Integrating Secure Multiparty Computation into Data Spaces

Veronika Siska[1][a], Thomas Lorünser[1][b], Stephan Krenn[1][c] and Christoph Fabianek[2][d]
[1]*AIT Austrian Institute of Technology, Center for Digital Safety & Security, Vienna, Austria*
[2]*Frequentis, Vienna, Austria*

Keywords: Secure Multiparty Computation, Data Spaces, Policy Definitions, Decentralized Trust.

Abstract: Integrating secure multiparty computation (MPC) into data spaces is a promising approach for enabling secure and trustworthy data-sharing in the future Data Economy. This paper systematically analyzes the integration challenges of MPC in data spaces and proposes a comprehensive approach to address these challenges. The authors evaluate various use cases to identify key challenges and gaps in existing research. They propose concrete methods and technologies to solve these challenges, focusing on areas such as authentication and identity management, policy description, node selection, global system parameters, and access control. The paper emphasizes the importance of standardization efforts to ensure interoperability among MPC-enabled data spaces. Overall, this work provides valuable insights and directions for further research in integrating MPC into dynamic data sharing environments.

1 INTRODUCTION

Data spaces are central to enabling sovereign, interoperable, and trustworthy data-sharing, crucial for the emerging data economy. Although certain techniques to support data sovereignty are inherent to data spaces, the use of modern cryptography beyond the state-of-the-art can propel the concept to the next level and unleash collaboration on sensitive data.

Multiparty computation (MPC) is an especially interesting technique for computing on encrypted data. MPC is a distributed protocol which naturally fits the federated architecture of data spaces and could therefore be an integrated part of it.

To the best of our knowledge no comprehensive analysis nor integration concept for MPC in data spaces exist, especially in support of modern collaborative use cases. In this paper we systematically analyze integration challenges for multiparty computation into data spaces. We evaluate a wide spectrum of use cases to identify a comprehensive set of challenges. Moreover, we propose a complete approach for the integration, propose concrete methods and technologies to solve the identified challenges, and identify gaps where further research is required.

This paper is structured as follows. Section 2 gives a short review of the concepts of data spaces and multiparty computation. In Section 3 we introduce three use cases and discuss them from a deployment perspective, extracting their key characteristics and challenges. In Section 4 we propose a first approach for an ubiquitous and comprehensive integration of MPC into data spaces. Based on that, potential technical solutions and research gaps for the identified challenges are discussed in Section 5. We conclude in Section 6.

1.1 Related Work

Related work that considers MPC in the context of data spaces is not extensive, since the latter is relatively young as a research field.

(Garrido et al., 2022) conduct a systematic review on the application of privacy-enhancing technologies (PETs) for internet-of-things (IoT) data markets, including MPC. They conclude that PETs are not frequently used in this setting, despite relevant use cases; and that there is no consensus on a general architecture, in particular regarding the usage of blockchain.

(Agahari et al., 2021) and (Agahari et al., 2022) offer a business perspective on MPC for data sharing, building on the business model for data marketplaces from (Spiekermann, 2019). They conduct semi-structured interviews in the privacy and security domain to study the perceived value propositions, ar-

[a] https://orcid.org/0000-0002-8057-1203
[b] https://orcid.org/0000-0002-1829-4882
[c] https://orcid.org/0000-0003-2835-9093
[d] https://orcid.org/0009-0002-4410-8796

Siska, V., Lorünser, T., Krenn, S. and Fabianek, C.
Integrating Secure Multiparty Computation into Data Spaces.
DOI: 10.5220/0012734600003711

In *Proceedings of the 14th International Conference on Cloud Computing and Services Science (CLOSER 2024)*, pages 346-357
ISBN: 978-989-758-701-6; ISSN: 2184-5042

chitecture and financial models (Agahari et al., 2021), as well as control, trust, and perceived risks (Agahari et al., 2022). They find that the value of MPC is seen in increased privacy, enhanced control and reduced need for trust, but that specific data sharing risks remain since the results may still reveal sensitive information. Different deployment scenarios are also described, such as the distributed, asynchronous setup that we present via data spaces in the current paper.

(Müller et al., 2022) focus on federated machine learning, with an application for the automotive industry via the project Catena-X[1]. They explore various cryptographic techniques, such as MPC, and identify usability challenges as the primary obstacle. They note that these technologies are lacking in user-friendliness and specialized libraries, and currently necessitate expert knowledge for specific use cases.

Besides the limited research on MPC integration into data spaces, some work on MPC on blockchain exists, with Secret Network[2] and Partisia[3] (described in Section 2.2.2) being the most prominent candidates. One important difference to data space integration is the lack of a registration procedure to establish trust relationships. Contrary to blockchain-based solutions, the MPC node pool in data spaces is open, but nodes and their attributes are certified e.g. via verifiable credentials (VCs). Thus MPC groups are also not necessarily random subsets, but can be chosen by attributes. Also, there is no need for complex broadcast protocols for arbitration, and contracts can be signed without involving a blockchain. Payment also does not necessarily need to flow through cryptocurrencies.

To the best of our knowledge, concrete integration of MPC into data spaces has not been discussed in the literature and we are the first to propose a general and comprehensive treatment. Data spaces require a fundamentally different approach to a pure blockchain based system, and can be more flexible, scalable and energy efficient compared to permissionless systems.

2 PRELIMINARIES

We next outline some fundamental concepts.

2.1 Data Spaces

A data space is "a distributed system defined by a governance framework that enables secure and trustworthy data transactions between participants while supporting trust and data sovereignty" (Data Spaces Support Centre (DSSC), 2023). The goal of data spaces is to share data and data-related services via a federated data marketplace (Zappa et al., 2022). This includes data-based services, such as storage, web servers, or algorithms operating on shared data. The latter is particularly relevant for privacy-preserving and/or distributed computing approaches that respect access and usage restrictions, such as MPC.

Data spaces were introduced in computer science as a shift from a central database to storing data at the source (Franklin et al., 2005). This new way of data management, where participants retain control over their own data, is now called data sovereignty (Otto et al., 2022). Data sovereignty is at the heart of the European data strategy and related regulations, in particular the General Data Protection Regulation (GDPR)[4], the Data Governance Act[5], and the Data Act[6]. The concept is also of international interest: by now, GDPR-like regulations exist in 17 countries and even more on the federal level (e.g. New York Privacy Act[7] and the California Consumer Privacy Act[8]); with some (e.g. South Korea's Personal Information Protection Act[9]) even pre-dating GDPR.

There are many initiatives supporting data space development. The International Data Spaces Association (IDSA) provided the initial concept, including the first reference architecture, the International Data Spaces Reference Architecture Model (IDS RAM). Gaia-X is taking the concept further and considers generic data products, also including services like storage or data analytics, to enable interoperability between different infrastructures. Gaia-X also develops a trust framework: a composition of policies, rules, standards and procedures based on standardized descriptions for participants and services. These are built using W3C Verifiable Credentials: cryptographically signed digital certificates that are thus tamper-proof and automatically verifiable.

The Data Spaces Business Alliance (DSBA), formed by BDVA, FIWARE Foundation, Gaia-X, and IDSA, aims to harmonize these efforts by providing a common technical framework (DOME) (Alliance, 2023). The Data Spaces Support Centre (DSSC) contributes with coordination efforts, including a glossary and building blocks, whereas simpl focuses on creating reusable data space software. Sector-specific

[1] https://catena-x.net/

[2] https://scrt.network/

[3] https://partisiablockchain.com/

[4] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[5] https://eur-lex.europa.eu/eli/reg/2022/868/oj

[6] https://eur-lex.europa.eu/eli/reg/2023/2854

[7] https://nyassembly.gov/leg/?bn=S00365

[8] https://oag.ca.gov/privacy/ccpa

[9] https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=213857&viewCls=engLsInfoR&urlMode=engLsInfoR

projects like Catena-X in the automotive industry or Manufacturing-X for manufacturing, exemplify the application of these frameworks. Promising open-source software components for data spaces are now also available, such as the Eclipse Dataspace Components (EDC), the Gaia-X cross-federation services or the Pontus-X ecosystem.

These collaborative efforts are laying the groundwork for a unified, efficient, and sovereign digital ecosystem, marking significant strides toward the realization of a comprehensive Data Economy.

2.2 Secure Multiparty Computation

Multiparty computation (MPC) is a technology for computing on encrypted data in a distributed setting, i.e., with multiple nodes holding only secure fragments of input data not learning anything from them. The concept appeared more than 30 years ago and has been the target of active research over the last 3 decades. For a long time, it was considered only theoretical, but progress in recent years led to many interesting applications which can be realized with practical efficiency, given a suitable deployment.

2.2.1 Basic Model

In principle, MPC can be used to decentralize systems where typically a central trusted authority is needed to execute a function on behalf of the users. With MPC, the function is evaluated jointly between multiple parties such that the correctness of the output is guaranteed and the privacy of the inputs of the individual parties is preserved; only the output of the computation is learned. Furthermore, information-theoretically secure MPC exist which makes it the ideal method if long-term security is needed.

We quickly present the generic model of MPC as introduced in ISO/IEC 4922[10][11]. Different roles are necessary in a generic MPC system in order to qualify as such. **Input parties** hold inputs for the secure computation which must be encoded and then sent to the compute parties. **Compute parties** run the multiparty protocol, which is executed among them as they jointly compute the intended function on the encoded inputs. The **intended function** to be computed is not kept secret and is defined according to the use case. The function is composed of basic operations available to the MPC protocol and typically composed of simple gates from a boolean or arithmetic circuit, depending on the encoding and protocols used. After the computation, the result is held by the compute parties in an encoded form and then sent to the **result parties**, which can reconstruct the result of the computation.

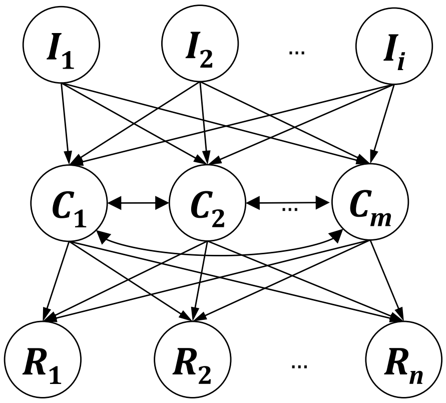

Figure 1: Generic MPC model: input nodes I_i encode data and send them to compute nodes C_i, which then execute the MPC protocol. After that, compute nodes hold the secret in encoded form, which is finally sent to result nodes R_i that recover the result in plaintext.

The main security properties are **correctness** and **input privacy**, and it is the latter which guarantees the confidentiality of the data. Depending on the protocol, the security parameters could hold against different kind of adversaries.

Certain additional, optional security guarantees are also possible, e.g., fairness, guaranteed output delivery or covert security. Fairness means, that malicious parties only receive their output if also the honest parties do so. With guaranteed output delivery the honest parties always receive their output. Contrary, in a covert security model, the protocol aborts in case of error and allows for cheater detection.

In summary, the overall concept is well understood and elaborated, i.e., many computations have been shown practical. However, the security assumptions are very different from traditional secret or public key cryptography. Here, security is mainly governed by the non-collusion assumption, which makes deployment of the technology challenging, especially in dynamic scenarios as we often find in emerging data markets and digital ecosystems with many stakeholders involved.

2.2.2 MPC as a Service

Due to the complexity and deployment challenges, potential users are often reluctant to use MPC. Thus, collaborative use cases are often prevented in data spaces if data privacy cannot be assured.

Leveraging the as-a-service paradigm could be a way out for this problem, but requires careful integration of the service to assure high security and prevent data leakage along the data life-cycle.

[10] https://www.iso.org/standard/80508.html

[11] https://www.iso.org/standard/80514.html

Moreover, additional integrity guarantees and data leakage prevention methods may be desirable depending on the sensitivity of the data and the use case. In particular, public verifiability could be of additional value for MPC-as-a-service (MPCaaS) and contribute to the trustworthiness of the service.

Publicly verifiable MPC can assure the correctness of computations even if all compute nodes are compromised, and although input privacy does not hold anymore. Typically, this is achieved by combining MPC protocols with compatible zero-knowledge proof (ZKP) systems to provide the best possible security guarantees for the outsourcing scenario of remote MPC, which is the case for the as-a-service usage. Yet, this is only to prevent from corrupt results in the worst case of a fully malicious MPC system, which can be prevented by careful selection of nodes.

The possibility for public audits of computation results have additionally benefits for data spaces, because it also allows for high assurance levels of computation results. If even third party stakeholders are able to verify the results of a computation, this could be used to establish end-to-end authenticity in data spaces. For example, (Kanjalkar et al., 2021) used this concept by combining MPC and zk-SNARKS (Chiesa et al., 2020) with universal setup to enable flexible verifiability for MPCaaS. The idea has also been shown to be useful in the manufacturing context (Lorünser et al.,).

Partisia is another example which uses blockchain to persist data and as a broadcast channel in combination with an event driven architecture[12]. Here, MPC node pools are built from available compute nodes, and each MPC service is randomly assigned to a subset of the nodes in the pool. Service buyers pay a pool to run a service, and the whole process is orchestrated via a smart contract, without the secret state appearing on the blockchain.

Although first proposals for MPCaaS exist, is is an open question how generic MPCaaS shall be integrated into data spaces to support a wide range of use cases, but without burdening complex configuration and deployment issues on the users of the system. In our work we systematically analyze this problem and propose relevant technologies to be used to realize the concept.

3 USE CASES

After introducing the basic technologies, we review three use cases where MPC could add value to data spaces. The use cases were selected to be highly complementary, in order to derive representative challenges and requirements.

3.1 Air Traffic Management

In air traffic management, the value attributed to individual flights can significantly vary. During peak periods, when demand exceeds available resources (e.g., due to bad weather or strikes), airlines have a vested interest in prioritizing flights that are of higher value to them. This need aligns with the economic interests of airports, which aim for optimal utilization of their infrastructure and a steady flow of passengers. Concurrently, air navigation service providers (ANSPs) are tasked with ensuring the safety of air travel, maintaining fairness and equality among all participants.

This scenario presents a multifaceted set of preferences and constraints, forming an optimization problem: determining the ideal sequence of flights for arrivals and departures. Each stakeholder – airlines, airports, and ANSP – has different needs, which include additional strict confidentiality requirements on which information to keep secret from other stakeholders. In a series of works, (Schuetz et al., 2021; Lorünser et al., 2022; Schuetz et al., 2022) proposed systems to optimize the use of airport capacities while taking all stakeholders' needs into consideration.

Their approach is built on MPC to satisfy the different confidentiality and integrity needs. In particular, verifiability of the computation is required, to minimize the risk of incorrect outputs resulting in a bias for or against a specific airline. More generally, fairness conditions are considered, to ensure that no specific airline is systematically privileged. Performance-wise, slot assignments are periodically computed for larger time intervals and the computation may take several minutes to succeed.

The considered approaches vary slightly: (Lorünser et al., 2022) output optimal solutions solving linear assignment problems, while (Schuetz et al., 2022) consider genetic algorithms that reach a near-optimal solution with high efficiency. Independent of the precise strategy, the necessary computations are agreed upon in advance by the various stakeholders and remain fixed over a high number of executions.

On the deployment side, air traffic management turns out to be a relatively static scenario, where a steady group of input providers (i.e., airlines) contributes their preferences, and all stakeholders (e.g., compute nodes, inputs providers, output consumers, etc.) are mutually known to each other.

[12] https://medium.com/partisia-blockchain/

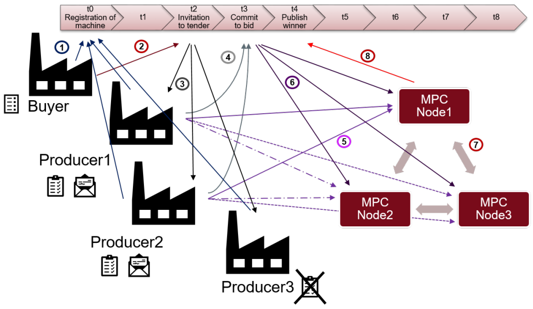

Figure 2: Manufacturing-as-a-Service Architecture.

3.2 Manufacturing as a Service

The sharing economy promises environmental benefits, innovation, and reduction of costs, but concerns persist over data sovereignty and trust. Also, centralization in large infrastructures raises economic alarms. Specifically for the manufacturing domain, (Lorünser et al.,) examine a platform where manufacturing site owners can enlist as producers, registering their machinery along with pertinent meta information such as configurations and quality standards. Customers can place orders, prompting producers to submit bids to secure the order. A high-level architecture and flow is depicted in Figure 2.

Their solutions heavily relies on secure multiparty computation to achieve the requirements posed by the different stakeholders. Specifically, producers need confidentiality to make sure that non-winning bids are not leaked, to avoid exposing internal cost structures or similar information to competitors. Both customers and producers are asking for integrity and verifiability, i.e., it needs to be ensured that the correctness of the computation can be publicly checked. This is achieved leveraging zero-knowledge proofs, providing integrity guarantees in the case that a majority of the MPC nodes acts maliciously during the computation. Finally, immutability of bids, to avoid adjustments depending on competing bids, is avoided using blockchain for securely storing encrypted bids and outcomes.

In the scenario of manufacturing as a service, the function to be computed is not entirely static, but may vary depending on the specific tender. For instance, while (Lorünser et al.,) consider first-price sealed-bid auctions, also alternative options like second-price (= Vickrey) auctions or multi-attribute auctions could be used. The precise model would be defined by the customer when publishing the tender.

From a deployment point of view, the compute nodes, selected by the auction platform provider but hosted by independent entities, can be assumed to be known a priori to all stakeholders in the default setting. However, in new versions, parties can also request to host nodes to be part of the MPC network, leading to dynamic configurations. Moreover, as anybody may act as a customer and/or producer, the users cannot be assumed to be static and known to each other, such that a permissioned setting requiring, e.g., a registration phase, need to be introduced in order to overcome challenges with rogue bids and offers.

3.3 Secondary Use of Data

Data is often generated for a specific purpose, e.g., for medical treatment or collecting GPS information for charging road usage. However, often this data would also be highly valuable in other contexts, e.g., medical studies in hospitals or road traffic planning for public authorities. This gives raise to the concept of data market places, which enable selling (computations on) data to customers.

Different approaches based on different cryptographic primitives have been proposed in the literature, e.g., using fully homomorphic encryption (Koutsos et al., 2022), or secure multiparty computation (Koch et al., 2020; Koch et al., 2022).

According to (Koch et al., 2022), confidentiality and privacy are paramount, ensuring that (computations on) data cannot be requested without consent. That is, data providers must have fine-grained control over data usage and sales, without relying on a single trusted entity. Furthermore, verifiability and authenticity are crucial: the marketplace operator should not be able to tamper with analysis outputs, and mechanisms are needed to prevent the sale of fake data to increase trustworthiness and value of data, without compromising privacy. Where possible, end-to-end guarantees on data integrity are desirable, spanning from data source (e.g., a sensor) to consumer.

In the context of data markets, it is also crucial to support high flexibility in the computation to be carried out. This is necessary to protect privacy and address the asynchronous nature of these ecosystems, where data providers may not be available at computation time. Therefore, data subjects must have the ability to define precise usage policies linked to their data, specifying constraints on computations, compute nodes, and the number of inputs involved. It is imperative that compliance with these policies is immutably documented for auditing purposes for each computation. Additionally, contractual agreements must be in place, e.g., to prevent the acquisition of previously independent compute nodes by the same entity before data deletion. Moreover, the tradeoffs between transparency and auditability on the one hand, and customer needs on the other hand, must be carefully considered. For instance, the mere interest of a customer in certain data may inadvertently disclose information about their business strategy.

In terms of deployment, data markets require a high level of flexibility. Data may be stored in various locations, and users may define different types of policies, such as geographical constraints on nodes. Consequently, in contrast to the previous use case domains, node selection becomes a complex task. It is also uncertain which nodes will require access to which shares during data creation and storage, necessitating the deployment of advanced encryption mechanisms and related key management procedures to support this dynamism. Furthermore, since data providers and consumers are typically unknown to each other, strong identity management mechanisms are essential. These mechanisms not only ensure that users' policies (e.g., "only medical research institutes may request computations on my data") are adhered to, but also mitigate the risks associated with rogue data. Finally, potential payments for data usage must be executed in a manner that preserves privacy.

3.4 Challenges

As illustrated by the application scenarios above, integrating MPC into complex federated scenarios such as data spaces comes with practical challenges that may directly influence system design. In the following we cluster the lessons learned from the considered use cases to obtain a set of challenge categories to be considered, which are also summarized in Table 1.

C1. Global System Parameters. In case that the protocols to be executed require global system parameters – such as a common reference string (CRS) – the security and trustworthiness of these parameters needs to be guaranteed. This may for instance apply when leveraging zkSNARKs to obtain public verifiability of the computation output.

C2. Authentication and Identity Management. Identity management is at the core of any security architecture: any confidentiality concerns are vacuous if the communication partner is not genuine. In the context of MPC, not only compute nodes that handle the data, but also data providers and receivers need to be authenticated. The former is required to increase trust in the input data and potentially achieve accountability, while the latter is needed to ensure that only eligible parties may request computations.

However, out-of-the-box authentication methods are not always applicable in certain scenarios, as the identity of data sources and data receivers may subject to data protection requirements. For example, it may be desired to determine only the eligibility to request a computation, but not the actual identity. Yet, in case of misuse, methods for accountability may be needed.

The situation is further complicated when the data is managed on behalf of the owner by a third party (data custodian); when the owner is not able or willing to manage their own data. In this case, authentication would also be handled by the data custodian, with the owner first granting the right to do so.

To support large scale adoption, compatibility with governmental identities such as the upcoming European eIDAS 2.0 regulation is also necessary.

C3. Data Usage Policies. Precise data usage policies play a critical role in increasing trust and achieving acceptance by end users, particularly when personal or confidential data is involved.

Such policies describe the permissible ways in which data can be utilized, encompassing aspects such as eligible groups of receivers, temporal restrictions, requirements on the MPC setup (e.g., threshold or geographical distribution of nodes), the computation to be carried out (e.g., certain statistics including the required sample size or validation mechanisms), or data retention.

However, formulating and enforcing effective data usage policies presents several challenges. These include striking a balance between maximizing data utility for innovation and safeguarding privacy rights, achieving high usability also for end users, addressing evolving technological advancements and data-sharing practices, and ensuring transparency and accountability. Additionally, changing legal and market situations need to be addressable, potentially without re-involving data subjects in asynchronous scenarios.

C4. Node Selection. The security of any MPC deployment crucially depends on the involved compute nodes, as well as the selected parameters (i.e., threshold and number of nodes).

In certain (mainly static) scenarios, the selection of these nodes can be done once and (almost) forever. However, the situation is very different in highly dynamic scenarios where data from many data sources is used as input, as each of them pose certain constraints on node selection. Furthermore, compute nodes may be offered on an "as-a-service" basis by market players, such that their availability may have temporal variety. Therefore, any mechanism for node selection needs to take these requirements into consideration.

In combination with the identity management challenges mentioned before, it further needs to be

Table 1: Comparison of challenges affected by different use cases.

	UC1: Air traffic	**UC2: Industry 4.0**	**UC3: Secondary Use**
C1. Global system parameters	CRS for end-to-end verifiability and integrity		
C2. Authentication and identity management	static, permissioned	semi-static, permissioned	dynamic, permissionless
C3. Data usage policies	static, fully defined from beginning		dynamic, meta-level specifications
C4. Node selection	static	dynamic	
C5. Access control	online input provisioning; early encoding	synchronous input; early encoding; audit info	asynchronous input; late encoding

guaranteed that the involved nodes are not (potentially indirectly) controlled by a single legal entity.

This immediately also poses the question who decides, which nodes to involve. If this process relies on a central entity, appropriate measures to minimise the required trust should be taken, e.g., by aiming for transparency of performed computations, or by having compute nodes verify usage policies without compromising privacy. On the other hand, if this process is performed in a federated way, a circular argument (who chooses the participants of this set of entities) should be avoided.

C5. Access Control. In static situations characterized by fixed computations and entities, it is often predetermined which inputs and outputs must be accessible to each party. In this case, data providers may, e.g., encrypt input shares directly for designated compute nodes, which in turn encrypt the output for the specified data recipient.

Yet, in dynamic environments, this predictability may not always hold true. Thus, if it is unknown upfront which (or how many) MPC nodes will execute a given computation – and nodes might engage in computations on the same data across different sessions – appropriate technologies must be implemented to ensure that the shares for these nodes can be derived as needed without compromising privacy.

A fundamental challenge lies in avoiding dependence on a single trusted entity or a single point of failure, necessitating careful design of key management procedures. Moreover, it is essential to guarantee that nodes cannot receive multiple consistent shares when the same input data is utilized in multiple computations involving the same node.

4 MPC INTEGRATION FOR DATA SPACES

We propose using data spaces as a basis to deploy secure multiparty computing in a dynamic scenario; that is, where some or all elements (stakeholders, input data, algorithm) are not known in advance. Our goal is to create an ecosystem where participants can offer MPC-related assets under well-defined conditions ("policies"): input datasets, compute nodes or algorithms (intended function to be computed). Other participants may consume these offers by running a computation on a chosen set of input datasets and compute nodes, while respecting the conditions set by the providers of these assets. We divide the deployment of such a system in three phases: onboarding (participants), (asset) setup, and the transaction phase, where a single computation is executed. The overall architecture is shown in Fig. 3.

4.1 Onboarding and Setup Phase

First, participants need to be onboarded to the system (data space), which includes checking their identity and issuing some form of a proof of membership. At this phase, the identity of participants may be checked, possibly connecting to external trust anchors (TAs), see also challenge C2.

Second, onboarded participants may publish assets in the data space. For MPC, these include input data, compute nodes or even intended functions, each described by asset-specific metadata and associated with an individual policy that describes how they can be used, cf. C3. Note that in a fully dynamic setting, both steps of the setup are also dynamic: participants and offers may be added, modified or removed during the lifetime of the data space.

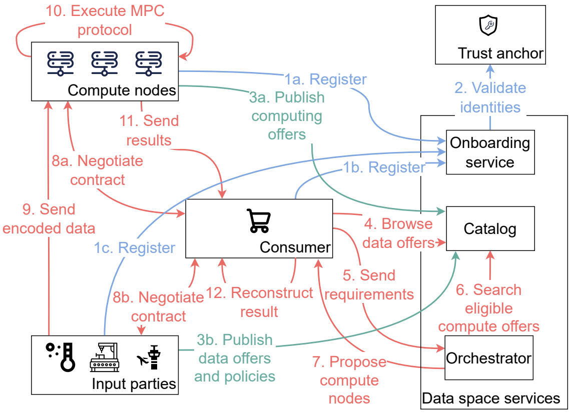

Figure 3: Components of the proposed data space-based deployment. **Blue**: onboarding phase; **green**: setup phase; **red**: transaction phase.

4.2 Transaction Phase

In the last phase, the actual transaction may occur.

Offer Selection and Contract Negotiation. First, participants (potential consumers) may browse available offers and select a combination of input data, compute nodes and a function they would like to evaluate. When selecting compute nodes, the consumer may pick offers explicitly or define conditions that nodes need to satisfy (e.g. not all nodes are hosted on the same server, all nodes are hosted in Europe). From a usability point of view, it could also be desirable to offer some preconfigured choices relevant in different domains (Framner et al., 2019) and from a performance standpoint also network or performance requirements could be included in the node selection, e.g., latency <30ms between nodes. However, if the MPC nodes are not concretely defined, the orchestrator service may pick a random selection of compute nodes on offer, to satisfy such criteria.

This request is sent to owner of the respective offers as a contract request, after which an automatic contract negotiation process takes place to validate that all requirements with regards to the policies are met. If this is the case, a contract between all parties is signed and the computation can be triggered. Validation of conditions may happen via a service ("MPC orchestrator") offered by the data space authority, which can be the same service orchestrating MPC computation, cf. also C4.

Input Provisioning. After all parties agreed to the transaction, the actual computation is started. Therefore, the input data has to be read by the compute parties in encoded form. Depending on the configuration, this step can be done either synchronously by the input parties sending the inputs to the compute nodes, but also asynchronously, if the data have been stored at a data custodian. In this case, for security reasons and following the zero trust principle, the data should only be stored in encrypted form. However, this is not trivial, if the receiving compute nodes are not known in advance, cf. also C5.

Furthermore, to be more flexible, it is also desirable to delay the time of encoding if possible. Thus, we distinguish immediate and late encoding.

Immediate encoding is the naive way to generate input data by encoding the data prior to encrypting it for storage at the data custodian. Then each compute node only has to decrypt his received data fragment during input processing. This is easier from a technological point of view, but less flexible and produces more overhead: as each share is encrypted individually, the total amount of data to be stored is large. Additionally, MPC system parameters and encoding scheme have to be defined in advance.

In *late encoding*, the plaintext is directly encrypted and stored at the data custodian. This significantly reduces the storage overhead and increases flexibility, as MPC parameters are decided during the transaction phase and not the setup phase. However, it is also technically more challenging, because some form of flexible threshold decryption is needed. A compromise would be to symmetrically encrypt the input data and then only encrypt the key with a threshold method. This would also save storage space, but require the compute nodes to first decrypt the data obliviously (Lorünser and Wohner, 2020).

Protocol Execution. During computation the agreed MPC protocol is executed among the agreed nodes to compute the intended function on the data. Although the step is rather straightforward, from a data space perspective it is important that the protocols available are standardized. Policies can only be practically enforced, if wide interoperability among MPC nodes available in the ecosystem is guaranteed and enough stakeholders publish offers. Additionally to executing the MPC protocols, plugins may also be of use. If verifiability is a requirement, an additional zero-knowledge proof has to be generated by the system, posing additional challenges for policy definition, the capabilities of the MPC nodes, and the trustworthiness of required parameters, cf. also C1, C3, and C4.

Additionally, calculating leakage of MPC computations which are intrinsic to the compute function by methods from differential privacy could also require for a plugin.

Post-Processing. Finally, after the computation the results are held by the compute nodes in encrypted form and has to be communicated (synchronously or asynchronously) with the result party. Post-computation validation, logging, and payment could then take place to finalize the transaction.

In summary, by our comprehensive integration proposal of MPC to data spaces, we have shown the complexity we are facing when we go beyond the naive approach where dedicated parties with profound technology knowledge run a specific instance of a protocol. However, this extra effort is necessary to make the system interoperable, being compatible with data spaces, and to leverage the MPC-as-a-service approach to lower the barriers for adoption.

5 TECHNICAL SOLUTIONS

In this section technical methods to solve the identified challenges are discussed. We identify gaps in the state-of-the-art, present potential avenues to address the challenges and highlight where additional research is needed.

Global Parameters. To minimize the necessary trust, global parameters should be setup in a way that does not give any sufficiently small set of entities the possibility to control the choice of parameters. Different approaches for this can be found in the literature.

One option are so-called *setup ceremonies*, where a group of entities jointly generates parameters that are later needed for cryptographic protocols, thus ensuring the trustworthiness of the outcome. Such ceremonies have been implemented for a variety of applications, including, e.g., the ZCash crypto currency[13].

Another active research field in cryptography is focusing on so-called subversion resilience, where at least partial security guarantees can also be achieved if, e.g., a common reference string (CRS) cannot be trusted, e.g., (Abdolmaleki et al., 2021).

Other works, e.g., (Baghery and Sedaghat, 2021), consider the updatable CRS model, where users can update the CRS at any time, provided they demonstrate the correctness of the update. The new CRS can then be deemed trustworthy (i.e., uncorrupted) as long as either the previous CRS or the updater was honest. If multiple users partake in this process, it's possible to obtain a sequence of updates by different individuals over time. If any update in the sequence is honest, the scheme remains sound.

Authentication and Identity Management. Authentication and identity management can differ between data spaces and may rely on traditional centralized (e.g. via a user database based on LDAP or Active Directory) or decentralized (e.g. using Decentralised Identifiers and Verifiable Credentials (VCs)) approaches. In any case, an onboarding process needs to be defined as part of data space governance, where the identity of participants is validated before granting them membership. The validation step normally relies on external trust anchors (e.g., eIDAS, DV SSL, GLEIF), with accepted trust anchors defined by the given data space's governance framework. As part of the onboarding process, participants may also record their public key and prove their control over it, providing a basis for a secure communication channel.

For instance, for Gaia-X (Gaia-X European Association for Data and Cloud AISBL, 2023), aspiring participants would submit their data as defined in the Trust Framework (e.g., ID, public key, address) to one of the Gaia-X Digital Clearing Houses (GXDCH) and receive a VC that they can use as proof. Internally, the GXDCH applies multiple validation checks, such as compatibility with the required metadata schema and validation via accepted trust anchors.

When a data custodian (ensuring data accessibility and security for a data owner) is also part of the system, the data owner first needs to authorize the custodian to act on their behalf. This can happen outside the data space context, via a separate contract between these parties, or is part of a data space service offering. The custodian then participates in the data space on behalf of the data owner. A more formalized and regulated instance of a data custodian is a Data Intermediary as defined in the Data Governance Act[14] - while a data custodian focus on the technical and security aspects of data management the data intermediary facilitates data sharing and usage in compliance with legal and regulatory frameworks.

While strong authentication may be required in many application cases, some scenarios require a delicate balance between privacy and authenticity, e.g., when an entity needs to fulfill a data usage policy but does not want to reveal its identity. This can be achieved, e.g., using attribute-based credentials (Camenisch and Lysyanskaya, 2002; Camenisch et al., 2015; Tessaro and Zhu, 2023) letting parties prove statements about their attributes without revealing them in the plain. In particular, this also covers selective disclosure as considered by W3C[15] or EBSI[16].

[13] https://zkproof.org/2021/06/30/setup-ceremonies/

[14] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R0868

[15] https://w3c-ccg.github.io/data-minimization/

[16] https://ec.europa.eu/digital-building-blocks/sites/disp

Furthermore, somewhat similar to direct anonymous attestation (DAA) (Brickell et al., 2004) or Intel's Enhanced Privacy ID (EPID)[17], in order to increase reliability in data without compromising security, concepts like privacy-enhancing group signatures (Krenn et al., 2019; Diaz and Lehmann, 2021) could be used. These let data sources such as sensors sign data to prove that it was generated using a genuine device, while keeping the precise identity of the device confidential. MPC over authenticated inputs is also considered by (Dutta et al., 2022).

Data Usage Policies. A significant challenge for MPC in data spaces is the enforcement of data policies. While the Open Digital Rights Language (ODRL)[18] offers a flexible mechanism for defining permissions, prohibitions, and duties concerning digital content and services, its effectiveness is limited in the context of MPC where data processing involves complex computations across multiple data owners. The enforceability of these policies becomes even more complicated when considering the simpler, yet enforceable, nature of Rego[19] within the Open Policy Agent (OPA) framework, which may not fully cater to the legal nuances required in MPC scenarios.

Moreover, the integration of MPC-as-a-service within data spaces necessitates a high degree of interoperability between different policy standards and legislative frameworks. The diverse landscape of standards like the Data Privacy Vocabulary (DPV)[20] for expressing policies related to personal data processing, and international standards such as ISO/IEC 29184 and ISO/IEC 27560 for online privacy and data sharing, must be seamlessly aligned to support the complex operations of MPC.

Compliance poses another challenge, especially with the introduction of legislative frameworks such as the Data Governance Act and the Data Act. These acts introduce new concepts like data intermediaries and data altruism, which, while enriching the data ecosystem, also add layers of complexity in ensuring that MPC services adhere to these regulations. Additionally, the empowerment of individuals through platforms like SOLID[21], granting them control over their data, intersects with the operational dynamics of MPC, requiring robust mechanisms to ensure that user consent and data usage terms are respected in a multi-stakeholder environment.

Incorporating also the Data Catalog Vocabulary (DCAT)[22] into the ecosystem of data spaces, to facilitate the discovery and interoperability of datasets, makes integrating usage polices even more challenging but also leads to a convergence of standards and practices for the participating stakeholders. By establishing a common framework, DCAT can serve as a tool in bridging the gap between different data policy standards. This convergence simplifies the process of managing and enforcing data usage policies across multiple platforms and jurisdictions, promoting a more unified and efficient approach to data sharing and processing.

Node Selection. In contrast to the permissionless systems prevalent in the blockchain world (e.g., Enigma, Partisia), data space services require registration, meaning they operate within a permissioned environment, thereby providing significant benefits with regards to node selection.

Nodes or node operators must be registered and each node will be assigned with attributes describing its abilities. Besides standards capabilities, like supported protocols, connection parameters like bandwidth, compute capabilities and other functional parameters, nodes must also be assigned with trust parameters. Every node must be assigned to an identity, geo location, and trust zones, to enable automatic matching of compute task policies and nodes.

The following sample settings illustrate policies that shall be supported in an MPC-ready data space:

- Nodes must be from 3 different entities in three different countries
- All nodes must be from the same country but from three different institutions or trust boundaries
- Nodes must have latency <10ms but be from different trust zones

It is also of interest to combine basic attribute-based matching with random assignment capabilities for additional robustness. Given the policy settings above, it should be possible to randomly assign nodes from all available combinations for different functions or even sub-functions, thereby also preventing sybil attacks.

Access Control. An integral aspect of data usage policies is the delineation of authorized users' access to specific datasets. While contractual enforcement suffices in numerous practical scenarios, there's

lay/EBSI/Selective+Disclosure%3A+An+EBSI+Improvement+Proposal

[17] https://www.intel.com/content/www/us/en/developer/articles/technical/intel-enhanced-privacy-id-epid-security-technology.html

[18] https://www.w3.org/TR/odrl-model/

[19] https://www.openpolicyagent.org/docs/latest/

[20] https://w3c.github.io/dpv/dpv/

[21] https://solidproject.org/

[22] https://www.w3.org/TR/vocab-dcat-3/

a growing preference for technical solutions. This approach, e.g., obviates the need for a data custodian to possess plaintext access to users' sensitive data.

In the following we sketch two options that realise this goal by leveraging advanced cryptographic methods beyond what was already discussed before.

One option following the late encoding approach could be to let data owners encrypt their data under their own public key using a so-called proxy re-encryption scheme (Blaze et al., 1998; Zhou et al., 2023). This allows the data custodian to transform ciphertexts under the user's public key into ciphertexts under a compute node's public key, provided that the user previously handed a so-called re-encryption key to the data custodian. In case that the encryption scheme supports a homomorphic operation on ciphertexts consistent with the secret sharing scheme, the data custodian could now derive the shares for the selected compute nodes ad-hoc, without ever requiring to access the plaintext. One drawback of this approach is, however, that the user has to derive individual re-encryption keys for all possible compute nodes, which may exclude nodes joining the ecosystem after the user making their data offer.

An alternative option based on early encoding leverages attribute-based encryption (ABE) (Sahai and Waters, 2005; Hohenberger et al., 2023). In an ABE scheme, each participant receives a secret key linked to some attributes (e.g., geographical location), while ciphertexts are linked to policies. A secret key can now only decrypt a ciphertext if the attributes of the secret key satisfy the policy of the ciphertext. For instance, users could encrypt their shares according to their requirements (e.g., each share with a specific country); while each compute node would receive a secret key linked to the country of its location. Assuming proper identity management, doing so could cryptographically enforce that only compute nodes located in specific countries could decrypt certain shares, thereby enforcing that nodes from different legislations participate in a computation. The main limitation of this approach is, that the master secret key, from which the individual secret keys are derived, needs to be administered securely and trustworthy within the MPCaaS ecosystem, e.g., by distributing it among several nodes which engage in an MPC protocol to derive novel keys for joining nodes. Furthermore, the encoding scheme required for the computations need to be known in advance.

6 CONCLUSION

This paper presents a comprehensive approach for integrating secure multiparty computation (MPC) into data spaces,to enable secure and trustworthy data-sharing in the future Data Economy. The authors address various challenges and their potential solutions, namely global parameters, authentication and identity management, data usage policies, node selection and access control. By adopting these solutions, organizations can enhance privacy and security while facilitating data sharing in dynamic environments.

However, the paper also highlights several research gaps that need to be addressed. Firstly, there is a need for more efficient and scalable MPC protocols that can handle large-scale datasets effectively. Additionally, techniques for dynamic and flexible access control in distributed environments are required. Privacy concerns arising from potential information leakage during protocol execution must also be addressed. Finally, standardized and interoperable frameworks are needed to support MPC-enabled data spaces across different domains and applications.

Addressing these research gaps will be crucial for fully realizing the potential of integrating MPC into data spaces and creating a secure and trustworthy Data Economy. Further research and development efforts are needed to overcome these challenges and ensure the successful adoption of this approach in practice.

ACKNOWLEDGMENTS

This work was in part funded by the European Union under the HORIZON SESAR JU Grant Agreement No. 101114675 (HARMONIC), and the Austrian Research Promotion Agency FFG within the PRESENT project. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funding agencies. Neither the European Union, FFG, nor the granting authority can be held responsible for them.

REFERENCES

Abdolmaleki, B., Lipmaa, H., Siim, J., and Zajac, M. (2021). On subversion-resistant snarks. *J. Cryptol.*, 34(3):17.

Agahari, W., Dolci, R., and de Reuver, M. (2021). Business model implications of privacy-preserving technologies in data marketplaces: The case of multi-party computation.

Agahari, W., Ofe, H., and de Reuver, M. (2022). It is not (only) about privacy: How multi-party computation redefines control, trust, and risk in data sharing. *Electronic Markets*, 32(3):1577–1602.

Alliance, D. S. B. (2023). Technical convergence. Technical report, Data Space Business Alliance.

Baghery, K. and Sedaghat, M. (2021). Tiramisu: Black-box simulation extractable nizks in the updatable CRS model. In *CANS*, volume 13099 of *LNCS*, pages 531–551. Springer.

Blaze, M., Bleumer, G., and Strauss, M. (1998). Divertible protocols and atomic proxy cryptography. In *EUROCRYPT*, volume 1403 of *LNCS*, pages 127–144. Springer.

Brickell, E. F., Camenisch, J., and Chen, L. (2004). Direct anonymous attestation. In *ACM CCS*, pages 132–145. ACM.

Camenisch, J., Krenn, S., Lehmann, A., Mikkelsen, G. L., Neven, G., and Pedersen, M. Ø. (2015). Formal treatment of privacy-enhancing credential systems. In *SAC*, volume 9566 of *LNCS*, pages 3–24. Springer.

Camenisch, J. and Lysyanskaya, A. (2002). A signature scheme with efficient protocols. In *SCN*, volume 2576 of *LNCS*, pages 268–289. Springer.

Chiesa, A., Hu, Y., Maller, M., Mishra, P., Vesely, P., and Ward, N. P. (2020). Marlin: Preprocessing zksnarks with universal and updatable SRS. In *EUROCRYPT, Part I*, volume 12105 of *Lecture Notes in Computer Science*, pages 738–768. Springer.

Data Spaces Support Centre (DSSC) (2023). DSSC Glossary Version 2.0.

Diaz, J. and Lehmann, A. (2021). Group signatures with user-controlled and sequential linkability. In *PKC*, volume 12710 of *LNCS*, pages 360–388. Springer.

Dutta, M., Ganesh, C., Patranabis, S., and Singh, N. (2022). Compute, but verify: Efficient multiparty computation over authenticated inputs. Cryptology ePrint Archive, Paper 2022/1648.

Framner, E., Fischer-Huebner, S., Loruenser, T., Alaqra, A. S., and Pettersson, J. S. (2019). Making secret sharing based cloud storage usable. *Information & Computer Security*, 27(5):647–667.

Franklin, M., Halevy, A., and Maier, D. (2005). From databases to dataspaces: a new abstraction for information management. *ACM SIGMOD Record*, 34(4):27–33.

Gaia-X European Association for Data and Cloud AISBL (2023). Gaia-X Framework.

Garrido, G. M., Sedlmeir, J., Uludağ, O., Alaoui, I. S., Luckow, A., and Matthes, F. (2022). Revealing the landscape of privacy-enhancing technologies in the context of data markets for the IoT: A systematic literature review. *Journal of Network and Computer Applications*, 207:103465.

Hohenberger, S., Lu, G., Waters, B., and Wu, D. J. (2023). Registered attribute-based encryption. In *EUROCRYPT, Part III*, volume 14006 of *LNCS*, pages 511–542. Springer.

Kanjalkar, S., Zhang, Y., Gandlur, S., and Miller, A. (2021). Publicly auditable MPC-as-a-service with succinct verification and universal setup. In *IEEE EuroS&PW*, pages 386–411.

Koch, K., Krenn, S., Marc, T., More, S., and Ramacher, S. (2022). KRAKEN: a privacy-preserving data market for authentic data. In *Data Economy*, pages 15–20. ACM.

Koch, K., Krenn, S., Pellegrino, D., and Ramacher, S. (2020). Privacy-preserving analytics for data markets using MPC. In *Privacy and Identity Management*, volume 619 of *IFIP AICT*, pages 226–246. Springer.

Koutsos, V., Papadopoulos, D., Chatzopoulos, D., Tarkoma, S., and Hui, P. (2022). Agora: A privacy-aware data marketplace. *IEEE TDSC*, 19(6):3728–3740.

Krenn, S., Samelin, K., and Striecks, C. (2019). Practical group-signatures with privacy-friendly openings. In *ARES*, pages 10:1–10:10. ACM.

Lorünser, T., Wohner, F., and Krenn, S. (2022). A verifiable multiparty computation solver for the linear assignment problem: And applications to air traffic management. In *CCSW*, pages 41–51. ACM.

Lorünser, T. and Wohner, F. (2020). Performance Comparison of Two Generic MPC-frameworks with Symmetric Ciphers:. In *ICETE 2020*, pages 587–594, France.

Lorünser, T., Wohner, F., and Krenn, S. A privacy-preserving auction platform with public verifiability for smart manufacturing. In *ICISSP*, pages 637–647. SciTePress. Backup Publisher: INSTICC.

Müller, T., Gärtner, N., Verzano, N., and Matthes, F. (2022). Barriers to the Practical Adoption of Federated Machine Learning in Cross-company Collaborations. In *ICAART (3)*, pages 581–588.

Otto, B., ten Hompel, M., and Wrobel, S. (2022). *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer Nature.

Sahai, A. and Waters, B. (2005). Fuzzy identity-based encryption. In *EUROCRYPT*, volume 3494 of *LNCS*, pages 457–473. Springer.

Schuetz, C. G., Gringinger, E., Pilon, N., and Lorünser, T. (2021). A privacy-preserving marketplace for air traffic flow management slot configuration. In *IEEE/AIAA DASC*, pages 1–9.

Schuetz, C. G., Lorünser, T., Jaburek, S., Schuetz, K., Wohner, F., Karl, R., and Gringinger, E. (2022). A distributed architecture for privacy-preserving optimization using genetic algorithms and multi-party computation. In *CoopIS*, volume 13591 of *LNCS*, pages 168–185. Springer.

Spiekermann, M. (2019). Data marketplaces: Trends and monetisation of data goods. 54(4):208–216.

Tessaro, S. and Zhu, C. (2023). Revisiting BBS signatures. In *EUROCRYPT, Part V*, volume 14008 of *LNCS*, pages 691–721. Springer.

Zappa, A., Le, C.-H., Serrano, M., and Curry, E. (2022). Connecting data spaces and data marketplaces and the progress toward the european single digital market with open-source software. In *Data Spaces : Design, Deployment and Future Directions*, pages 131–146. Springer International Publishing.

Zhou, Y., Liu, S., Han, S., and Zhang, H. (2023). Fine-grained proxy re-encryption: Definitions and constructions from LWE. In *ASIACRYPT, Part VI*, volume 14443 of *LNCS*, pages 199–231. Springer.