**Consideration for Privacy-Preserving Technologies for Official Statistics**

**Keywords:** Privacy-preserving technologies, secure multiparty computation, differential privacy, homomorphic encryption, trusted execution environments

# **Introduction**

In recent years, the widespread interest of privacy-preserving technologies (PPTs) have soared and the many forms of PPT have found use in official statistics. However, considerations of both technical and regulatory nature remain a challenge for practitioners wanting to explore the applicability of PPTs in their area of work. While it is nearly impossible to form a security or privacy checklist to assure practitioners total peace of mind, this work instead endeavours to identify a range of important considerations and questions the practitioner should take into account as a baseline for further inspection. In doing so, we endeavour to untangle much of the common jargon used within the cryptography and security research domain and to increase the ability of practitioners to assess the fitness-of-purpose and potential risks of a deployment within the context of official statistics. We endeavour to do this through a technical overview, regulatory synopsis and a selection of relevant case studies.

The findings in this paper have been deducted from the work of the UN Privacy-Preserving Task Team [1] and use cases are taken from active research at national statistical institutes and international organizations. National statistical institutes typically get access to data needed for compiling and publishing official statistics because there is a legal requirement, for example due to national statistical law, or due to a trusted relationship with the data providers. However, there are some scenarios when this is not the case and traditional data acquisition approaches cannot be used to secure access to data, for example:

* Untrusted relationship, such as with private data holders who are not obliged to provide data to the Government or international organization,
* Third-party involvement in which the two external parties do not trust each other, for

example if data is provided from competing organisations or corresponds to commercially sensitive information,

* Legal barriers, such as in the case of particularly privacy-sensitive information, for example medical data.

Privacy preserving technologies offer solutions that satisfy these trust and compliance restrictions with minimal impact on statistical utility. They allow for macro-level statistical analysis while offering guarantees about the level of micro-level information pertaining to an individual or subset of sensitive records.

Further, privacy preserving techniques can also bring benefits in scenarios where a trusted relationship already exists, however, improved privacy protection could increase respondent participation. This is the case, for example, of data collection from smartphone sensor data during statistical surveys performed via respondent’s mobile devices (citation).

In this paper, we investigated real-world case studies of data sources that are particularly sensitive and can therefore significantly benefit from the use of PPT. In particular, examples of such sensitive data include; mobile phone data from cellular network operators, medical data from hospital and medical research studies, commercially sensitive trade data and sensor data from smart devices. Case studies with these selected data sources have provided the basis for selection of techniques, included in the scope of this paper, as well as focus on most relevant technical and legal aspects of these techniques. Importantly, this paper is not intended to have a complete overview of all privacy-preserving techniques and their application in official statistics but rather the more common use cases and techniques, critical for statistical organizations to be familiar with.

# **Methods**

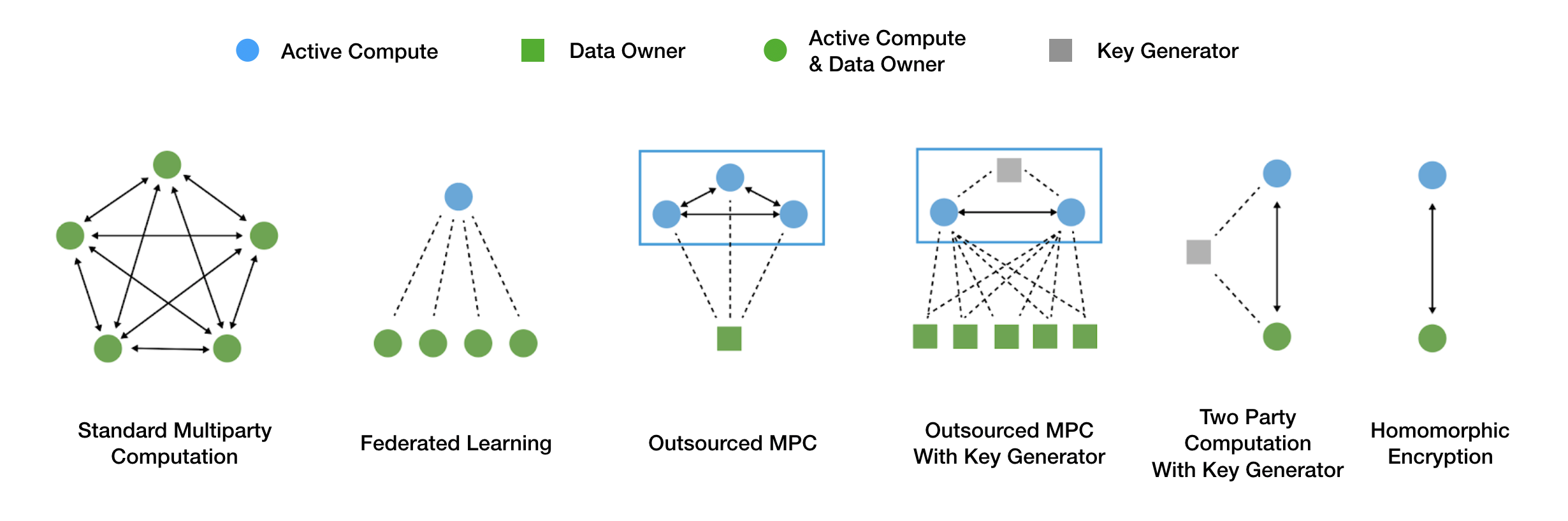
The term privacy-preserving technologies encapsulates a wide variety of complementary technologies which solve specific types of privacy concerns and forms of analysis. Specifically in this work we focus on those most relevant to official statistics. In the following subsections, we give a high level overview of these PPT, leaving their technical limitations, relevant legal and compliance considerations for the main paper.

## Output Privacy Techniques: Differential Privacy

Differential privacy is a mechanism used to ensure individual/record level anonymity of aggregate level statistics and inference. By definition, differentially private algorithms make it epsilon-hard to distinguish between any individual/record being a member of the database or not. Typically, this is achieved by introducing additive noise, usually Laplacian noise, to the output of the query. However, the introduction of this noise in particular is not what defines differential privacy, but rather the relationship between the query, the underlying data and the type/scale of the noise introduced. This relationship becomes complex when nonlinear functions are applied as part of the query and currently the range of queries that are well understood from a differential privacy perspective is limited. Extending the breadth of differentially private functionalities is still a domain of active research within the academic community.

## Input Privacy Techniques: Secure Multiparty Computation, Homomorphic Encryption & Federated Learning

Secure multiparty computation is a broad domain of cryptography which encompasses a variety of approaches and settings. In general, it is defined as cryptographic protocols that allow two or more parties to jointly compute a function from their independent inputs while keeping their inputs secure (secret). The figure below helps break down commonly studied and deployed multiparty computation schemes, differing based on who are the data owners, who are active in the computation and who acts in an offline capacity within the secure handshake.



Each variant of secure multiparty computation (SMPC) holds its respective trade-offs with respect to who inputs information, what combination of actors would need to collaborate in order to cheat, the ability for honest actors to identify when cheating takes place by others within the protocol, the range of functionalities that can be applied and the efficiency of applying them. In terms of efficiency, typically the total communication between parties, the number of rounds of communication between the parties and the latency of independent operations are taken into account. These differences and their respective impact on the deployment of their systems are discussed in detail within the full paper.

## Trusted Execution Environments

Trusted execution environments (TEEs) are hardware where code is executed and data is accessed in isolation and protected in terms of confidentiality and integrity. Confidentiality in this context means that the data is kept private from other processes, while integrity means that the code is unable to be modified by other processes. TEEs are incorporated into a variety of hardware such as smartphones, personal computers and games consoles. That said, TEEs remain a relatively new technology where there is a lack of standardization between implementations and usage depending on the hardware vendors. Popular uses of on-device TEEs are typical for authentication purposes, payment services and digital rights management. However, there are some application uses for the handling of government data which we will outline in the case studies.

## Other Supportive Mechanisms

While the former three technologies are the core PPTs which we will focus on in this work, there are a range of other technologies which support these in various ways throughout the case studies. Blockchain and smart contracts, for instance, were used to provide transparency and to enforce query restrictions in the context of one of the use cases. Each supportive technology is highlighted and its role within the context of the case studies elaborated. One important note is that PPTs exist in conjunction to classic computer security techniques such as encryption at rest, in transit, firewall and proxies services and so on.

# **Case Studies for Official Statistics**

In this paper, we consider a range of case studies that leverage PPTs at national and international level. Below is a short synopsis of a subset of the case studies considered in this work:

**Statistics on Healthcare Data:** Healthcare information is often perceived as one of the most sensitive forms of information. In this case study, we highlight the considerations taken by Statistics Netherlands (CBS) as they work with healthcare providers, technology companies and national research organisations. The technologies leveraged include multiparty computation and homomorphic encryption supported by blockchain and smart contracts. Differential privacy is also under consideration.

**Statistics on the Relationship Between Disease and Environmental Factors**: In this work, CBS is collaborating with a university and medical databank in order to identify relationships between disease, socio-economic and environmental factors. In this work trusted execution environments were leveraged in combination with pseudonymization, smart contracts and statistical output controls.

**The ESS Smart Data Platform**: This project involves the collaboration of 12 EU member states and EUROSTAT who are investigating the use of PPTs, through the use of smartphone and smart sensor data collection, to perform European social surveys such as the Household Budget Survey and Time Use Survey. Secure multiparty computation and homomorphic encryption are investigated in order to realise these goals.

**International Economic Data Sharing and Analysis**: Currently under active research is the UN and European Commission's endeavours to put in place systems for international collaborative sharing of economic information. This is being looked at from a range of perspectives, including the use of secure multiparty computation, homomorphic encryption and trusted execution environments.

**Official Statistics and Pandemic Response Using GSM Data**: Many national organizations have wrestled with the prospect of using GSM telecom information for official statistics. In some settings, such as in Indonesia and South Korea, national laws have been pasted that specifically allow for this data to be leveraged. However, for many European countries, national data protection commissions have prevented this sensitive information from being used. We examine how the use of PPTs may enable the potential use of this data while balancing the concerns of the data protection offices in a wider context internationally.

**References**

**[1]** UN Handbook on Privacy-Preserving Computation Techniques, United Nations Big Data Global Working Group for Privacy Preserving Data, 2019