Ensuring Personal Data Anonymity in Data Marketplaces through Sensing-as-a-Service and Distributed Ledger Technologies

Mirko Zichichi[1,2,3], Michele Contu[2], Stefano Ferretti[2], and Víctor Rodríguez-Doncel[3]

[1] Law, Science and Technology Joint Doctorate - Rights of Internet of Everything
mirko.zichichi@upm.es
[2] Department of Computer Science and Engineering, University of Bologna, Italy
michele.contu@studio.unibo.it, s.ferretti@unibo.it
[3] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
vrodriguez@fi.upm.es

Abstract

Personal data has undoubtedly assumed a great value with the advancements on technologies able to gather it and infer from it. The businesses that operate in a data-driven economy offer services that rely on data collected about their users and usually they store this personal information in "silos" that impede transparency on their use and possibilities of easy interactions. The introduction in EU of the General Data Protection Regulation (GDPR) moves this economy towards a user-centered vision, in which individuals have rights for their data sovereignty and the free portability of it. However, more efforts are needed to reach both transparency and balance between privacy and data sharing. In this paper, we present a solution to promote the development of personal data marketplaces, exploiting the use of Distributed Ledger Technologies (DLTs) and a Sensing-as-a-Service (SaaS) model, in order to enhance the privacy of individuals, following the principles of personal data sovereignty and interoperability. Moreover, we provide experimental results of an implementation based on IOTA, a promising DLT for managing and transacting IoT data.

1 Introduction

The importance of data in our life is not a news anymore. Although this situation has clearly brought enormous benefits to our society, the development of new advanced techniques for managing data and inferring new information has also had a significant impact on individuals' privacy. With the growth of the Internet of Things (IoT) and the ubiquitous connectivity with mobile phones, the information that is possible to collect on behalf of individuals will become always more detailed. This includes demographic data, medical data, tweets, emails, photos, videos, as well as location information. Such personal data is defined, indeed, as the piece of information that can identify or be identifiable to a natural person. Smart cities, smart transportation systems, smart health, smart agrifood and all remaining smart environments have in common the ability to transform the data into meaningful information needed by the liveness of the ecosystem they generate, in order to provide services that are becoming more and more targeted towards individuals. It is commonly known that this information is used to recommend opportunities to individuals and make their life easier, but entities that control this

data do not always operate with the aim of social good. Indeed, world largest companies make consistent profits operating in this data-driven economy and this is now public aware due to the frequent scandals on the abuse of personal data, such as the Facebook-Cambridge Analytica revelations[1] or the most recent Google's Nightingale revelations[2].

Many internet businesses rely on data collected about their users, usually storing this personal information in corporate databases and transacting it to other parties with no transparency for individuals. The privacy of European Union's citizen has been empowered in the process of renewing a series of rights regarding their data protection and portability through the General Data Protection Regulation (GDPR) [6]. This is a response to these businesses that collect and control individuals' personal data, i.e. data controller, that are usually in contrast with the process of free data portability and with the possibility of economic exploitation that individuals should have. IoT devices vendor, social network sites, internet service providers are some examples of data controllers that are driving towards a condition in which individuals are simple source of data, concentrating the entire decision-making and operational power over personal data on them.

To ensure to individuals the sovereignty of their personal data and the possibility of an appropriate data interoperability, hence moving towards the use of personal data for open data markets and for social good, we use the personal databox model [7]. The databox is a data store model that acts as a virtual boundary, where individuals can control how, when and what data is shared with external parties [21]. In the context of GDPR, databoxes suit with the right of individuals to see the data collected about them and the right to transfer data to other service providers. There is, indeed, an economic incentive for data controllers to provide access to personal data to their users, since combining their own data with other sources in the databox makes the data much more valuable.

Meanwhile, between the many technologies that regard general-purpose data management and storage, Distributed Ledger Technologies (DLTs) are rising as powerful tools to avoid the control centralization. The current use of DLTs is in financial (i.e. cryptocurrencies) and data sharing scenarios. In both cases, there are several parties that concur in handling some data, there is no complete trust among parties and often these ones compete to the data access/ownership. Such features suit perfectly with the process of moving the data sovereignty towards users and releasing them more influence over access control, while allowing anyone else to be able to consume this data with transparency. This can be made possible through smart contracts, the new concept of contract that brought a second blockchain revolution.

In this work, we present a possible high-level solution for the personal data marketplace involving the use of DLTs and Semantic Web technologies for the design of a GDPR compliant databox, then we focus on describing the process of personal data anonymization through Sensing-as-a-Service (SaaS). Indeed, privacy assumes the main role among all the challenges presented by the personal data marketplace, as it represents the biggest concern for individuals [14] and recent data abuse scandals can only increase this. Since, for most service providers, gathering aggregated data from a single source (or a few) may result more worthwhile and less expensive, than gathering "raw" data from multiple sources, the use of a SaaS [22] model can be a sustainable, scalable and powerful solution for the creation of a personal data marketplace. The SaaS model allows utilising resources efficiently, so limited resource can be used to accommodate large numbers of consumers. With the support of anonymization techniques, the SaaS model allows data subjects (i.e. the individuals whose personal data is being collected, held or

[1] https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

[2] https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790

processed) to control and protect their privacy. For instance, k-anonymity (shown in section 2.2) can be applied to anonymize such data [25].

The remainder of this paper is organized as follows. Section 2 provides the background needed for personal data management, data protection techniques and DLTs used to perform the experimental study. Section 3 describes the personal data market model considered to perform the study. Section 4 we present the details of the sensing service architecture. In Section 5 we provide results of the experimental evaluation we conducted. Finally, Section 6 provides some concluding remarks.

2 Background and Related Work

In this section we will review some background aspects that are needed regarding personal data management, data protection methods and DLTs and some related works.

2.1 Personal Data Management and Interoperability

The databox must be conceived as a concept that describes a set of storing and access control technologies that allows a user to have direct access to his data and to share this through the definition of some policies and preferences (that comply with GDPR in the case of personal data). In [5, 7] the databox is defined as a platform that provides means for individuals to *reflect* on their online presence, restore *agency* over their data, and enable a process of *negotiation* with other parties concerning personal data. It must be a *trusted platform*, providing facilities for *data management* of data at rest for the data subjects as well as *controlled access* by other parties wishing to use their data, and *supporting incentives* for all parties. The barriers of adoption of such platform are i) the trust in the infrastructure itself, both by users and data controllers; ii) the system usability, since dealing with all type of data and user policies is a complex matter (e.g. users cannot be overloaded by information); iii) the costs of maintaining such platform.

The databox model is largely symbolic at the time of writing, but is not a theoretical model. An undirected link to this model, that puts in practice the concept of data sovereignty through this model, is the Solid project [26]. Led by the creator of the Web Tim Berners-Lee, the project was born with the purpose of giving users their data sovereignty, letting them choose where their data resides and who is allowed to access and reuse it. Solid involves the use of distributed technologies and Semantic Web integration in social networks. Semantic Web technologies are used to decouple user data from the applications that use this data. Data is, indeed, stored in an online storage space called Pod, a Web-accessible storage service, which can either be deployed on personal servers or on public servers.

The storage itself can be conceived in different manner, while the use of Semantic Web represents to us the core element that eases data interoperability and favours reasoning over individuals' policies. Semantic Web standards bring structure to the meaningful contents of the Web by promoting common data formats and exchange protocols. The RDF (Resource Description Framework) is the most diffused paradigm to represent information. It consists in resources identified by URIs and described with collections of triples. OWL (Web Ontology Language) is used to formally establish the precise meaning of each resource. An ontology is a formal representation of knowledge through a set of concepts and a set of relations between these concepts, within a specific domain. The advantages consist in the fact that these ontologies are recommended by the W3C and thus universally understood, and that reasoning with the

information represented using these data models is easy because they are mapped in a formal language.

2.2 Data Protection Techniques

GDPR requires data controllers to put in place appropriate technical and organizational measures to implement the data protection principles and to use the highest-possible privacy settings by default. Datasets must not be publicly available without explicit, informed consent, and cannot be used to identify a subject without additional information. The confidential storage of information about users, their identity protection and the untraceability of their actions are not trivial operation to be accomplished. Additionally, the correlation of different information sources can lead to leakage of information and de-anonymization [20, 27]. In literature, data protection techniques have been broadly classified in two main categories [9]:

- Syntactic - techniques aimed at satisfying a syntactic privacy requirement, e.g. each release of data must be indistinguishably related to no less than a certain number of individuals in the population.
- Semantic - techniques aimed at satisfying a property that must be satisfied by the mechanism chosen for releasing the data, e.g. the result of an analysis carried out on a released dataset must be insensitive to the insertion or deletion of a tuple in the dataset.

The k-anonymity proposal was introduced in [25] and it is considered one of the most popular syntactic privacy definitions, developed for protecting a released dataset against identity disclosure. A data release is said to have the k-anonymity property if the information for each individual contained in the release cannot be distinguished from at least $k-1$ individuals, whose information also appear in the release. Generally, k-anonymity can be achieved applying two operations over data that can directly or indirectly (by linking it with external information) identify individuals: generalization and suppression. Generalization consists in substituting the original values with more general values, for instance the date of birth can be generalized by removing the day and the month of birth, while suppression consists in completely remove the information. k-anonymity is studied in different contexts, that comprehend personal data; surely, location privacy is one among the most concerning topics [2].

Differential privacy [12] is a semantic definition where the protection of individuals' privacy is achieved by releasing a dataset by which recipients learn properties about the population as a whole. Differential privacy is traditionally enforced by adding noise to the released data, typically using the *Laplace distribution*. Obviously, knowing the noise distribution allows to compensate it when analyzing the released dataset, thus enabling to the retrieve the complete data. Differential privacy is used in different research areas and also by companies such as Apple[3] and Google[4].

2.3 Distributed Ledger Technologies

In this section, we review some of the main characteristics of DLTs, with specific focus on smart contracts and IOTA. A DLT is a software infrastructure maintained by a peer-to-peer network, where the network participants must reach a consensus on the states of transactions submitted to the distributed ledger, to make the transactions valid. A distributed ledger is often described as decentralized because it is replicated across many network participants,

[3] https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
[4] https://europe.googleblog.com/2015/11/tackling-urban-mobility-with-technology.html

each of whom collaborate in its maintenance and in the high availability of the system. The decentralization and collaboration are powerful attributes that mirror the way entities and organizations exchange goods and services in the real world. The information recorded in a DLT, indeed, is append-only and cannot be modified, hence allowing every operation to be transparent.

DLTs were firstly introduced to support cryptocurrencies in a form where the ledger is a linked list of blocks, i.e. blockchain. Among the others, the Ethereum blockchain has brought a new concept of decentralized computation in the form of smart contracts. These contracts provide a new paradigm where unmodifiable instructions are executed in an unambiguous manner during a transaction between two (or more) parts. In this work we focus our studies on smart contracts and IOTA [24], a DLT specifically designed for working in the IoT landscape.

2.3.1 Smart Contracts and DAOs

Smart contract is a new paradigm of contract that does not completely embodies the same features of a legal contract, but can act as a self-managed structure able to execute code that forces agreements between two or more parts. These are fundamental components of Ethereum, that reside on the blockchain and are triggered by specific transactions [13]. Moreover, smart contracts can communicate with other contracts and even create new ones. The use of these contracts grants to build Decentralized Applications (dApps) and Decentralized Autonomous Organizations (DAOs) [30].

A DAO is a virtual entity managed by a set of interconnected smart contracts, where various actors maintain the organization state by a consensus system and are able to implement transactions, currency flows, rules and rights within the organization. Members of a DAO are able to propose options for decision in the organization and also discuss and vote those through transparent mechanisms.

2.3.2 IOTA

IOTA is a lightweight, permissionless DLT that aims to solve the problems of scalability, control centralization and post-quantum security issues, not always addressed in other DLTs. IOTA nodes (called full nodes when they maintain the whole ledger) are organized as a peer-to-peer overlay, in which monetary transactions are made and message are exchanged.

The IOTA ledger is structured as a Direct Acyclical Graph (DAG) called the Tangle [24], where graph vertices represent transactions and edges represent approvals: to issue a new transaction it is necessary to approve two previous tip transactions. A transaction is called tip when it has not been approved yet. The process of selecting two random tip transactions from the ledger is termed "tip selection". Finally, in order to attach a novel transaction to the Tangle, a node must perform a Proof of Work (PoW), i.e. a computation to obtain a piece of data which satisfies certain requirements and which is difficult (costly and time-consuming) to produce but easy for others to verify. The purpose of PoW is to deter denial of service attacks and other service abuses. IOTA is an interesting choice to support data management services because it has been designed to offer fast validation without requiring any fee. Thanks to these features IOTA is not only able to provide transactions between two parties, but also messaging mechanisms.

Masked Authenticated Messaging (MAM) is a second layer data communication protocol, which adds functionality to emit and access an encrypted data stream over the Tangle. MAM channels take the form of a linked list of transactions ordered in chronological order, i.e. a transaction points to the next one. Only the channel owner, i.e. the one who maintains the

private key used to sign each transaction, is able to publish encrypted messages in the form of transactions. Any encryption key possessor can subscribe to the channel and access to messages. In other words, MAM enables users to subscribe and follow a stream of data, generated by some devices. The data access to new data may be revoked simply by using a new encryption key.

2.3.3 Related Work with DLTs

The use of DLTs to manage data has been deeply studied in literature, even including various proposals focused on personal data and GDPR, but in this case implementations are rare [28]. For what concern privacy, the degree of trust that DLTs may provide in developing technological innovations is backed up by the security and privacy properties provided by design [29]. For these reasons, related works combine DLTs (mainly to monitor access control) and off-chain storage solutions to provide privacy in data [19, 32]. Many proposals also combine the use of privacy properties such as k-anonymity with DLTs, especially in the context of Location Privacy [17].

3 Personal Data Marketplace

Data generated by IoT devices or by third parties systems on behalf of individuals are often private in nature, but sharing them can be beneficial in terms of economic profit and social good. The challenge is to allow the access to it under the conditions that data subjects find acceptable and under law regulations (GDPR). Digital data marketplaces allow connecting data providers and consumers, ensuring high quality, consistency, and security. Data providers are recognized as owners in the marketplace and receive benefits (mostly economical) from the data sets or streams they supply. Consumers pay for data they acquire and may provide new information back to the marketplace. A marketplace could be created in a way such that it automates the negotiation data subjects and consumers, providing advantages for both parties.

Our vision of this personal data marketplace is based on the use of a databox that supports the right of individuals to the protection of their personal data, data interoperability, economic exploitation and social good. This vision follows some principles that can be achieved using specific technologies: i) the system stores and transacts personal data in a controlled, transparent and non-centralized manner and, for this purpose, the use of DLTs would grant data validation, access control, no central point of failure, immutability, and most importantly traceability; ii) in order to favour personal data interoperability, the use of data models that adapt the W3C specifications for the semantic web allow every party to operate in the same environment; iii) to let the individual define both high-level goals and fined-grained preferences for what regards the access to his data, smart contracts can be used to represent and reason with policies.

In this work, we use SaaS to provide a model that is able to enhance the privacy of data subjects following the principles mentioned above. SaaS has been introduced as a solution based on IoT infrastructures, with the capability to address the challenges in smart environments [22]. It is proposed as a service that enhances the concept of local mediator [8] to a more broad scope. SaaS is usually implemented as a middleware that aggregates data coming from multiple sources, following specific policies [11, 1]. In our work, we focus on the privacy policy, and in particular on anonymization.

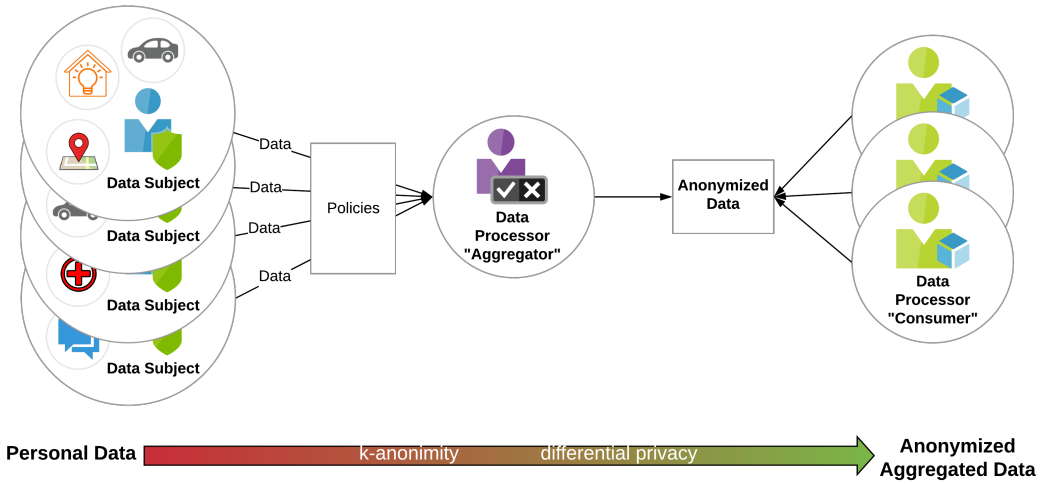

Figure 1: SaaS in Personal Data Marketplace

3.1 Anonymizing Data by Aggregation

In the marketplace we envision, data subjects maintain their personal data in databoxes and register data they want to sell along with standardised descriptions of what they measure, i.e. ontologies. On the other end, data consumers query the system to find data their application needs, for instance companies, research institutions, marketing agencies and similar. Personal data may be produced by devices such as smartphones, smart vehicles, smart watches, smart house appliances and similar or may be coming from third parties that interact with the databox, such as social network sites data. It is important to stress the definition of data processor: an entity that deals with personal data as instructed by a controller for specific purposes and services offered to the controller that involve personal data processing. In this case the controller corresponds to the data subject and the processor corresponds to the consumer.

In our model, another entity takes the role of data processor, standing between providers and consumers: the "aggregator" fulfils the sensing service, gathering data from individual data subjects and producing anonymized aggregated data, ready to be acquired by consumers. Figure 1 helps to explain the process. Data subjects provide data to the aggregator and both parties agree on predefined policies regarding their rights and obligations. Once the aggregator has received the access to data coming from k subjects, he is able to perform the data aggregation producing a dataset that presents properties of k-anonymity and differential privacy. This dataset can be then acquired by data consumers in a process where every participant to the dataset creation is rightfully rewarded.

This model is specifically thought to protect the privacy of data subjects and produce anonymized data, but the advantages of this service are many. For example, it may be difficult for a consumer to gather large quantities of the same kind of data if the process consists in acquiring it from several independent sources. And of course another advantage is that acquiring a unique aggregated data is usually cheaper than acquiring data from multiple sources. For the same reasons, a data subject among many others may never have the opportunity to stand out from the crowd in the market and it is more convenient for him to group with others, since the

probability of selling his data increments by doing this.

4 Sensing Service Architecture

A personal data market can be modeled using the infrastructure proposed by Zichichi et al. [31]. This work focuses on data produced in Smart Transportation Systems, but the same principles can be applied to any type of personal data. An infrastructure based on decentralized storage, such as IPFS [4], provides an almost complete availability, if well incentivized [3]. IOTA MAM channels can be used both for storage and for validation, in fact using a DLT in this context allows us to benefit from the properties of data traceability and immutability while maintaining privacy through encryption and also pseudonymization. On top of these, it is possible to have smart contracts that refers to specific parts of data and grant access control based on subjects preferences. Following this simple layer architecture is possible to create a market simply putting a price tag to data in smart contracts. Access to data consists in the process of releasing an encryption key and it is performed by Authentication Services, i.e. servers that maintain the keys and provide them to consumers following the directives of smart contracts [19]. These services make the infrastructure "less decentralized" but are currently required to transact secrets. A possible workaround that increments data subjects' privacy consists in the use of *dynamic threshold encryption* [15].

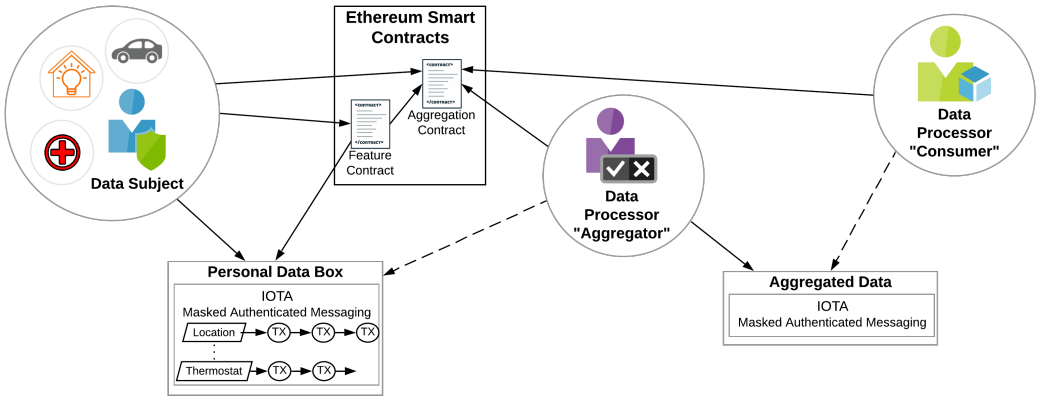

Figure 2: Smart Contracts and MAM Channels Architecture

Figure 2 shows the interaction between the actors in the sensing service architecture. Data subjects maintain personal data in a databox, implemented as MAM channels, that contains data or refers to data in IPFS. A "Feature Contract" is a smart contract owned by the data subject that points to a particular kind of data that it is possible to invoke to get access permissions. It indicates who has the right to access to a specific kind of data or a single datum itself and the history of these accesses. Access to data in IOTA (dashed arrow) is provided by Authentication Services that check on the related smart contract. The feature contract may refer to another smart contract owned by the aggregator and at the center of our architecture, the "Aggregation Contract". This smart contract indicates, but most importantly *forces* the policies that constitute the sensing service. In is also used by data consumers to acquire, i.e. get access to, the anonymized aggregated data. Its functionality is described in the next subsection.

4.1 Smart Contracts

In the context of data protection regulations (GDPR), the interesting aspect of smart contracts is that an algorithm executed in a decentralized manner allows two parties, e.g. data subject and processor, to reach an agreement in the process of the data flow. In this work we show how the combination of two kind of contracts leads to an increment of privacy, leaving traces of all operations and giving incentives to all the actors to correctly behave.

4.1.1 Feature Contract

This contract, already presented in detail in [31], allows its owner to indicate an Ethereum account or to reference another contract simply by using an address. It mainly consists in an Access Control List (ACL) that associates an Ethereum address to a bundle of data, i.e. addresses of IOTA transactions that contain a single datum or entire MAM channels. This is made possible in order to provide access to any portions of data managed by the data subject. In this particular instance, the contract owner, i.e. data subject, that is interested the sensing service indicates that part of his data (depending on which task the aggregator is performing) is accessible by the aggregator, referencing the Aggregation contract.

4.1.2 Aggregation Contract

The aggregation contract is at the core of the model and embodies the SaaS concept. It is owned by an aggregator that has the incentive to behave as the protocol imposes to him. The control over his action is taken by the DAO composed by the data subjects that participate to the sensing. The anonymized aggregated data that results from the sensing operation is, then, sold (i.e. put in the marketplace) through this contract, leaving traces of the operations carried out and repaying rightfully every participant. The operations (that take the form of contract methods) executed in this contract are, in order:

1. **Call for Data** - Initially, the operation for a single service session starts with a "Call for Data" where the aggregator indicates the kind of data he is interested in. This operation could be also requested by someone else to the aggregator. This step shows the importance of using the same standards and protocols, possibly based on the exchange of messages with well-defined semantics, e.g., RDF data supported by OWL ontologies.
    - Subjects interested in participating must provide a *Proof of Sensing* (PoSen). It consists in giving to the aggregator access to some of their data, e.g. two MAM messages in a channel. This is needed by the aggregator to determine the pertinence and quality of subject's data.
    - The call duration can depend on a prefixed time of closing or on the number of participants
2. **k-DAO formation** - The call is successfully closed only when $k \geq m$ data subjects that took part had been selected by the aggregator (using PoSen as discriminator). Then, a k-DAO is formed where the members are the k data subject chosen.
    - m is the minimum number of participant required to provide "reasonable" anonymity and it is fixed before the call.
    - The aggregator stakes a safety deposit, used to limit his malicious behavior. The k-DAO indeed, in every moment can decide to vote to redeem this deposit if the aggregator misbehaves.

3. **Aggregated data production** - The work of the aggregator is then to produce new data in form of anonymized aggregated data, providing k-anonymity by design and differential privacy.

    - The data can be uploaded in another databox in form of MAM channels or IPFS storage.
    - Multiple configuration of aggregated data can be produced.
    - The aggregator must provide in the smart contract the exact quantity of data used from each subject's dataset. This is achievable maintaining subjects privacy through Merkle trees. The contract must contain only the root of a Merkle tree that contains all the data packets hashes used as leaves. k-DAO members can validate data used requesting (off-chain) leaves to the aggregator.

4. **Aggregated data sale** - New data is treated as all the other kind of data in the marketplace. Data consumer can access to it through the contract.

    - The payment is proportional to the contribute produced by each participant (e.g. aggregator = 55%, $u_1 = 20\%$, $u_2 = 10\%$, $u_3 = 15\%$)
    - Up to $n < k$ (with n predefined) can ask for a Proof of Sensing to the aggregator in order to check quality of data

Of course other mechanisms at the level of marketplace can protect both data subjects and aggregators from malicious behaviors, for instance a reputation mechanism (also implemented on DLT [10, 18]). However, the processes that stream the data flow in this sensing service are governed by the k-DAO, where members have all the interest to reach consensus.

5 Evaluation

The performances and the scalability of the model are the main issues to evaluate since, in the current state of the art, DLTs are known for these two important aspects (at least for the public permissionless ones). The proposed model, indeed, is expected to attend a large number of client that publish data to be put on the marketplace. The main blockchain technologies, Bitcoin and Ethereum, are presenting promising technology advancements for both performance and scalability, e.g. State Channels, however, currently these can achieve limited throughput. Our interest is in Ethereum smart contracts but the implementation of these can be thought also for other "faster" technologies [16]. Ethereum, indeed, achieves around 15 transactions per second and 15 seconds of blocktime, i.e. the time interval between the insertion of two consecutive blocks in the blockchain. In permissioned blockchains, additional permission control ensures that a majority of nodes are trusted allowing the use of consensus mechanisms that result in an higher throughput. However, the different kind of trust enabled by these, present properties that are not in the scope of this work for what concerns data protection.

For these reasons our evaluation concerns the performances of publishing data on the marketplace using IOTA. This technology cannot be considered in his infancy anymore, but it is still dealing with some structural issues that limit the vision of fast and feeless transactions. However, the current investments on this technology aim towards a suitable solution for IoT ecosystems.

5.1 Test Design

We simulated a number of devices that issue data, sensed by an IoT sensor, to the DLT. The communication scenario is simple: the device senses the measurement, creates a related transaction, and then sends the transaction to a IOTA full node, asking for the insertion of this transaction in the IOTA Tangle (main network). The devices waits for an answer before transmitting another transaction, in order to avoid message bombing to the full node. In tests, nodes are chosen randomly between 71 public nodes that provide the service for publishing a transaction.

To test the publication of multiple data packets from multiple IoT devices, we designed and implemented a system that emulates them, rather than using physical devices. In Linux operating systems the use of GNU Parallel[5] is useful for the purpose of executing jobs in parallel across any number of CPU cores. In order to emulate a single IoT device (or a cluster of these) it is necessary to assign to every emulated device (or a cluster) a different job, which is processed simultaneously with the other instances running in the system. In this way it is possible to emulate the data flow produced by a large number of IoT devices, evaluating how many of these affects the performances of IOTA nodes in the overlay network.

Each block of N emulated devices requires N socket connections to a IOTA node and each of these are made in parallel for each block. For instance, in figure 3 shows the specific case of emulation of 100 devices sending 100 request each (the computational heaviest case for IOTA nodes). For each job it is emulated the data flow of 5 devices and 5 socket connections are opened between clients and IOTA nodes.

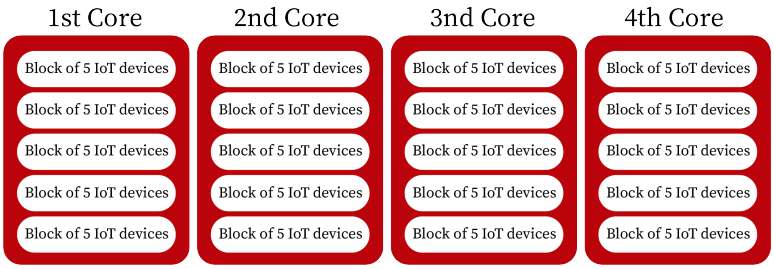

Figure 3: An example of IoT devices emulation in blocks across 4 CPU cores

This test design is used to verify what is the status of the network under the proposed workload, how reliable the network is, how many transactions will be lost in the process, and how many devices can be served by full nodes at the same time. As mentioned in Section 2.3.2, the process of submitting a transaction in IOTA is mainly composed by two operations: tip selection and PoW. Transactions submitted are MAM messages published in specific MAM channels (that are actually stored in the main Tangle). We measure the latency for executing tip selection followed by PoW i.e. the time elapsed between the moment in which the process of submitting the message starts and the moment in which the full node replies with a result.

In detail, tests have been conducted sending firstly 10 and then 100 messages from N IoT devices, where in different steps N assumes the following values $1, 10, 100$.

[5] https://www.gnu.org/software/parallel/

5.2 Results

Tests results give us a general picture with respect to what is the average time in recording a message under different sub-network stress configurations. Figure 4 shows average latencies for publishing a message and relative errors. Errors corresponds to messages not published to MAM channels because of a full node error. At first glance, the best case seems the one in which messages are sent from single devices. In fact, messages are recorded within an average of 10-15 seconds and there are no lost messages.

Moving to results obtained by multiple devices, it can be seen that the validation times are much longer, compared to those with single devices. Yet, in the case in which 10 devices send 10 messages, we can see that average latencies are three times higher than sending the same number of messages from a single device and the validation time of an average message is around 47 seconds. However, there are no errors in this case. Meanwhile, in the case of 10 devices sending 100 messages, not published messages largely exceed the published ones: 61,7% of errors. In this case, the average time for publishing a message is 70 seconds.

Considering tests carried out with 100 devices, recorded average latency is doubled, with respect to tests carried out with 10 devices. In the first case (100 devices sending 10 messages each), the average time is 140 seconds, while in the second case (100 messages), the average time is 124 seconds. The percentage of messages not published is respectively 14,9% and 36,16%.

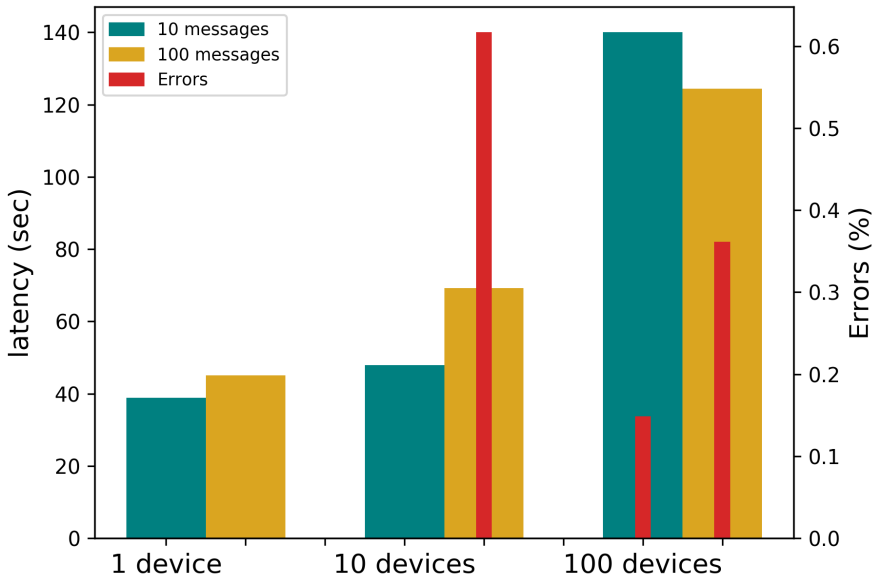

Figure 4: Average latency and errors for publishing a message in a MAM channel

Figure 5 shows the empirical cumulative distributions of latencies obtained from tests, that give a measure of what is the probability that a message is published within certain time. In the case with a single device, each message is published within 20 seconds on average; when 10 messages have been sent, the 95% confidence interval is (5.86, 18.16) sec, while (12.83, 15.8) sec is the interval for 100 messages sent.

If we consider the tests with multiple devices, we can see that for 10 devices publishing 10 messages each, all messages are recorded in an interval ranging from 10 to 100 seconds. 56% of these are recorded within 50 seconds, while the remaining 44% take more than 50 seconds. The

95% conf. int. is (43.71, 52.21) sec. We can find similar features in tests where 10 devices send 100 messages each: mainly, these are published in an interval ranging from 4 to 100 seconds and only the 20% of messages (79) are have a greater latency. The 95% conf. int. is (65.24, 73.1) sec. In this case the errors impact affects the results obtained.

The publication of messages sent by 100 devices, on the other hand, shows a substantial increase in time. In the case of 10 messages sent by each devices, it can be seen that the 66% of these is sent within 100 seconds, while the remaining 44% is published in greater times. For 2% of these, recording times of more than 600 seconds are observed. The last empirical cumulative distribution function refers to test conducted sending 100 messages from 100 different devices. In this case the 53% of messages is published within 100 seconds, while the 47% with greater times. 1.7% of these are validated in times of more than 500 seconds.

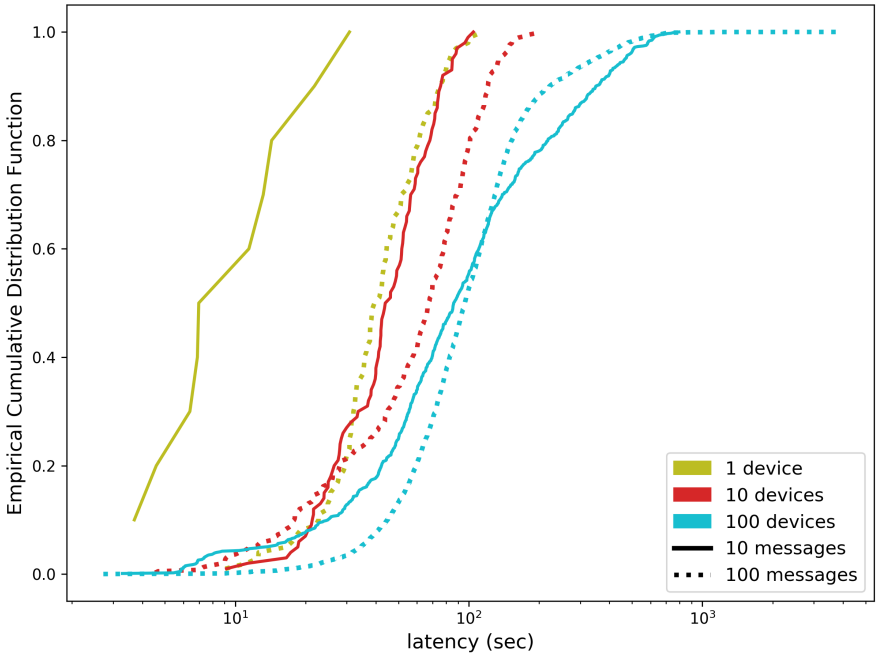

Figure 5: Empirical cumulative distributions of latencies obtained from tests simulating 1,10 and 100 devices.

5.3 Discussion

These results clearly shows that, currently, the response to scalability by the IOTA network is not sufficient enough. However, it is important to stress the fact that full nodes involved in testing perform PoW for each request received. Hence, the more requests nodes receive in a time frame, the more the latency increases for messages in that time frame. This is shown in both figures 4 and 5, where the increase of number of messages sent simultaneously implies a latency increase. Another important result to focus on is the fact that when accumulating multiple requests the node tendency to fail increases. Fail in this case means that a node has not replied to a message submit request, thus it can be attributed to an internal error or to a mechanism of protection to DOS attack. The solution to these nodes' resources overloading

could be an algorithm that balances the number of requests to send to these, based on their response time. At the time being, public full nodes IOTA network able to perform both tips selection and PoW range around the 70 units and not all of these perform equally (see for instance the 10 devices sending 100 messages case in Figure 4). Indeed, further tests with only one simulated single device show very variable latency averages depending on the full node randomly chosen, between 21 and 116 sec.

Probably, using a proprietary IOTA full node that execute PoW for messages sent by a limited set of devices would enhance the performances, e.g. a gateway able to do PoW in an edge computing architecture. However, a performing solution would require costs that users may not want to spend and additionally the throughput would be limited by PoW in any case. Nevertheless, these evaluation are ongoing works.

6 Conclusion

In this paper, we presented a model architecture where different actors are incentivized to ensure the personal data anonymity when this is offered in a marketplace. Regulations such as GDPR are moving the sovereignty of personal data towards subjects that are the source of this data, thus allowing them to assume a central role in the tradings of their personal data in a data marketplace. In our vision, personal data must be managed through a databox, a virtual boundary where data is stored and individuals can control how, when and what data is shared. To this aim, we exploit different technologies: i) IOTA and IPFS are used as the backbone to store and share personal data; ii) the interoperability in the market is assured by W3C standards such as RDF and OWL; iii) smart contracts allow to control data access and authorization.

To ensure anonymity two data protection techniques are used to provide k-anonymity and differential privacy properties in a Sensing-as-a-Service model. Through the use of a specific smart contract where parties agree on predefined policies regarding their rights and obligations, an "aggregator", i.e. the one that provides the sensing service, receives the access to data coming from k subjects. Then, this entity produces an anonymized dataset that can be acquired by data consumers in the marketplace and each actor who participated to the dataset creation is rightfully rewarded. The aggregator is incentivized to correctly behave because of a k-DAO, composed by data subjects participating in the sensing, that manages a safe deposit staked by this one.

We provided some results on a test scenario where multiple IoT devices publish data packets on the IOTA DLT, in order to investigate if this DLT represents a bottleneck in the model. Results show latencies higher than 15 sec up to 136 sec on average, which is quite high depending on the use case. This may be attributed to some structural issues in the IOTA peer-to-peer infrastructure and moving PoW requests from the full nodes elsewhere might improve the performances in latency. On the other end, the Ethereum blockchain is commonly known for its issues in scalability, thus in future works our approach based on smart contracts will be focused towards the use of new proposals such as Plasma [23]. Meanwhile, a new DLT is taking hold for its promise of high scalability and may be the solution to both Ethereum and IOTA problems, namely Radix [16].

Nevertheless, the main component that is worth investigating consists in the ability of smart contracts to express policies that satisfy legal requirements and individual's preferences that meet data requests, that is still an open question in computer science and law.

Acknowledgment

This work has been partially funded by the EU H2020 MSCA project LAST-JD-RIOE with grant agreement No 814177 and PROTECT with grant agreement No 813497.

References

[1] Muhamed Alarbi and Hanan Lutfiyya. Sensing as a service middleware architecture. In *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 399–406. IEEE, 2018.

[2] Claudio A Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, and Pierangela Samarati. An obfuscation-based approach for protecting location privacy. *IEEE Transactions on Dependable and Secure Computing*, 8(1):13–27, 2009.

[3] J Benet and N Greco. Filecoin: A decentralized storage network. *Protoc. Labs*, 2018.

[4] Juan Benet. Ipfs-content addressed, versioned, p2p file system. *arXiv preprint arXiv:1407.3561*, 2014.

[5] Amir Chaudhry, Jon Crowcroft, Heidi Howard, Anil Madhavapeddy, Richard Mortier, Hamed Haddadi, and Derek McAuley. Personal data: thinking inside the box. In *Proceedings of the fifth decennial Aarhus conference on critical alternatives*, pages 29–32. Aarhus University Press, 2015.

[6] Council of European Union. Regulation (eu) 2016/679 - directive 95/46.

[7] Andy Crabtree, Tom Lodge, James Colley, Chris Greenhalgh, Kevin Glover, Hamed Haddadi, Yousef Amar, Richard Mortier, Qi Li, John Moore, Liang Wang, Poonam Yadav, Jianxin Zhao, Anthony Brown, Lachlan Urquhart, and Derek McAuley. Building accountability into the internet of things: the iot databox model. *Journal of Reliable Intelligent Environments*, 4(1):39–55, Apr 2018.

[8] Nigel Davies, Nina Taft, Mahadev Satyanarayanan, Sarah Clinch, and Brandon Amos. Privacy mediators: Helping iot cross the chasm. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications*, pages 39–44. ACM, 2016.

[9] Sabrina De Capitani Di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy: definitions and techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(06):793–817, 2012.

[10] Richard Dennis and Gareth Owen. Rep on the block: A next generation reputation system based on the blockchain. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 131–138. IEEE, 2015.

[11] Salvatore Distefano, Giovanni Merlino, and Antonio Puliafito. Sensing and actuation as a service: A new development for clouds. In *2012 IEEE 11th International Symposium on Network Computing and Applications*, pages 272–275. IEEE, 2012.

[12] Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.

[13] S Ferretti and G. D'Angelo. On the Ethereum blockchain structure: A complex networks theory perspective. *Concurrency and Computation: Practice and Experience, Wiley*, in press.

[14] Jessica Groopman and Susan Etlinger. Consumer perceptions of privacy in the internet of things: What brands can learn from a concerned citizenry. *Altimeter Group: San Francisco, CA, USA*, pages 1–25, 2015.

[15] Amir Herzberg, Stanisław Jarecki, Hugo Krawczyk, and Moti Yung. Proactive secret sharing or: How to cope with perpetual leakage. In *Annual International Cryptology Conference*, pages 339–352. Springer, 1995.

[16] Dan Hughes. Radix-tempo. Technical report, Radix DLT, Sep. 2017.[Online], 2017.

[17] Hui Li, Lishuang Pei, Dan Liao, Gang Sun, and Du Xu. Blockchain meets vanet: An architecture for identity and location privacy protection in vanet. *Peer-to-Peer Networking and Applications*, 12(5):1178–1193, 2019.

[18] Zhaojun Lu, Qian Wang, Gang Qu, and Zhenglin Liu. Bars: a blockchain-based anonymous reputation system for trust management in vanets. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 98–103. IEEE, 2018.

[19] Damiano Di Francesco Maesa, Paolo Mori, and Laura Ricci. Blockchain based access control. In *IFIP international conference on distributed applications and interoperable systems*, pages 206–220. Springer, 2017.

[20] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large datasets (how to break anonymity of the netflix prize dataset). *University of Texas at Austin*, 2008.

[21] Charith Perera, Susan YL Wakenshaw, Tim Baarslag, Hamed Haddadi, Arosha K Bandara, Richard Mortier, Andy Crabtree, Irene CL Ng, Derek McAuley, and Jon Crowcroft. Valorising the iot databox: creating value for everyone. *Transactions on Emerging Telecommunications Technologies*, 28(1):e3125, 2017.

[22] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. Sensing as a service model for smart cities supported by internet of things. *Transactions on emerging telecommunications technologies*, 25(1):81–93, 2014.

[23] Joseph Poon and Vitalik Buterin. Plasma: Scalable autonomous smart contracts. *White paper*, pages 1–47, 2017.

[24] Serguei Popov. The tangle. *https://iota.org/IOTA$_W$hitepaper.pdf*, 2015.

[25] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, technical report, SRI International, 1998.

[26] Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. Solid : A platform for decentralized social applications based on linked data. 2016.

[27] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

[28] Nguyen Binh Truong, Kai Sun, Gyu Myoung Lee, and Yike Guo. Gdpr-compliant personal data management: A blockchain-based solution. *arXiv preprint arXiv:1904.03038*, 2019.

[29] Rui Zhang, Rui Xue, and Ling Liu. Security and privacy on blockchain. *arXiv preprint arXiv:1903.07602*, 2019.

[30] Mirko Zichichi, Michele Contu, Stefano Ferretti, and Gabriele D'Angelo. Likestarter: a Smart-contract based social DAO for crowdfunding. In *Proc. of the 2st Workshop on Cryptocurrencies and Blockchains for Distributed Systems*, 2019.

[31] Mirko Zichichi, Stefano Ferretti, and Gabriele D'Angelo. A distributed ledger based infrastructure for smart transportation system and social good. In *IEEE Consumer Communications and Networking Conference (CCNC)*, 2020.

[32] Guy Zyskind, Oz Nathan, et al. Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops*, pages 180–184. IEEE, 2015.