Multi-Party Secure Private Computing and Quality: a prospective alliance

Fabio Ricciato

Eurostat, Luxembourg, Fabio.RICCIATO@ec.europa.eu

Abstract

Multi-Party Secure Private Computing (MP-SPC) solutions represent a sub-group of Privacy Enhancing Technologies that enable two or more organisations holding confidential data sets to cooperate and compute statistical results based on the integration/combination of their respective datasets without revealing their input data to each other or to an external third party. Systems based on such technologies may facilitate cooperation among statistical organisations and with external data providers, and contribute to overcoming the barriers – particularly concerning legal compliance and public acceptance – to the (secondary re)use for statistical purposes of data collected primarily for non-statistical purposes by other entities. In this contribution, starting from the definition of statistical "quality" as defined in the European Statistics Code of Practice (CoP) and the Quality Assurance Framework of the European Statistical System (QAF), we elaborate on the relationship between MP-SPC and statistical quality. We show that such relationship is one of mutual reinforcement and alliance: MP-SPC effectively contributes to strengthen statistical quality along several dimensions, while a proper consideration of the quality dimensions spelled out in CoP and QAF may guide the design and implementation of effective MP-SPC systems, both at the technical and organisational level.

Keywords: Privacy-Enhancing Technologies, Secure Private Computing, Quality Assurance Framework, European Statistics Code of Practice, Data Protection

1. Introduction

Multiple innovation trends in the field of Official Statistics concur to increase the demand for statistics based on the integration of multiple data sets held by different organisations. When the desired output statistics requires the joint processing of multiple input data sets, the traditional solution foresees that the involved data holder share (i.e., transmit a copy of) their input data sets with each other or with a so-called "trusted third-party". In so doing, some mild form of protection may be adopted, e.g., removal of unique identifiers or other simple forms of pseudonymisation. In this scenario, the receiving entity take some legally binding commitments, for instance: (1) to use the data exclusively for the agreed-upon purpose; (2) to keep the data stored only for the time that is strictly necessary to produce the desired statistics; (3) to keep the data safe and protect them from external intruders, and so on. The transmitting entity must trust the receiving entity to live up to these commitments, i.e., that the receiving entity has both the *intention and the capability* to keep the data safe and prevent possible misuses, but in general it has no technical means of controlling or verifying the actual fulfilment

of the agreed-upon conditions. Therefore, from the perspective of the transmitting data holder, sharing confidential data with another entity increases the exposure of data to potential risks. However, the traditional “data sharing” approach is one, not the only possible approach, and hereafter we consider possible alternatives based on recent technologies.

We restrict our attention to computation functions that involve the kind of parsimonious low-dimensional statistical methods and models that are common in the “classical” statistical production (e.g., counting, averaging, regression in low-dimensional subspaces). We do not consider here high-dimensional computational models, as common in large-scale Machine Learning and deep learning, that we believe require a separate treatment when it comes to data protection aspects.

2. Privacy Enhancing Technologies for Multi-Party Computation

The term Privacy Enhancing Technologies (PET) is used to refer collectively to a range of technological approaches having in common the quest for mitigating, if not resolving, the tension between data (re)use and data protection. We focus in this paper on the problem of enabling the computation of some desired statistics, according to a pre-defined method or function, requiring the integration of multiple input data sets held by different entities (“input parties”). When applied to the problem of computing across data held by multiple parties (“multi-party computation” problem), the solution approaches offered by the PET family can be divided into two large groups.

The first group is based on so-called “Input Privacy” technologies. These are based on concepts and primitives from the field of cryptography (e.g., secret sharing, homomorphic encryption) and seek to enable the computation of the exact desired output statistics *without exposing the input data in intelligible form* to any different entity other than the originally data holder. A combination of technological and organisational measures are put in place to prevent the extraction of input data elements and any information derived thereof other than the result of the agreed-upon computation task. Input Privacy technologies are also referred by the term Secure Private Computing (SPC), whereby the term “Secure” links to the cryptographic nature of the core system components, and “Private” signifies that the input data are never disclosed to any single entity other than the original data holder. When applied to the computation of statistics based on multiple data sets held by different and mutually distrusting parties, we refer to such solutions with the term Multi-Party Secure Private Computing (MP-SPC).

Like all security technologies, MP-SPC systems must be designed based on a set of requirements and assumptions about the capabilities of potential adversaries against which the system should offer protection, that collectively represent the “adversary model”.

Robustness against higher levels of adversary capabilities correspond to higher levels of sophistication and complexity (and ultimately cost) of the MP-SPC system.

The second group of PETs is based on so-called "Output Privacy" methods. These methods seek to transform each original input data set into some "sanitised" approximated version that is (i) sufficiently different from the original data to prevent reidentification risks when the data are disclosed to untrusted parties, but at the same time (ii) sufficiently close to the original data to enable computation of an acceptable approximation of the desired output result. These methods may be considered extensions of and derivations from traditional Statistical Disclosure Control (SDC) techniques. However, while classical SDC methods seek to protect *aggregate* data with considerably lower granularity and information detail than the original from micro-data sets (e.g., tables), in the multi-party computation scenario at hand they seek to deliver transformed sets of detailed micro-data with the same level of granularity as the original data. The applied transformations typically involve explicitly or implicitly some form of randomisation or pseudo-random perturbation. Differential Privacy and generation of (pseudo-)Synthetic Data based on deep learning models are among the proposed approaches in this area. Despite the hyping attention given to such group of methods, we have serious doubts about the effectiveness and theoretical viability of these approaches in the multi-party computation scenario considered here (see e.g. the critical examination by Stadler and Troncoso (2022) and references therein). These approaches are not considered further in the remaining part of the paper and any reference made hereafter to "PET" is to be interpreted as being referred exclusively to Input Privacy approaches.

3. A shared Multi-Party Secure Private Computing system

In 2021-2022 Eurostat started to elaborate the concept of a shared MP-SPC system for the European Statistical System (ESS) as part of the work conducted in the context of the UNECE HLG-MOS project on Input Privacy Preservation[1]. The idea was to build a single system, based on the most advanced available MP-SPC technologies, to offer secure computation services *on demand* to the ESS members and their partners. The envisioned system, initially termed MPSPC-as-a-Service, would be developed, built, owned, and used by the ESS members. The motivations and the main characteristics of such a system were presented in a recent paper by Ricciato (2024). Therein we indicated that the **GDPR principles should serve as high-level design requirements** for the development and specification of the envisioned MP-SPC system. In the following sections we complement that view by suggesting that also the quality

[1] The Final Report of the project is available from https://zenodo.org/records/10400296

dimensions encoded in the European Statistics Code of Practice (CoP) and Quality Assurance Framework of the European Statistical System (QAF) may provide additional inspiration and useful guidance for the specification of the envisioned system.

4. Quality motivations for adopting a shared MP-SPC system

In this section we elaborate on the costs and benefits of adopting a shared MP-SPC solution (such as the one elaborated conceptually in Ricciato, 2024) from the perspective of the "quality" dimensions defined by the CoP. Towards this aim, we shall refer to a hypothetical scenario where two organisations hold two sets of confidential micro-data and are considering producing a new statistical indicator based on their integration. This scenario may represent for example on National Statistical Institutes (NSI) and a private data holder, or two NSIs from different countries. The following alternative options are considered:

A. The desired statistics is produced with a **shared MP-SPC solution** developed by the ESS and made available to all ESS members and partners.
B. The desired statistics is produced with a **non-shared MP-SPC solution** developed and deployed ad-hoc for this specific computation task by the involved NSI.
C. The desired statistics is produced based on **plain data transmission** to some trusted party (traditional data sharing).
D. No micro-data set integration takes place: an approximation of the desired statistics is produced based on aggregate data computed from individual data sets.
E. The reuse of the available micro-data set is abandoned, and a new data collection is launched (e.g., a new survey).

It should be made clear that all these options are perfectly legitimate and each of them may preferred over the others depending on the specific scenario at hand. In other words, no single option is “best” in all scenarios, and the choice must be done case-by-case considering multiple contextual aspects. Our contribution here is to map the relevant dimensions to the “quality” principles defined in the CoP, and based on those propose a sort of “decision tree” to guide the decision (in the text below we indicate each principle by the number assigned in the CoP).
In most practical scenarios it may be expected that, when some “reusable” data are already available, option E will be readily dismissed due to the increased **burden on respondents (Principle 9)** and **cost effectiveness (Principle 10).**

The next candidate option to consider is D. With this option, some local pre-processing takes place on each individual micro-data set at the premises of each data holder, resulting in some intermediate aggregate data that is considerably less sensitive than the original full data set (or not at all sensitive at all) in terms of confidentiality risk. The intermediate aggregate data

from each data holder are then exchanged and combined into the final indicator. This approach is not always feasible and anyway may lead to a coarse approximation of the desired statistics. In fact, unless the intermediate aggregate data represent a sufficient statistic (in the mathematical sense), the "factorisation" of the computation procedure introduces a certain error, thus impinging on the **accuracy and reliability (Principle 12).** If the error can be safely quantified and is guaranteed to remain below an acceptable threshold, then option D should be preferred. Conversely, when appropriate levels of accuracy cannot be ensured, then option D should be dismissed and the remaining options for micro-data integration are then evaluated. Option C represents the traditional approach of plain data sharing *in intelligible form*: individual data records are transmitted in a form that is immediately interpretable intelligible by the receiving party in charge of executing the computation. Information that is not strictly necessary for the computation of the desired indicator should be stripped away before transmission in accordance with the data minimisation principle of GDPR. Direct identifiers (e.g., name or social security number) should be removed at least replaced by less informative *pseudonyms*. from a legal standpoint both the removal and the replacement of direct identifiers constitute *pseudonymisation, not anonymisation*. Data minimisation and data pseudonymisation contribute to reduce the confidentiality risk and should be adopted whenever possible (if they *can* then they *must* be adopted). Depending on various contextual elements, vanilla pseudonymisation may or may not suffice to achieve acceptable levels of **Statistical Confidentiality and Data Protection (Principle 5)**. Several contextual factors play a role in the assessment, including (but not limited to) the nature, granularity, and content of the input data (depth) and the number of data subjects represented in the data set relative to the whole population (breadth). To illustrate, take for example the following two kinds of data set: (*I)* a set of micro-data records containing the place of residence and the place of employment for a sample of 0.1% of the total population in a certain reference year, and *(II)* another set of "nano-data" (following the definition given in Ricciato et al., 2020) containing the precise location of the whole population recorded at every minute for the whole year: both data sets constitute "personal data" and should be regarded as confidential, but they clearly involve different level of risk and therefore require different levels of protection. In this toy example, Option C may be perfectly adequate to handle data sets of type *(I)* but may not suffice to deal with data of type *(II)*. Notably, the latter entail higher *risk* in terms of both higher *impact* of a hypothetical attack, as more detailed information would be revealed to the successful attacker, and higher *probability* of attack, as the perspective of acquiring a richer data set would attract more numerous and more powerful potential attackers.

In general, the assessment as to whether the adopted protection measures are proportionate to the risk must be carried out case by case in the form of a Data Protection Impact Assessment, as required by GDPR and explicitly recalled in the QAF under Indicator 5.5. In other words, ensuring appropriate levels of confidentiality and data protection is considered also a matter of *quality* assurance.

The transmission of confidential data *in intelligible form*, as per option C, involves risks on both the data transmitter and data receiver. Any breach of data confidentiality or data misuse on the side of the entity receiving the data would probably have consequences also for the transmitting entity, at least in terms of reputation damage, if not legal liability. As a matter of fact, the transmitting entity must trust not only the *intentions* of the receiving entity to exclude deliberate misuse of the acquired data, but also its *technical capabilities* to prevent intrusions and attacks against its IT infrastructure by other rogue actors.

In the light of the above considerations, we must consider scenarios where option C is deemed insufficient and higher levels of data protection are required at the data processing stage. As statistical institutions tend to expand the pool of data sources serving official statistics, also in the direction of privately held data, and develop newer and more sophisticated multi-source indicators, we may expect a proliferation of scenarios where some of the involved parties require stronger data protection guarantees, beyond what can be attained with plain data sharing agreements. This is where MP-SPC solutions come into play.

The choice between option A and option B, or equivalently between shared and dedicated MP-SPC solutions, is essentially a matter of **cost effectiveness (Principle 10).** Even if certain Input Privacy technologies are already relatively mature, they are far from being commoditized. Statistical offices, like other potential adopters, need to mobilise experts from different areas, including technology specialists and legal experts, to identify the technological components that are best suited for their needs and understand how to introduce them into the organisation. On top of all such knowledge capital, they need to mobilise also financial resources for the development, deployment, and operation of the MP-SPC system. Mobilising all such human and financial resources is unlikely to be justified by the computation of a single statistical indicator by a single statistical office, as per option B. The resource limitations may induce the potential adopter to opt for less robust technological components and shacky design, but this would come in contradiction with the goal of increasing the level of data protection. In other words, with option B the statistical office would have to compromise between the quality dimensions of **cost effectiveness (Principle 10)** and **statistical confidentiality and data protection (Principle 5)**. Instead, option A would allow to pool resources (human and

financial) and develop a shared system with top-level features at acceptable costs, in this way fulfilling both quality dimensions.

5. Quality considerations for designing a shared MP-SPC system

In the previous section we have seen how the decision to opt for a shared MP-SPC system may be entirely justified (in certain scenarios) based on quality considerations hooked to the CoP principles. In this section we take a step further and provide initial hints as to how a well-designed shared MP-SPC system for the ESS may offer additional opportunities for improving on additional quality aspects as detailed in QAF.

The adoption of a shared MP-SPC system by the ESS should ideally lower the barriers against the (re)use of non-statistical data for statistical purposes, and in this sense, it would directly contribute to improve *Indicator 2.4 (Access for statistical purposes to other data, such as privately held data, is facilitated, while ensuring statistical confidentiality and data protection)* as well as *Indicator 8.6 (Agreements are made with holders of administrative and other data which set out their shared commitment to the use of these data for statistical purposes).*

The technical specifications of the shared MP-SPC system could be published to enable and even encourage scrutiny by qualified external experts. Setting in place a formal system for reporting and reacting to possible glitches would then qualify as a measure for continuous quality improvement, thus contributing to *Indicator 4.2 (Procedures are in place to plan, monitor and improve the quality of the statistical processes, including the integration of data from multiple data sources*) as well as to *Indicator 8.3 (Statistical processes are routinely monitored and revised as required)*. At the same time, as the specification of a shared MP-SPC system include both technical and organisational measures, their publication would qualify as a contribution to *Indicator 5.4 (Guidelines and instructions are provided to staff on the protection of statistical confidentiality throughout the statistical processes. The confidentiality policy is made known to the public).*

The operation of any MP-SPC system, including a shared one, requires that the computation method is encoded in some programming languages (e.g., a script), and then vetted by the entities in charge of authorizing ex-ante the execution of the computation task (e.g., the statistical office and the other data holder(s)). The system should also produce logs to enable ex-post audit. For the sake of transparency as specifically included *Indicator 6.4 (Information on data sources, methods and procedures used is publicly available*) the scripts and even part of the logs may be made available to the public or at least to other NSIs, e.g. in the context of peer-reviews.

6. Summary and Outlook

In this contribution we have tried to establish an initial bridge between Input Privacy and the ESS quality framework. We have shown how the choice of adopting a shared MP-SPC system by the ESS may be viewed through the glasses of the CoP and QAF and interpreted as a sort of "quality optimisation". In other words, MP-SPC may be seen as an instrument to improve statistical quality and reinforce the implementation of the quality framework. At the same time and in the reverse direction, we have shown that the quality framework may help to improve the specification of a shared MP-SPC system. We have provided initial hints as to how valuable design elements can be derived from specific QAF items.

Looking ahead, we expect that MP-SPC technologies will eventually make their way into *regular* statistical production. At that point the role of MP-SPC in the quality framework will need to be articulated explicitly in the future versions of the CoP and QAF. We anticipate that the concept of "*data governance*" may play a useful role in liaising between MP-SPC and (the future versions of) the quality framework: while this term never appears in the current version of CoP and QAF, as matter of fact several elements therein relate to data governance aspects within and across different organisations. And ultimately *the role of any MP-SPC solution is to enforce technologically data governance policies* defined at the organisational level.

In April 2024 Eurostat launched the new project **JOCONDE** (**J**oint **O**n-demand **Co**mputation with **N**o **D**ata **E**xchange) to advance towards the specification and demonstration of a shared MP-SPC system designed specifically for the ESS[2]. The project is conducted in collaboration with an Estonian company specialised in SPC solutions, selected based on an open call for tender. The project has a planned duration of 24 months and will terminate in March 2026.

References

Ricciato, F., (2024). Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics. *Journal of Official Statistics*, *40*(1), 3-15. https://doi.org/10.1177/0282423X241235259

European Statistics Code of Practice – revised edition 2017 (CoP). https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf

Quality Assurance Framework of the European Statistical System – version 2.0 (QAF). https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework

Stadler, T., and C. Troncoso (2022). Why the Search for a Privacy-Preserving Data Sharing Mechanism is Failing. *Nature Computational Science* 2: 208–10. https://doi.org/10.1038/s43588-022-00236-x.

Ricciato, F., A. Wirthmann, and M. Hahn (2020). Trusted Smart Statistics: How New Data Will Change Official Statistics. Data & Policy 2. https://doi.org/10.1017/dap.2020.7

[2] Project page: https://cros.ec.europa.eu/joconde