EasyChair Preprint

№ 5760

Steps Toward Fostering Peer-to-Peer Blockchain-Based Data Markets

José Parra-Moyano, Marcel Bühler, Michel Avital and Karl Schmedders

June 7, 2021

Steps Toward Fostering Peer-to-Peer Blockchain-Based Data Markets

Completed Research Paper

José Parra-Moyano
Copenhagen Business School
Frederiksberg, Denmark
jpm.digi@cbs.dk

Marcel Bühler
ETH Zurich
Zurich, Switzerland
buehlmar@ethz.ch

Michel Avital
Copenhagen Business School
Frederiksberg, Denmark
michel@avital.net

Karl Schmedders
IMD
Lausanne, Switzerland
karl.schmedders@imd.org

Abstract

The emergence of data markets will provide an arena for massive data trading. A large-scale implementation and adoption of data markets can have significant economic and societal implications. Currently, however, two obstacles hinder the implementation and broader development of data markets—the lack of guarantees of data provenance and the lack of incentives for honest participation in the market. Using the Design Science Research approach, we develop a blockchain-based artifact that can help mitigate these two obstacles. The evaluation of the artifact demonstrates that blockchain technology could be instrumental in the design of new, decentralized incentive structures that motivate the honest participation of all the involved agents in a data market while establishing the provenance of the data.

Keywords: Blockchain, data markets, Design Science Research, economics of IS

Introduction

It has long been acknowledged that data is part and parcel of modern firms. For example, firms use data to improve managerial decisions (Chen, Chiang, and Storey 2012), optimize the prices of products and services, and create targeted advertising (McAfee and Brynjolfsson 2012). Some firms—the so-called AI factories—use data to design smart connected products, which can satisfy consumers' needs only if fueled with abundant data (Porter and Heppelmann 2015). While these are merely a few examples of data usage by firms, they imply that data has become a necessary asset if firms are to compete (Parra-Moyano, Schmedders, and Pentland 2020). Therefore, we argue that the emergence and broad use of data markets will result in an increase in economic efficiency. Furthermore, ultimately data will be recognized by firms as a resource that is essential for value generation.

Data is, however, usually contained within the firm that generates it (as a by-product of the firm's activity) and confined within its boundaries. While some markets for data already exist, they are still in a very incipient phase, and firms do not yet have the means to transparently and efficiently trade data with one another (Koutroumpis, Leiponen, and Thomas 2020). This constitutes an economic inefficiency that is curbing competition and hindering innovation and growth.

Consequently, researchers and practitioners alike are paying close attention to data markets. Studying different architectures and economic incentives schemes allows a broad implementation of such markets (Dubovitskaya, Sunyaev, and Novotny 2020). Thus far, two aspects have been identified as major obstacles to the adoption of data markets. First, data sellers' difficulties in demonstrating the provenance (i.e., the origin) of the data (Koutroumpis, Leiponen, and Thomas 2020). Second, the inadequate incentives for buyers and sellers in data markets deter the honest participation of data hosts and data producers (Raskar, Vepakomma, Swedish, and A. Sharan 2019). Moreover, some data is private, and the willingness of the data owners to share data is therefore limited and context dependent (De Montjoye, Shmueli, Wang, and Pentland 2014). Hence, any incentives to participate in such markets need to address the privacy concerns of the data owners.

Scholars in the IS field have already suggested that blockchain technology can be used to overcome the current architectural and governance challenges regarding digital platforms and infrastructures, as well as to transform existing value-creating interactions, solving misaligned incentive structures and trust problems currently faced by digital platforms (Constantinides, Henfridsson, and Parker 2018). In this paper we address the challenge of how to leverage the unique properties of blockchain technology in order to implement an incentive structure that motivates the honest participation of all the agents involved in a data market while guaranteeing the provenance of the data and mitigating the privacy concerns of the data owners. Specifically, we pursued the following research question: *How can the use of blockchain technology provide an incentive structure that motivates honest multilateral participation in a data market while demonstrating data provenance?*

Successfully addressing this question can help to pave the way toward establishing new market structures for the trade of novel assets such as data. Moreover, our study also shows how blockchain technology can be used to instrumentalize new incentive schemes, which contribute to the articulation of new market structures, linking blockchain technology with principles of economic theory.

We designed, implemented, and evaluated a prototype of a blockchain-based, decentralized data market using the Design Science Research (DSR) approach. Designing prototypes using the DSR lens to study blockchain use cases and to generate transferable knowledge is a common approach in the IS field (Regner, Schweizer, and Urbach 2019).

The study offers two useful contributions to the IS literature. First, we show how the properties of blockchain can in fact be used to design an incentive structure that motivates agents' honest behavior in a data market. Second, we illustrate how blockchain technology can be used to validate the provenance of the data. Hence, the study offers tangible ways to apply blockchain technology in the implementation and adoption of data markets.

The remainder of this paper is structured as follows. In Section 2 we present a literature review on data markets and blockchain technology. In Section 3 we outline the DSR methodology. In Section 4 we

describe the resulting artifact and present its software architecture and the associated incentive structure. In Section 5 we evaluate the artifact to generate tangible knowledge about the incentive schemes and the provenance of the data. In Section 6 we discuss the implications of our results, and in Section 7 we conclude.

Literature Review

This section provides a brief literature review on data markets and blockchain technology. While we acknowledge that this review is not exhaustive, it provides the necessary foundation to motivate the problem and anchor it in the relevant discourse.

Data Markets

Data markets are being studied by scholars in different fields. Thus far, scholars have considered different definitions of what a data market is. Some consider any organization that offers the exchange of, or access to, data to be a market for data (Koutroumpis, Leiponen, and Thomas 2020). Others define a data market as any organization carrying out data trading as its core activity (Parmar et al. 2014, Thomas and Leiponen 2016). We define a data market as the commercial exchange of private data informed by some type of price mechanism (Parmar et al. 2014).

One of the major challenges that sellers in data markets are currently facing is their ability to prove the provenance of the data that they sell (i.e., the origin of the data) (Koutroumpis, Leiponen, and Thomas 2020). Since data quality and legality are difficult to assess, data consumers use data provenance as a proxy for the quality of data. This means that instead of verifying the data directly, data consumers tend to rely on the reputation and legal liability of the party from which they buy the data. For this reason, demonstrating data provenance is crucial for the widespread implementation of data markets (Evans 2014, Catalini and Gans 2016). However, existing data markets suffer from deficient provenance and are thus hampered by the strategic behavior of participants. Large-scale multilateral data markets are therefore unlikely to succeed without additional governance innovations that strengthen the provenance of data for all parties (Koutroumpis, Leiponen, and Thomas 2020).

Beyond the challenge of validating the provenance of data, there are other challenges with regard to the implementation and governance of data markets, such as how to define and implement a market design that enforces participation and honest behavior between a market's participants, how to incorporate and to incentivize data hosts to participate in the market, and how to establish a consistent governance structure that efficiently conducts arbitration between data purchasers and data sellers (Borgman 2012; Koutroumpis, Leiponen, and Thomas 2017; Raskar, Vepakomma, Swedish, and A. Sharan 2019).

Given the nature of data markets, blockchain technology has been identified as a technology that can open up a path toward the design of more sophisticated incentive structures. Specifically, scholars in the IS field have identified blockchain technology as having the potential to solve misaligned incentive structures and trust problems currently faced by digital platforms (Constantinides, Henfridsson, and Parker 2018).

Blockchain and Smart Contracts for Data Management

The idea of using a proof-of-work (PoW) system to implement a distributed timestamp server that prevents the double spending problem in peer-to-peer networks gave rise to blockchain technology (Nakamoto 2008). Blockchain technology enables the design of decentralized systems that do not need a central authority to validate information and generate trust.

Since the introduction of blockchain and of Bitcoin (Nakamoto 2008), which was the first system to use of blockchain technology to enable a fully decentralized peer-to-peer currency, many contributions have been made to the field. Vitalik Buterin, for example, introduced the idea of blockchain-based smart contracts, systems that automatically move digital assets according to arbitrary, pre-specified rules (Buterin 2019).

Blockchain has the potential to disrupt existing business models, and blockchain infrastructure holds the promise of increased speed of exchange, a reduction in the number of intermediaries and in associated costs, improved security, digitized assets, wider access for disadvantaged groups (especially in emerging economies), and improved regulatory compliance (Constantinides, Henfridsson, and Parker 2018). Consequently, interest in the application of blockchain to manage data is growing rapidly.

However, existing blockchain-based systems can offer only limited capabilities and solutions from technical, legal, and social perspectives. For this reason, there is a call for IS scholars to study elements such as "specific use-case scenarios", "technical aspects of privacy and security for data management", the "compliance of the systems with standards and regulations", and the "social perspectives of the technology" in order to ease the mass adoption of blockchain-based data management (Dubovitskaya, Sunyaev, and Novotny 2020).

The fact that there are two specific aspects of data markets that require further investigation (namely, how to make use of technology to implement adequate incentive structures that motivate the honest participation of all the agents involved in a data market, and how to use innovation to strengthen data provenance) motivates our paper. Moreover, and acknowledging the potential adequacy of blockchain technology to address issues related to managing data, we aim to present a "specific use-case scenario" for blockchain technology, one that enables us to learn about the "technical aspects of privacy and security for data management".

Research Method

DSR is an approach that involves a rigorous process of designing, creating, and evaluating artifacts with the purpose of generating relevant and generalizable knowledge (March and Smith 1995, Peffers et al. 2007). DSR is inherently a problem-solving process (Hevner et al. 2004) that focuses on the creation and evaluation of innovative IT artifacts that enable organizations to address important information-related tasks. Seven guidelines have been defined in order to aid scholars in the IS field to implement DSR-based projects (Hevner et al. 2004).

First, DSR requires the creation of an innovative, purposeful artifact. Second, this artifact needs to be created to solve a problem in a specified domain. Third, the thorough evaluation of the artifact must yield utility for the specified problem. Fourth, the artifact needs to be innovative, and to contribute to solving an existing problem. Fifth, the artifact itself must be rigorously defined, formally represented, coherent, and internally consistent. Sixth, the process by which the artifact is created, and often the artifact itself, needs to incorporate or enable a search process whereby a problem space is constructed, and a mechanism posed or enacted to find an effective solution. Seventh, the results of the DSR process must be communicated effectively both to a technical audience (researchers who will extend them and practitioners who will implement them) and to a managerial audience (researchers who will study them in context and practitioners who will decide if they should be implemented within their organizations).

In undertaking this project, we have taken both a reactive and a proactive approach to technology: we have taken the state of the technology as it is today (reactive) and use it to design an artifact (proactive) that solves the market inefficiencies present today in terms of the exchange of data. Given that designing and implementing a consistent incentive structure for a decentralized data market and having it ensure data provenance is one of the open challenges that existing data markets are facing, we have rigorously defined (Guideline 5) an innovative incentive structure and blockchain-based architecture (Guideline 1) to solve problems in the specific domain of data markets (Guideline 2). Moreover, we have evaluated the artifact (Guideline 3) by presenting it to different scholars and practitioners in seminars and workshops. This has led to the incorporation of their feedback in a recursive search process (Guideline 6) that has helped us to find an effective solution to the previously defined problem.

In this evaluation and search process we have selected six groups of experts to provide feedback at different points of the development phase, and to recursively and iteratively improve the design to meet its objectives and generate generalizable knowledge.

Moreover, in order to demonstrate the utility of the artifact, we present in this paper an initial evaluation of the prototype, comparing illustrative scenarios of the artifact in a synthetic situation as well as

conducting descriptive evaluations. These two evaluation methods have been deemed as adequate ways of evaluating prototypes using a DSR approach (Peffers et al. 2007 Regner, Schweizer, and Urbach 2019). Finally, our aim in writing this paper is to communicate our results effectively both to a technical and to a managerial audience (Guideline 7).

The aim of the project at this early stage is to demonstrate how blockchain technology enables the implementation of an incentive structure that induces adequate behavior of agents in a decentralized data market, while proving the provenance of data. Since we are interested in studying the aspects of blockchain technology that help in implementing an adequate incentive structure to enforce adequate behavior in the market, our case is limited in scope. It is important to note that the ultimate goal of DSR is to theorize about an IS artifact that needs to be designed. This means that a theoretical contribution is the result. Given that markets for data need to be designed with wider adoption firmly in mind, and that a deeper theoretical understanding of such markets could foster their implementation, we consider the DSR methodology as the most adequate for studying our research question. These reasons make our case suited to DSR prototype building.

System Design

This section introduces the design principles and describes the development of the artifact.

Blockchain and Smart Contracts

From our literature review it emerges that there are (at least) two obstacles that hinder the broad adoption of data markets. The first is the lack of an incentive structure that motivates the honest participation of all engaged agents and explicitly addresses any privacy concerns of data owners. The second obstacle is the difficulty of ensuring data provenance. Data provenance is essential if data consumers are to assess the value of the data that they might potentially buy. The aim of our project is to design an artifact that helps to generate knowledge about possible pathways toward the mitigation of the abovementioned two obstacles. Additionally, the artifact shall use blockchain technology in order to embed the incentive structure that guarantees data provenance into a decentralized architecture. This will help us to understand how the use of this technology can contribute to the design of data markets that do not suffer from the abovementioned two obstacles.

Following the relevance cycle (Hevner 2007), we defined the design objectives and necessary criteria for the evaluation of our artifact (Hevner 2004). We specify three design objectives for the prototype. First, it needs to guarantee data provenance. This means that any agent in the system needs to be able to trace the data that it is receiving to the original source of that data. In order to evaluate if this design objective is fulfilled, we conduct a descriptive evaluation that ensures that any agent involved in a transaction can demonstrate the source from which the data originated. Second, the prototype needs to have a decentralized architecture, such that no central instance has control over the whole system. While the need for decentralization might not seem obvious at the first glance, it is required if the potential privacy concerns of the data owners are to be addressed. A decentralized architecture prevents an unnecessary concentration of the data with a central instance. Moreover, it disincentivizes attacks on a central instance that gathers the data of all participants. Hence, while a centralized architecture would indeed be easier to implement, it would be associated with privacy concerns that would probably continue to hamper the implementation and adoption of data markets. The fulfillment of the decentralization objective is also assessed by means of a descriptive evaluation. Third, each of the three participating agents—namely, the consumer who buys the data, the producer who generates the data, and the host who hosts the data—needs to be incentivized to participate in the system in an honest manner. This means both that the participation of all the agents needs to be incentivized, and that the dishonest behavior of any agent shall be deterred (i.e., each agent shall expect no positive outcome from any dishonest behavior). We evaluate the fulfillment of this objective by means of an illustrative scenario comparison. Table 1 displays the design objectives and evaluation criteria to which we adhere.

Objective	Evaluation
Guarantee data provenance	Descriptive evaluation
Decentralized architecture	Descriptive evaluation
Motivate honest behavior	Illustrative scenario comparison

Table 1. Design Objectives and Evaluation

In line with the design objectives previously defined, we develop a prototype that aims to solve some of the problems that existing data markets are facing. We have conducted the development of the prototype by following the DSR cycle. The prototype presented hereafter is the result of several improvement cycles that incorporate the evaluation and feedback of several stakeholders. Therefore, while it is still a basic and modest prototype, it has already gone through a recursive refinement process. At this stage it is important to recall that the goal of the prototype's development and evaluation is to generate transferable knowledge that contributes both to scholars and practitioners. Hence, this prototype should be seen as a means of knowledge generation, and not as a finalized, fully deployable system.

System Principles

The purpose of the prototype is to enable the consumers to purchase data from the producers, while compensating the hosts for hosting the data and making the deal possible. The system is conceived as a platform in which a consumer makes offers to the producers to acquire their data for a specific price. The offers are made through the host, which is the economic agent hosting the data of the producers. Since the host co-creates, stores, and maintains the data, we consider it a fundamental agent in the exchange. As such, the host needs to be properly incentivized. Thus, the host receives a fee each time a data producer accepts a consumer's offer. This is a monetary incentive that shall motivate the participation of both the producer and the host.

The system's flow works as follows. First, the consumer and the host negotiate the terms of the deal. This negotiation defines the data that the consumer is willing to acquire, the price that it will pay to each producer for this data, and the fraction of this price that the host will receive for each producer that accepts an offer. Moreover, this negotiation also defines the minimum number of producers that need to accept the offer of the consumer for the deal to become effective. Since we assume that only a significant amount of data will serve the consumer's purpose, the deal is conditional on the number of producers accepting the offer. Once the consumer and the host agree on the terms of the deal, they sign a smart contract containing all these terms.

Moreover, the consumer pays a pledge to this smart contract, which covers the transactions costs and ensures that it has enough funds to compensate the producers and the host. Additionally, this pledge contains a deposit, which motivates honest behavior from the consumer. Once the smart contract is signed by both the consumer and the host, it is the host who communicates with its users (the producers) and makes the terms of the offer available to them. Producers can either accept or reject the deal by means of an interface. They accept the deal by signing the smart contract. The smart contract records the permissions given by the producers to the consumer, as well as all the relevant details (rights and obligations) of the contract.

Once enough producers have accepted the deal, the consumer locks the deal, such that no more producers can accept it. The consumer does this by locking the smart contract. This prevents any further interaction with the contract.

Once the smart contract is locked, the host grants the consumer access to the producers' data using an API endpoint. This API is linked to the smart contract, such that the consumer only gets access to the data of those producers who have agreed to it. Once the consumer receives the data, it releases the pledge stored in the smart contract. By doing so, it receives its deposit back, and the producers and host receive the agreed payment. The pledge prevents the host from sending corrupted data or data different to that agreed upon. Should the data not fulfill the requirements stored in the contract, the consumer can

decide to not release the pledge, and hold it as a proof when seeking external arbitration. The process and all possible resulting scenarios, which depend on the honest or dishonest behavior of the agents, is explained in detail in the next subsection.

Process Description

We have designed a transaction process that encompasses all possible interactions between the three agents. This process can result in four different scenarios depending on the actions taken by each agent. These actions can either be honest or dishonest. Each of the four scenarios results in a particular outcome for each of the agents. The process and the prototype are designed such that dishonest actions by any given agent result in a scenario with a negative outcome for that agent. Thus, dishonest actions are deterred. The process and the resulting scenarios are depicted in Figure 1.

The process starts by a consumer making a request to a data host. This request describes the data it is willing to purchase from the producers that are storing data at the particular host, as well as the terms under which it is willing to buy this data. The terms include the price it is willing to pay for the data of each producer, as well as the fraction of this price that the host will receive as a fee. Moreover, the contract defines the minimum and maximum number of producers from which the consumer would need data in order for the contract to be effective. Hence, if too few producers subsequently accept the proposal, such that the consumer would end up with insufficient data for the generation of the insights it requires, the deal is not effective. Should more than enough producers be willing to accept the proposal, meanwhile, the amount that the consumer will have to pay is limited. If the host agrees with the terms proposed by the consumer and is willing to pass the offer to the producers whose data is being hosted in the host's database, then it will ask the consumer to write these terms in a smart contract.

This smart contract is called Dataquery. Dataquery comes into effect when the consumer pays the corresponding pledge. The pledge covers the fees for the conduct of the contract, the maximum amount that the consumer would pay if all the producers would agree on the deal, as well as an additional amount that serves as a deposit that will only be returned once the transaction is closed. The pledge incentivizes the consumer to pursue the deal until its end. In order to later obtain authorization to query the data, the consumer also creates a password. The password is encrypted with the public key of the host. This password will later allow the consumer to claim its right over the data of the producers who have agreed to the terms of the deal.

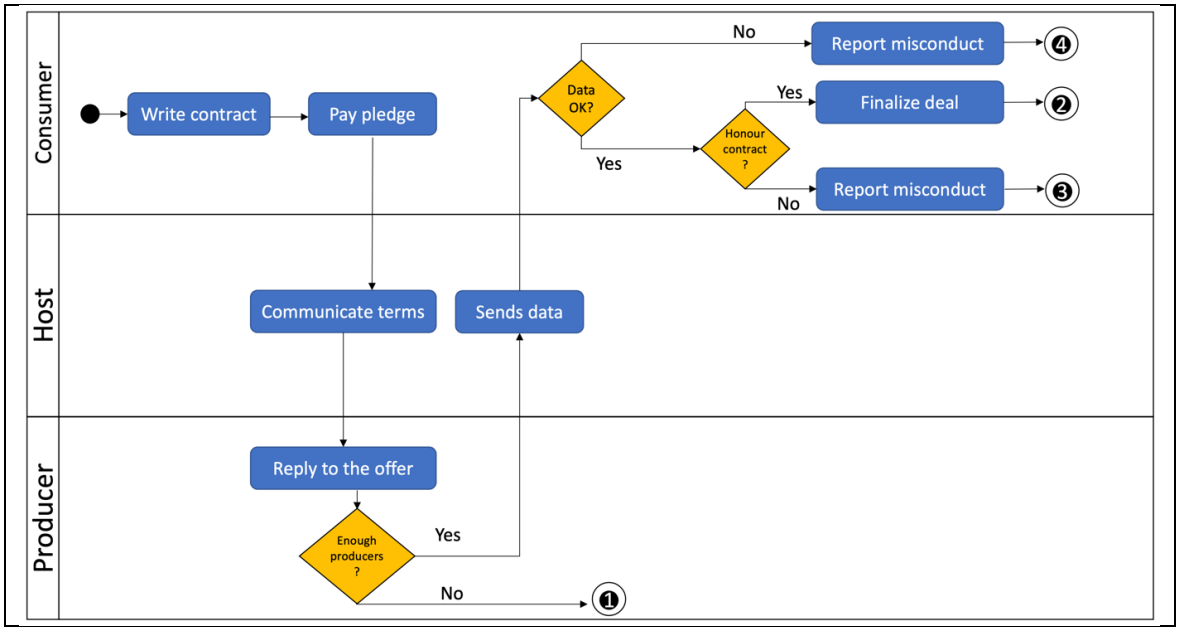

Figure 1. Four Different Process Outcomes

Once both the consumer and the host have signed the smart contract and therefore are adhering to the previously defined terms, the host informs the producers about the deal. The process of informing the producers occurs outside of the blockchain, by means of an external communication channel. Producers can either accept or reject the proposed deal.

By accepting the contract, producers give their permission to the consumer to read their data from the host in exchange for the proposed compensation. This permission and the resulting right to claim the compensation from the consumer are also recorded in the smart contract. Likewise, by rejecting the deal a producer is publicly recording in the smart contract that this consumer has no right to read its data from this host.

The consumer has the right to lock the deal at any time. Locking the deal prevents more producers from accepting or rejecting it. Should too few producers have accepted the deal, then the smart contract sends the pledge minus the transaction costs back to the consumer. This outcome is illustrated as Scenario 1 in Figure 1. In this scenario, the transaction fails due to too little interest on the side of the producers. In this case, the host gets no compensation, and the consumer ends up paying the transactions costs without receiving any data.

Should enough producers accept the deal, then the corresponding permissions are stored in the blockchain. At this stage, the host needs to provide API access to the consumer, such that the latter can read the data of the corresponding producers. Recall that the data is stored in the host's database. In order to read the data, the consumer sends the unencrypted ("plain text") password that it used to create the deal to the host via an encrypted request (SSL). In order to verify the validity of the deal, the host queries the encrypted password from the blockchain and decrypts it with its own private key. If the received "plain text" password matches the decrypted password from the blockchain, the deal is considered verified, and the consumer is granted access to the data of the corresponding producers.

Once the consumer reads the data, we consider two possible events: either the data fulfills the agreed terms, or it does not. If the data fulfills the agreed terms, the consumer is meant to finalize the deal. By finalizing the deal with its signature in the smart contract, it liberates the pledge. The pledge is used by the smart contract to pay the producers and host. The deposit is returned to the consumer. This outcome is illustrated as Scenario 2 in Figure 1.

Should the consumer not be willing to finalize the deal, even though the data actually fulfills the agreed requirements, then no payment is exchanged, and external arbitration will be necessary. In such a case, and until the conflict is resolved, no party receives their compensation. This outcome is illustrated as Scenario 3.

If the data does not fulfill the agreed terms, then the consumer can report misconduct and ask for external arbitration. In this scenario, no party would receive their compensation until the conflict was resolved. This is illustrated as Scenario 4.

System Architecture

The data of the producers is hosted in the host's database. The three agents interact by means of different interfaces that are connected to two smart contracts. While the actual data transfers occur between the host and the consumer outside of the blockchain and by means of an API, the contracts, agreements, payments, and permissions to read are managed by means of smart contracts in the blockchain.

Specifically, the system is implemented on the Ethereum blockchain. Ethereum is a blockchain with a built-in Turing-complete programming language that allows anyone to write smart contracts and decentralized applications following their own arbitrary rules, transaction formats, and state transition functions. The system consists of two smart contracts coded in Solidity, and of a javascript/html front end. These two components communicate via a web3.js API. In order to connect these components to the blockchain network we use INFURA. The prototype is deployed to the Rinkeby test network. Figure 2 illustrates this system architecture.

The back end consists of two smart contracts coded in Solidity. The first smart contract, Dataquery, manages data deals and their permissions. The second smart contract, Bank, manages temporary

accounts for deals and holds the deposits from consumers. The smart contract Bank is only used to temporarily store deposits and payments from consumers. For more details, we refer the reader to our anonymous repository: https://anonymous.4open.science/r/b144110c-af8e-4fb2-a9d2-2d4891e3c24e/.

The front end of the system consists of a javascript/html webapp, which the three different types of actors (consumers, hosts, and producers) use to log in and manage their operations.

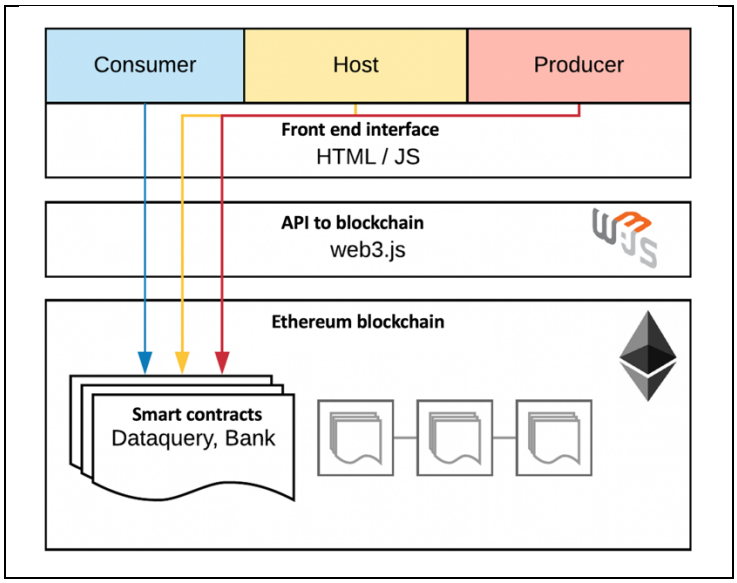

Figure 2. Overview of Components, Agents, and their Interaction with the Artifact

Evaluation

This section provides an evaluation of how the proposed prototype artifact meets the objectives from Table 1. The prototype of the artifact has been evaluated by six expert groups, as depicted in Table 2, which describes the experts that were selected for the evaluation phase, as well as the reason for selecting them.

ID	Short Description	Reason for Selection
1	Scholars at the blockchain center of a large English university	To validate the correct use of blockchain technology
2	Scholars at the blockchain lab of a large Swiss university	To validate the correct use of blockchain technology
3	Scholars at the ethics lab of a large Swiss technical university	To validate the social perspective of the artifact
4	Economists participating in a workshop at a US market design conference	To validate the market logic and efficiency of the artifact
5	Developers at a large Spanish telecommunication provider developing a data market	To validate the actual use of the artifact for practitioners in the EU
6	Developers at a large Swiss telecommunication provider developing a data market	To validate the actual use of the artifact for practitioners in environments outside the EU

Table 2. Experts for the Evaluation

In the following subsections we indicate, in brackets, the expert group that provided the idea or validated the concept in the sentence that describes the idea or the concept.

Data Provenance

The data is originated by interaction between producer and host. For the data to be sold to a third party (the consumer), both the producer and the host need to agree and sign the corresponding smart contract. By signing the smart contract, a trace of this transaction is stored in the blockchain [expert groups #1 and #2]. The smart contract contains the description of the data, the public keys of the agents involved in the transaction, a timestamp, and other information related to the payments. This smart contract serves as a certification to trace the data to its original source (the producer and the host) [expert groups #1 and #2]. This leaves a trace of the origin of the data, the consumer that has access to this data, the time and date at and on which the permission to read the data was granted, etc. This feature of the system enables a tree-structure tracing path for any data and any party engaged in a deal. Thanks to this feature, any party seeking to prove its participation in a deal can do so by providing access to the smart contract to anyone, who—by hashing the smart contract—can verify that it is the true contract that is actually stored in the blockchain [expert groups #5 and #6]. Should a consumer wish to enrich and resell the data to another party, it could prove that it originally acquired the data and from what source (by showing the original smart contract) [expert group #4]. By including the hash of the original smart contract in a new smart contract, the consumer, which is now reselling a potentially enriched version of the data, can prove and leave a trace of the data provenance. The prototype displays in its front end a log of all the transactions and signatures made in the smart contract.

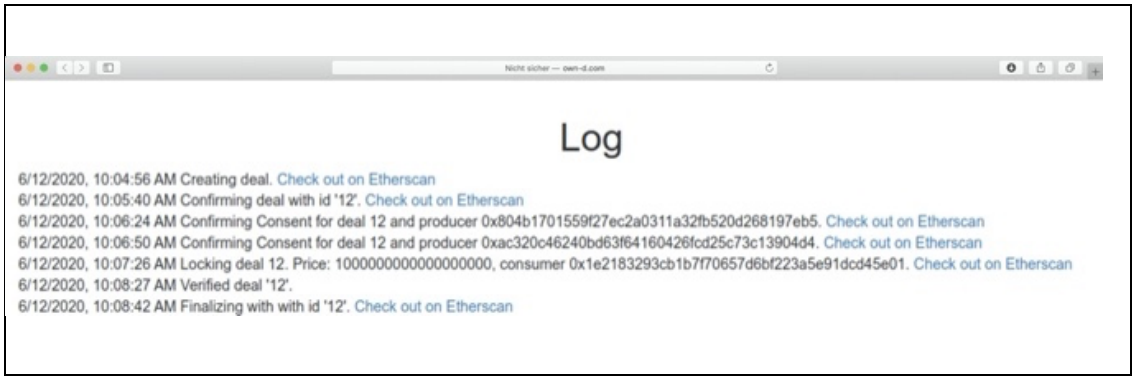

Figure 3. Log of Information Related to a Transaction

This log, depicted in Figure 3, reflects how all relevant aspects of each transaction are traced [expert groups #1 and #2]. Beyond these features, more sophisticated versions of this prototype could incorporate conditions under which the data can be resold by a consumer in a private manner [expert group #3]. Such conditions could determine potential compensation to the host and producer for a transaction made by the consumer, by the consumer's consumer, etc. By storing the smart contract with the details of the transactions in the blockchain, tracing the data to the original source becomes possible. It is thanks to the hash tree structure that provenance is guaranteed [expert groups #1, #2, and #4].

Decentralized Architecture

The architecture of the artifact is decentralized. There is no central instance that has control over the system. All transactions occur between peers. The terms of any agreement are stored in a smart contract on the Ethereum blockchain. This setup guarantees the desired decentralization [expert groups #1, #2, #3, #5, and #6]. Moreover, it is only thanks to the features of blockchain technology that we can implement the incentive structure designed here [expert groups #4, #5, and #6]. Without the features of blockchain technology, we would need a third party to implement the incentive structure, to manage the permissions, conduct the payments, store the data, etc. This would imply a central instance collecting

private data, which might be perceived by the data owners as an insurmountable obstacle to their participation in the market [experts # 4, #5, and #6].

Motivation for Honest Participation

As depicted in Figure 1, the process has four possible outcome scenarios. In Scenario 1, the deal is aborted due to a lack of interest from the producers. If the number of interested producers does not reach the minimum established by the consumer, the deal is locked. This results in the consumer paying the transaction costs but getting no data [expert group #4]. In such a scenario, no transaction occurs. Such a scenario could, for example, occur when the consumer offers too low a price to the producers. In such a case, too many producers have been offered too small an incentive to participate in the transaction. It is assumed, though, that increasing the price offered to the producers shall decrease the likelihood of this outcome [expert group #4]. Our artifact is flexible enough to let the consumer and host agree on the price paid to the producers and the fee received by the host [expert groups #5, and #6].

In Scenario 2, all the agents behave honestly (i.e., the host provides the data that it agreed to provide in the contract, and the consumer acknowledges the quality of the data and liberates the funds for final payment to all other involved agents). In such a scenario, all agents end up with a positive outcome: the producers receive the compensation they agreed to; the host receives the fee for every producer that has accepted the deal; and the consumer receives the data, for which it pays what it initially offered (receiving, additionally, the amount of the deposit back). This scenario represents the desired outcome [expert group #4]. All participating agents behave as expected, and all end up with a positive reward [expert group #4].

In Scenario 3, it is the consumer that behaves in a dishonest manner. Specifically, while it receives the data as has been agreed in the contract, it refrains from liberating the payment to pay the corresponding fees to the host and the producers. While the consumer is free to do this, this would imply that it would not receive the deposit back. Moreover, both the host and the producer could ask for external arbitration (e.g., a legal procedure), which could result in additional costs for the consumer. Hence, if the consumer receives the data as agreed in the contract, it is in its interests to liberate the funds and behave adequately [expert groups #4, #5, and #6].

In Scenario 4, it is the host that behaves in a dishonest manner. Specifically, the host gives access to data that does not fulfill the requirements established in the contract. Should this occur, then it is in the interests of the consumer to not liberate the payment and to use the incomplete transaction as a grievance when asking for external arbitration. As a result, the host—so, the agent acting in a dishonest manner—would not receive any payment. Hence, it is in its interest to honor the contract and give access to the data in the form described by the contract [expert groups #4, #5, and #6]. One design feature that could reduce the frequency with which Scenario 4 might occur would be the introduction of a reputation system based on reviews. A reputation system would help consumers to report situations in which the host has acted in a dishonest manner. This would leave a trace regarding every transaction and could be read as a signal by data consumers. A reputation system would therefore make dishonest behavior more costly for the host. Given the traceability of all transactions thanks to the use of blockchain technology, every review could be associated with factual metrics about the transaction, including data volume, the date of the transaction, etc. This would contribute to increasing the credibility of the review (Cheung, Sia, and Kuan 2012)

Of the four scenarios, only Scenarios 3 and 4 emerge as a result of a dishonest action. In these two scenarios, the dishonest agent ends up with a negative outcome. This should deter misbehavior in the system [expert group # 4]. Scenario 1, which results in no transaction, can be avoided by the consumer and host by setting up a higher payment to the producers [expert group # 4].

Discussion

Building on a DSR approach, we demonstrate how blockchain technology can be used to design smart contracts that ease the implementation of data markets. The way in which smart contracts open up a path toward the adoption of such markets is twofold. First, these smart contracts enable the creation of

decentralized incentive structures that motivate the honest participation of all the agents in the market, deterring misbehavior from the engaged parties by making such misbehavior costly and therefore unattractive. This opens up a path toward the implementation of new governance innovations that strengthen the provenance of data for all parties in data markets (Koutroumpis, Leiponen, and Thomas 2020). Second—and as demonstrated by our evaluation of the artifact—the characteristics of blockchain technology (specifically its ability to generate hash trees that relate smart contracts to one another) help to prove the provenance of data, which has been identified as a crucial obstacle to the adoption of such markets. This contributes to solving the issues of data provenance intrinsic to existing, centralized data markets (Evans 2014, Catalini and Gans 2016).

We contribute to theory by demonstrating how a blockchain-based solution can provide participants in a data market with appropriate incentives. Thus, we generate descriptive and prescriptive knowledge for the incipient research domain of data markets and decentralized incentive structures. This lays ground for further research in the field (Gregor 2006). Our results are generalizable and therefore transferable to other blockchain-based systems that need to implement decentralized incentive structures. The evaluation of the prototype reveals that only one of the four scenarios that can possible emerge is the desired one. It also shows that the consumer has the ability to offer higher compensation to producers. And that the host has the ability to motivate the participation of data producers. Our artifact therefore emerges as an infrastructure for implementing a new incentive structure, which can only be implemented thanks to the characteristics of blockchain technology (decentralized yet trusted smart contract architecture). Thus, this technology can be used to apply novel economic mechanisms that enable fair, transparent, and efficient markets for data.

Moreover, our results can be used by practitioners designing and seeking to establishing data markets. We illustrate how an adequate incentive structure can contribute to aligning the objectives of all the participants, and show how such an incentive structure can be implemented, by means of blockchain technology, to solve potential misalignments of the incentives of participants in such markets (Borgman 2012, Koutroumpis, Leiponen, and Thomas 2017, Raskar, Vepakomma, Swedish, and A. Sharan 2019).

While our paper offers interesting results, it also suffers from some limitations. First, the implementation of the described system in a corporate context is not easy and is accompanied by potentially high costs. Second, the prototype reveals that in the system misbehavior, while deterred, is still possible. In such cases, external arbitration is required and the blockchain part of the system "just" serves as a tracer of the contracts signed between the engaged parties. Additionally, the prototype is still incipient, and a large-scale implementation would require much more advanced development.

Future research could focus on studying how the size of the payment affects the producers' participation; how more sophisticated incentive structures could enable the reselling of enriched data while ensuring the continued engagement of the original data producers and/or host; how the host's fee alters the producers' willingness to participate; how to incorporate reputation mechanisms; how to tune the incentive knobs to make misbehavior more or less painful for users; and how the transaction costs affect the system.

Conclusion

This paper illustrates how blockchain technology can be used to incentivize the honest participation of all the agents involved in a decentralized data market while validating the provenance of the data. It also addresses some privacy-related issues and provides a specific technical example of the implementation of such a market. Evaluating the prototype that results from our application of the DSR methodology, we conclude that blockchain technology enables the design of new types of incentive structures to open up a path toward the adoption of decentralized data markets, and show one, specific, validated example of such a structure. In a context in which data has become a necessary asset if firms are to compete, and in which data markets might become crucial for the unfolding of the digital economy, the study of sophisticated incentive structures that motivate the honest participation of all the agents involved in a data market might prove crucial.

Acknowledgements

We are grateful to Dave Brooks for excellent editorial support.

References

Borgman, C. L. 2012. "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology* (14:4), pp. 1059-1078.

Buterin, V. 2008. "A next generation smart contract and decentralized application platform," Online; accessed 18 March 2019.

Catalini, C. and Gans, J.S. 2016. "Some Simple Economics of the Blockchain," *NBER Working Papers* 22952, National Bureau of Economic Research, Inc.

Chen, H. C., Chiang, R., and Storey, V. 2012. "Business intelligence and analytics: From big data to big impact," *MIS Quarterly* (36:4) pp. 1165-1188.

Cheung, C.M.Y., Sia, C.L. and Kuan, K.K., 2012. "Is this review believable? A study of factors affecting the credibility of online consumer reviews from an ELM perspective," *Journal of the Association for Information Systems*, 13(8), pp. 618-635.

Constantinides, P., Henfridsson, O., and Parker, G. G. 2018. "Platforms and infrastructures in the digital age," *Information Systems Research* (29:2) pp. 381-400.

De Montjoye, Y. A., Shmueli, E., Wang, S. S., and Pentland, A. S. 2014. "Protecting the privacy of metadata through safeanswers". *PloS one*, 9(7), e98790.

Dubovitskaya, A., Sunyaev, A., and Novotny, P. 2020. "Introduction to the Minitrack on Blockchain-based Intelligent Data-Management for Healthcare (BID4Health)," on *Proceedings of the 53rd Hawaii International Conference on System Sciences*. 2020.

Evans, D. 2014. "Economic aspects of bitcoin and other decentralized public-ledger currency platforms," *SSRN Electronic Journal*.

Gregor, S. 2006. "The nature of theory in information systems," *MIS Quarterly* (30:3) pp. 611-642.

Hevner, A. 2007. "A three-cycle view of design science research," *Scandinavian Journal of Information Systems* (19:2) pp. 87-92.

Hevner, A., March, S., Park, J., and Ram, S. 2004. "Design science in information systems research," *MIS Quarterly* (28:1) pp. 75-105.

Koutroumpis, P., Leiponen, A., and L.D., T. 2017. "The (unfulfilled) potential of data marketplaces," *ETLA* Working Papers No 53.

Koutroumpis, P., Leiponen, A., and Thomas, L. 2020. "Markets for data." *Industrial and Corporate Change* (29:3) pp. 645-660.

March, S. and Smith, G. 1995. "Design and natural science research on information technology," *Decision Support Systems* (15:4)2 pp. 51-266.

McAfee, A. and Brynjolfsson, E. 2011. "Predictive analytics in information systems research," *MIS Quarterly* (35:3) pp. 553-572.

Nakamoto, S. 2008. "Bitcoin: A peer-to-peer electronic cash system," Online; accessed 18 March 2019.

Notheisen, B., Cholewa, J. B., and Shanmugam, A. P. 2017. "Trading real-world assets on blockchain," *Business & Information Systems Engineering* (59:6) pp. 425-440.

Parmar, R., Mackenzie, I., Cohn, D., and Gann, D. M. 2014. "The new patterns of innovation," *Harvard Business Review*, (92:3) pp. 86-95.

Parra-Moyano, J., Schmedders, K., and Pentland, A. 2020. "What managers need to know about data exchanges," *MIT Sloan Management Review* (61:4) pp. 39-44.

Porter, M.E., and Heppelmann J. 2015. "How smart, connected products are transforming companies," *Harvard business review* (93:10) pp. 96-114.

Peffers, K., Rothenberger, M., Tuunanen, T., and Vaezi, R. 2012. "Design science research evaluation," In *Proceedings of the 7th International Conference on Design Science Research in Information Systems: Advances in Theory and Practice*, Berlin, Heidelberg. Springer-Verlag.

Peffers, K., Tuunanen, T., Rothenberger, M., and Chatterjee, S. 2007. "A design science research methodology for information systems research," *Journal of Management Information Systems* (24:1) pp. 45-77.

Raskar, R.,Vepakomma, P., Swedish, T., and Sharan, A. 2019. "Data markets to support AI for all: Pricing, valuation and governance." *CoRR*.

Regner, F., Schweizer, A., and Urbach, N. 2019. "NFTs in practice – non-fungible tokens as core component of a blockchain-based event ticketing application," In *Proceedings of the Fortieth International Conference on Information Systems*, Munich, Germany pp. 1-17.

Thomas, L. and Leiponen, A. 2016. "Big data commercialization," *IEEE Engineering Management Review*, (44:2) pp. 74-90.

Tsoi, K., Hung, P. and Poon, S. 2021. "Introduction to the minitrack on big data on healthcare application," in *Proceedings of the 54th Hawaii International Conference on System Sciences* (p. 3389).