**Decentralized Data Marketplace to Enable Trusted Machine Economy**

Zan-Jun Wang‡, Ching-Hua (Vivian) Lin†, Yang-Hao Yuan§, Ching-Chun (Jim) Huang∗

∗†Department of Computer Science and Information Engineering, National Cheng Kung University

‡Department of Computer Science and Information Engineering, National Taiwan University

§BiiLabs, Co., Ltd.

∗†No.1, University Road, ‡No.1, Sec. 4, Roosevelt Road

∗†Tainan City, Taiwan (R.O.C.), ‡Taipei City, Taiwan (R.O.C.)

∗[jserv@ccns.ncku.edu.tw,](mailto:jserv@ccns.ncku.edu.tw) †[jkrvivian@gmail.com,](mailto:jkrvivian@gmail.com) ‡[twzjwang@gmail.com,](mailto:twzjwang@gmail.com) §[yanghau@biilabs.io](mailto:yanghau@biilabs.io)

**Abstract**

Transacting IoT data must be different in many respects to build much-needed trust in data marketplaces, trust that will be key to their sustainability. Data generated internally to an orga- nization is usually not enough to remain competitive, enhance customer experience, and improve strategic decision-making. In this paper, we propose a novel approach to construct IoT-enabled data marketplace by decentralized and trustless architecture, posting the trade records but also including the transaction process on the distributed ledgers. It can efficiently enhance the degree of transparency, since all interactions with smart contracts will be written on-chain. Storage via an end-to-end encrypted message channel allows emitting and accessing trusted data stream over the distributed ledgers regardless of the size or cost of the device, simultaneously making a verifiable Auth-compliant request to the platform. Furthermore, the platform will complete matching, trading and refunding process without the human intervention which also protects the rights of data providers and consumers through a trading policy written on the smart contract, and finally apply evolutionary game theory to the machine economy.

**Key words:** streaming data, crowd sensing, data marketplace, decentralization

**Introduction**

The growth of data marketplaces is an inevitable result of the IoT (Internet of Things) revolution. As physical assets such as ships, factories, vehicles, farms and buildings become digital, their digital twins will gradually act as secure data exchanges. [[1]](#_bookmark15) [[2]](#_bookmark16) As data streams surge across silos and carry value across organizations, traditional value chains will transition into a web of value. This paradigm shift will be more complex to administer, forcing businesses to rethink their competitive play as part of these ecosystems. Data marketplaces will emerge as a means to exchange data, monetize data streams and provide the basis of new business models. We refer to this new wave of value creation, for the Internet of Everything, as the ”Economy of Things.” There are three main barriers to achieving data marketplace:

1. Data owners do not have much control over their data and their data is locked in silos managed by products and services companies.
2. Data owners only have access to their own data which has little value when it comes to knowledge discovery.
3. Data owners do not know how to discover knowledge from raw data.

To overcome these barriers, we implemented IoT-enabled data marketplace and sensor data submission functionalities which are intended to be very lightweight and capable of running on embedded devices. They will

only need to perform Tangle operations (e.g., producing and consuming secure chan- nels) and communicate with decentralized facilities, which do not rely on single-source network infrastructure. This proposed reference architecture includes functions that could be mapped to different stakeholders, and multiple functions can be implemented by the same administrative stakeholder in a given operational deployment.

1. Data Sellers are entities that deploy an IoT infrastruc- ture, for example smart energy meters. These entities are interested in selling the collected data or subsets of that data.
2. Managed Data Lakes would typically store a massive amount of data and metadata to enable data discovery.
3. Data Buyers consuming data streams or downloading data sets are interested in the additional value that external data can bring to their internal data.

Take the use case of the Airbox [[3]](#_bookmark17) for instance. Every household with an Airbox device can collect air quality records automatically and autonomically, rather than passively accept the outcomes from the centralized authorities. To protect privacy, data should be encrypted before on-chain. As for data reliability, we extended the backbone design of the Airbox to interoperate smart-contract-oriented structure to write down every transaction. At the same time, data on-chain will send a verifiable request to mark itself as ”being tradeable.” This step enables buyers to review and bargain at will. Last but not least, payments will be stored on the smart contracts until the transactions are confirmed. The entire sequence is illustrated in Figure [1.](#_bookmark0)

**Related Work**

As the economic value of huge amount of data emerges, several researchers have started to explore the design of data marketplace. The Third Party Auditors (TPAs)-based frameworks [[4]](#_bookmark18) is far from being satisfactory due to the unstan- dardized data format dynamic nature of IoT data. Therefore a decentralized data integrity validation and trading process has been proposed recent years, and distributed ledger technology (DLT) is considered a solution, which has the immutable data storage feature that the existence of data can be trusted and no longer rely on any third party authority.

Data Integrity as a Service (DIaaS) is a blockchain based framework for data integrity proposed by Liu et al. [[5]](#_bookmark19) which is a Cloud Server Service (CSS) which allows both data provider and consumer to validate data integrity by comparing hashes on Ethereum smart contracts [[6]](#_bookmark20) and

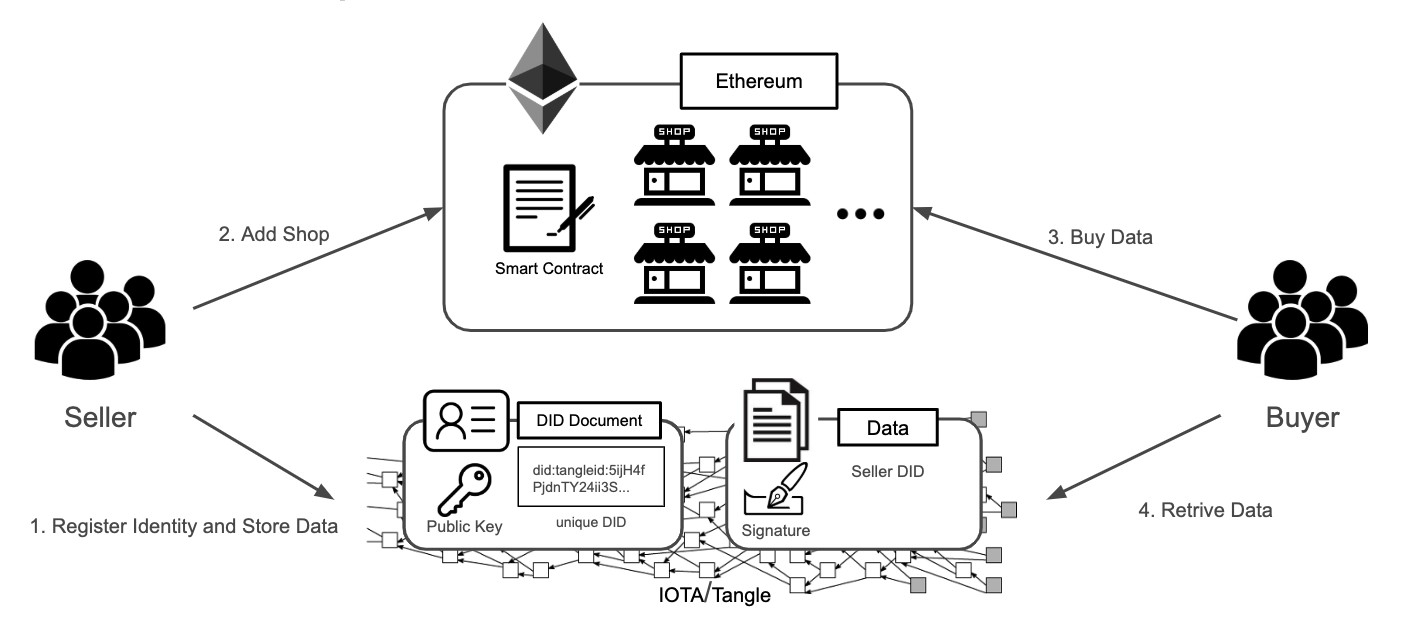


Figure 1. IoT-enabled personal air quality assistant.

cloud servers. However, the architecture of DIaaS needs high trust on the CSS. If the CSS is malicious or has crashed, losses to both providers and consumers will result. Also, performance analysis shows that IoT devices have low efficiency interacting with Ethereum due to the time-consuming Proof-of-Work(PoW) process.

To address the inefficiency of the Blockchain, IOTA [[7]](#_bookmark21) is a cryptocurrency for the IoT industry based on a revolu- tionary distributed ledger technology, the Tangle, which en- ables zero-transaction. On top of that, Masked Authenticated Messaging(MAM) [[8],](#_bookmark22) is provided for a second layer data communication protocol which allows transmission, access, and verification of an encrypted data stream over the Tangle where privacy and integrity meet.

Based on Tangle and MAM, the IOTA foundation launched a data marketplace [[9]](#_bookmark23) which is suitable for IoT streaming data that not only allows data providers to put data on Tangle without any trusted cloud services, but also allow providers and consumers to trade on Tangle with privacy protected and assurance of data integrity from source with MAM. Nevertheless, the platform design is centralized, new devices require manual approval from the IOTA foundation to be visible in the marketplace. Also, interacting with Tangle is still the bottleneck for low-level devices especially in an unstable network or electrically noisy environment.

A different framework design proposed by Gupta, S.Kanhere and Jurdak [[10]](#_bookmark24) could reduce the burden of low- level devices as mentioned. The infrastructure is a 3-tier decen- tralized data marketplace architectural design with Ethereum smart contracts which consist of providers, consumers and brokers. The broker is a trustless and highly resourced device that will facilitate the trading of data between the consumer and providers. However, the data integrity and authentication of each participant is still forthcoming.

The sustainability of IoT economic system was under dis- cussion by Dusit Niyato et al. [[11].](#_bookmark25) They evaluated the utility structure of data trading and presented game theory model. Data marketplace organizers can determine their policies to ensure a sustainable system with the Nash Equilibrium found by game theory and the utility structure. However, refund and subscription economy were not mentioned in their work.

Such work proposed different solutions to specific issues of data marketplaces. In this paper, we had an overall design of decentralized data marketplace that handles data integrity and trading procedures such as buyout and subscriptions to future data on a distributed ledger.

**System Architecture**

Our proposed data marketplace framework is a 3-tier de- centralized architecture with a registrar who is responsible for marketplace registration in order to post participants’ infor- mation on distributed ledgers for validation, data providers who publish data, data consumers who search for interested products and issue a new trade, and brokers who interact between data providers and consumers, including data pub- lishing, product metadata generation and trading process.

1. *Participants*

There are four major roles in the decentralized data market- place (Figure [2).](#_bookmark1)

* 1. *Registrar*: The registrar is responsible for creating a

Registration Contract, which maintains a lookup table of participants, and has the authority to control the participant lists of the decentralized data marketplace.

* 1. *Data Provider*: Data providers, who generate and pre-

serve streaming data, are willing to sell streaming data to consumers and receive subscription fees from consumers, which can be used to improve the quantity and accuracy of their device or service.

* 1. *Consumer*: Consumers aspire to obtain streaming data to

promote the value of their service. However, it is a significant challenge for most consumers to collect the desired data by themselves. So they look to purchasing the streaming data from data providers.

* 1. *Broker*: Brokers represent data providers and consumers

to perform computing tasks as brokers are expected to have higher resource. Some trustworthy brokers who pass proce- dures for conformity assessment are added to the decentralized data marketplace. Once a qualified data provider requests to launch a new product, the broker is requested to deal with the trading process and publish the providers data streams to the MAM channel. Brokerage fees for each product will be charged by the broker.

1. *Components*
   1. Masked Authenticated Messaging: IOTA [[7]](#_bookmark21) is a feeless and scalable cryptocurrency while MAM is the second layer data communication protocol built on top of Tangle.  
       MAM resolves the challenge to publish encrypted streaming data to distributed ledgers. It publishes messages encrypted with a session key to channel, which is a form of transactions that each address can be derived from the previous one. Therefore, with channel root, the first message of a channel, and the session key, all data on the channel is accessible.  
       One can create multiple channels as possible, however, the size of a channel is fixed which is decided before creation, thus data providers should decide how to distribute data product into MAM channels. While the creation of a MAM channel is time-consuming, brokers are also responsible for channel creation, encrypted data publishing and session key certification in our system. See Figure [3](#_bookmark2)
   2. TangleID: The TangleID [[12]](#_bookmark26) is a self-sovereign identity system based on IOTA that do not require any third-party authority to verify an identity and its digital footprint. With TangleID, the digital footprint is converted into digital assets under the principle of Decentralized Identifiers (DIDs) [[13]](#_bookmark27) defined by W3C.

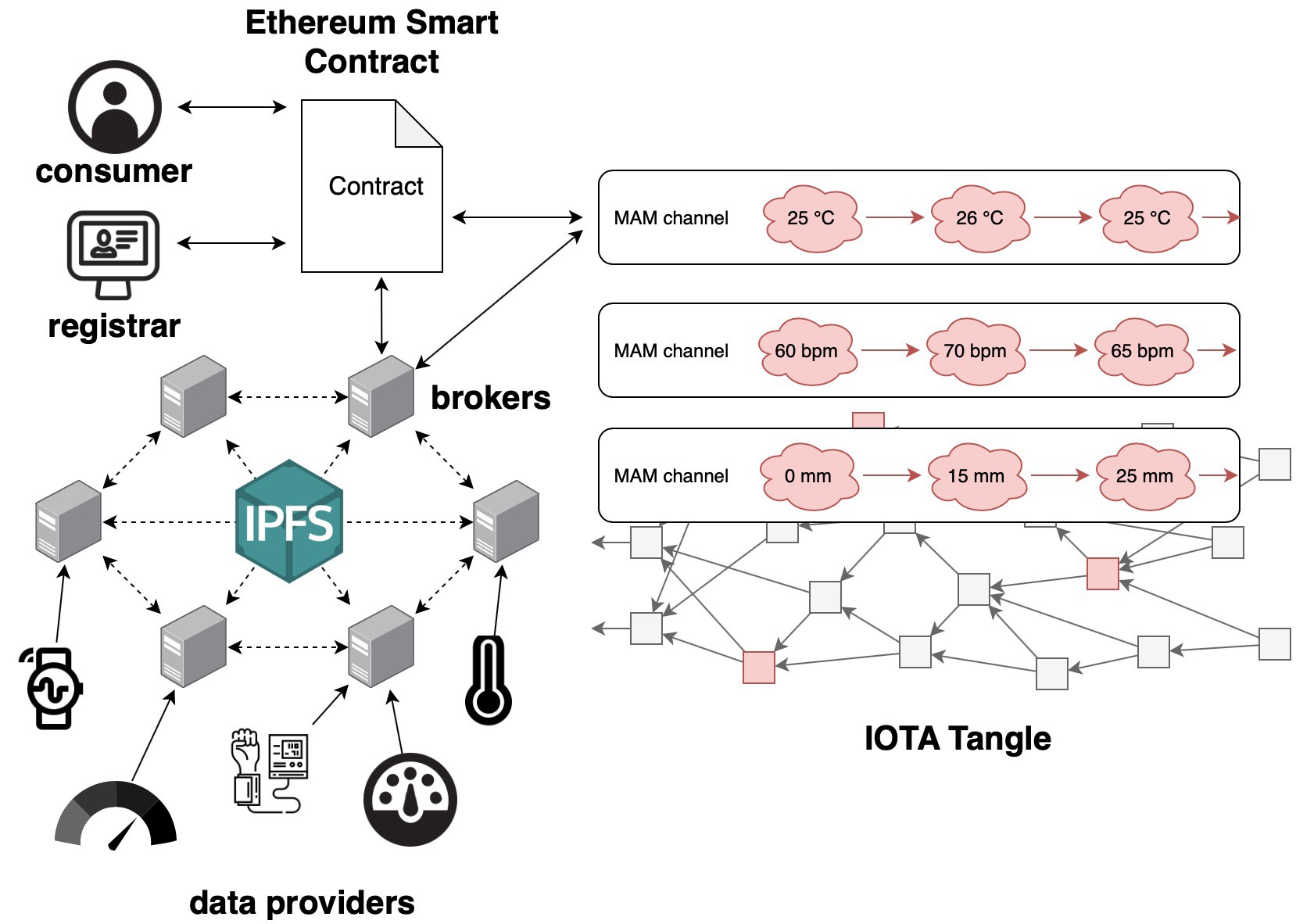


Figure 2. The system design of a decentralized architecture which consist of a registrar, data providers, consumers and brokers.

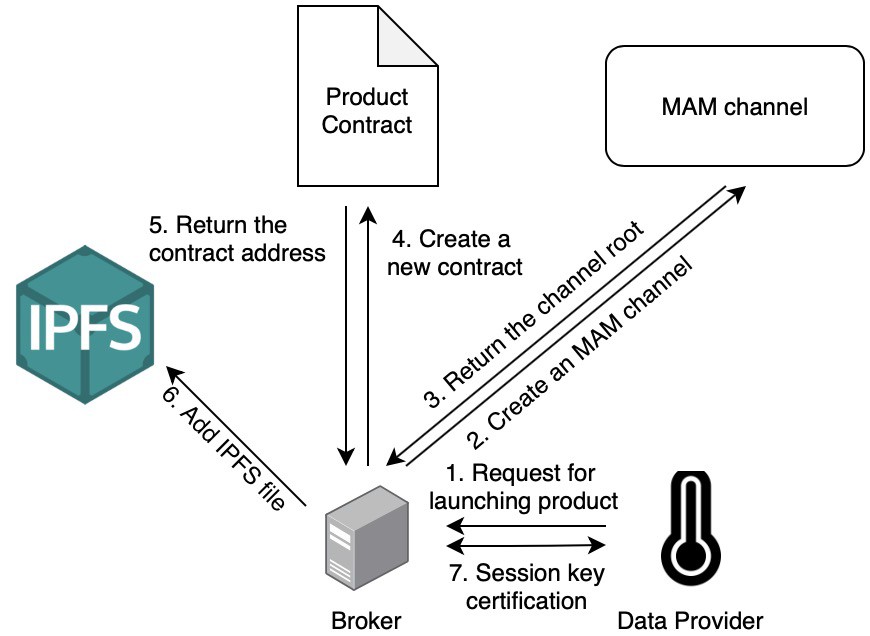


Figure 3. The process of launching a product.

Posting DID documents on MAM makes TangleID a GDPR-Complianced system. Every participant in the data marketplace registers on TangleID, hence one can easily verify data providers’ identity to ensure the data persistency from data sources.

* 1. *Ethereum Smart Contract:* Smart contract is a protocol for formulating agreement on a blockchain that provides ver- ification and execution of the contract. The code in the smart contract can interact with other contracts, make decisions, store data and transfer cryptocurrency. All conditions and states established in the contract are transparent and with enforcement. The appearance of smart contracts makes trading more flexible, and achieves more complex trading patterns in reality.
  2. *Blind Signature:* It is a risk revealing session keys to brokers since contents may be copied by brokers, which would result in data providers’ losses. Therefore, a blind signature is used to prevent session key copying for such circumstances. Blind signature [[14]](#_bookmark28) is a form of digital signature where the message is first ”blinded” by a random ”blinding factor”, then passed to a signer to sign. The resulting message, along with the blinding factor, can be later verified with the signer’s public key. In our system design, brokers would perform blind signature during the process of adding new products for data providers, in order to send the secret key of the MAM channel to smart contract without knowing it.
  3. IPFS: Inter-Planetary File system(IPFS) [[15]](#_bookmark29) is a peer- to-peer network for storing and accessing files, websites, applications, and data in a distributed file system which is not maintained with certain nodes or entities but all IPFS users. In our proposed architecture, brokers are responsible to publish the metadata of products, including title, data provider information and data preview to the IPFS in order to provide users with search capabilities to meet the consumers’ need.

**Trading Model**

In the following, we describe the data trading process in detail and we would use game theory to ensure the sustain- ability of our tading model. To participate a data marketplace, data providers and consumers have to register first. Then data provider can launch its product on the marketplace. Once a product is launched, it is searchable and can be subsequently traded. The whole trading and refunding process is defined in smart contracts which are easily traceable and irreversible.

The consumer will pay for the data, only when the data sold by the data provider sufficiently accurate, for which we called such data ”decent data.” Once the accuracy is lower than a certain threshold, which we call ”unacceptable data,” the consumers would then consider this data provider as a low quality data provider, and stop buying data from this data provider.

Refunds are a major issue in our research. However, we don’t have to take refund as a factor in our game theory model, since in our decentralized data marketplace we use smart contract to store the subscription fee which will be paid to buy the future data. The subscription fee will be paid as new data is transmitted to consumers. In other words, data provider doesn’t need to take any procedure to transfer subscription fee from his/her own account to consumers’ account. For the same reason, data provider has no responsibility on the refunding processing fee charged by the smart contract as well.

1. *Participants Registration*

At the beginning, the registrar creates a Registration Con- tract, which maintains all participants’ information, including their DID documents and public keys. Everyone can query participants’ public keys. Those who would like to sell or pur- chase data may register to become a data provider or consumer. The registrar has the authority to agree with applications. After that, their identities are available in the Registration Contract and the launching and trading processes can begin.

1. *Launching and Searching Products*

To sell streaming data, a data provider needs to launch a new product on the data marketplace in advance as shown in Figure [3.](#_bookmark2) The data provider determines a trusted broker and asks the broker to create a new MAM channel and Product Contract. Each product has a Product Contract to record the participants, subscription price brokerage fee, quantity of data and trading process.

Brokers certify session keys as well. Only one session key can be signed in each product, so data providers cannot fake a session key to deceive consumers. Figure [4](#_bookmark3) shows the certification process. When a data provider asks a broker to certify new session key k, the data provider uses the broker’s public key, which is available in the Registration Contract

after the broker’s registration, as a blinding factor, and the session key is blinded. Then the data provider sends the blinded session key Blind(k) to the broker. The broker signs the message and returns the signature Sign(Blind(k)) to data provider. The data provider removes the blinding factor and obtains the broker’s signature of the session key, Sign(k), which is verifiable by consumers.

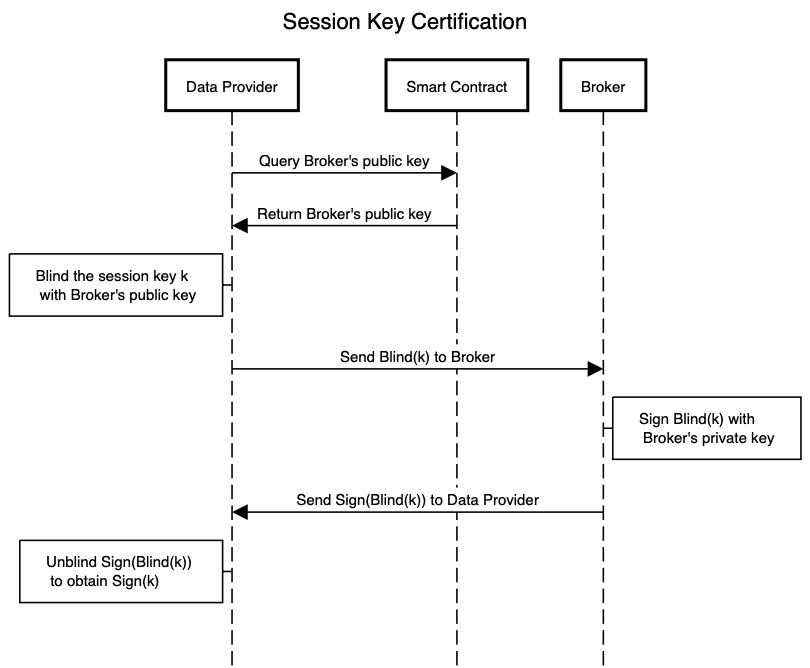


Figure 4. Session key certification process with blind signature.

The contract address and product description will be stored in a file which is uploaded to IPFS. Consumers can search the desired product by keywords or tags. The consumer then evaluates the product and initiates the trading with the data provider if the consumer is interested in subscribing to the data.

1. *Trading*

The entire trading process is as shown in Figure [5.](#_bookmark4) Once a consumer who wants to subscribe certain streaming data that is generated in the future pays a subscription fee to the Product Contract, it is added to the consumers list automatically by the smart contract.

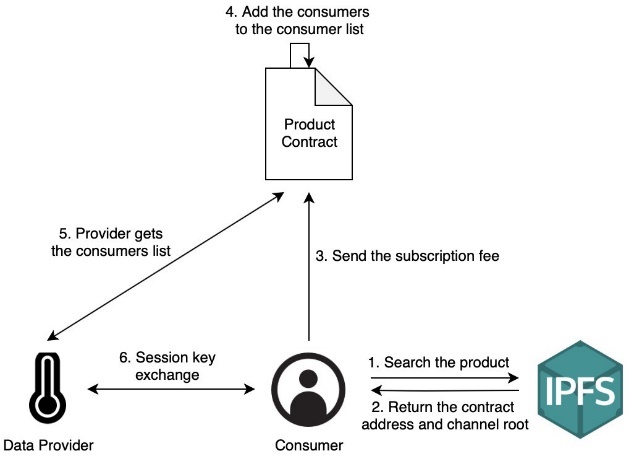


Figure 5. The process for the product trading.

Next, session key k should be exchanged between the data provider and consumers as shown in Figure [6.](#_bookmark5) The data provider can obtain public keys of each consumer from the

Registration Contract. For each consumer, the data provider encrypts the session key and broker’s signature with the consumer’s public key and sends the ciphertext Encrypt(k + Sign(k)) to the Product Contract. Consumers listen to the smart contract event which is triggered when the ciphertext is updated, and decrypt the ciphertext to obtain the session key k and signature Sign(k).

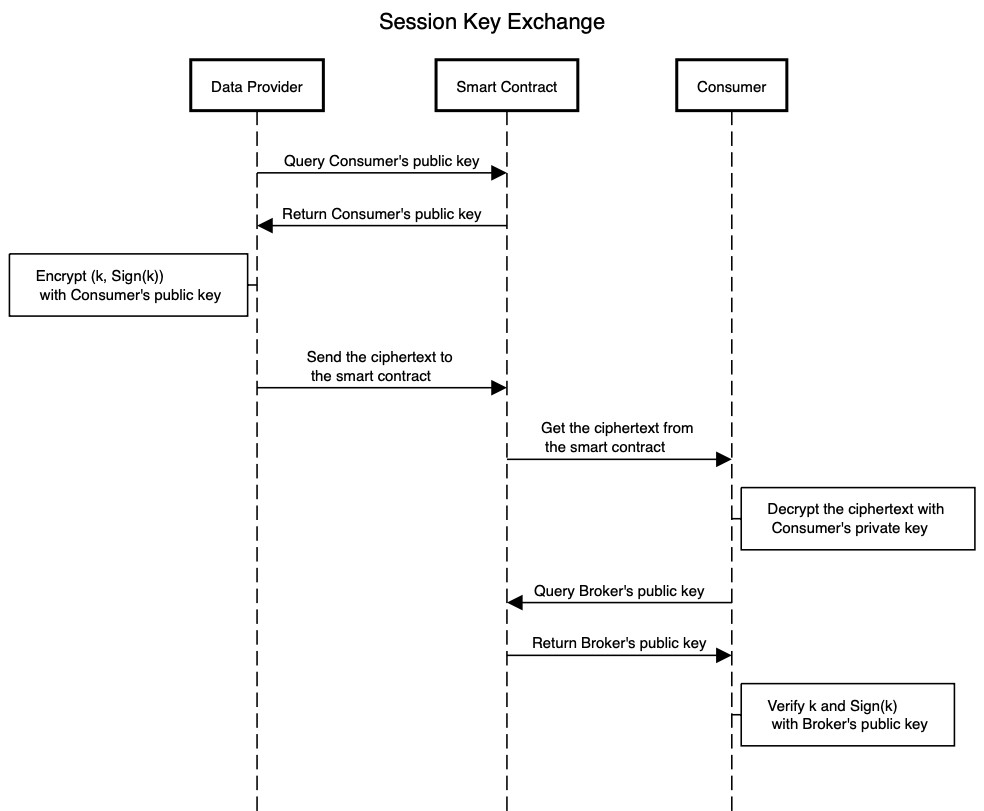


Figure 6. Session key exchange process between the data provider and consumer.

Consumers can obtain the broker’s public key on the Reg- istration Contract as well, so they can verify that the signature is valid and the session key is the only one that is certified by the broker. Afterward, encrypted data is published to the MAM channel, and consumers can obtain and decrypt it with the session key.

1. *Refunding*

It is probable that the streaming data sources are delayed or even interrupted after the consumers pay the subscription fee. To protect consumer rights, the subscription fee are not transferred to the data provider until data is generated and published to the MAM channel. If the expected data is not available, consumers can request refunds. We assume that a very small percentage of consumers in the Product Contract are irrational and/or malicious. Each consumer can vote for a refund at any time. If the ratio of consent votes of refunding is higher than the threshold at the ith piece of data, the subscription fee is proportionally transferred to the data provider, broker and every consumer. The subscription fee can be prorated as below:

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |

where price is the subscription price, M is the number of expected data samples, Fb (%) is the brokerage fee which

is expressed as a percentage, Ft is the transaction fee of the smart contract, N is the number of consumers in this contract.

To refund or withdraw subscription fee from the smart con- tract, data provider, broker, and consumer send a transaction to execute the smart contract and are responsible for transaction fee. We assume that only a half of the expected records are published to the MAM channel. When Fb is 5%, the data provider and broker can withdraw a half of total subscription fee from the smart contract and 5% of the subscription fee belongs to the broker while the remaining belongs to the data provider. For consumers, they can get half of the subscription fee refunded which should be deducted the transaction fee.

On the other hand, we would consider the situation that one of the consumers requests a refund in our future work. When the consumer is disappointed with the data quality, it may request a refund. Its permit should be cancelled while other consumers are unaffected.

1. *Game Theory Evaluation*

Game Theory is a methodology to discuss strategic interaci- ton among game players. If we can ensure Nash Equilibrium of decentralized data marketplace exists at certain acceptable range, then we can promise the sustainability of decentralized data marketplace. Fan Liang et al [[16]](#_bookmark30) listed several different types of game theory models which are applied on data pricing. We employed repeated game to build our game theory model. Repeated game consists of several repetitions of the same base game which meets the scenario of data subscription.

Figure.7 shows the decision tree to depict the the repeated game we used. Each level in this decision tree represents each round of data transmission from data provider to consumers, and the consumers would pay subscription fee, , each person. The sum of all the subscription fee pays to data provider is denoted as .

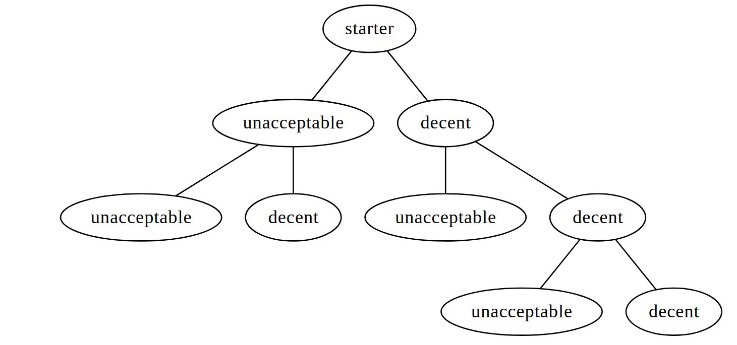


Figure 7. Decision Tree

For every round, the data provider would pay to enhance the sensors that data provider owns. We call this as a maintain period. Moreover, the value of data and the cost of maintenance will decrease as time pass, so we introduce a discounted factor to depict this phenomena.

We have these following assumptions:

1. We assume consumers are rational
2. Each data provider monopoly of the product they produced.
3. Subscription fee only depends on data quality.
4. At least 51\% of consumers are in the same group  
     
     
     
   which has complete information exchange.
5. One of the consumers who has complete information exchange with other 51\% buyers has the ability to examine the data quality.

According to [assumption(1)](#_bookmark7), we learn that few consumers would take malicious actions in this system, since all the consumers are rational.

[Assumption(2)](#_bookmark8) implies there is no other data providers provide the same product in Data Marketplace. In this way, we can consider each provider’s behavior independently. That means this decision tree represent the behavior of only one data provider.

From [assumption(4)](#_bookmark9), we derive one thing that once one of any consumer in that group launches voting for stopping buying data (in other words, the comsumer annonces the data quality is not acceptable), they would succeed in rejecting the processing subscription.

[Assumption(5)](#_bookmark10) implies if unacceptably low accuracy data are delivered to consumers, the examiner will spread this information out.

Based on the description above, the discounted sum of payoff for our repeated game is

|  |  |
| --- | --- |
|  | (4) |

Al-Fagih et al. presented [[17]](#_bookmark31) *Ps* is sigmoid function of *Pr* , and based on rule of thumb *Cmaintain* is about a exponential function of *Pr* . We can substitute *Ps* and *Cmaintain* with *Pr* into Eq. [(4),](#_bookmark11) then we can find out the Nash Equilibrium of repeated game.

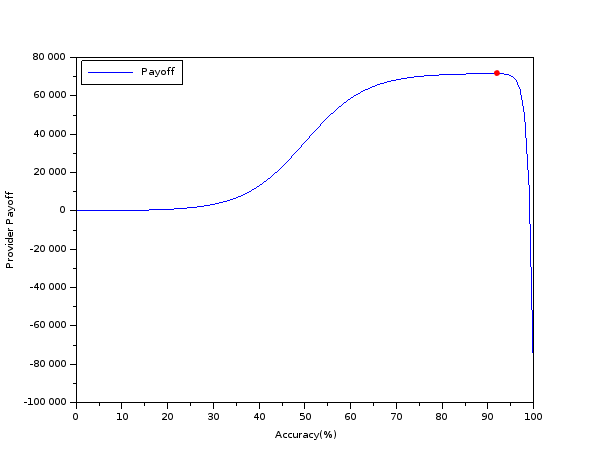


Figure 8. Payoff Function

1) *Numerical Example:* To evaluate the behavior of the model we presented, first, we would express the function of Ps and Cmaintain as function of Pr respectively. Therefore, the sum of subscription fee of all subscriber can be expressed as

|  |  |
| --- | --- |
|  | (5) |

where a and b are tuning factors, and R is the maximun  
  
  
  
  
 data value. And we would express the cost of maintenance as

|  |  |
| --- | --- |
|  | (6) |

Where c and d is tuning factor.

Substitute Eq. [(5)](#_bookmark12) and Eq. [(6)](#_bookmark14) into Eq. [(4).](#_bookmark11) We can derive

|  |  |
| --- | --- |
|  | (7) |

Let tuning factors , a = 0.15, b = 50, R = 10000, c = 1, d = 88, and there is only one maintenance period has happened and one maintenance period consists of 12 rounds which means m = 1, n = 12. Based on these parameters, we can plot the payoff function as function of accuracy, which is [Figure.8.](#_bookmark13) The maximun point occurs at accuracy ≈ 92% which we point it out with red dot.

[Figure.8](#_bookmark13) illustratres the market mechanism of the subscription trading policy we used under the decentralized data marketplace. First, data providers have weak motivation to operate their sensor in low accuracy, since the incentive of low quality data is mush less than moderate quality data. Second the great deficit at high accuracy is mostly drived by the rapid increasing of maintenance cost. Thus, if the maintenance fee to achieve decent data quality is under a fair price range, then data provider will automatically provide data with high enough data quality under our assumptions.

**Future Work**

The data auction process is a data trading mechanism and an economically-driven scheme that establishes corresponding prices of data products through bidding process between consumers and providers. While there has been many auction models [[18]](#_bookmark32) in several areas, only a few of research focus on third-party auction platform. Our proposed decentralized data marketplace protects the privacy of participants which reveals the minimum information for validation and reduces the suspicion of trust to centralized parties or auctioneers. However, the auction process between multiple participants and analysis of potential threats are still critical issues that need to be resolved. The game theory model presented in this work, we assume each data provider’s behavior is independent. However, normally there would be multiple data providers provide similar product (substitution). Only one provider is taken into considerations at present.

**Reference**

[1] K. Borodulin, G. Radchenko, A. Shestakov, L. Sokolinsky, A. Tchernykh, and R. Prodan, “Towards digital twins cloud platform: Microservices and computational workflows to rule a smart factory,” in Proceedings of the10th International Conference on Utility and Cloud Computing, ser. UCC ’17. New York, NY, USA: ACM, 2017, pp. 209–210.

[2] S.-W. C. Shao-Hua Wang and C.-C. J. Huang, “Puyuma: Linux-based rtos experimental platform for constructing self-driving miniature vehicles,” in Intelligent Computing. Springer International Publishing, 2019, pp. 985–994.

[3] S. Mahajan, W.-L. Wu, T.-C. Tsai, and L.-J. Chen, “Design and implementation of iot-enabled personal air quality assistant on instant messenger,” in Proceedings of the 10th International Conference on Management of Digital EcoSystems, ser. MEDES ’18. New York, NY, USA: ACM, 2018, pp. 165–170.

[4] C. Liu, C. Yang, X. Zhang, and J. Chen, “External  
 integrity verification for outsourced big data in cloud and iot: A big picture,” Future Generation Computer Systems, vol. 49, pp. 58 – 67, 2015. [Online]. Available:   
  
  
  
http://www.sciencedirect.com/science/article/pii/S0167739X14001551

[5] B. Liu, X. L. Yu, S. Chen, X. Xu, and L. Zhu, “Blockchain based data integrity service framework for iot data,” in IEEE 24th International Conference on Web Services. Berlin, DE: Springer, 2017, pp. 3–17.

[6]ethereum/wiki.[Online].Available:https://github.com/ethereum/wiki/wiki/White-Paper

[7] S. Popov. Iota whitepaper. technical white paper. [Online]. Available: https://iota.org/IOTA Whitepaper.pdf

[8] Masked authenticated messaging. [Online]. Available: https://blog.iota.org/introducing-masked-authenticated-messaging-e55c1822d50e

[9] Iota data marketplace. [Online]. Available: https://data.iota.org

[10] R. J. Pooja Gupta, Salil S. Kanhere, “A decentralized iot data marketplace,”2019.

[11] D. Niyato, D. T. Hoang, N. C. Luong, P. Wang, D. I. Kim, and Z. Han, “Smart data pricing models for the internet of things: a bundling strategy approach,” IEEE Network, pp. 18–25, 2016.

[12] Tangleid. [Online]. Available: https://tangleid.github.io

[13] Decentralized identifiers. [Online]. Available: <https://w3c>  
-ccg.github.io/did-spec/

[14] D. Chaum, “Advances in cryptology proceedings of crypto,” 1983.

[15] J. Benet. (2014) Ipfs - content addressed, versioned, p2p file system.

[16] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, “A survey on big data market: Pricing, trading and protection,” IEEE Access, pp. 15 132 – 15 154, 2018.

[17] A. E. Al-Fagih, F. M. Al-Turjman, W. M. Alsalih, and H. S.  
 Hassanein, “A priced public sensing framework for heterogeneous iot architectures,” IEEE Transactions on Emerging Topics in Computing, pp. 133–147, 2013.

[18] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, “A survey  
 on big data market: Pricing, trading and protection.” IEEE, 2018,  
 pp. 15 132–15 154.