GDPR-Compliant Personal Data Management: A Blockchain-based Solution

Nguyen Binh Truong, *Member, IEEE,* Kai Sun, *Senior Member, IEEE,* Gyu Myoung Lee, *Senior Member, IEEE,* and Yike Guo, *Fellow, IEEE*

***Abstract*— The General Data Protection Regulation (GDPR) gives control of personal data back to the owners by appointing higher requirements and obligations on service providers (SPs) who manage and process personal data. As the verification of GDPR-compliance, handled by a supervisory authority, is irregularly conducted; it is challenging to be certify that an SP has been continuously adhering to the GDPR. Furthermore, it is beyond the data owner's capability to perceive whether an SP complies with the GDPR and effectively protects her personal data. This motivates us to envision a design concept for developing a GDPR-compliant personal data management platform leveraging the emerging blockchain (BC) and smart contract technologies. The goals of the platform are to provide decentralised mechanisms to both SPs and data owners for processing personal data; meanwhile empower data provenance and transparency by leveraging advanced features of the BC. The platform enables data owners to impose data usage consent, ensures only designated parties can process personal data, and logs all data activities in an immutable distributed ledger using smart contract and cryptography techniques. By honestly participating in the platform, an SP can be endorsed by the BC network that it is fully GDPR-compliant; otherwise any violation is immutably recorded and is easily figured out by associated parties. We then demonstrate the feasibility and efficiency of the proposed design concept by developing a profile management platform implemented on top of a permissioned BC framework, following by valuable analysis and discussion.**

***Index Terms*—Blockchain, Data Management, GDPR, Personal Data, Smart Contract.**

I. Introduction

THE General Data Protection Regulation legislation came into force in May 2018 in all European Union (EU) countries. The GDPR is a major update to the data privacy regulations released in 1995, which is before the proliferation of cloud platforms and social media, let alone the scale of today's data usage. The provision of the GDPR is to ensure that personal data "can only be gathered legally, under strict conditions, for a legitimate purpose"; also, to bring full control back to the data owners[1].

As the GDPR requirements are highly abstract, it is open to interpretation. In fact, each organisation has its own way to satisfy the new regulations; and to demonstrate the compliance. Supposedly, each EU member state provides a Supervisory Authorities (SA) who is responsible for monitoring the GDPR-compliance. Organisations are required to demonstrate the compliance only in case of suspicion of a violation or when a Data Subject (i.e., the owner of data, denoted as DS) lodges a complaint with the SA. In this regard, the challenge of complying with the GDPR is not because of lacking technical solutions for tackling down the GDPR requirements nor providing required mechanisms; it is because such solutions are designed and implemented under a centralised client-server architecture mindset. Due to the irregular verification of GDPR compliance, critical concerns on the lack of transparency have been imposed accordingly. In particular, it is unachievable for a Service Provider (SP) to prove that it has been continuously adhering to the GDPR using existing centralised solutions. Moreover, it is beyond the DS's capability to perceive whether an SP fully complies with the GDPR and effectively protects her personal data. For these reasons, GDPR-compliant personal data management is a well-suited scenario for blockchain (BC) to come into play. A BC platform employed Smart Contracts (SCs) is expected to be a promising solution to deal with such challenges thanks to its advanced features of decentralisation, transparency, tamper-resistance and traceability.

In this article, we propose a design concept for developing a GDPR-compliant personal data management platform, along with a detailed implementation of a platform in a specific use-case. The goal of the design concept is to preserve advanced features of BC and SCs in personal data management by leveraging distributed ledger and public-key cryptography technologies for complying with the manifold legal requirements of the GDPR [1]. By following the proposed design concept, a personal data management platform ensures that only designated DSs and Data Controllers (DCs) are permitted to create, update and withdraw consents; and only authorised Data Processors (DPs) can process personal data respecting rules defined in corresponding data usage policy agreed between the DSs and the DPs. The platform not only provides mechanisms for DS rights, but also plays as a role of a DC for handling personal data processing and demonstrating data accountability. By honestly participating in the BC-based personal data management platform, an SP can be endorsed by the BC network that it is GDPR-compliant. Otherwise any violations are recorded in an immutable distributed ledger as a record of the infringements, which can be then used for the GDPR compliance investigation by SAs.

We demonstrate the feasibility and effectiveness of the proposed design concept by developing a platform, exactly

N.B. Truong, K. Sun and Y. Guo are with Data Science Institute, Department of Computing, Imperial College London, London, SW7 2AZ United Kingdom. E-mail: n.truong@imperial.ac.uk, k.sun@imperial.ac.uk, y.guo@imperial.ac.uk

G.M. Lee is with Department of Computer Science, Liverpool John Moores University, Liverpool, L3 3AF United Kingdom. E-mail:g.m.lee@ljmu.ac.uk

[1]https://gdpr-info.eu/

following the instructions in the design concept, for managing personal profiles. The platform, which is built on top of the Hyperledger Fabric (HLF) permissioned BC framework[2] and cooperates with an honest Resource Server (RS) for data storage, plays as a profile management system for a social networking service (SNS) provider. The proposed platform provides SNS clients rights as well as facilitates the SNS provider's obligations, following by analysis and discussion on the GDPR-compliance, threat models and system performance. It is affirmed that the SNS is fully compliant with the GDPR requirements. We believe the proposed approach is a promising solution not only for a GDPR-compliant personal data management but also for digital assets governance.

The rest of the article is organised as follows. Section II presents background and related work. Section III describes challenges and motivation. The design concept is proposed in Section IV following by the implementation of the profile management platform in Section V. Section VI provides the analysis and discussion about the platform. The last section concludes our work and outlines future research.

II. Background and Related Work

In this section, relevant background knowledge on GDPR and BC and related work are presented. Table I depicts some of the notions frequently used throughout this article.

Table I: NOTATION TABLE WITH ENTRIES IN ALPHABETICAL ORDER

Notation	Description
API	Application Programming Interface
BC	Blockchain
BFT	Byzantine Fault Tolerance
C-ID	Complex Identity
CA	Certificate Authority
CRUD	Create-Read-Update-Delete operations
DBMS	Database Management System
DC	Data Controller
DP	Data Processor
DS	Data Subject
GDPR	General Data Protection Regulations
HLF	Hyperledger Fabric permissioned Blockchain framework
IdM	Identity Management
MSP	Membership Service Provider
OSN	Ordering Service Node
RS	Resource Server
SA	Supervisory Authority
SC	Smart Contract
SP	Service Provider

A. The GDPR in a Nutshell

The full GDPR regulations are described in detail across 99 articles covering all of the technical and admin principles around how commercial and public organisations process personal data. GDPR lays out the means by which personal data is to be protected which are founded on a set of six core data processing principles: Lawfulness, Fairness and Transparency; Purpose Limitation; Data Minimisation; Accuracy; Storage Limitation; Integrity and Confidentiality[3]. To preserve such principles, the GDPR clearly differentiates three roles (i.e., DS, DC and DP) and explicitly specifies associated rights and obligations under the EU data protection law. The goal of the GDPR legislation is to provide a DS full control over her personal data by specifying variety of rights[4]. The GDPR requires that personal data should be managed by a DC employing mechanisms to ensure the rights of the DS. Such mechanisms enable the DS to impose consents and to arbitrarily withdraw the consents whenever needed. The DS is also able to trace back all activities on her data including who, what, why, when, and how the data is processed. Valid legal consents must be given by the DS to the DC for processing her personal data. The DC then takes appropriate measures to provide the rights of the DS; meanwhile determines the purposes for which, and the method in which, the personal data is processed by DPs. Being compliant with the GDPR is not enough, DCs should also be able to demonstrate the compliance to SAs once required (when a SA has suspicion of a violation or when a DS lodges a complaint with the SA). In this case, the SA shall establish and make public a list of processing operations subjected to Data Protection Impact Assessment and the Privacy Impact Assessment requirements[5]; then file a report of infringements if it is the case.

B. Blockchain Technology

A BC is a distributed immutable database constituted from a continuous growing list of blocks. The BC plays as the role of a distributed ledger as it records all transactions between entities in a network. By nature, a BC is inherently resistant to data modification. Once recorded, information in any given block cannot be altered retroactively as this would invalidate all hashes in the previous blocks in a BC; and break the consensus among nodes in the network. The concept of BC was introduced in Bitcoin in 2008 [2]. Bitcoin is the first cryptocurrency that not only transacts digital currency in a securely manner, but also resolves the long-standing problem of "double spend" without the need for a trusted third-party. BC underpins Bitcoin, but BC is not only Bitcoin. Its usage goes far beyond [3]–[5].

In a BC network, a consensus protocol needs to be implemented to ensure any disruptive action from an adversary will be negated by a majority of participants [2]. The protocol is to decide which player among the participants in the BC network has permission to append a new block; other participants are able to verify the permission and update their local ledgers accordingly; which establishes consensus over the network [6], [7]. Proof of Work (PoW) is the most common consensus model used in public BCs. Unfortunately, PoW is computation-intensive, as it requires powerful nodes (i.e., miners) dedicate to solve a computationally intensive puzzle (i.e., mining), in order produce a new block to the chain [8]. To overcome latency and throughput bottlenecks of PoW, alternative consensus models have been proposed, including Proof of Stake (PoS) [9], [10], Byzantine fault-tolerant (BFT) variants [11], Proof of Elapsed Time (PoET)[6], and Algorand [12]. Nonetheless, such consensus protocols impose their own

[2]https://www.hyperledger.org/projects/fabric
[3]https://gdpr-info.eu/art-5-gdpr/
[4]https://gdpr-info.eu/chapter-3/
[5]https://gdpr-info.eu/issues/privacy-impact-assessment/
[6]https://sawtooth.hyperledger.org/docs/core/releases/latest/index.html

disadvantages which results in limited usage in the real-world compared to the PoW-variant mechanisms [7].

C. Smart Contracts

A SC is a computer program deployed onto a BC network. It automatically executes "actions" when necessary "conditions" are met, specifying business logic of a service that participants have agreed to [13]. As a mutual agreement, content of the SC is accessible to all participants [14]. A SC is a form of decentralised automation that facilitates, verifies, and enforces an agreement in a transaction and records the results (i.e., state changes) into a ledger. All BC frameworks have built-in mechanisms for executing SCs from a simple stack-based scripting system (e.g., Bitcoin) to a Turing-complete system (e.g., Ethereum and Hyperledger). Ethereum is among the first BCs offering Turing-completeness. Its SCs are written in either Solidity, Serpent or LLC, before being compiled to bytecodes and executed in an Ethereum Virtual Machine (EVM) [15]. The EVM keeps track of resources consumed by the execution (i.e., gas) and charges to the sender's account as an incentive for miners. Hyperledger does not have its own bytecode for SCs. Instead, its SCs are language-agnostic programs which are then compiled into native code, packed, installed and executed inside Docker containers [16]. As a result, this language-agnostic design supports multiple high-level programming languages such as Go and JavaScript [17].

D. Related Work

Besides cryptocurrencies, the use of BC in other areas has been intensively carried out over the last few years. Specifically, prominent features of BC such as immutability, traceability, transparency and pseudo-anonymity can be preserved for a wide range of decentralised applications (DApps), especially for managing and accounting digital assets. For instance, several projects have utilised BC in supply-chain and logistics to provide provenance tracking mechanisms for products leveraging its immutability and traceability features [18]–[20]. The immutability and transparency features have also been utilised in a cloud data provenance platform called ProvChain [21] in which all data operation history was transparently and permanently recorded into a BC.

Furthermore, a BC framework employed SCs can provide autonomous functionalities executed in a decentralised manner for a wide range of domain services. Blockstack [22] took advantages of BC for managing domain names in order to replace the traditional centralised Domain Name System. This work introduced pivotal functionalities including identity and discovery mechanisms deployed on top of the Namecoin platform [23] and integrated with an off-chain storage service. In Blockstack, domain name registration and modification operations were implemented in BC whereas payload and digital signatures were stored in a Kademlia[7] Distributed Hash Table (DHT), which was connected to a virtualchain that separated off-chain storage and BC operations. Only hashes of "name-data" tuples and state transitions were recorded on-chain. This design of decoupling the storage layer from the BC has paved the way to other studies, particularly in large-scale Internet of Things (IoT) data management [24], [25]. In these studies, data generated from IoT devices was stored in a DHT system and only keys of the data were recorded onto a BC. DHT nodes, responsible for managing IoT data, are required to join the BC network and listen to transactions for sending/retrieving data to/from legitimate IoT devices. BigchainDB [26] further provided a mechanism to balance between on-chain and off-chain storage to achieve advanced features from both BC and distributed databases by using Tendermint[8], a weak synchronisation BC engine built on a BFT consensus.

Besides general-purpose data storage, BC-based accounting and management mechanisms (e.g., IdM, authorisation, access and permissions control) have also been proposed in a variety of scenarios. Lee proposed a BC-based cloud ID service for IdM [27], which used public-key cryptography for pseudo-identity and a distributed ledger for recording public keys. This study introduced a concept of mutual authentication by combining signatures from a client and an SP for granting access to a service. A fast security authentication scheme based on permissioned BC was proposed by Chen *et al.* in a 5G ultra-dense network [28] by using an optimised Practical BFT (PBFT) consensus protocol called APG-PBFT. APG-PBFT propagated authentication results embedded in BC among a group of access points, resulting in reducing the authentication frequency. In [29], a distributed access control in the IoT was proposed, with operations embedded in a SC on a public BC (i.e., Ethereum). However, most of these studies only presented high-level system design, without technical details to demonstrate the feasibility of their proposed solutions. Some platforms (e.g., [29]) relied on a set of management nodes to play as a hub for access control, which in fact turns into the scenario of centralised management.

Only few studies in the literature concerning BC-based personal data management, particularly on supporting SPs to comply with the new GDPR legislation. In [30], Wang *et al.* proposed a fine-grained access control scheme deployed in the Ethereum framework, for personal files stored a distributed file system called Interplanetary File System (IPFS) [31]. It customised an attributed-based encryption, but the dependency of a centralised trusted private key generator is eliminated by leveraging BC. The main limitation of this system is data owners were responsible for all required tasks, from secret key generation, file encryption, to the establishment of a secure channel for communicating with another party. The Ethereum framework was just used as a medium to execute SCs in which crypto-artifacts were embedded for identity authentication. Zyskind and Nathan [32] proposed another access control scheme for a privacy-preserving personal data sharing platform, taking advantages of immutability and public-key cryptography in BC for identity verification and authorisation mechanisms. Similar ideas were proposed for Electronic Heath Records (EHRs) access control using Ethereum [33], [34] or a permissioned BC [35]. In these works, EHRs were stored off-chain in secure data custodians whereas access control

[7]https://en.wikipedia.org/wiki/Kademlia

[8]https://tendermint.com

was carried out on a BC using a digital signature scheme. Neisse *et al.* [36] proposed a BC-based approach for data accountability, resulting in GDPR-compliance. They discussed different design choices respecting to who create and manage data usage SCs. Similar ideas can be found in [37], [38].

However, in these studies, only conceptual approach were presented; technical details on platform development were missed out. The challenges including ledger data models and functionalities in SCs have not been addressed.

III. Challenges and Motivation

In this paper, we propose a comprehensive design concept with detailed technical aspects for the implementation of a BC-based GDPR-compliant personal data management platform. We consider scenarios that a personal data management mechanism is implemented under a centralised client-server architecture (Fig. 1). This specifies three roles as follows:

- *End-user*: the client of a service who owns the personal data (i.e., a DS in the GDPR terminology).
- *Service Provider (SP)*: an entity that collects and manages personal data (i.e., a DC) for operational and business-related purposes (i.e., a DP). An SP stores personal data in an RS, which is either a service run by the SP or an independent service. An SP may share collected data with third-parties for its benefits. In the context of GDPR, an SP plays both roles as a DC and a DP.
- *Third-party (TP)*: an entity that processes personal data for its own service (i.e., a DP in the GDPR terminology). TP relies on a SP' infrastructure to acquire desired personal data by calling APIs provided by the SP.

As illustrated in Fig. 1, the procedure of granting data access for an SP and a TP is in four steps:

1) A user starts to use a service provided by an SP. The SP asks the user for permission to collect her personal data.
2) The end-user grants a set of permissions to the SP for personal data collection and processing.
3) The TP asks the end-user to access her personal data which is collected and managed by the SP.
4) End-user logs into the service provided by the SP and consents a set of permissions to the TP

Once the permission is granted, the data access procedure is in the fifth and the sixth steps in Fig. 1:

5) The SP authenticates and authorises the TP for accessing the data and provides an access token to the TP.
6) The TP then calls associated APIs using the provided token in step-5 to obtain the desired data.

Current approaches used by SPs to meet GDPR requirements are based on the client-server architecture, resulting in limited transparency and a lack of trust. For instance, a majority of SPs follow the $OAuth2$[9] standard for access delegation, which includes IdM, authentication, authorisation, and access control mechanisms that allows end-users to share their personal data with single sign-on in a simplified and secure manner [39]. However, the centralisation of the current approaches poses severe concern [40]: it fully relies on the truthfulness of the SP (or a delegated authentication server) as it is the only authority to *(i)* authenticate and authorise participants; and *(ii)* control data access and provenance, as illustrated in Fig. 1.

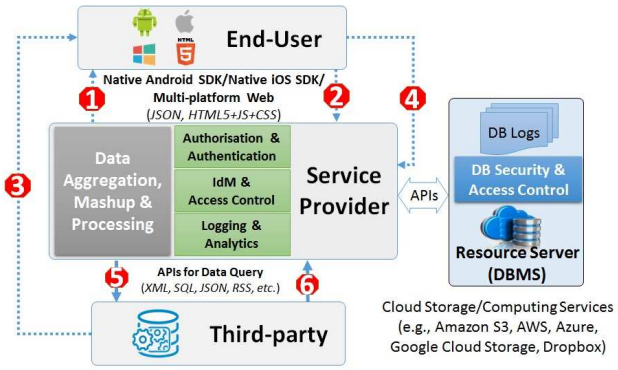

Fig. 1: System model of a personal data management and sharing scheme using the conventional client-server architecture

From an end user's prospective, this leads to lack of transparency and accountability of data management and raise risks of personal data leakage. As all data management mechanisms are operated in a centralised system and under the SP's control, the SP may still be able to hand over personal data to an unauthorised TP without the end-user's knowledge, as far as it is not investigated by SAs. From an SP's prospective, as investigation from SAs is occasionally carried out, it is challenging for an SP to declare that it has been continuously, securely and legally processing all personal data as required. This is of paramount importance for any SP to build trust with prospective clients. Furthermore, delegated permissions among end-users, SPs and TPs on personal data are not flexible. In most cases, end-users do not have a fine-granular access control to impose their preferences on data usage except simple conditions predefined by SPs. Indeed, many SPs provide only options to either "accept all" or opt-out.

Motivated by such challenges, our ultimate goal is to develop a GDPR-compliant personal data management platform by leveraging the state-of-the-art BC and SC technologies. The use of BC with SC provides autonomous operations securely executed in a decentralised manner. Furthermore, the prominent features of the BC technology, namely immutability, traceability, transparency and pseudo-anonymity, can be effectively utilised to manage personal data fully complying with the GDPR legislation.

IV. Design Concept

In this section, we propose a design concept for a GDPR-compliant personal data management platform, including a high-level system architecture, design guidelines and detailed functionalities and algorithms.

A. Conceptual Model and System Architecture

1) Assumption: The design of a BC-based platform depends on the security models of the parties involved. In this article, we assume that a RS is "honest-but-curious" whereas SPs follow a malicious model. This means the RS executes required protocols honestly, even though it might be curious

[9]https://oauth.net/2/

about the results it receives after the operations. If an SP correctly follows the required protocols; it will be compliant with the GDPR; otherwise violations will be logged in an immutable ledger as a record of GDPR infringements.

2) High-level System Architecture: A conceptual model of the proposed platform is illustrated in Fig. 2. The inclusive idea is that mechanisms which are related to GDPR compliance are ported to a BC network from a traditional centralised server. In particular, the Authorisation and Authentication, IdM and Access Control; and Logging and Provenance components are implemented in form of SCs employed in a BC network. If a BC framework offers Turing-completeness (e.g., Ethereum and Hyperledger Fabric), GDPR-related mechanisms can be conveyed by SCs. As depicted in Fig. 2, all activities on personal data are authenticated and authorised by the proposed BC platform (step 1 and 2). An authorised SP receives an access token from the platform (step 2) and use it to request the data from the RS (step 3). The RS interacts with the BC platform to validate the granted access (step 4 and 5) before returns the requested data (step 6). The validation ensures the granted access is still valid and honestly used by the corresponding authorised party.

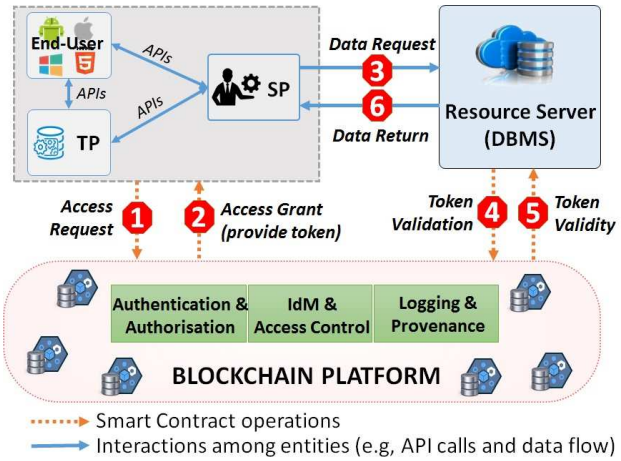

Fig. 2: High-level System Architecture of the design concept for a BC-based personal data management platform. The operation flow consists of 6 steps, among which step 1, 2, 4, and 5 are dedicated to granting and validating permissions operated through SCs. Step 3 and 6 operated via API calls and data-flow from/to an RS.

B. Design Guidelines

1) IdM, Authentication and Authorisation mechanisms: IdM, authorisation and authentication mechanisms are of paramount importance in any data management system since they are directly related to security and privacy of the system. In the design concept, an entity in a BC network should be uniquely identified using a public-key (or hash of the public-key) in an asymmetric cryptography key-pair; authentication and authorisation processes should be implemented leveraging public-key cryptography techniques (e.g., digital signatures and encryption). In case of permissioned BC, an additional access control layer is consolidated by using a Certificate Authority (CA) and a Membership Service Provider (MSP).

2) Design of Distributed Ledgers: Content of a distributed ledger reflects historical and current states of information recorded in the ledger maintained by the BC network. A personal data management platform should clarify what information and associated data model to be stored in the ledger.

(i) Information required to be tamper-resistant, transparent and traceable should be recorded in a distributed ledger.
- Any personal dataset should be specified by both DS and DC using digital signatures in a distributed ledger;
- Data Usage Policy should be clearly specified and recorded in a distributed ledger;
- Data activities should be logged in a distributed ledger. The logs should contain information about 'who', 'why', 'when', 'what' and 'how' personal data was processed;
- Hash of personal data can be recorded in a distributed ledger for data integrity checking.

(ii) The design of a distributed ledger must ensure:
- Designated nodes in the BC network are able to verify whether an entity is the DS or the DC of a dataset;
- Designated nodes in the BC network should be able to verify whether an entity's activity satisfies the data usage policy as recorded in a distributed ledger

3) Data Usage Policy: The policy specifies data governance measures including rights, permissions and conditions. The usage policy should be defined in a fine-grain and expressive way using a policy language such as eXtensible Access Control Markup Language (XACML) and Model-based Security Toolkit (SecKit) designated for the IoT domains [41].

4) Off-chain Data Storage: Personal data should be stored off-chain for better scalability and higher efficiency. Moreover, storing personal data directly onto BC, even in an encrypted form, could pose potential privacy leakage and result in non-compliance with the GDPR [42]. Depending on specific scenarios, a conventional DBMS (e.g., Oracle or MongoDB), a storage cloud service (e.g., S3, AWS or Azure), or a distributed storage system (e.g., IPFS [31] or Storj [43]) can be used for data storage. Only reference to the data is stored on-chain (i.e., stored in distributed ledgers). The reference is called $data\_pointer$ that can be a hash[10], a connection string, an absolute path, or an identifier referring to a dataset; depending on specific off-chain storage system used in the platform.

C. Functionalities, Ledgers Data Model and Algorithms

1) Identity Management: We introduce *complex-identity*, denoted as *c-ID*, to specify a digital asset associated with two or more parties. A *c-ID* can be considered as an extension of asymmetric keys. In the context of the personal data management, a *c-ID* of a dataset m comprises an asymmetric key pair of the DS, an asymmetric key pair the DC, and an asymmetric key pair of the data pointer (denoted as p_m) of m. As the data usage policy depends on the requester's role (i.e., DS, DC, or DP), the way we define *c-ID* specifies the entities associated with m, and simplifies the process of verification. A digital signature scheme can be used to generate and manage the *c-ID*, which is formally defined as a triple of probabilistic polynomial time algorithms $(\mathcal{G}, \mathcal{S}, \mathcal{V})$:

- $\mathcal{G}$: a key generator that creates a public-private key pair (pk, sk).

[10]Hash is a type of the $data\_pointer$ used in a content-addressed storage system such as DHT, IPFS, and Stoij.

- $\mathcal{S}$: a signing algorithm that takes sk and a message x as inputs and produces a signature $t = S(sk, x)$ as the output.
- $\mathcal{V}$: a signature verifying algorithm that takes pk, x, t as inputs, and outputs $accept$ or $reject$. For all x and (pk, sk), $V(pk, x, S(sk, x)) = accept$.

A *complete c-ID* is defined as a 6-tuple as follows:

$$c\text{-}ID^{comp}_{DS,DC} = (pk_{DS}, sk_{DS}, pk_{DC}, sk_{DC}, pk_{enc}, sk_{enc}) \quad (1)$$

where (pk_{DS}, sk_{DS}), (pk_{DC}, sk_{DC}) and (pk_{enc}, sk_{enc}) are asymmetric key-pairs of DS, DC and p_m, respectively. The *c-ID* is externally observed by nodes in a BC network as a 3-tuple:

$$c\text{-}ID^{ext}_{DS,DC} = (pk_{DS}, pk_{DC}, pk_{enc}) \quad (2)$$

The *c-ID* is observed by the DS (or DC) as a 5-tuple:

$$c\text{-}ID^{DS}_{DS,DC} = (pk_{DS}, sk_{DS}, pk_{DC}, pk_{enc}, sk_{enc}) \quad (3)$$

$$c\text{-}ID^{DC}_{DS,DC} = (pk_{DS}, pk_{DC}, sk_{DC}, pk_{enc}, sk_{enc}) \quad (4)$$

When a DS grants consent to a DP to access m, the private key sk_enc of p_m is shared to the DP through a secure channel. The DP then observes the *c-ID* as a 4-tuple:

$$c\text{-}ID^{DP}_{DS,DC} = (pk_{DS}, pk_{DC}, pk_{enc}, sk_{enc}) \quad (5)$$

The $c\text{-}ID^{DP}_{DS,DC}$ includes the key-pair (pk_{enc}, sk_{enc}) used to encrypt and decrypt sensitive information, including the data pointer p_m. Thus, only designated nodes are able to decrypt the ciphertext using the shared private key sk_{enc}. As a result, the information is protected from all other players in the system. Normally, RSA (Rivest-Shamir-Adleman) is used for the public-key encryption scheme, formally defined as a 4-tuple ($\mathcal{G}$,$\mathcal{D}$,$\mathcal{E}$,d): the key generator, key distribution scheme, encryption and decryption schemes, respectively.

2) Distributed Ledgers Data Model: In the proposed design concept, ledgers are in form of key-value pair, which is widely used in BC frameworks including Ethereum and HLF. For complex business logic, extra tasks might be required for mapping high-level data structures into key-value pairs. A state is a snapshot of a ledger at a specific time whereas state transitions are a result of transactions for creating, updating or deleting key-value pairs. A ledger contains full history of state transitions recorded in a BC, thus it is timestamp-sequenced, immutable and tamper-resistant. With the key-value data format, all information can be obtained by referring to the latest state of the ledger, which is written in the most recent block of the BC. Some frameworks duplicate the latest state of a ledger (i.e., world-state) from a BC to a DBMS for better performance and for supporting advanced query capability (e.g., rich query). For example, either CouchDB[11] or LevelDB[12] are used in the HLF for its world-state database.

Following the design guidelines for distributed ledgers, we specify data models for two separate ledgers used in personal data management: $3A\_ledger$ (Listing 1) and log_ledger (Listing 2). The $3A\_ledger$ is used in authentication, authorisation and access control whereas the log_ledger is used for access validation and logging. Both ledgers are in key-value format in which $keys$ in the $3A\_ledger$ and log_ledger are $c\text{-}ID_{DS,DC}$ and $c\text{-}ID_{DS,DC,DP}$, respectively. The $value$ in both ledgers contains information being used in the personal data management and provenance operations.

[11]http://couchdb.apache.org
[12]http://leveldb.org

```
{"3A_ledger": {
    "key": {
      "owner": pk_DS,
      "controller": pk_DC
    }
    "value" {
      "en_pointer": 3erwf3ese6d5c4...,
      "policy": {
        "rule": {Effect},{Condition},
        "action": "read, update",
        "target": "{pk_1, pk_2, ...}"
      },
      "pk_enc": "fMA0GCSqGSIb3...",
      "hash": "369f2e3e69dc40543...",
      "timestamp": 1549480378
}}}
```

Listing 1: A state of the $3A\_ledger$ in JSON format. Content of the ledger includes $en\_pointer$: ciphertext of a data pointer; pk_enc: public key used to encrypt the $en\_pointer$; $policy$: data usage policy, and $hash$ of the data.

```
{"log_ledger": {
    "key": {
      "owner": pk_DS,
      "controller": pk_DC,
      "processor": pk_DP
    }
    "value" {
      "access_token": "aAD0Gdfs234S3...",
      "issued_at": 1549480378,
      "status": "approved",
      "operation": op,
      "scope": []ops,
      "expires_in": 3600,
      "refresh_count": 1,
}}}
```

Listing 2: A state of the log_ledger in JSON format. Content of the ledger includes $status$: either $approved$ or $rejected$, $operation$: an activity a DP used to process the data such as CRUD, $scope$: a set of allowed permissions, $expires\_in$ and $refresh\_count$: dedicated to controlling the $access\_token$.

Note that content of the ledgers can be seen by corresponding nodes in the BC network, either honest or malicious ones. Therefore, sensitive information should be protected using appropriate methods. For instance, asymmetric cryptography is used for pseudo-anonymous identity; and reference to a dataset (i.e., data pointer p_m) is encrypted (Eq. 6).

$$en\_pointer = \mathcal{E}(pk_{enc}, p_m) \quad (6)$$

3) Authentication, Authorisation and Access Control: Public-key cryptography has been commonly used in BC-based systems to authenticate participants involved in a variety of tasks from consensus protocol participation to SC operations. In our design concept, the authentication is achieved by using the algorithm $\mathcal{V}$ in the 3-tuples digital signature scheme $\mathcal{G}, \mathcal{S}, \mathcal{V}$ based on any RSA/DSA-variants. The authorisation in personal data management is to specify access control (e.g., consent and usage policy); and data provenance tracking is to log data activities in an immutable and tamper-free ledger.

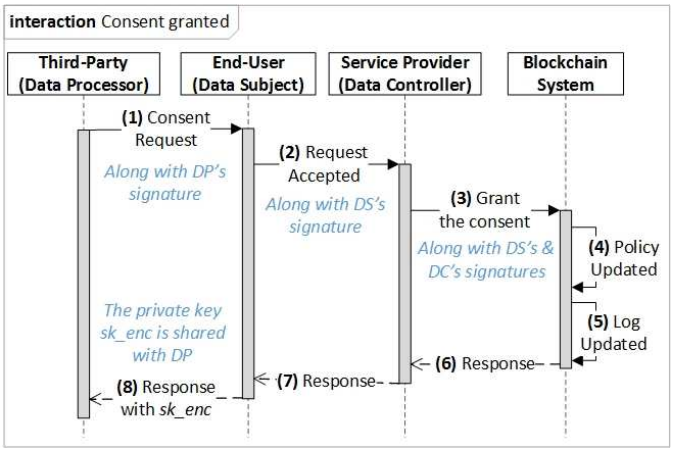

Fig. 3: Process of granting consent for a DP.

In the initial step (i.e., $Registration$ function), a DS grants consent to a DC for managing her personal data along with a shared key-pair (pk_{enc}, sk_{enc}). A new record is appended into the $3A\_ledger$ specifying a new key-pair for the personal dataset with default settings granting DS all permissions (e.g, CRUD operations) specified in the $policy$. The $policy$ can be considered as an access control list/rules for a dataset, updated when a consent is granted or revoked. The $hash$ and the $en\_pointer$ in the record are then updated once the DS upload her data to an RS by calling $DataUpload$ function. In our pseudo-codes, interactions with BC is through either $GetState$ or $PutState$ function provided by built-in APIs.

Alg. 1: $GrantConsent$ grants a consent for a DP

Input : c-ID ci, signature t_{DS}, signature t_{DC}, public-key pk_{DP}, signature t_{DP}, permission op
Output: out
1 **Initialisation:** $rec \leftarrow null$, $out \leftarrow error$
2 $s1 \leftarrow \mathcal{V}(ci.pk_{DS}, t_{DS})$
3 $s2 \leftarrow \mathcal{V}(ci.pk_{DC}, t_{DC})$
4 $s3 \leftarrow \mathcal{V}(pk_{DP}, t_{DP})$
5 **if** $(s1 \wedge s2 \wedge s3)$ **then**
6 $policy \leftarrow$ GetState($3A\_ledger$).GetPolicy(ci)
7 PutState($3A\_ledger$).Update(ci, $policy$, $\{pk_{DP}, op\}$)
8 $rec \leftarrow$ JSON.Marshall($\{ci, pk_{DP}\}$, $\{scope[]$+=op, $access\_token$=rand(), $issue\_at$=Time.now(), $status$="approved"$\}$);
9 PutState(log_ledger).Append(rec);
10 $out \leftarrow success$
11 **Return** out

Fig. 3 depicts a sequence diagram of granting a consent for a DP. The consent is granted if both DS and DP accept the request by providing their digital signatures t_DS and t_DC in step (2) and (3). Step (4) and (5) are carried out by the $GrantConsent$ function (Alg. 1). Authentication is achieved by using verification function $\mathcal{V}$ for all DS, DC and DP (line 2-4). If the authentication is accepted (line 5), access control is then carried out by reflecting the permission into $policy$ in the $3A\_ledger$. As depicted in Alg. 1, the $GrantConsent$ firstly grants permissions (i.e., requested operation op) by updating policy with op in the $3A\_ledger$ (line 6, 7). Secondly, the $GrantConsent$ appends a new record into the log_ledger (line 9), which is used for validating and logging whenever the DP accesses the data. The $access\_token$ with other metadata is generated as $value$ in the $key - value$-format record (line 8). Technically, $access\_token$ is a string of random-looking characters referring to a collection of metadata in the log_ledger. A multi-signature technique is also used in the algorithm to ensure a consent is granted by both DS and DC.

$RevokeConsent$ function is to revoke a permission previously granted to a DP. As depicted in Alg. 2, it is only executed by either DS or DC. Similar to $GrantConsent$ function, $RevokeConsent$ appends an updated policy excluded the revoked permission op to the $3A\_ledger$ (line 4, 5) and updates the log_ledger accordingly (line 6,7).

Alg. 2: $RevokeConsent$ revokes a permission previously granted to a DP

Input : c-ID ci, signature t, public-key pk_{DP}, permission op
Output: out
1 **Initialisation:** rec = $null$, out = $error$
2 $s \leftarrow (\mathcal{V}(ci.pk_{DS}, t) \vee \mathcal{V}(ci.pk_{DC}, t))$
3 **if** s **then**
4 $policy \leftarrow$ GetState($3A\_ledger$).GetPolicy(ci)
5 PutState($3A\_ledger$).Update(ci, $policy$, $\{pk_{DP}, -op\}$)
6 $rec \leftarrow$ GetState(log_ledger).GetRecord(ci, pk_{DP})
7 $rec \leftarrow$ PutState(log_ledger).Update(rec, $\{scope[]$-=op, $access\_token$=rand(), $issue\_at$=Time.now()$\}$);
8 $out \leftarrow success$
9 **Return** out

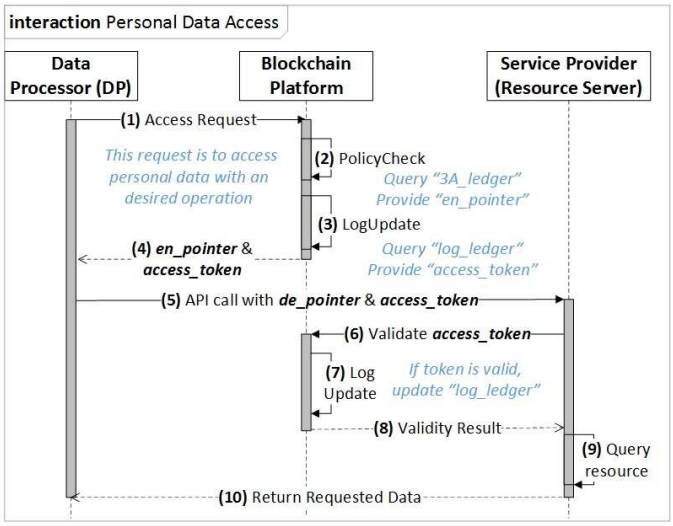

Fig. 4: Sequence Diagram of accessing data stored in an RS by a DP

Once consent is grant, the operation flow of accessing personal data is demonstrated in Fig. 4. Whenever DP desires to access personal data (step (1)), it invokes a corresponding SC with the $DataAccess$ function (Alg. 3). As can be seen in Fig. 4, after checking eligibility of the call (i.e., step (2) and (3) executed by line 2, 3 in Alg. 3), the SC returns two outputs $en\_pointer$ and $access\_token$ to the DP (step (4)), executed by line 6-9 in Alg. 3. The DP then uses the shared private key sk_{enc} (already obtained from step (8) in Fig. 3) for decrypting the $en\_pointer$. The decrypted ciphertext (i.e., $de\_pointer$) is the $datapointer$ for the desired dataset. Both $de\_pointer$ and $access\_token$ are used as parameters for an API call to process the data (step (5)).

Alg. 3: $DataAccess$ returns $en\_pointer$ and $access\_token$ for an eligible request

Input : c-ID ci, public-key pk_{DP}, signature t_DP, permission op
Output: out
1 **Initialisation:** $rec \leftarrow null$, $out \leftarrow rejected$
2 $s \leftarrow (\mathcal{V}(pk_{DP}, t_{DP})$
3 **if** s **then**
4 $policy \leftarrow$ GetState($3A\_ledger$).GetPolicy(ci)
5 **if** $(policy \subset (pk_{DP}, op))$ **then**
6 $en\_pointer \leftarrow$ GetState($3A\_ledger$).GetPointer(ci);
7 $access\_token \leftarrow$ GetState(log_ledger).GetToken(ci, pk_{DP});
8 $out \leftarrow (en\_pointer, access\_token)$
9 **Return** out

A function called $TokenValidation$ is dedicated to double-checking the validity of the $access\_token$ and updates the log_ledger. In Alg. 4, line 4 is to obtain metadata associated with the $access\_token$ from the log_ledger; if the request is from DS or DC then there is no need to validate the $access\_token$; only log_ledger is updated (line 5-7). Otherwise, the validation is then conducted by inspecting the metadata (line 9-12) before updating the log_ledger (line 13). The $TokenValidation$ is performed to ensure that only API calls with valid an $access\_token$ leads to an execution of the call (step (9)). Step (7) safeguards that all valid API calls are autonomously logged in the log_ledger. It is worth to mention that the honest-but-curious RS assumption plays a key role in the success of our platform because the RS must follow the authorisation process (i.e., double-check API calls from DPs with the BC system) before executing the calls.

Alg. 4: $TokenValidation$ double-checks the validity of an $access\_token$ and update the log_ledger

Input : Token $access\_token$, public-key pk, signature t permission op
Output: out
1 **Initialisation:** $rec \leftarrow null$, $out \leftarrow rejected$
2 $s \leftarrow (\mathcal{V}(pk, t)$
3 **if** s **then**
4 $rec \leftarrow$ GetState(log_ledger).Query($access\_token$)
5 **if** $((rec.owner = pk) \vee (rec.controller = pk))$ **then**
6 $rec \leftarrow$ PutState(log_ledger).Update(rec, {$expires\_in$-=Time.now(), $issue\_at$=Time.now()});
7 $out \leftarrow accepted$
8 **else**
9 **if** $(rec.processor = pk) \wedge (rec.scope \subset op) \wedge$
10 $(rec.expires\_in > 0) \wedge (rec.operation = op) \wedge$
11 $(rec.status = approved) \wedge ...)$ **then**
12 $rec \leftarrow$ PutState(log_ledger).Update(rec, {$expires\_in$-=Time.now(), $issue\_at$=Time.now()});
13 $out \leftarrow accepted$
14 **Return** out

V. Platform Deployment in Permission Blockchain

In this section, we implement a platform following the proposed design concept for managing personal profiles for a SNS. The choice of using a permissioned BC framework in the demonstration does not imply that a public one is less appropriate implementing the proposed design concept. Instead HLF is chosen due to its business-oriented architecture offering better adaptation to the use-case; also, thanks to its readily existing software components for a rapid development cycle of our platform. Detailed technical solutions and implementation of the platform are presented. Source-code of the demonstration can be obtained from Github[13].

A. HLF Platform Setup

HLF is the most popular permissioned BC framework used by big enterprises such as IBM and Microsoft. As being permissioned, a node involved in an HLF network is associated with an identity and permissions provided by a CA and an MSP, respectively. Nodes in HLF take up one of three roles: *Client*, *Peer* and *Ordering Service Nodes* (OSNs). In our demonstration, the HLF network consists of 3 OSNs running in $Kafka$ cluster mode for providing the ordering service, 5 peers, and 10 clients (4 as DS, 4 as DP, 1 as SP and 1 as RS). All 5 peers endorse both SCs (i.e., chaincodes in HLF terminology), namely $3A\_cc$ and log_cc. That means these two SCs are locally installed, instantiated and executed in all 5 peers to interact with the two ledgers $3A\_ledger$ and log_ledger, respectively. These two ledgers are exactly following the data models described in Section IV.D. As the two distributed ledgers are being used and HLF allows only one ledger per channel[14], two HLF channels are created, namely $3A\_channel$ and $log\_channel$. All Peers and OSNs belong to both channels; the $3A\_cc$ and the log_cc SCs are operated in the $3A\_channel$ and the $log\_channel$, respectively. As a result, all 5 peers endorse the two SCs separately corresponding to different local ledgers. The BC local ledger is stored in Linux filesystem whereas the world-state database is duplicated in CouchDB.

All 10 clients are implemented using the Fabric Client SDK (for NodeJS) for interacting with the HLF network. As illustrated in Fig. 5, a client constructs a transaction proposal to invoke either $3A\_cc$ or log_cc SCs (step-1) and sends to all endorsing peers (i.e., endorsers). These peers verify the proposal and locally execute the $3A\_cc$ or log_cc to produce an endorsement signature (i.e., transaction results with the peer's signature) (step-2) and pass back to the client (step-3). Once receiving endorsement signatures, the client assembles the endorsements into the transaction and broadcast it to the OSNs, running $Kafka$ mode (step-4). The OSNs validate and commit the transaction (step-5), then broadcast a message to all peers to update their local ledgers (step-6). In case the transaction is not successful, and the ledgers are not updated but the proposal is still logged for audit.

A built-in CA called *Fabric CA* is used to generate $X.509$ certificates and keys supporting Elliptic Curve (ECDSA); and to sign them (i.e., providing digital signatures). The Fabric CA server is initialised using Docker which hosts an HTTP server on the default port 7054 that offers REST APIs. All entities have to enrol and register with the CA server before participating the network. Once an entity is enrolled and

[13]https://github.com/nguyentb/Personal-data-management

[14]Channel is a terminology in HLF technically referring to a private blockchain overlays which offers data isolation and transaction confidentiality.

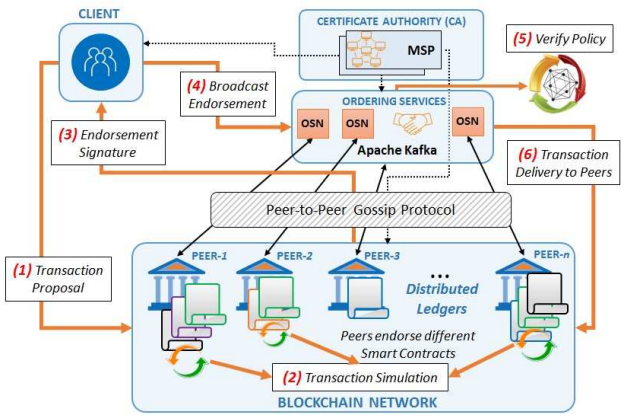

Fig. 5: High-level System Architecture and Transaction Flow of the HLF framework

registered, an enrolment certificate ($ECert$), corresponding private key and CA certificate are stored in PEM files in the subdirectories of the entity's directory. MSP is a configuration file identifying trusted CAs. The CAs then define members of a trust domain by either *(i)* listing identities of the members or *(ii)* identifying authorised CAs that issue valid identities for members. The latter is used in the demonstration.

B. Personal Profile Management Use-case

We consider a use-case that an SNS processing profile data stored in a separate RS. This RS follows the honest-but-curious model anticipating the BC as an HLF client and honestly executing required protocols (i.e., interacting with the BC network for token validation). For the purpose of complying with the GDPR, the SP participates in the proposed BC-based platform (Fig. 6). For the demonstrate the use-case, we build the RS as a profile management web-service based on REST architecture[15] for parties to process profile data through calling corresponding RESTful APIs. Profile information is stored in JSON-like documents using MongoDB[16], a document-oriented database system. The profile data model follows the Friend-Of-a-Friend (FOAF) ontology for describing person which is normally used in social networks[17]. Processing a profile includes $\{create, read, update, delete\}$ CRUD operations by making a request to a corresponding API provided by the RS.

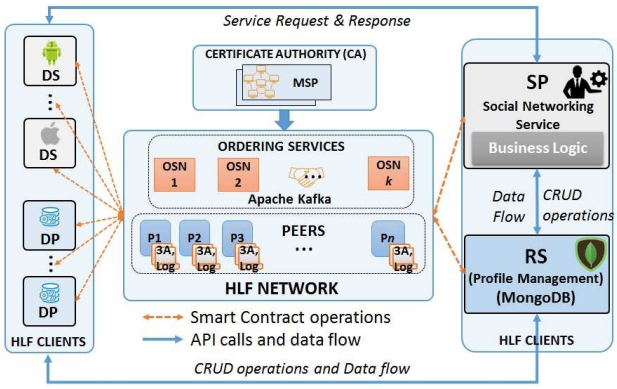

Fig. 6: System Architecture of a GDPR-compliant social networking service with the RS for personal profiles using HLF

[15]https://en.wikipedia.org/wiki/Representational_state_transfer
[16]https://www.mongodb.com/
[17]http://xmlns.com/foaf/spec/

A request to a RESTful API contains 6 parameters: $(1)API\text{-}Endpoint$, $(2)REST\text{-}Endpoint$, $(3)Method$, $(4)Header$, $(5)Params$, $(6)Payload$ in which the first fours are required. A RESTful request is as follows:

```
POST localhost:8080/ProfileManagement
-H 'Content-Type:application/json'
pubkey=pk&
signature=t&
token=access_token
&operation=read
```

where $Method$ is $POST$, $REST\text{-}Endpoint$ is $localhost$:8080, $API\text{-}Endpoint$ is $/ProfileManagement$, Header is $Content\text{-}Type$:$application/json$ following by $Params$ including the public-key pk with the signature t, the $access\_token$, and the requested $Read$ operation.

C. Smart Contracts Implementation

There are two chaincodes implemented in the HLF network: *(i)* the $3A\_cc$ for authentication, authorisation and access control, operating with the $3A\_ledger$; and *(ii)* the log_cc for access validation and logging, operating with the log_ledger. Theoretically, a contract can be written in any programming language; and in the demonstration Go language is used. The two chaincodes inherit the built-in $shim$ package[18], which provides a variety of APIs to interact with distributed ledgers such as accessing state variables, transaction context and call other chaincodes. As Fabric CA adopts a traditional Public Key Infrastructure (PKI) hierarchical model, an client ID (which is a $X.509$ certificate used as the identifier) is only guaranteed to be unique within a MSP. Therefore, an IdM for the deployment based on HLF is necessarily designed insuring an identity is unique across the HLF network. A simple solution used in the demonstration is that we concatenate the $X.509$ certificate with the MSP identifier to form a client ID.

$$ClientID = mspID + X.509Certificate \tag{7}$$

Specifically, the IdM solution is implemented utilising the client identity chaincode library cid[19] in HLF which is illustrated by the pseudo-code as follows:

```
function ClientID(stub shim.ccAPI) ci *clientID {
    hlfId = ci.New(stub);
    mspID = hlf.GetMSPID();
    cert = hlf.GetX509Certificate();
    return &clientID{mspID, cert};
}
```

Listing 3: Definition of a global identity for HLF client from $mspID$ and $X.509$ certificates utilising the cid library in HLF

Regarding the distributed ledgers, $en\_pointer$ is the ciphertext of an identifier of a data object (i.e., $profile.ID$) using the encryption function $\mathcal{E}$ with the encryption key pk_{enc}:

$$en\_pointer = \mathcal{E}(pk_{enc}, profile.ID) \tag{8}$$

A party who permitted access a profile has a shared private key sk_enc to decrypt $en\_pointer$ in order to obtain the

[18]https://godoc.org/github.com/hyperledger/fabric/core/chaincode/shim
[19]https://github.com/hyperledger/fabric/blob/release-1.1/core/chaincode/lib/cid/README.md

$profile.ID$, which is then passed as a parameter for a RESTful API to access the desired profile information:

$$profile.ID = (sk_{enc}, en\_pointer) \tag{9}$$

The $policy$ in the $3A\_ledger$ is simply defined as an access control list matching each of the CRUD operation to a list of granted parties as follows:

```
"policy" {
    "Create": {pk_DS, pk_DC, ..},
    "Read": {pk_DS, pk_DC, pk_DP1, pk_DP2, ..},
    "Update": {pk_DS, pk_DC, pk_DP3 ...},
    "Delete": {pk_DS, pk_DC, pk_DP3, pk_DP4, ..}
}
```

Listing 4: Data Usage Policy defined as an Access Control List

Based on the identity scheme and detailed information for the two ledgers, core functions in personal data management such as $GrantConsent$, $RevokeConsent$, $TokenValidation$ and $DataAccess$ are then implemented exactly following the algorithms described in Section III.D.

VI. Analysis and Discussion

This section provides analysis and discussion on the platform deployed in Section V, including GDPR-compliance applicability, threat models and system performance.

A. Trust Assumption

Besides an honest-but-curious RS, a must assumption is that a large portion of peer nodes in the HLF network are honest. Technically, HLF v1.x offers multiple ordering techniques including a variety of BFT-based approaches such as PBFT and Simplified BFT. Such BFT-variant protocols are able to conditionally tolerate $\lfloor \frac{N-1}{5} \rfloor$ (e.g., in Ripple [44]) to $\lfloor \frac{N-1}{2} \rfloor$ (e.g., in crash-fault tolerance) simultaneously faulty nodes. However, such BFT-variants only guarantee consistency despite any number of crash-faulty or partitioned replicas, with at most $\lfloor \frac{N-1}{3} \rfloor$ faulty nodes [45]. Unfortunately, such protocols are under development for the HLF framework, only Apache Kafka is provided as a reference implementation, which supports some levels of fault-tolerant (e.g., crash-faulty) but not BFT failures.

The cryptographic primitives (i.e., digital signature schemes) are assumed to be secure. As HLF is used in the platform, the built-in PKI and the Fabric CA, which are responsible for the distribution of management of digital certificates, are assumed to be secure and honest. Regarding key management, we assume that private keys obtained by the key generator $\mathcal{G}$ are effectively protected from adversaries. As the personal data management platform is built on top of the HLF framework, existing solutions in enterprise systems can be readily integrated. However, this is the weak assumption and is considered as a security threat in the next section.

B. GDPR-Compliance

From an applicability perspective, the proposed platform provides SPs (e.g., the SNS) mechanisms to fully comply with the GDPR regulations. This is due to the following reasons:

1) Full Control back to Data Owners: As following the design concept, the platform provides DSs:

- "Right of access" and "right of rectification": This is because DS is eligible to do all CRUD operations to her personal data as specified in the default policy when ledgers are initialised; and no one can change these rights.
- "Right of restricted processing" and "right of data portability": This is because DSs have full permissions to manage data usage policy (e.g., to grant or revoke consent anytime/anywhere by invoking the $GrantConsent$ and $RevokeConsent$ functions in the $3A\_cc$).
- "Right to be informed": This is because the platform always requires DS's signature for data collection or for granting consent.
- "Right to be forgotten": As personal data is stored off-chain, a RS is able to erase the data as requested from DS. However, a question is posed when leveraging BC for personal data management: "whether a BC platform complies with the GDPR as distributed ledgers are immutable; meaning that the ledgers, theoretically, will never be erased?". Therefore, if a piece of personal information is recorded in a ledger, the platform will violate the "right of forgotten". In the design concept, sensitive information is encrypted before writing into a ledger (e.g., $data\_pointer$). The "right of forgotten" is then ensured by throwing decryption keys. Whether this remedy fully satisfies the GDPR is still an open question [42], [46].

2) Security, Transparency and Accountability: By following the design concept, the platform insures that:

- Security of the identity, authentication and authorisation mechanisms, which depends on the security of the cryptographic primitives, is assumed to be secure.
- Operations (e.g., grant or revoke a consent, update usage policy, verify access token, and CRUD) are authenticated, authorised and autonomously executed only by invoking corresponding SCs deployed in the HLF network. This ensures system procedures are executed in a transparent and not compromised by any individuals.
- Information about management operations and CRUD activities on personal data, including who/what/when/why/ and how, are immutably recorded in the log_ledger.

Consequently, the proposed platform forces SPs, who participate in the system, to be responsible for complying with the GDPR; otherwise any unauthorised or malicious transactions initiated by a corresponding SP can be always figured out. Furthermore, the investigation for GDPR-compliance is empowered as all activities logged in the ledgers can be traced back. The signalling of a non-compliant activity could trigger official investigation and auditing of a SP by a SA. The decisions could be made based on whether a malicious activity recorded in the log_ledger exists that respects the associated data usage policies in the $3A\_ledger$. In this regard, the two distributed ledgers can be considered as legal grounds for the GDPR compliance. As a result, the platform is able to demonstrate the GDPR compliance. Therefore, the proposed BC-based platform provides efficient measures to meet the requirements of data accountability. For those reasons, the

SNS provider, which utilises the platform for its personal data management tasks, fully complies with the GDPR.

C. Threat Models

The advanced capability of the BC framework plays a key role in providing a secure and trustworthy platform for complying with the GDPR. However, certain aspects of the contemporary BC and SC technologies present limitations imposing threats resulting in non-compliance with the GDPR.

1) Security Threats: Given the aforementioned assumptions, the decentralised nature of the BC ensures that an adversary cannot corrupt the BC network to unauthorisedly change the content of the ledgers as that would imply majority of the network's resources are compromised. Also, the adversary cannot impersonate an authorised party as a digital signature cannot be forged. Security threats are, thus, from two sources: *(i)* an internal malicious party acting in a Byzantine way, who has been granted to access personal data; and *(ii)* an honest party whom both private key and decryption key sk_{enc} are disclosed to an external adversary; thus, the adversary could pose itself as the party. In such scenarios, the $TokenValidation$ function is of paramount importance since it plays as a role of a gatekeeper to reassure that any $access\_token$ expires after amount of time and needs to be refreshed (i.e., re-authenticated and re-authorised). As a result, the $TokenValidation$ mitigates the risk of a long-lived $access\_token$ leaking, similar to the use of both $access\_token$ and $refresh\_token$ used in the standard OAuth2 specification[20].

Admittedly, it is inevitable that at an adversary is able to access the data in the time-frame window of the $access\_token$ (defined by the $expires\_in$ parameter in the log_ledger). During this period, it is unachievable to prevent the adversary from accessing data unless the security breach is detected. Once being detected, DS is able to revoke the consent by updating the ledgers to remove all permissions related to the adversary. The remedy is straight-forward in case of the first scenario - the party is malicious. However, it turns to a complex situation when an honest party leaks its private key to the adversary. This party is never able to get granted again as its identity is compromised, which is unreasonable. A key management with an account recovery scheme could be an applicable solution to deal with this situation although it is expected to be much complicated to integrate the recovery scheme with a BC system [47]. Another security threat comes from poor quality code in SCs which exposes vulnerabilities to be exploited. For example, an attacker stole 3.6M Ether (worth $50M at that time) in DAO[21] attack exploiting a concurrency bug in DAO's SCs. As a BC framework supporting Turing-complete SCs, software bugs are painful to avoid. Thus, SCs must be written in high quality standards and follows strict security specifications [17], [48].

2) Privacy Threats: The openness of distributed ledgers, which allows parties to inspect, violates the idea of privacy. Even in a permissioned BC in which transactions take place between authenticated parties, some privacy threats still remain as any participants could be malicious. In the proposed design concept, measures to tackle privacy leakage are to both: *(i)* constitute anonymity for parties' identities using public and private key-pairs in transactions; and *(ii)* encrypt sensitive information recorded in the ledgers. The first measure indeed provides pseudo-anonymity since there is possibility to link between public addresses with physical identification of the users by using variety of de-anonymisation techniques [49]. Literally, the risk of revealing real-world identity by an adversary can be significantly reduced in a permissioned BC compared to a public one thanks to an additional permission access control layer [14], [50]. As a trade-off, anonymity is sacrificed as it requires more identity materials for stringent privacy requirements. The second measure encrypts the $data\_pointer$ (i.e., $profile.ID$. Although a $profile.ID$ does not contain personal information, it is used as a parameter in API calls for accessing a personal dataset. Thus, it should only be visible to designated parties, reducing the risk of leaking the information to adversaries.

3) Performance and Scalability: As the proposed platform is expected to serve a large number of clients accessing data simultaneously, performance and scalability of the platform is necessarily evaluated. At the moment, public BCs can only achieve limited throughput (e.g., Bitcoin gets 7 transactions per second (tps) whereas Ethereum reaches around 15 tps[22]). In permissioned BCs, additional permission control ensures that a majority of nodes are trusted; this allows the use of BFT-variant consensus, theoretically resulting in higher throughput. For instance, FabricCoin deployed on top of the HLF framework can achieve more than $3,500$ tps at a second latency [50]. We use the BLOCKBENCH benchmarking framework for performance evaluation of various BC systems including HLF [51], [52]. Fig. 7 interprets performance and scalability of HLF ver.0.6 running PBFT consensus protocol.

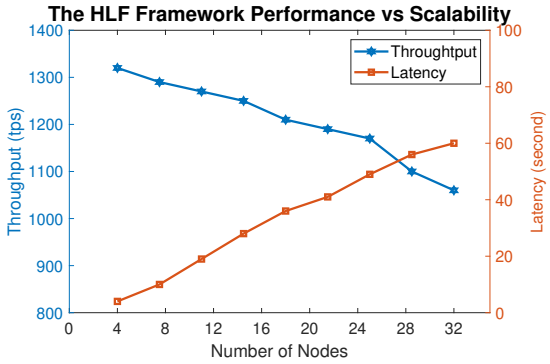

Fig. 7: Performance vs Scalability in the HLF framework with 10 clients intensively performing high workload.

The demonstration consists of 10 concurrent clients intensively incurring workload to the HLF system in 5-minute period, and a number of peer nodes varied from 4 to 32. Although the BLOCKBENCH framework has not been customised for our platform, we believe the results are relatively similar as it measures the performance of the HLF framework whereas overhead of an application built on top is neglected. As

[20]https://tools.ietf.org/html/rfc6749

[21]https://ethereum.org/dao

[22]https://blockchain.info/charts/n-transactions

depicted in Fig. 7, HLF fails to support high performance and scalability since the throughput significantly decreases and the latency dramatically increases when the BC network scales up. This is due to overhead messages exchanged between nodes, and the wait for endorsement messages before broadcasting a message to update a distributed ledger. At about $1000\ tps$ with 60-second latency for a setup of 8 concurrent clients and 32 peers, the system is far from usable real-world applications.

VII. Conclusion and the road ahead

In this article, a design concept for a GDPR-compliant BC-based personal data management platform is proposed. Following the guidelines from the design concept including system architecture, ledger data models, and SC functionalities, a BC-based platform is implemented on top of the HLF framework. The platform interplays among an honest RS, an SNS, DPs, and DSs ensuring that all processing activities over profile data stored in the RS are compliant with the GDPR. The feasibility and effectiveness of the design concept are, therefore, successfully demonstrated.

As a future work, performance evaluation of the HLF-based personal data management platform will be carried out utilising the BLOCKBENCH framework, dedicated to the HLF ver.1.x framework. The second task is to deploy the design concept in a public BC (e.g., Ethereum) with an RS using distributed storage (e.g., IPFS and Storj). In this regard, the RS is not trustworthy as some storage nodes might be malicious. Thus, more mechanisms need to be implemented to resolve the lack of a trusted centralised RS. As a reward, the system is truly decentralised. Another future work is to develop a fine-grain expressive data usage policy as in the demonstration, the simple access control list is used. A policy generator deployed in SCs that autonomously acquires data usage policy depending on specific contexts is also a promising research direction. Additionally, pricing and incentives models for the cost of data storage and BC operations should be carried out to finalise a complete system.

As the processing of personal data refers to CRUD operations which is under the mindset of data storage, an ambitious research direction is to provide computational capability on a BC network [32]. This means an SP directly runs computation on the network and obtain results using secure Multi-Party Computation (MPC)[23]. This approach is much securer as the SP does not directly observe raw data. We believe our work acts as a catalyst to open a variety of research directions regarding the use of BC and SCs in decentralised authorisation and access control, which plays a crucial role in digital assets management, particularly in personal data regulations.

Acknowledgment

This research was supported by the HNA Research Centre for Future Data Ecosystems at Imperial College London.

References

[1] M. Walport *et al.*, "Distributed ledger technology: Beyond blockchain," *UK Government Office for Science*, vol. 1, 2016.

[2] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[3] M. Crosby, P. Pattanayak, S. Verma, V. Kalyanaraman *et al.*, "Blockchain technology: Beyond bitcoin," *Applied Innovation*, vol. 2, no. 6-10, p. 71, 2016.

[4] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2084–2123, 2016.

[5] N. B. Truong, T.-W. Um, B. Zhou, and G. M. Lee, "Strengthening the blockchain-based internet of value with trust," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–7.

[6] V. Gramoli, "From blockchain consensus back to byzantine consensus," *Future Generation Computer Systems*, 2017.

[7] W. Wang, D. T. Hoang, P. Hu, Z. Xiong, D. Niyato, P. Wang, Y. Wen, and D. I. Kim, "A survey on consensus mechanisms and mining strategy management in blockchain networks," *IEEE Access*, 2019.

[8] A. Gervais, G. O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, and S. Capkun, "On the security and performance of proof of work blockchains," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. ACM, 2016, pp. 3–16.

[9] A. Kiayias, A. Russell, B. David, and R. Oliynykov, "Ouroboros: A provably secure proof-of-stake blockchain protocol," in *Annual International Cryptology Conference*. Springer, 2017, pp. 357–388.

[10] I. Bentov, C. Lee, A. Mizrahi, and M. Rosenfeld, "Proof of activity: Extending bitcoin's proof of work via proof of stake." *IACR Cryptology ePrint Archive*, vol. 2014, p. 452, 2014.

[11] A. Miller and J. J. LaViola Jr, "Anonymous byzantine consensus from moderately-hard puzzles: A model for bitcoin," *Available online: http://nakamotoinstitute.org/research/anonymous-byzantine-consensus*, 2014.

[12] Y. Gilad, R. Hemo, S. Micali, G. Vlachos, and N. Zeldovich, "Algorand: Scaling byzantine agreements for cryptocurrencies," in *Proceedings of the 26th Symposium on Operating Systems Principles*. ACM, 2017, pp. 51–68.

[13] V. Buterin, "White paper: A next-generation smart contract and decentralized application platform," *April. https://www. ethereum. org/pdfs/EthereumWhitePaper. pdf*, 2014.

[14] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 839–858.

[15] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, pp. 1–32, 2014.

[16] C. Cachin, "Architecture of the hyperledger blockchain fabric," in *Workshop on distributed cryptocurrencies and consensus ledgers*, vol. 310, 2016.

[17] H. A. WC. (2018) Hyperledger architecture-volume ii-smart contracts. [Online]. Available: https://www.hyperledger.org/wp-content/uploads/2018/04/Hyperledger_Arch_WG_Paper_2_SmartContracts.pdf

[18] K. Markus and G. Chung, "Blockchain in logistics," DHL Trend Research, Germany, 2018.

[19] F. Tian, "An agri-food supply chain traceability system for china based on rfid & blockchain technology," in *2016 13th international conference on service systems and service management (ICSSSM)*. IEEE, 2016, pp. 1–6.

[20] N. Hackius and M. Petersen, "Blockchain in logistics and supply chain: trick or treat?" in *Proceedings of the Hamburg International Conference of Logistics (HICL)*. epubli, 2017, pp. 3–18.

[21] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla, "Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability," in *Proceedings of the 17th IEEE/ACM international symposium on cluster, cloud and grid computing*. IEEE Press, 2017, pp. 468–477.

[22] M. Ali, J. Nelson, R. Shea, and M. J. Freedman, "Blockstack: A global naming and storage system secured by blockchains," in *Annual Technical Conference (USENIX/ATC-16)*, 2016, pp. 181–194.

[23] H. A. Kalodner, M. Carlsten, P. Ellenbogen, J. Bonneau, and A. Narayanan, "An empirical study of namecoin and lessons for decentralized namespace design." in *WEIS*. Citeseer, 2015.

[24] H. Shafagh, L. Burkhalter, A. Hithnawi, and S. Duquennoy, "Towards blockchain-based auditable storage and sharing of iot data," in *Proceedings of the 2017 on Cloud Computing Security Workshop*. ACM, 2017, pp. 45–50.

[25] R. Li, T. Song, B. Mei, H. Li, X. Cheng, and L. Sun, "Blockchain for large-scale internet of things data storage and protection," *IEEE Transactions on Services Computing*, 2018.

[26] T. McConaghy, R. Marques, A. Müller, D. De Jonghe, T. McConaghy, G. McMullen, R. Henderson, S. Bellemare, and A. Granzotto,

[23] https://en.wikipedia.org/wiki/Secure_multi-party_computation

"Bigchaindb: a scalable blockchain database," *White paper, BigChainDB*, 2016.
[27] J.-H. Lee, "Bidaas: Blockchain based id as a service," *IEEE Access*, vol. 6, pp. 2274–2278, 2018.
[28] Z. Chen, S. Chen, H. Xu, and B. Hu, "A security authentication scheme of 5g ultra-dense network based on block chain," *IEEE Access*, vol. 6, pp. 55 372–55 379, 2018.
[29] O. Novo, "Blockchain meets iot: An architecture for scalable access management in iot," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1184–1195, 2018.
[30] S. Wang, Y. Zhang, and Y. Zhang, "A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems," *IEEE Access*, vol. 6, pp. 38 437–38 450, 2018.
[31] J. Benet, "Ipfs-content addressed, versioned, p2p file system," *arXiv preprint arXiv:1407.3561*, 2014.
[32] G. Zyskind, O. Nathan *et al.*, "Decentralizing privacy: Using blockchain to protect personal data," in *2015 IEEE Security and Privacy Workshops*. IEEE, 2015, pp. 180–184.
[33] L. A. Linn and M. B. Koo, "Blockchain for health data and its potential use in health it and health care related research," in *ONC/NIST Use of Blockchain for Healthcare and Research Workshop. Gaithersburg, Maryland, United States: ONC/NIST*, 2016.
[34] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," in *2016 2nd International Conference on Open and Big Data (OBD)*. IEEE, 2016, pp. 25–30.
[35] M. J. M. Chowdhury, A. Colman, M. A. Kabir, J. Han, and P. Sarda, "Blockchain as a notarization service for data sharing with personal data store," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications(TrustCom)*. IEEE, 2018, pp. 1330–1335.
[36] R. Neisse, G. Steri, and I. Nai-Fovino, "A blockchain-based approach for data accountability and provenance tracking," in *Proceedings of the 12th International Conference on Availability, Reliability and Security*. ACM, 2017, p. 14.
[37] B. Faber, G. C. Michelet, N. Weidmann, R. R. Mukkamala, and R. Vatrapu, "Bpdims: A blockchain-based personal data and identity management system," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.
[38] C. Wirth and M. Kolain, "Privacy by blockchain design: a blockchain-enabled gdpr-compliant approach for handling personal data," in *Proceedings of 1st ERCIM Blockchain Workshop 2018*. European Society for Socially Embedded Technologies (EUSSET), 2018.
[39] D. Hardt, "The oauth 2.0 authorization framework," Tech. Rep., 2012.
[40] T. Lodderstedt, M. McGloin, and P. Hunt, "Oauth 2.0 threat model and security considerations," Tech. Rep., 2013.
[41] R. Neisse, G. Steri, I. N. Fovino, and G. Baldini, "Seckit: a model-based security toolkit for the internet of things," *computers & security*, vol. 54, pp. 60–76, 2015.
[42] M. Berberich and M. Steiner, "Blockchain technology and the gdpr-how to reconcile privacy and distributed ledgers," *European Data Protection Law Review*, vol. 2, no. 422, 2016.
[43] S. Wilkinson, T. Boshevski, J. Brandoff, and V. Buterin, "Storj a peer-to-peer cloud storage network," 2014.
[44] D. Schwartz, N. Youngs, A. Britto *et al.*, "The ripple protocol consensus algorithm," *Ripple Labs Inc White Paper*, vol. 5, 2014.
[45] S. Liu, P. Viotti, C. Cachin, V. Quéma, and M. Vukolić, "{XFT}: Practical fault tolerance beyond crashes," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 485–500.
[46] C. R. Meijer. (2018) Blockchain versus gdpr and who should adjust most. [Online]. Available: https://www.finextra.com/blogposting/16102/blockchain-versus-gdpr-and-who-should-adjust-most
[47] H. Zhao, P. Bai, Y. Peng, and R. Xu, "Efficient key management scheme for health blockchain," *CAAI Transactions on Intelligence Technology*, vol. 3, no. 2, pp. 114–118, 2018.
[48] L. Luu, D.-H. Chu, H. Olickel, P. Saxena, and A. Hobor, "Making smart contracts smarter," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 254–269.
[49] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A fistful of bitcoins: characterizing payments among men with no names," in *Proceedings of the 2013 conference on Internet measurement conference*. ACM, 2013, pp. 127–140.
[50] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich *et al.*, "Hyperledger fabric: a distributed operating system for permissioned blockchains," in *Proceedings of the Thirteenth EuroSys Conference*. ACM, 2018, p. 30.
[51] T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi, and J. Wang, "Untangling blockchain: A data processing view of blockchain systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1366–1385, 2018.
[52] T. T. A. Dinh, J. Wang, G. Chen, R. Liu, B. C. Ooi, and K.-L. Tan, "Blockbench: A framework for analyzing private blockchains," in *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 2017, pp. 1085–1100.

Nguyen Binh Truong Dr. Nguyen B.Truong is currently a Research Associate at Data Science Institute, Department of Computing, Imperial College London, United Kingdom. He received his PhD. and Master degrees from Liverpool John Moores University, United Kingdom and Pohang University of Science and Technology, Korea in 2018 and 2013, respectively. He was a Software Engineer at DASAN Networks, a leading company on Networking Products and Services in South Korea from 2012 to 2015. His research interest is including, but not limited to, Security, Privacy and Trust for IoT, Blockchain, Personal Data Management, Fog, Edge and Cloud Computing.

Kai Sun Dr. Kai Sun received the BEng degrees in Computer Science from Harbin Institute of Technology and the University of Birmingham in 2009. She received the MSc degree and the PhD degree in Computing from Imperial College London, in 2010 and 2014, respectively. From 2014 to 2017, she was a Research Associate at the Data Science Institute at Imperial College London. She is currently the lab manager of the HNA Centre of Future Data Ecosystem. Her research interests include translational research management, network analysis and decentralised systems.

Gyu Myoung Lee Dr. Gyu Myoung Lee received his BS degree from Hong Ik University and MS, and PhD degrees from the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1999, 2000 and 2007, respectively. He is currently a Reader at Department of Computer Science, Liverpool John Moores University, UK. He is also with KAIST as an adjunct professor. His research interests include Future Networks, IoT, and multimedia services. He has actively contributed to standardization in ITU-T as a Rapporteur, oneM2M and IETF. He is chair of the ITU-T Focus Group on data processing and management to support IoT and Smart Cities & Communities.

Yike Guo Dr. Yike Guo (FREng, MAE) received the BSc degree in Computing Science from Tsinghua University, China, in 1985 and received the PhD in Computational Logic from Imperial College London in 1993. He is a Professor of Computing Science in the Department of Computing at Imperial College London, as well as the founding Director of the Data Science Institute at Imperial College. He is a fellow of the Royal Academy of Engineering. His research interests are in the areas of data mining for large-scale scientific applications including distributed data mining methods, machine learning and informatics systems.