of individuals, differential privacy randomises query computation to produce query results in the form of safe aggregates.

Pseudonymisation under the EDPB's use case 2 would thus require the implementation of sophisticated 'anonymisation techniques.' Although these techniques can have serious utility implications, the best solutions available on the market are able to prioritise the most useful data attributes for the analysis and optimise utility for a given set of queries.

2. Split or multi-party processing

The second supplementary technical measure worth mentioning is what the EDPB calls 'split or multi-party processing' in its use case 5. However, it is difficult to understand from the EDPB's description when such a split or multi-party processing would make sense.

(Continued on page 8)

(Continued from page 7)

For context, secure multi-party computation ('SMC') is a branch of cryptography concerned with designing protocols that enable somewhat adversarial parties to jointly perform a computation, while keeping their respective inputs secret from each other. For example, two hospitals may wish to determine which patients they treat in common without revealing to each other the names of other patients. As another example, a government may wish to enact a voting system whereby all members vote to elect a leader while keeping their votes secret from each other.

Roughly speaking, the parties participating in these protocols are only able to learn whatever is jointly inferable from the knowledge of their own input together with the output, if permitted to see the results. Typically, all parties learn the output of the joint computation. However, protocols can be designed so that only some subset of the parties learn the results. This is of particular relevance here, as we are directly concerned with what is inferable by parties in other jurisdictions.

Use case 5 seems to envision a scenario in which the data exporter is in possession of personal data to be outsourced for processing. Before discussing this in detail, let's outline two scenarios where cryptographic computation may be useful in light of the *Schrems II* decision:

- a data exporter wishes to utilise the computational resources operating in another jurisdiction; and
- a data exporter wishes to engage in joint processing with contributing collaborators residing in other jurisdictions, and party members would like their input to remain secret from the data exporter and/or the jurisdiction thereof. Here, contributing means that the collaborator will contribute input data to the joint processing activities, as opposed to simply carrying out a computation on behalf of other parties. In other words, some of the input data comes from data which resides in this jurisdiction.

The first scenario can be addressed by fully homomorphic encryption ('FHE'), which enables the data exporter to provide encrypted inputs to processors in untrusted jurisdictions who can then, without decryption, compute the encryptions of results. In turn, those results can only be decrypted, and therefore read, by the data exporter.

The second scenario can be addressed by SMC. The parties utilise SMC to perform the desired processing over their private inputs. Upon completion of the protocol, the parties learn nothing more than what is jointly inferable from the processing results and their respective inputs. If desired, the protocol may be designed such that only the data exporter learns the results.

As written, the 'split or multi-party processing' scenario seems to have elements of both Items 1 and 2. However, in adopting mutually exclusive requirements from both, it fails to satisfy either.

The scenario description appears to suggest that the data exporter is in possession of the entire input, which they split and distribute to processors in other jurisdictions. This is curious for a few reasons:

- if the data exporter is already in possession of the entire input (specifically in the sense that they do not require the private contributions of extra-jurisdictional party members), then there is nothing preventing the data exporter from simply performing the processing themselves; and
- if the split portion of the data received by a party is encrypted beyond their means to manipulate it, and they possess no data of their own to contribute, then this situation is equivalent to FHE with the additional and artificial constraint that the evaluation be distributed across jurisdictions.

Instead, the split appears intended to force alignment to a setting reminiscent of SMC. This is made clear in the list of conditions under which the EDPB would consider split processing an effective supplementary measure. The first condition states: 'a data exporter processes personal data in such a manner that it is split into two or more parts each of which can no longer be interpreted or attributed to a specific data subject without the use of additional information', while a later condition continues 'the processors optionally process the data jointly, e.g. using secure multi-party computation, in a way that no information is revealed to any of them that they do not possess prior to the computation.'

It should be noted that neither of the above conditions explicitly require encryption; only that the data not be attributable to a specific data subject without the use of additional information. This rules out the possibility for participating extra-jurisdictional processors to augment their received portion of the data split with additional personal information, simply because joining to the received data implies being able to match on de-identified records, which in turn implies the existence of an extra-jurisdictional means of re-identification.

In the absence of requiring party members to keep their inputs confidential, the problem is trivially solvable outside of this use case by having these collaborators send their inputs to the data exporter's jurisdiction for processing.

To make sense of use case 5, we need to assume all parties are contributing parties with confidential input, and that all extra-jurisdictional partners are unable to learn anything new about the exporter's data from their participation, which is an extremely narrow use case. What is more, there also appears to be a gap between satisfaction of the use case criteria and requirements for safe implementation of SMC.

In short, the EDPB's legitimate data transfer use cases are likely to appear too narrow to many.

Sophie Stalla-Bourdillon and Alfred Rossi
University of Southampton and Immuta
sstalla-bourdillon@immuta.com
a.rossi@immuta.com