UNECE HLG-MOS project on Input Privacy-Preservation for Official Statistics

Open technical consultation on *Towards a trustworthy Multi-Party Secure Private Computing-as-a-service infrastructure for official statistics*

About this consultation

The UNECE HLG-MOS launched in 2021 a project on Input Privacy-Preservation for Official Statistics[1] with the aim of encouraging the participating institutions to familiarize with privacy-preserving techniques and explore their relevance for the production of future official statistics. As part of this work, upon initial idea by Eurostat, the project team formulated the concept of a **shared infrastructure based on Multi-Party Secure Private Computing technologies serving the needs of official statistics.** The motivation and general terms of the concepts are described briefly in the rest of this document.

The practical implementation of this concept requires addressing a number of challenges and open issues. In order to ensure that nothing is missed, the project team decided to launch an **informal and open technical consultation among experts and stakeholders.** The consultation is mainly targeted at:

- Privacy and security experts from both the technical and legal sides.
- Potential users of the envisioned MPSPC infrastructure, including but not limited to statistical authorities, public bodies and private companies.
- Digital activists and representative of civil society (e.g., citizen associations).
- Researchers and developers in relevant fields.

The consultation was open in mid-October 2022 and the deadline for responses is ~~set on 30 November 2022~~ extended to 15 January 2023. After this date the project team will summarize the main outcome of the consultation in a public report.

Motivations and context

The traditional model of statistical production assumes that a single organization, namely the statistical authority, collects the whole **input data** and from there computes the desired output information, i.e. the final *statistics*, according to some data analysis methodology. Whenever the desired output information requires the **integration/combination of different input data sets held by different organizations**, the traditional solution is to arrange for an exchange of input data, either directly between the concerned institutions or with a Trusted Third Party (TTP). In so doing, the receiving party commits to certain terms of use (e.g., to use the data to extract solely the agreed information for the agreed purpose, to delete the data immediately afterwards, to secure the data against intrusions, etc.). The transmitting party and any other involved stakeholder, if any, must *trust* the receiving entity that it will abide by the agreed terms of data use because they have no technical means to enforce and verify the actual respect of these terms. This approach requires a strong trust relationship between the transmitting and receiving entities. It also amplifies the risks, since it increases the number of copies of the data and the number of actors that have access to the data.

It is important to remark that **exchanging the input data is a means** towards the goal of computing the desired output, not a goal in itself. Furthermore, **data exchange is not the only means** available today: alternative solutions based on Privacy Enhancing Technologies (PET), and specifically technologies for **Multi-Party Secure Private Computing (MPSPC)**, allow today to compute the output statistics without necessarily disclosing the input data to any entities other than their respective data holders.

The appeal for MPSPC technologies in official statistics stems from the fact that several innovation trends in this domain point towards the need to combine/integrate data sets held by different organizations. For example, the prospective extension of official statistics towards "non-traditional" data sources relies on the possibility to (re)use new types of data generated for non-statistical purposes by other organizations, including public

[1] Webpage of the project: https://statswiki.unece.org/display/IPP/Input+Privacy-Preservation+project

administrations and private companies[2]. In another direction, improving the quality of statistics referring to intrinsically cross-border phenomena (e.g., migration, international trade) requires the integration of data from different countries. These trends concur to increase the appetite for integrating/combining data held by multiple actors. Responding to such increasing demand with the traditional paradigm of data exchange may not be the most effective option in all cases, since any a new copy of the data that is passed to another organization creates additional risks and calls for additional protection costs. This motivates the search for alternative models to execute inter-organization computation that do not require direct data exchange.

A shared MPSPC-as-a-service infrastructure *by* and *for* statistical authorities

Setting up a robust MPSPC solution requires investments, capacity and also specialized skills *on the side of potential adopters*. Not all statistical institutions may have the internal resources and/or the necessary knowledge to develop, deploy and maintain their own solutions, and anyway the costs might be disproportionate compared to the expected benefit. The cost factor may discourage adoption wholly or drive towards adoption of sub-optimal solutions with less-than-maximum levels of security and robustness. Furthermore, interoperability may not be guaranteed among solutions developed independently by different institutions.

The concerns about costs, robustness and interoperability led the project team[3] to elaborate the vision of a **shared MPSPC infrastructure**, developed and operated by a network (or consortium) of statistical institutions and then made available *on demand* to execute computation based on the MPSPC paradigm. As in many other areas of Information Technologies, the basic idea is to decouple the *development* (and maintenance) from the *utilization* of the prospective MPSPC infrastructure. This allows to pool together resources and expert knowledge during the development phase, increasing cost-effectiveness and ultimately enabling the achievement of very high levels of robustness and security guarantees, based on state-of-the-art technologies and design criteria.

The shared MPSPC infrastructure developed in this way could then be used *on demand* by statistical institutions and by their partners (e.g. external data providers). This model will be referred hereafter as **MPSPC-as-a-service** (MPSPCaaS for short) in order to highlight that what is provisioned to potential users is a (multiparty, secure, private) *computation service* rather than a *computation infrastructure*. The '*servitisation*' of MPSPC is instrumental in providing a cost-effective ready-to-use alternative to direct data exchange, thus accelerating the widespread adoption of the MPSPC paradigm in the field of official statistics, i.e., statistical authorities and their partners[4].

Example of MPSPCaaS operation

In an exemplary usage scenario, two organizations *Px* and *Py* have agreed to execute a particular operation *f(Dx,Dy)* on their respective input data sets *Dx* and *Dy* and let organization *Pz* learn the result *Dz= f(Dx,Dy).* In this simple example, *Px* and *Py* play the role of *input parties* while *Pz* is the *output party*. In practical cases, the same organization may play the role of input party and output party at the same time, i.e. *Pz* might coincide either with *Px* or with *Py* (but not with both, as otherwise the whole set of users would reduce to a single entity in control of all the input data and output result, with no necessity to consider MPSPC solutions). In the field of official statistics, the input data sets *Dx* and *Dy* often take the form of confidential micro-data and the domain of the function *f* lies in the union or intersection between the two input data sets *Dx* and *Dy*. Notably, as far as applications in official statistics are concerned, the function *f* is defined in advance, as part of the adopted statistical methodology, and does not constitute a business secret – an aspect that simplifies the operation compared to other business sectors where the function (model, algorithm) *f* is itself a confidential component. Also, we assume that the output party (typically a statistical authority) is entitled to receive the computation result *Dz*, regardless of whether or not it still contains privacy-sensitive or business-sensitive information. MPSPC allows performing such computation without requiring the input parties to share their data sets with any

[2] See e.g. the Final Report of the Expert Group on Facilitating the use of new data sources for official statistics, June 2022. Available from https://ec.europa.eu/eurostat/documents/7870049/14803739/KS-FT-22-004-EN-N.pdf

[3] Preliminary versions of this concept were presented by Eurostat at international conferences and workshops – all public presentations ae available from https://ec.europa.eu/eurostat/cros/content/privacy-enhancing-technologies-official-statistics-pet4os_en

[4] Differently from other business sectors where the service provider is typically a private business company, in the context of official statistics it is natural to expect that the MPSPC infrastructure will be built and operated by a network or consortium of mutually independent public institutions. The MPSPC infrastructure governance is intimately connected with the trust model underlying the infrastructure design, an aspect that is full in the scope of the present consultation.

other single entity, be it the other input party, the output party or any other individual third party. What we have described here is a particular *MPSPC task* with parameters *[Px,Py,Pz,Dx,Dy,f]* to be configured and executed by the MPSPCaaS infrastructur along with – and independently from – other parallel tasks.

In the envisioned scenario, the institutions playing the roles of input parties *Px*, *Py* and output party *Pz* represent the group of *users* for this particular *MPSPC task*. In the envisioned MPSPCaaS, they would rely on the MPSPC functionalities made available by the shared infrastructure in order to let the computation result *Dz=f(Dx,Dy)* flow towards the output party *Pz*, with no other information disclosed to any other party. In practice, the group of users would *connect* to the MPSPCaaS infrastructure and configure a new MPSPC task taking advantage of the functionalities offered by the infrastructure. In this way, the marginal cost of configuring a new MPSCP task would be much smaller than the cost of setting up an ad-hoc MPSPC infrastructure dedicated to this specific task.

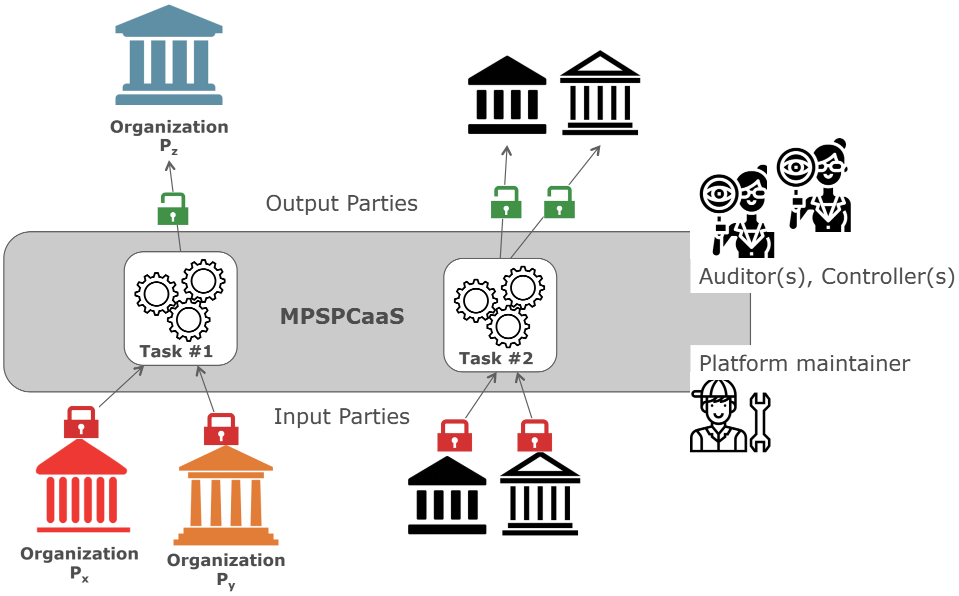

Figure 1 - Shared MPSPC-as-a-service infrastructure

Multi-party = no single point of trust

At an abstract level, the MPSPC infrastructure intermediating between the input and output parties may be seen as replacing a centralized Trusted Third Party (TTP), as shown in Figure 1. Indeed, if operation of the infrastructure would be such that a single entity would be *technically* able to control the whole computation process, the central controller would represent the *single point of trust* corresponding *de facto* to a TTP. In other words, a Secure Private Computing solution with centralized control would not be fundamentally different from the traditional model of data sharing with a TTP. In order to avoid that, at the heart of the MPSPC paradigm lays the requirement that no single entity should ever be technically capable to take over control of the process. Therefore, MPSPC operation must be designed so as to **avoid any single point of trust**. That means process control must be split (or divided) among a multiplicity of K>1 parties, which will be referred hereafter as *processing parties*[5]. In principle, K=2 processing parties would suffice to meet this formal requirement, but for increased robustness we will assume hereafter a minimum number of processing parties equal to K=3 or higher. Furthermore, in addition to the K processing parties with *active* control over the processing operation, additional entities may be foreseen to act as *passive controllers*, in order to increase the overall level of security and trust.

By definition, the K processing parties are in charge of *jointly* controlling the computation process, and therefore they are to be **trusted collectively, not individually**. The MPSPC infrastructure shall operate according to a set of policies centered around the principle that no computation task (thereby including simple queries) may be executed on the data without preliminary explicit approval of *all* K processing parties. The MPSPC infrastructure

[5] The abstract notion of *processing party* introduced here may possibly, but not necessarily correspond to the role of *computing parties* in secret sharing schemes. In fact, secret sharing is one among several possible schemes of choice for MPSPC operation. In multi-key encryption schemes, where the equivalent of a single decryption key is split among multiple key holders, the notion of processing party may correspond to key holders.

shall be engineered based on state-of-the-art technologies that are able to strictly enforce these policies. The robustness of the overall design shall therefore depend jointly (1) on the choice of the processing parties; (2) on the strength of the policies that define the operation of the processing parties; and (3) on the strength of the technologies that enforce these policies at hardware and/or software levels.

Conceptually, we may think of the MPSPC infrastructure as a **multi-party safe environment**, i.e., a locked safe where the key is split into K shares held by K different processing parties. In order to unlock a new computation task, all key-shares must be inserted into the lock[6], therefore all K processing parties have to agree to it. The implemented policies and technologies determine the strength of the safe, but the overall level of trust depends also on the choice of the K key-share holders, i.e., on their *collective* level of trustworthiness.

Moving from the traditional 'single key' paradigm to 'multiple key-shares' is the first innovation of MPC over TTP, as depicted in Figure 2. This allows replacing a single external TTP with multiple Processing Parties (PP). The next step is to let the entities serving as Input and/or Output Parties play the role of Processing Parties themselves, as depicted in Figure 2. When cast into the MPSPCaaS model, where the set of Input/Output Parties varies from one computation task to another, the distinction as to whether the entities in charge of acting as Processing Parties (PP) correspond or not to the Input/Output Parties leads to two different flavors of MPSPCaaS operations:

A) Fixed-PP model, where the set of Processing Parties (PP) is fixed and does not change from one MPSPC task to another;
B) Mixed-PP model, where the PP set varies from one MPSPC task to another in order to let some of the Input/Output Parties play the PP role for that specific task.

Both models are in principle applicable to official statistics use-cases, and each of them may be preferred in different context. It is clear that the Mixed-PP model is more demanding than the Fixed-PP model for the MPSPCaaS users that are willing to take on the PP role (e.g. in terms of computational and organizational resources).

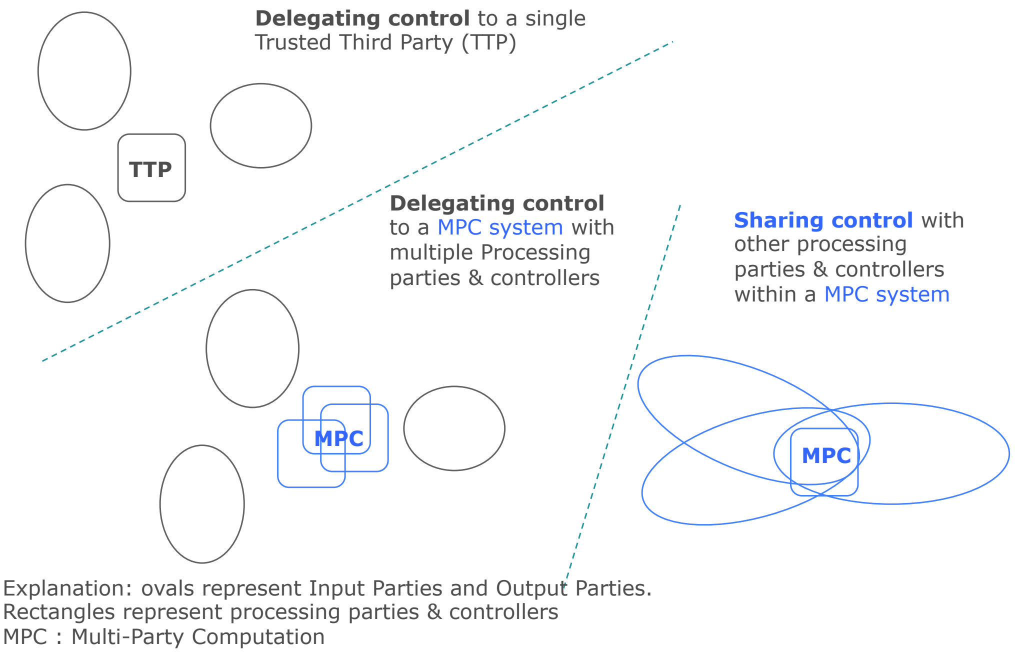
Delegating control to a single Trusted Third Party (TTP)
TTP
Delegating control to a MPC system with multiple Processing parties & controllers
Sharing control with other processing parties & controllers within a MPC system
MPC
MPC
Explanation: ovals represent Input Parties and Output Parties.
Rectangles represent processing parties & controllers
MPC : Multi-Party Computation

Figure 2 – Abstract representation of processing control distribution in the different paradigms. In the Trusted Third Party model a single entity is delegated full processing control (left). Multi-Party Secure Private Computation technologies enable multiple processing parties to share processing control. Therefore the processing parties must be trusted collectively, not individually. The input/output parties may delegate processing control to an external set of multiple processing parties (middle) or share processing control directly with the other parties (right).

[6] The term *Multi-Party* is used here in the general sense, with no intention at this stage to focus on any particular scheme. Different mechanisms and technologies may be adopted (and composed) to build a *multi-party safe environment* like the one outlined here, including but not limited to secret sharing, multi-key homomorphic encryption, trusted execution environment with multi-party authorization and possibly others.

List of questions

1. **General feedback**. What do you think about the envisioned concept of a shared MPSPCaaS infrastructure operated *by* and *for* statistical offices? What are the main points of strength and the main points of concern? Write down your thoughts and comments. [This Question is directed to all]

2. **Use-cases**. The initial design of (a first version of) the envisioned MPSPCaaS infrastructure would likely focus on supporting a **set of selected use-cases**. Could you provide examples of the kind of use-cases that you consider important and you would recommend to be considered as test cases? [This is mainly directed at potential adopters]

3. **Requirements and design criteria**. What should be in your opinion the main **technical requirements** and **design criteria** of the envisioned MPSPCaaS infrastructure in order to provide the strongest possible security and privacy guarantees?

4. **Technologies**. What kinds of technologies, or combinations thereof, you would consider as the most suitable building blocks for the envisioned MPSPCaaS infrastructure? [This is mainly directed at potential providers]

5. **Processing Parties and Controllers.** What criteria should drive the identification of Processing Parties and Controllers in order to maximize trustworthiness of the envisioned MPSPCaaS infrastructure? As for the Fixed-PP model, which organizations in your opinion are best qualified to serve as Processing Parties? And which one(s) as Controller(s)? How they can be incentivised to participate?

6. **Governance**. Beyond the technical privacy and security guarantees, are there additional governance processes required to ensure the safe and trustworthy operation of all parties involved in the MPSPCaaS operation?

7. **Testing. H**ow should the system be tested to for its performance, accuracy, robustness of its security and privacy and guarantees? How such guarantees should be verified?

8. **Public trust and acceptance.** Assuming that a robust MPSPCaaS infrastructure has been built and deployed, what additional actions should be taken in order to build public trust and acceptance into the proposed model?

9. **Free suggestions.** You are invited to provide below any specific suggestion or comment that does not fall in any of the previous items (free text).