Decentralised Data Markets

Don Quijote de la Mancha, Sancho Panza, Dulcinea y Rocinante

I. Introduction

A. Preamble

In recent years there has been a shift in many industries towards data-driven business models [31]. Namely, with the advancement of the field of data analytics, and the increased ease in which data can be collected, it is now possible to use both these disruptive trends to develop insights in various situations, and to monetise these insights for financial gain. Traditionally, users have made collected data available to large platform providers, in exchange for services (for example, web browsing). However, the fairness and even ethics of this business modes continues to be questioned, with more and more stake-holders arguing that such platforms should recompense citizens in a more direct manner for data that they control [33] [17], [2], [3]. Apple has recently responded to such calls by introducing changes to their ecosystem to enable users to retain ownership of data collected on their devices. to At the time of writing, it was recently reported that these new privacy changes have caused the profits of Meta, Snap, Twitter and Pinterest plummet (losing a combined value of $278 billion since the update went into effect in late April 2021[1]). The privacy change introduced by Apple allows users to mandate apps not to track their data for targeted advertising. This small change has been received well amongst Apple users, with a reported 62% of users opting out of the tracking [4]. Clearly this change will have a profound impact on companies relying on selling targeted advertisements to the users of their products. Users can now decide how much data they wish to provide to these platforms and they seem keen to retain data ownership. It seems reasonable to expect that in the future, companies wishing to continue to harvest data from Apple will need to incentivise, in some manners, users or apps to make their data available.

The need for new ownership models to give users sovereignty over data is motivated by two principal concerns. The first one concerns fair recompense to the data harvester by large corporations. While it is true that users receive value from companies, in the form of the services their platforms provide (e.g., Google Maps), it is not obvious that this is a fair exchange of value. The second one arises from the potential for unethical behaviours that are inherent to the currently prevailing business models. Scenarios in which unethical behaviour have emerged arising out of poor data-ownership models are well documented. Examples of these include Google Project Nightingale [2] [3], where sensitive medical data was collected of patients that could not opt out of having their data stored in Google Cloud servers. The scale of this project was the largest of its kind, with millions of patient records collected for processing health care data. Another infamous case study was the Cambridge Analytica (CA) scandal that was released in 2015. Personal data of 87 million users was acquired via 270,000 user giving access to a third party app that gave access to the users' friend network, without these people having explicitly given access to CA to collect such data [4] [15]. CA has a vast portfolio of elections they have worked to influence, with the most notorious one being the 2016 US presidential elections that Donald Trump won [38], [22].

It is important to understand that these cases are not anecdotal. Without the adequate infrastructure to trade ownership, cases like the ones outlined above, with mass privacy breaches, have the potential to become more frequent. Apple's actions are an important step in the direction of giving individuals ownership over their data and potentially alleviating such issues, however, one may correctly ask why users should trust Apple, or any other centralised authority, to preserve their privacy and not trade with their data. Motivated by this background, and by this latter question, the authors argue for the shift towards a more decentralised data ownership model, where this ownership can be publicly verified and audited. The authors are interested in developing a data-market design that is hybrid in nature; *hybrid* in the sense that some non-critical components of the market are provided by trusted infrastructure, but where the essential components of the market place, governing ownership, trust, data veracity, etc., are all designed in a decentralised manner. The design of such markets in not new and there have been numerous attempts to design marketplaces to enable the exchange of data for money [32]. This, however, is an extremely challenging

[1] https://www.bloomberg.com/news/articles/2022-02-03/meta-set-for-200-billion-wipeout-among-worst-in-market-history

[2] https://www.bbc.co.uk/news/technology-50388464

[3] https://www.theguardian.com/technology/2019/nov/12/google-medical-data-project-nightingale-secret-transfer-us-health-information

[4] https://www.nytimes.com/2018/04/04/technology/mark-zuckerberg-testify-congress.html

endeavour. Data cannot be treated like a conventional commodity due to certain properties it possesses. It is easily replicable; its value is time-dependant and intrinsically combinatorial; and dependent on who has access to the data set. It is also difficult for companies to know the value of the data set a priori, and verifying its authenticity is challenging [1] [5]. These properties make marketplace models difficult to design and have been an emergent research area.

In what follows, the authors describe a first step in the design of a marketplace where data can be exchanged, and where certain guarantees can be provided to the buyer and seller alike. More specifically, the goal is to rigorously define and address the challenges related to the tasks of selling and buying data from unknown parties, whilst compensating the sellers fairly for their own data. As mentioned, to prevent monopolisation, a partially decentralised setting will be considered, focusing on the important case of data rich environments, where collected data is readily available and not highly sensitive. Accordingly, this work focuses on the automotive industry data collection, in order to provide a specific case study that, the authors hope, might represent a first step towards more general architectures.

B. Specific motivation

As discussed in the preamble I-A, the authors focus on a class of problems where there is an oversupply of data but where there is a lack of adequate ownership methods to underpin a market. One example of such a situation is where agents collaborate as part of coalitions in a crowd sourced environment to make data available to potential buyers. It is applications of this nature that will be the focus in this paper. More specifically, the interest is placed in the context of a city where drivers of vehicles wish to monetise the data harvested from their car's sensors. An architecture is proposed that enables vehicle owners to sell their data in coalitions to buyers interested in purchasing their data.

While this situation is certainly a simplified example of a scenario in which there is a need for data market, it is of great interest for two reasons. Firstly, situations of this nature prevail in many application domains. Scenarios where metrics of interest can be aggregated to generate a data rich image of the state of a given environment are of value to a wide range of stakeholders, which, in the given context, could include anyone from vehicle manufacturers, mobility and transport companies to city councils. Secondly, this situation, while simplifying several aspects of the data-market design, captures many pertinent aspects of more general data-market design: for example, detection of fake data; certification of data-quality; and resistance to adversarial attacks.

The authors see the context of automotive data collection as ripe for opportunity to develop a decentralised data market. The past decade has seen traditional vehicles transition from being a purely mechanical device to a cyber-physical one, having both a physical and digital identity. From a practical viewpoint, vehicles are quickly increasing their sensing capabilities, especially given the development of autonomous driving research. Already, there is an excess of useful data collected by the latest generation vehicles, and this data is of high value. According to [20] *"car data and shared mobility could add up to more than $ 1.5 trillion by 2030"*. Such conditions prevail not only in the automotive sector; for example, devices such as smart-phones; smart watches; modern vehicles; electric vehicles; e-bikes and scooters; as well as a host of other IoT devices, are capable of sensing many quantities that are of interest to a diverse range of stakeholders. In each of these applications the need for such marketplaces is already emerging. Companies such a Nissan, Tesla, PSA and others have already invested in demonstration pilots in this direction and are already developing legal frameworks to develop such applications[5] in anticipation of opportunities that may emerge. As mentioned, the main issue in the design of such a data market lies in the lack of an adequate ownership method. Who owns the data generated by the vehicle? The answer is unclear. A study by the Harvard Journal of Law & Technology concludes that most likely, it is the company that manufactured the car who owns the data, even though the consumer owns the smart car itself [40]. According to the authors of the study, this is because the definition of ownership of data is not congruent to other existing definitions of ownerships such as intellectual property, and therefore the closest proxy to owning a data set is having the rights to access, limit access to, use, and destroy data. Most importantly consumers do not have the right to economically exploit the data they produce. Nonetheless, EU GDPR laws expressly state that users will be able to transfer their car data to a third party should they so wish. According to [34] "The data portability principle was expressly created to encourage competition". However, if the data is owned by the automobile company, how can consumers verify who has access to their data? Placing trust assumptions on the car manufacturer should be justified. Given the lack of verifiability of a centralised authority, such as a car manufacturing company, the authors propose exploring decentralised or hybrid alternatives.

[5]https://www.aidataanalytics.network/data-monetization/articles/tesla-automaker-or-data-company

The author's objective in this paper is directly motivated by such situations and by the issues described so far. However, as previously discussed, rather than enabling manufacturers to monetize this data, the authors are interested in situations where device owners, or coalitions of device owners, own the data collected by their devices, and wish to make this data available for recompense. This is fundamentally different to the situation that prevails today, whereby users make their data freely available to platform providers such as Google, in exchange for using their platform. Nevertheless, the recent actions of Apple suggest that this new inverted (and emancipated) business model, whereby providers compete and pay for data of interest, could emerge as an alternative model of data management, and also whereby users are able to control and manage the data that they reveal. Given this background context, we are interested in developing a platform whereby data owners can make available, and securely transfer ownership of data streams, to other participants in the market. To do this a data-market architecture is proposed that borrows technology elements from distributed ledgers, and merges these with consensus algorithms and game theory.

C. Related Work

1) Decentralised vs Centralised Data Markets: Numerous works have proposed decentralised data markets. While many of these proposals use Blockchain architectures for their implementations, many simply utilise the underlying technology and fail to address Blockchain design flaws as they pertain to data markets [24] [14] [37]. For example, Proof-of-Work (PoW) based Blockchains reward miners with the most computational power. Aside from the widely discussed issue of energy wastage, the PoW mechanism is itself an opportunity, hitherto that has not been utilised, for a data-market to generate work that can be useful for the operation of the marketplace. As we shall shortly see, the PoW mechanism can itself be adapted to generate work that is useful for the operation of the marketplace. In addition, Blockchain based systems, also typically use commission based rewards to guide the interaction between users of the network, and Blockchain miners. Such a miner-user interaction mechanism is not suitable in the context of data-markets, effectively prioritising wealthier users' access to the data-market. In addition, miners with greater computational power are more likely to earn the right to append a block, and thus earn the commission. This reward can then be invested in more computational power, leading to a positive feedback loop where more powerful miners become more and more likely to write blocks and earn more commissions. Similarly, the wealthier agents are the ones more likely to receive service for transactions of higher monetary value. This could cause traditional PoW-based Blockchains to centralise over time [6]. Indeed, centralised solutions to data markets already exist, such as [7], which namely focus on implementing methods to share and copy data, and certain rights to it, such as read rights. Considering the aforementioned properties of PoW-based Blockchains, the authors explore other distributed ledger architectures to implement a decentralised data market.

2) Trust and Honesty Assumptions: Another possible categorisation of prior work relates to the trust assumptions made in the system. The work in [25] assumes that upon being shared, the data is reported truthfully and fully. In practise, this assumption rarely holds, and a mitigation for malicious behaviour in shared systems must be considered. This assumption is justified in their work by relying on a third party auditor, which the authors of [37] also utilise. However, introducing an auditor simply shifts the trust assumption to their honest behaviour and forgoes decentralisation.

In [1], it is identified that the buyer may not be honest in their valuation of data. They propose an algorithmic solution that prices data by observing the gain in prediction accuracy that it yields to the buyer. However, this comes at the cost of privacy for the buyer: they must reveal their predictive task. In practise, many companies would not reveal this IP, especially when it is the core of their business model. The work of [21] is an example of a publicly verifiable decentralised market. Their system allows for its users to audit transactions without compromising privacy. Unfortunately, their ledger is designed for the transaction of a finite asset: creating or destroying the asset will fail to pass the auditing checks. For the transaction of money this is appropriate: it should not be possible to create or destroy wealth in the ledger (aside from public issuance and withdrawal transactions). However, for data this does not hold. Users should be able to honestly create assets by acquiring and declaring new data sets they wish to sell. Furthermore, their cryptographic scheme is built to transfer ownership of a single value through Pedersen commitments. A value corresponding to the amount of assets one agent wishes to transfer to another is chosen, and a series of mathematical operations are performed with it, following the procedure of Non-Interactive Zero-Knowledge Proofs.

There is a need to have trust assumptions in components of the data market, whether it be centralised or decentralised. However, the authors believe that the users of the data market should agree on what or who to trust. A consensus mechanism is a means for a group of agents to agree on a certain outcome. For users to trust the consensus mechanism, they must have a series of provable guarantees that it was executed correctly. Users should be able to audit and verify the functioning of the consensus mechanism. It is not sufficient for the consensus mechanism to function correctly, it should also

prove this to the users. Indeed, it can be observed how, in binding governmental elections, voters have means to verify that the election was honest, and if this is not the case, voters have grounds to contest the outcome of the election.

The authors advocate for placing the trust assumptions in consensus mechanisms that can be verified. In other words, the users of a data market should have a means to agree on what they trust, and they should have a means to verify that this agreement was reached in a correct, honest manner.

In fact, this verification should be decentralised and public. Shifting the trust to a third-party auditing mechanism to carry out the verification can lead to a recursion problem, where one could continuously question why a third, fourth, fifth and so on auditing party should be trusted, until these can generate a public and verifiable proof of honest behaviour.

3) Consensus Mechanisms: I need to finish this section and discuss the algorithms mentioned here as well: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8629877 this section is still in progress Consensus in the context of a distributed mechanism can be mapped to the fault-tolerant state-machine replication problem [26]. Each machine/node in the distributed system maintains a view of what is its overall state. The nodes must then agree on which one is the unique accepted state, in a scenario where it is unknown which nodes are honest or faulty. (ie Byzantine agreement problem/ consensus in a Byzantine environment).

Casper (149] Bitcoin-NG(1531 PeerCensus[154) Hybrid Consensus Protocol Tendermint [158]. Proof of authority (159] delegated proof of stake [160] Algorand[161]

PAXOS, PBFT, Randomized Byzantine Agreements, FPC, Tendermint, Hotstuff, Algorand.

Algorand is an example of a DLT that achieves consensus through voting. They use a simple voting mechanism. There must be at least 2/3 of the committee size that vote for a block for it to get approved. (Because its Byzantine) The lower the committee the more liveness but the less secure. To manage this trade off, the smaller the committee size the more votes they require for there to be consensus. Ultimately, Algorand uses majority rule. There is a nuance to notice here however. This consensus mechanism only has two possible outcomes: accept the transaction or not (timeout or temporary consensus). In the context of the data market, this would not suffice. The consensus mechanism is used to determine which agents are the most trusted to compute the average or median of the data collected of a certain location. In other words, the consensus mechanism is a means to delegate a computation to someone based on how much they are trusted. Agents may be more, or less trusted, with some being preferred over others. These preferences may be stronger or weaker too. Using a majority voting method that only yields two possible options fails to encapsulate this information and is known to exclude minorities. [citation-needed]

Algorand exploits this property because they desire to retain the voting power amongst the agents with the most stake. This is based on the assumption that the more stake an agent has in the system, the more incentive they have to behave honestly.

This assumption cannot be made in the data market. Owning more cars does not equate to being more trustworthy, and therefore should not increase an agent's voting power, or their chances of participating in decision making. In fact, owning more vehicles could be an incentive to misbehave in the data market and upload fake data, whether this be to mislead competitors or to force a favourable outcome for themselves as a malicious actor. Purposely reporting fake data in the context of mobility has been described in [27] and [36], where collectives reported fake high congestion levels to divert traffic from their neighbourhoods [6] This attack is known as *data poisoning* and is a known attack of crowdsourcing applications, usually mounted through a Sybil attack. Owning more vehicles makes mounting this attack easier. Therefore in this context, having more stake in the market cannot be assumed to be an incentive to behave honestly, and indeed can offer increased ease to behave maliciously.

Furthermore, the supra-majority voting scheme used Algorand means that minorities have less participatory power. Given that Minorities do not stand a chance to get elected in a majority voting scheme that competes against these types of companies. We don't want another data-opoly.

D. Structure of the Paper

Firstly the authors introduce a series of desirable properties that the market must satisfy. These are outlined in the design criteria section II. Subsequently, a high level overview of the working components of the data market are presented in section III, as well as describing how each functional component contributes to achieving the desired properties described in the preceding section.
Then the authors proceed to formalising definitions used in each component of the data market, as well as the assumptions made in V. This section describes in detail how each component of the data market works.
Finally, in section X, the authors describe the set of

[6]https://www.washingtonpost.com/local/traffic-weary-homeowners-and-waze-are-at-war-again-guess-whos-winning/2016/06/05/c466df46-299d-11e6-b989-4e5479715b54_story.html

attacks considered, and in section XII the robustness of the components of the data market are evaluated.

II. Design Criteria for the Data Market

Having discussed issues that pertain to and arise from poor ownership models, the authors present a series of desirable criteria that the data market should achieve. More specifically, the work here proposed, begins to address the following research questions that are associated with data market designs:

How to protect the market against fake data or faulty sensors?
Given an oversupply of data, how to ensure that everybody receives a fair amount of write access to the market?
How to enable verifiable exchange of data ownership?
How to select data points from all those available to add most value to the marketplace?
How to protect the marketplace against adversarial attacks?

Following directly from these open questions, the desirable criteria are defined as:

- *Decentralised Decision Making:* The elements of the marketplace pertaining to trust, ownership and veracity are decentralised and do not rely on trust from third parties .

- *Verifiable centralisation:* The infrastructure on which the data market relies that is centralised, can be publicly verified. The reader should note that this ensures trust assumptions are placed on components of the data market that can be publicly verified.

- *Generalised Fairness:* Access to the data market is governed by the notion of the potential value that a data stream brings to the market (as defined by a given application). This will determine which agents will get priority in monetising their data. Agents with equally valuable data must be treated the same and agents with data of no value should receive no reward. Further notions of fairness are considered and formalised by [30] under the defintion of Shapley Fairness, and described in V.7. These are the definitions of fairness that the authors use for this data market proposal.

- *Resistant to duplication of data:* The datamarket must not allow malicious attackers to earn reward by duplicating their own data. Precisely, a distinction must be made between preventing the monetisation of duplicated data, versus preventing data duplication, given that the latter is an open challenge.

- *Resistant to fake data and faulty sensors:* The data market must be resilient to data poisoning attacks, wherein adversaries collude to provide fake data to influence the network. Congruently, the datamarket must be resilient to poor quality data from honest actors with faulty sensors. Formally, the data for sale on the market must not deviate by more than a desired percent from the ground truth. For the purpose of this work, the ground truth is defined as data measured by centralised infrastructure. Given that this measurement is publicly verifiable (any agent in that location can verify that this measurement is true), this is considered an acceptable assumption.

- *Resistant to spam attacks:* The data market is not susceptible to spam attacks, and malicious actors are not in the condition of flooding and congesting the network with fake or poor quality data.

III. Architecture of the Data Market

Figure 1 depicts the functional components of the proposed data market. As can be observed in the figure, some components of the marketplace are decentralised, and some of the enabling infrastructure is provided by an external, possibly centralised, provider.

In a given city, it is assumed that a number of agents on vehicles are collecting data about the state of the location they are in. They wish to monetise this information, but first, it must be verified that they are real agents. They present a valid proof of their identity and location, as well as demonstrating that they information is timely and relevant.

The agents that successfully generate the aforementioned receive a validity token that allows to form spatial coalitions with other agents in their proximity. They then vote on a group of agents in their coalition that will be entrusted to calculate the agreed upon data of their corresponding location. The chosen agents do this by aggregating the coalition's data following a specified algorithm.

This procedure happens simultaneously in numerous locations of the city. At a given point in time, all the datasets that have been computed by a committee in a spatial coalition then enter the access control mechanism. One can consider the data a queue and the access control mechanism the server. Here, they are ranked in order of priority by determining which data provides the greatest increase in value to the datamarket. The coalitions with the most valuable data perform the least amount of work.

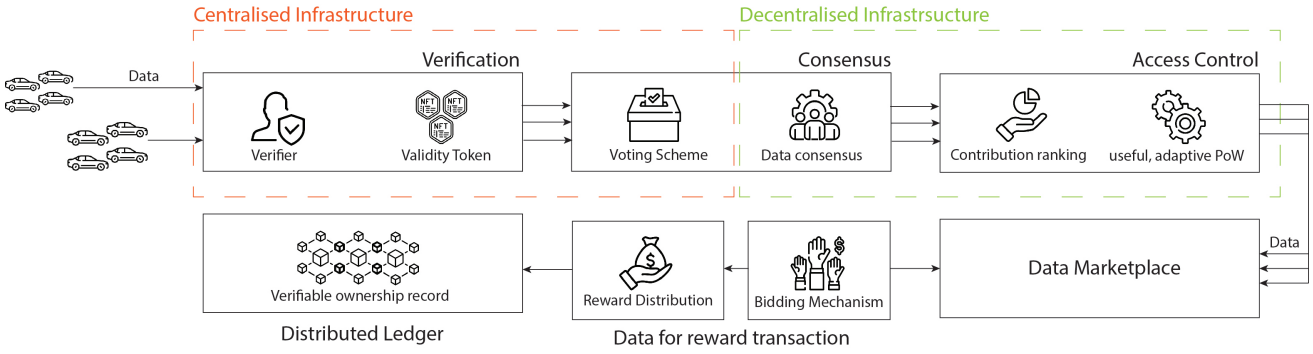

Figure 1. Data Market Architecture

Coalitions wishing to sell their data must complete a useful proof of work that is inversely proportional to their added value to the market. This PoW entails calculating the added value of new data in the queue. Once this work is completed, the data can be sold. The buyers are allocated to the sellers through a given bidding mechanism BM, and the reward of this sale is distributed amongst the sellers using the reward distribution function RD. Successful transactions are recorded on the distributed ledger, and the data corresponding to the successful transactions are removed from the data market.

IV. Preliminaries

The Architecture in Figure 1 makes reference to several technology components that may be new to the reader. To ease exposition these are now briefly described.

The Access Control Mechanism I : Agents must succeed in passing a verification stage to assert they meet the criteria to participate in the marketplace. This first verification stage prevents spam attacks.

The Consensus Mechanism: In a decentralised environment, agents must agree on what data is worthy of being trusted and sold on the data market. Agents express their preferences for who they trust to compute the most accepted value of a data point in a given location. The consensus mechanism will output a set of most trusted agents, who will then be tasked with computing the accepted datapoint of a given location using a specified algorithm. A series of options on how to compute this value are further discussed in section [?] and their properties are compared.

Access Control Mechanism II: Once datapoints are agreed upon for the given locations, it is necessary to regulate which ones should receive priority when being sold. Having an excess of available datapoints requires the system to determine which ones should enter the data marketplace to be sold first. Shapley value adaptive proof of work. Here the work carried out is useful work as it enables the functioning of the market. The work assigned is proportional to how valuable a datapoint is deemed, with respect to a given value function. This value function is the objective function of the collective. The more a datapoint contributes to achieving this objective, the less work the owner of said datapoint should have to do. The Shapley value is a means of assessing the marginal contribution of that data point towards increasing the combined worth of the marketplace

Distributed ledger :

Smart contract :

A. Distributed Ledger Technology

A distributed ledger technology (DLT) will be used to store the transactions of wealth in exchange for access (or any other given right) to a dataset.

A DLT is a decentralized database of transactions where these transactions are timestamped and accessible to the members of the DLT. They are useful to allow agents to track ownership, and are decentralized. Compared to a centralised storage system, this provides a geographically distributed, consensus-based, and verifiable system which is immutable after data has been written and confirmed. It is also a more resilient alternative against failure points.

There are many types of DLT structures, but they all aim to provide a fast, reliable, and safe way of transferring value and data. Namely, DLT's strive to satisfy the following properties: have only one version of the ledger state, are scaleable, make double spending impossible, and have fast transaction times.

One example of a DLT is the IOTA Tangle, shown in figure 2. In this DLT, all participants contribute to approving transactions, and the transactions are low to zero fee and near-instant. Further, decentralisation is promoted through the alignment of incentives of all actors [28].

In the Tangle by IOTA, a vertex represents a transaction

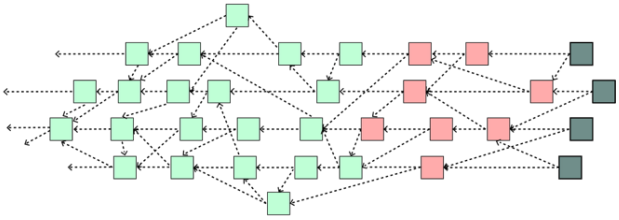

Figure 2. IOTA Tangle). credit: IOTA Foundation

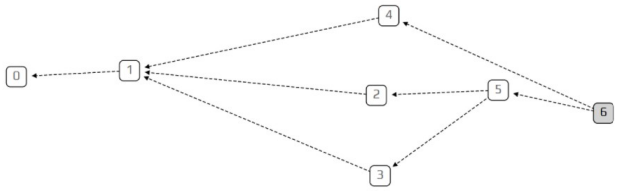

Figure 3. Directed Acyclic Graph (DAG). credit: IOTA Foundation

and an edge indicates that the transaction vertex at the tip of the edge has been verified by the transaction vertex at the tail of the edge. Therefore, there is a natural distinction between verified and unverified transactions, the latter are called tips. For a new transaction to be added to the Tangle, it needs to verify two transactions, which in case they were tips makes them verified transactions. In the figure below transaction 6 arrives and in order to be added it verifies the tips 4 and 5, and becomes itself a tip [12].

B. Access control mechanism

Because DLTs are decentralised, they need a method to regulate who can write information to the ledger and who cannot. An access control mechanism is necessary to protect the distributed ledger from spam attacks. One way is by using Proof-of-Work (PoW) as it is done in the Blockchain, where computationally intense puzzles need to be solved to be able to write to the ledger. In this case, users with more computational power earn the right to access the ledger. An alternative is Proof-of-Stake where nodes can stake tokens to gain the access rights proportional to their amount of staked tokens [13].

Another alternative is the mana based system used in the IOTA Tangle which, in combination with PoW to make Denial-of-Service (DoS) attacks expensive. Consequently, the PoW in the Tangle tends to be computationally much less expensive than the one used in PoW-based Blockchains and thus more resource efficient [41].

C. Useful Adapive Proof of Work

A big criticism of Bitcoin is its vast energy consumption. To put it in perspective, the Bitcoin network uses more than seven times as much electricity as all of Google's global operations.[7] By far the biggest share of this energy is used to solve the computational puzzels in its PoW algorithm. This motivates the use of useful adaptive PoW. It means that the carried out work itself is of use for the network and it is variable in difficulty. The question of what is useful work is context dependant and can for example calculating the mean be. To illustrate the varying difficulty can simply mean that one agent with higher difficulty has to compute multiple means whereas one with lower can complete the PoW by calculating only one mean.

We're interested in exploring alternatives to the existing access control mechanisms. Neither of the presented options above represent a logical reason to be granted more rights to access a ledger. This right should be earned irrespective of wealth in the ledger or computational power. We believe this right should be proportional to the agent's contribution in furthering the collective goal of the datamarket. This goal can be characterised mathematically, and we can evaluate each agent's marginal contribution and grant them rights proportionally.

D. Governance and Consensus

In the context of a decentralised system, there is a need to find a suitable method to aggregate the opinions or preferences of individuals to reach a consensus of the collective. Social choice theory provides a means to analyse how this aggregation of preferences is made [29], and voting mechanisms are tools designed to achieve this [43]. Indeed, what the authors are striving for is to create a data market where the participants govern themselves. The participants of the data market should have a means to deliberate and decide what data is most worthy for the collective, and which data is most trusted. This is not a novel idea by any means. In fact, the term *democracy* was first coined in Ancient Greece, and has its etymological origins in the combination of the words 'people' and 'rule' [10]. Since then, it has developed significantly, but the essence of the idea remains the same: it is a form of government where individuals elect representatives to make decisions on their behalf. Crucial to this system is the act of electing representatives. The amount of voting mechanisms available are vast, and there is no objectively best voting mechanism. Numerous voting scheme criteria have been formalised and one must first decide which properties are of most importance or relevance for the application in question, and then select a voting scheme that adheres to as many of these criteria as possible. In later sections the authors will discuss which of these are of most importance given the context of this work, and provide a series of options that best satisfy them.

1) :

[7] nytimes.com - bitcoin electricity consumption

2) Maximum Entropy Voting: Earlier in section I-C2 we discuss trust assumptions and how these should be allocated. Trusted Authorities (TAs) should either be able to publicly verify their honest behaviour, or there should be a consensus mechanism that elects said TAs, that can also be publicly verified.

Voting systems are a tool to achieve consensus, and can be roughly divided into deterministic and probabilistic systems. Maximum Entropy Voting (MEV) belongs to the latter. One important and desired property of probabilistic voting systems is 'Representative Probability' (RP). It states that the probability of the output distribution should be equal to the fraction of the population's voting, i.e. if candidate A is preferred over candidate B by the majority of voters then candidate A has to have higher chances to be elected than candidate B [19]. MEV aims to provide a voting outcome that adheres to the principle of RP while revealing the the minimum amount of information of the vote ordering [19].

E. Maximum Entropy Voting

Voting systems are a tool to achieve consensus, and can be roughly divided into deterministic and probabilistic systems. Maximum Entropy Voting (MEV) belongs to the category of probabilistic voting systems. In general, probabilistic voting means that according to the cast votes a probability is assigned to a set of voting outcomes. The winner of the election is then elected through sampling from the probability distribution. One important and desired property in that case it 'Representative Probability' (RP). It states that the probability of the output distribution should be equal to the fraction of the population's voting, i.e. if candidate A is preferred over candidate B by the majority of voters then candidate A has to have higher chances to be elected than candidate B [19]. MEV provides an voting outcome that adheres to the principle of RP which implies to match pairwise preferences, given the cast votes. At the same time it aims to find the outcome with the most uniform probability over the set of possible voting outcomes. This minimises the information and subsequently the made assumptions inherent to this voting method to the necessary minimum. [19] This implies that the randomness in this voting scheme is maximised as much as possible. As a result of this, it can be argued that it is difficult to predict the exact outcome of a vote and therefore it is secure against timed attacks because it is costly to have high confidence of success.

V. Building Blocks of the Data Market

A. Context

We present a case study with cars driving in a given city. We focus on Air Quality Index (AQI) as our metric of relevance, which can be measured through a car's air quality sensor. Many car manufacturers include an air quality sensor in the car's air conditioning mechanism. To illustrate the function of the data proposed data market, we divide the city into a grid with constant sized quadrants. As the cars drive across the city they measure AQI values at different quadrants of the city. Only agents with a valid license plate are granted access to collect and sell data on the marketplace. This acts as a defence mechanism against sybil attacks.

B. Assumptions

1) For each potential data point that can be made available in the marketplace, there is an over-supply of measurements.
2) Competing sellers are interested in aggregating (crowd-sourcing) data points from the market to fulfil a specific purpose.
3) Competing buyers only purchase data from the regulated market, and that each data point in the market has a unique identifier so that replicated data made available on secondary markets can be easily detected by data purchasers.
4) There is an existing mechanism that can verify the geographical location of an agent with a certain degree of confidence, and thus the provenance of the aforementioned agent's data collected. Several works have been carried out that corroborate that this is a reasonable assumption to make [18], [42] [9] [39] [8].
5) Following from 4 a Proof of Position algorithm is defined in V.13. Furthermore it is assumed that agents cannot be in more than one location at the same time. When an agent declares a measurement taken from a given location, we can verify this datapoint, the agent's ID and their declared position using V.13.
6) Following from assumption 1 and 2, the cases when a buyer wishes to purchase data from a geographical location where there is no data available are not accounted for. Note that there is the possibility to relax this constrain in future works through estimating for data-poor areas using neighbouring data, for example, with the KNN algorithm framework presented in [16].

C. Definitions

Definition V.1 (Datapoint)**.** *A datapoint is defined as $x \in X$ where X denotes the dataset of each seller.*

Definition V.2 (Location quadrant)**.** *The set of all possible car locations is defined as $\mathcal{L}$. The location quadrant q, is an instance of $\mathcal{L}$, where $q \in \mathcal{L}$.*

Definition V.3 (Buyer)**.** *A buyer is defined as m, where $m \in M$ and M is the set of agents looking to purchase*

ownership (or any other given rights) of the datasets that are available for sale on the marketplace.

Definition V.4 (Agent)**.** *An agent is defined as $a_{i,s} \in A$ where A is the set of all agents competing to complete the marketplace algorithm to become sellers. The index $i \in N$ where N is the total number of agents on the algorithmic marketplace at a given time interval $t \in T$. The index s denotes the stage in the access control mechanism that agent $a_{i,s}$ is in, where $s \in \{1, 2\}$. For example, agent $a_{5,2}$ is the agent number 5, currently in stage 3 of the access control mechanism in the data marketplace.*

Definition V.5 (Spatial Coalition)**.** *A spatial coalition is defined as a group of agents in the same location quadrant q, where these agents aggregate their individual data to provide an agreed upon dataset of that quadrant, D_q. The coalition is denoted as C_q where q is the location quadrant where the coalition is formed.*

Definition V.6 (Value Function)**.** *$v(S) = y$ where y is the number of location quadrants in the city with registered data measurements in them. S is a coalition of agents with corresponding datapoints. The function is strictly non-negative and satisfies $v(\emptyset) = 0$.*

Definition V.7 (Shapley Value)**.** *The Shapley Value is defined in [30] as a way to distribute reward amongst a coalition of n-person games. Each player i in the game receives a value ψ_i that corresponds to their reward. The Shapley Value satisfies the notions of Shapley fairness which are:*

1) *Balance:*
$$\sum_{a_i=1}^{A} \psi_m(a_i) = 1$$
2) *Efficiency: The sum of the Shapley value of all the agents is equal to the value of the grand coalition of agents $[A]$:*
$$\sum_{a_i=1}^{A} \psi_{a_i}(v) = v(A)$$
3) *Symmetry: If agents a_i and a_j are equivalent in the coalition of agents S such that both agents are providing data of the same value where $v(S \cup \{a_i\}) + v(S \cup \{a_j\})$ for every subset S of A which contains neither a_i nor a_j, then $\psi_{a_i}(v) = \psi_{a_j}(v)$*
4) *Additivity: If we define a coalition of agents to be $k = \{a_i, a_j\}$ then $\psi_k = \psi_{a_i} + \psi_{a_j}$*
5) *Null agent: An agent a_i is null if $v(S \cup \{a_i\}) = v(S)$. If this is the case then $\psi_{a_i} = 0$.*

Therefore formal definition of the Shapley value of an agent a_i that is in a set of A players is

$$\psi_(a_i) = \sum_{S \subseteq A \setminus \{a_i\}} \frac{|S|!(|A| - S - 1)!}{|A|!}(v(S \cup \{a_i\}) - v(S))$$

The Shapley Value is the unique allocation ψ that satisfies all the properties of Shapley fairness, described above.

Definition V.8 (Smart Contract)**.** *A smart contract is a program that will automatically execute a protocol once certain conditions are met. It does not require intermediaries and allow for the automation of certain tasks [11] [35]. In our context, a smart contract will be executed by agent $a_{i,s+1}$ to compute the Shapley value of agent $a_{j,s}$'s dataset.*

The outputs will be the Shapley value of agent $a_{j,s}$'s dataset and a new smart contract for agent $a_{i,s+1}$. Calculating the new smart contract generated serves as the proof of agent $a_{j,s}$'s useful work.

Every agent's smart contract will also contain a record of the buyer IDs and the permission that they have purchased from the agent. These could include permission to read the dataset, to compute analytics or to re-sell the dataset.

Definition V.9 (Bidding Mechanism)**.** *Following 6, there is a set of buyers M_q for each $q \in \mathcal{L}$ wishing to purchase data from that quadrant. A Bidding Mechanism is defined, BM, as a function that returns a buyer m that will purchase the dataset D corresponding to q, such that $m \in M_q$. Consequently, for all $q \in \mathcal{L}$: $m \leftarrow BM(M_q)$.*

Definition V.10 (Reward Distribution Function)**.** *The reward associated with the datapoint of a specific quadrant is defined as $v(C_q)$. In other words, the value that the spatial coalition C_q provides with their agreed upon datapoint D_q, of the location quadrant q. Each agent in C_q receives a coefficient $\alpha = \frac{1}{|D_q - d_i|}$, where d_i is the agent's individual datapoint. Consequently, the value $v(C_q)$ is split amongst all the agents in C_q as follows: for each agent, they receive $||\frac{v(C_q)}{|C_q|} \times \alpha||$*

Definition V.11 (Commitment)**.** *A commitment to a datapoint d_i, location quadrant q and ID i of an agent $a_{i,0}$ is defined as $c \leftarrow \text{Commitment}(a_{i,0}, d_i, q)$*

Definition V.12 (PoID)**.** *Let* PoID *be an algorithm that verifies the valid identity of an agent $a_{i,0}$, with ID i. In the context presented, this identification will be the license plate. The algorithm will return a boolean α that will be $True$ if the agent has presented a valid license plate and $False$ otherwise.*

Then PoID *is defined as the following algorithm:*
$\alpha \leftarrow \text{PoID}(i, c)$
This algorithm is executed by a central authority that can verify the validity of an agent's identity.

Definition V.13 (PoP)**.** *Let* PoP *be an algorithm that is called by an agent $a_{i,0}$, with ID i. The algorithm takes as inputs the agent's committment c, and their location quadrant q. We define* PoP *as the following algorithm:*
$\beta \leftarrow \text{a}_{\text{i,0}}^{\text{PoP}}(q, c)$
where the output will be a boolean β that will be True *if the position* q *matches the agent's true location and* False

otherwise. This algorithm is executed by a central authority that can verify the validity of an agent's position.

Definition V.14 (TimeCheck)**.** *The function* TimeCheck *takes in three arguments, the timestamp of the datapoint provided, t, the current time at which the function is executed, $timeNow$, and an acceptable range of elapsed time, r. The output of the function is γ. If $t - timeNow < r$, γ takes value $True$ and $False otherwise$.*
$\gamma \leftarrow \text{TimeCheck}(t, timeNow, r)$

Definition V.15 (Verify)**.** *Let* Verify *be an algorithm that checks that outputs of* PoID *and* PoP*. It will return a token $Token$ that will take the value $True$ iff α, β and γ are all $True$, and $False$ otherwise.*
$Token \leftarrow \text{Verify}(\alpha, \beta, \gamma)$
This algorithm can be executed by any agent in the data market, given that the proofs of PoID*,* PoP *and* TimeCheck *are publicly verifiable.*

new stuffffffffff

Definition V.16 (VotA)**.** *The voting algorithm $VotA$ takes as input the preference of each agent $S(a_i)$, as defined in section VII, and outputs an aggregated ranked order over all agents $S(C_j)$. It builds on what is explained in section **??**, where a hybrid-based reputation system is used to determine the agent's individual preferences. These preferences are written onto the smart contract by the agent and when each vote is cast, the aggregate is compute by averaging over all votes. This averaged ranking is then the output of the algorithm.*

$S(C_j) \leftarrow VotA(\{S(a_i) : a_i \in C_j\})$

Definition V.17 (MaxEntO)**.** *The Maximum Entropy Optimisation $MaxEntO$ takes the aggregated vote of a coalition and outputs a distribution over a set of possible vote outcomes, as introduced in section **??** with the reduced $\mathcal{T}$ and slack variable S_free. For efficiency reasons, the optimisation itself is carried out by a trusted third party, i.e. cloud computing service. Every agent can undertake checks if the resulting probability distribution is adhering to the constraint.*

$\pi_j(t) \leftarrow MaxEntO(S(C_j))$

Definition V.18 (CSA)**.** *The Committee Selection Algorithm CSA uses a decentralised random number generator to determines the committee within a coalition. It takes as input the probability distribution over the possible voting outcomes $\pi_j(t)$ and outputs a randomly chosen committee given $\pi_j(t)$.*

$c_j \leftarrow CSA(\pi_j(t))$ *where* $c_j = \{a_i : a_i \subset C_j\}$

VI. The Data Market

A. Stage 1: The Verification Algorithm

Algorithm 1: Verifying Algorithm ($a_{i,0}$, d_i, q, t, r)

$c \leftarrow \text{Commitment}(a_{i,0}, d_i, q, t)$;
$\alpha \leftarrow \text{PoID}(i, c)$;
$\beta \leftarrow \text{PoP}(q, c)$;
$\gamma \leftarrow \text{TimeCheck}(timeNow, t, r)$;
$Token \leftarrow \text{Verify}(\alpha, \beta, \gamma)$;
return $Token \leftarrow \{True, False\}$;

The validity of the data submission must be verified. This is done before the data reaches the data marketplace, to avoid retroactive correction of poor quality data. This is done through the VerificationAlgorithm. Firstly, an agent provides an immutable commitment of their datapoint, location quadrant, timestamp and unique identifier. Next, the agent submits their unique identifier to a centralised authority that verifies that this is a valid and real identity. In practise, for this context, this identifier will be the agent's vehicle license plate. Subsequently, the agent generates a valid proof of position. Following from assumption 2, an agent can only provide one valid outcome from algorithm V.13 at a given time instance t. Then, the datapoint is checked to ensure it is not obsolete through TimeCheck. Finally, the outputs of all previous functions are verified to ensure the agent has produced a valid proof. If and only iff all of these are $True$, the agent is issued with a unique valid token, that allows them to participate in the consensus mechanism.

B. Stage 2: Governance through Maximum Entropy Voting

One big challenge when creating a data market is the fact that data possess certain propterties that make it unlike most other commodities. As mentioned in section **??**, the fact that fake data is easily generated and replicated, its value is time-dependant, and the verificaion of the authenticity of data is challenging [1] [5], creates the need for new market structures and designs to cope with these circumstances. In what follows, different ways to tackle this challenge of faulty and fake data are presented, namely through the use of context and reputation. One can also call this approach source criticism, using context and past experience to verify and categorise information is not a new concept. Humans do this naturally everyday, potentially even without being aware of it, i.e. if a person is told news, they are more likely to believe it if the messenger is a well known friend (past experience) or multiple independent messenger present the same news (context).

VII. Reputation and Maximum Entropy Probabilistic Voting

In the following both topics will be introduced and combined to one which will be used as functional building block in the the data market design.

A. Reputation

In the general sense, reputation can be seen as a performance or trustworthiness metric that is assigned to an entity or group which is based on its attributes and previous behaviour. There are several categories of reputation [**?**] and one way to categorise reputation is by distinguishing between system-based and individual-based reputation. The former is when the reputation is assigned to an agent by the system. Applied to the case study this means that the data market keeps track of seller's previous behaviour and their properties and assigns a reputation score based on these. For the case of individual-based reputation, the agents in the system keep a track record of each other and assign a reputation score based on that. Additionally, there can be other factors that influence the reputation score i.e. when agents are in the same coalition. For the maximum entropy voting building block in this work, hybrid-based reputation assigned to an individual agent is of interest. In this context *Hybrid* can be understood as a mix of the aforementioned; that the reputation assigned to an agent is determined by both the system and the other agents in the system. This means the reputation score of an agent is individualised for each agent and additionally depends on the set of other agents.

B. Maximum Entropy Voting

Voting systems are a tool to achieve consensus, and can be roughly divided into deterministic and probabilistic systems. Maximum Entropy Voting (MEV) belongs to the category of probabilistic voting systems. In general, probabilistic voting means that according to the cast votes a probability is assigned to a set of voting outcomes. The winner of the election is then elected through sampling from the probability distribution. One important and desired property in that case it 'Representative Probability' (RP). It states that the probability of the output distribution should be equal to the fraction of the population's voting, i.e. if candidate A is preferred over candidate B by the majority of voters then candidate A has to have higher chances to be elected than candidate B [19]. MEV provides an voting outcome that adheres to the principle of RP which implies to match pairwise preferences, given the cast votes. At the same time it aims to find the outcome with the most uniform probability over the set of possible voting outcomes. This minimises the information and subsequently the made assumptions inherent to this voting method to the necessary minimum. [19] This implies that the randomness in this voting scheme is maximised as much as possible. As a result of this, it can be argued that it is difficult to predict the exact outcome of a vote and therefore it is secure against timed attacks because it is costly to have high confidence of success.

VIII. Verification through Repuation-based Probabilistic Maximum Entropy Voting

After having the two topics introduced, a more rigorous mathematical framework will be built and presented upon which are foundational for the later sections **??** and **??**. Let's assume a set of agents $a_i \in \mathcal{A}$ with cardinality $|\mathcal{A}| = N$, where every agent a_i has a datapoint x_i and a reputation $r_{i,k}$ for every other agent $a_k \in \mathcal{A} \backslash a_i$. The set of measurement $x_i \in \mathcal{X}$ are for example temperature or air quality measurements of an agent which they want to submit and sell. The reputation $r_{i,k} \in \mathbb{R}^+$ is a value agent a_i associates with agent a_k which expresses the individualised reputation of agent a_k from the perspective of agent a_i given the system they are part of, with the data market being the system. It can be based on recent behaviour, i.e. $r_{i,k}$ can be influenced by the the experience of agent a_i with agent a_k during their last engagements, or based on the fact that they know each other and are in coalition. Additionally, it can be influenced by the overall track record of agent a_k in the data market. $r_{i,k} \in \mathcal{R}$ could be seen like a graph and constructed like a social network. Every agent has an impression of how trustworthy every other agent in the network is and bases it on their previous interaction and the overall track record of an agent in the system. To draw the analogy to a social network of a community. I will consider people trustworthy who are my friends and people who I have had positive interactions with. Additionally, if I see proof that a person has done much good so far, I will also consider this person trustworthy. Finally, for now, assume that this reputation $r_{i,k}$ is given and its influence on the outcome will be discussed in later sections, namely **??** and **??**.

To combine maximum entropy voting and reputation, a key step is to move from reputation to a preference score $c_{i,j}$ which is assigned to agent a_j's measurement x_j by agent a_i. The higher the score the higher ranks an agent another agent's measurement. In general, this is a function of both agents $c_{i,j} = f(a_i, a_j)$. For this case study equation (1) will be used. In this form is the preference score $c_{i,j}$ will be positive and increase with reputation r_{ij}. Additionally, the it scales inversely to the absolute value of the relative difference between the two agents' measurements x_i and x_j. The idea behind is that an agent prefers agents with higher reputation who have similar measurements to their own measurement.

$$c_{i,j} = \frac{1 + |x_i| \cdot r_{i,j}}{1 + |x_i - x_j|} \tag{1}$$

Further, assume that agent i prefers agent j over agent k's measurement iff $c_{i,j} > c_{i,k}$ holds. Iff $c_{i,j} = c_{i,k}$ agent i prefers agent j's measurement equally as agent k's measurement. In this way for every agent a_i a sorted ranking of all other agents $a_j \in \mathcal{A}$ a_i can be constructed according to agent a_i's preference. The resulting order vector t_i of size N indexes as first element agent a_i's most preferred measurement up to the last element which is the least preferred measurement. To represent the order vector t as a relative preference matrix $S(t)$, the following definition for $S(t)$'s entries $s_{i,j}(t)$ is made to display the pairwise preference between each pair of measurements.

$$s_{i,j}(t) = \begin{cases} 1 & \text{if } c_i \text{ is preferred to } c_j \text{ and } i \neq j \text{ under } t \\ 0.5 & \text{if } c_j \text{ is equally preferred as } c_i \text{ and } i \neq j \text{ under } t \\ 0 & \text{if } c_j \text{ is preferred to } c_i \text{ and } i \neq j \text{ under } t \\ 0 & \text{if } i = j \end{cases} \tag{2}$$

To compile the preference of all agents into one object, assume a tensor z_{ijk}, where i, j, and k go from 1 to N. For every index i the entries are defined in analogy to $S(t_i)$ where t_i is the order vector representing the preference of agent a_i. The idea here is that the tensor z_{ijk} represents for every index i the relative preference matrix of agent i.

Further to aggregate and average $z_{i,j,k}$, let $Z_{j,k}$ be as follows. $Z_{j,k}$ is the empirical preference matrix because it is empirically found and contains the preference of every agent.

$$Z_{j,k} = \frac{1}{N} \sum_{i}^{N} z_{i,j,k} \tag{3}$$

Finally, coming to a determining piece which strongly defines the computational complexity of the problem is to define the set of considered orderings $t \in \mathcal{T}$ over which the maximum entropy optimisation runs. The broadest possible definition of $\mathcal{T}$, which is at the same time the original one, is to consider all strong total orderings of permutations on the set of all agents $\mathcal{A}$. The resulting cardinality of $\mathcal{T}$ is thus $|\mathcal{T}| = N!$. Another way to define the set of $\mathcal{T}$ which reduces the cardinality of $\mathcal{T}$ and thus the computational complexity of the optimisation problem relies on the key idea to look at the problem from a perspective of unordered sampling without replacement, i.e. the set of agents is seen as pool of size N from which a subset $\{\mathcal{P} \subset \mathcal{A} : |\mathcal{P}| = K\}$ of preferred agents is drawn without replacements. Within both resulting sets, the set of preferred agents $\mathcal{P}$ and the set of not preferred agents $\{\mathcal{NP} : \mathcal{A} \backslash \mathcal{P}\}$, the ordering of the agents does not matter and they are thus equally preferred. When considering all possible K-element subsets of $\mathcal{A}$, the number of considered ordering and thus the size of $|\mathcal{T}| = \frac{N!}{K!(N-K)!}$. For reasonable small K, this leads to a to a dramatic reduction of the computational complexity of the optimisation problem. Lastly, before presenting the maximum entropy lineare optimisation program, let π be a vector of size $|\mathcal{T}|$ and π_t denote the probability of S(t).

$$\max_{\pi} -\sum_{t} \pi_t \log \pi_t \tag{4}$$

$$\text{s.t.} \sum_{t}^{T} \pi_t S(t) = Z_{j,k} \tag{4a}$$

$$\sum_{t}^{T} \pi_t = 1 \tag{4b}$$

$$\pi_t \geq 0 \qquad \forall t \tag{4c}$$

Note that it is possible for every agent to check that the resulting π satisfies the conditions by simply summing and comparing. Therefore, centralising the solving of the optimisation problem does not require trust assumptions on the executing central authority.

The resulting probability distribution π fulfils the previously described properties of RP and minimal assumptions given the cast vote. It can then be used to sample orderings or sets, depending on the definition of $\mathcal{T}$, which represent then the elected agents.In the context of this work, these can be the input for another block in the linked chain of algorithms in the data market, like verification through context which will be introduced in the next section.

IX. Verification through Context

The core idea is that in an oversubscribed environment crowd-sourcing can be used to estimate a central measure which ideally reflects the ground truth as close as possible. The assumption is that every agent measures the same source and should therefore have the same results within margins of measurement precision. The only reasons why there can be deviations is either the used sensor is faulty or the agent is intentionally submitting incorrect results. Therefore, by comparing agents' measurements against each other the aim is to sort out faulty and incorrect results.

There are different ways to approach this and in order to characterise them two concepts will be introduced, namely k-anonymity and the breakdown point.

A. K-Anonymity

One definition is "*A data release is said to satisfy k-anonymity if every tuple released cannot be related to fewer than k respondents, where k is a positive integer set by the data holder, possibly as the result of a negotiation with other parties.*" [**?**][8] In other words, if a central

[8]latanyasweeney.org; k-anonymity

measure includes multiple measurements and is assigned a k-anonymity with k=2, it is not possible to identify a single measurement without a second party revealing their measurement. aida; pierre; pietro; freddy page;

B. Breakdown Point

The breakdown point characterises the robustness of an estimator and can be defined as follows. *The breakdown point is the smallest fraction of bad observations that may cause an estimator to take on arbitrarily large aberrant values* [**?**]. It is usually dependent on the sample size, n, and can be written as a function of n.

C. Mean as Central Measure

The mean $\bar{x}$, arithmetic mean, in its simplest form is defined in equation (5).

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{5}$$

where x_i are the individual measurements and n the sample size.

The mean can be calculated in a decentralised and privacy preserving manner [23]. In theory it can additionally be calculated under homomorphic encryption. [**?**] The k-anonymity of the mean results then in k = n-1. The breakdown point of the mean is $\frac{1}{n}$ or in other words, a single measurement can cause the mean to take on arbitrarily high or low values. This can be mitigated with domain knowledge, i.e. restraining the range for valid measurements, for example when considering temperature in London to [-30°C - 50°C]. However, even with this counter measure in place, a larger coalition of malevolent agents is still able to influence the mean significantly. This in combination with the face that malevolent agents can be expected in a data market context, it may be that the mean is not sufficiently robust for most use cases.

D. Median as Central Measure

The median is a is the value separating the higher half from the lower half of a data sample. It can be defined for a numerically ordered, finite sample of size n, as follows.

$$\text{median}(x) = \begin{cases} x_{(n+1)/2} & \text{if n is odd} \\ \frac{1}{2} \cdot (x_{(n/2)} + x_{(n/2)+1}) & \text{if n is even} \end{cases} \tag{6}$$

This definition is invalid for an unordered sample of measurements. In order to compute the median for such a sample, the measurements need to be sorted numerically first, at least partly. Given a multi agent setting, this can be done in a distributed way by using an selection algorithm that finds the k^{th} smallest element(s). In detail, when a comparison between two values is required by the algorithm, this can be executed in a P2P manner where two agents exchange their measurements and report which is bigger. However, This cannot be done without revealing the measurements, and thus the privacy of the data cannot be protected when using the median to estimate a centrality measure. The breakdown point of the median characterises it as one of the most robust estimators and is for the worst case given with $\frac{1}{2}$.

E. Mean Median Algorithm as Central Measure

In an adversarial environment, the high robustness of the median is desirable, however, often protection of privacy is also of concern. Therefore, the Mean Median Algorithm was designed to have an algorithm that estimates a central measure in a robust and privacy-preserving way. It must be said that it is a compromise and this alogorithm is not as robust as the median and privacy-preserving as the mean when compared individually.

To illustrate the algorithm, the earlier in section VII introduced notation will be used. The first step is to randomly assign every agent to a group in such a way that there are r groups with at least s agents each. The way the parameters r and s are chosen determine the anonymity and robustness properties of the algorithm and will be discussed in the simulation section **??**. The next step is to calculate the mean within each group. The resulting mean is at least of k-anonymity with k = s-1. As there are r groups, there are r means of which the median is chosen. This gives a breakdown point given in equation 7.

$$\text{Breakdown point of } \text{meanmedian}(x) = \frac{r}{2n} \tag{7}$$

The relationship between s, r and the number of agents n is given with the inequality (8)

$$n \geq s \cdot r \tag{8}$$

F. Bootstrapping as Central Measure

The last idea we will discuss in this context is bootstrapping. It also combines the first two approaches in such a way that repeatedly agents are sampled randomly with sample size s and the mean is taken over their measurements. After r repetitions the median is taken from the set of sampled means. This way the privacy of the measurements can be preserved while also being more robust than the mean. It is similar to the mean median algorithm with the difference that agents can be sampled again for different means. Note that as a result inequality (8) does not hold for this approach.

After having chosen a central measure the next step is to assess every agent's measurement based on that central measure. Here there are two approaches we will discuss.

- Relative Difference
- Similarity

The first approach is more suitable for the one-dimensional case and we define it as follows.

$$\nu(C) = \frac{CM - \frac{1}{|C|}\sum_i^C m_i}{|CM|} \tag{9}$$

where C is the set of agent that form a coalition.

The second approach is more suitable for the multi-dimensional case and we define it as follows.

$$\nu(C) = cos(CM, \frac{1}{|C|}\sum_i^C m_i) \tag{10}$$

where $cos(.,.)$ is the cosine similarity.

Using these $\nu(C)$ can then be used to estimate Shapley Values for each agent to account for their contribution to the CM.

G. Stage 4: Access Control Mechanism to the Data Market

The previous stages 1, 2 and 3 run simultaneously in numerous rounds, as vehicles sense data and form coalitions to provide an agreed up value for a given location quadrant. By stage 4, there is an excess of datapoints for different locations and these datapoints are on a queue to enter and be sold on the datamarket.

This section outlines the access control mechanism to sell this oversupply of datapoints on the market, and in what order these should be prioritised to enter the datamarket. This access control mechanism can be considered to have two intermiediary steps: firstly, all datapoints are assigned a priority; and secondly, proportionally to this priority, the coalition owning that datapoint must perform an adaptive, useful proof of work.

1) Stage 4.1: Shapley Value: At a given time t, a new set of datapoints will be submitted to a queue, to ultimately enter the datamarket. Let this set be $\mathbf{D_t} = \{D_{q1}, D_{q2}, D_{q3}...\}$ where each item of the set is the datapoint computed by a coalition C_q, of a given location quadrant, where $\{q1, q2, q3...\} \in \mathcal{L}$. For each element in $\mathbf{D_t}$, the Shapley value $\psi(C_q)$ is calculated. Note that each element in $\mathbf{D_t}$ is a datapoint that corresponds to a spatial coalition C_q. The grand coalition in this case is considered to be the union of all coalitions that have datapoints already for sale on the datamarket, denoted as $\mathbf{S}$. Each datapoint in $\mathbf{D_t}$ is assessed using the Shapley value, which determines what datapoints would increase the overall value of the datamarket, with respect to the defined value function, should they be added to the grand coalition $\mathbf{S}$. In other words, the datapoints that receive a higher Shapley value, would contribute more towards increasing the combined value of the data already for sale on the market. In this fashion, the Shapley value is used as a metric to rank the most valuable datapoints with respect to a value function.

2) Stage 4.2: adaptive, useful proof of work: Subsequently, once the datapoints in $\mathbf{D_t}$ have each received a Shapley value, they are then assigned a proof of work they must complete. This proof of work is inversely proportional to the Shapley value. The more valuable a datapoint is deemed for the datamarket, the less proof of work the coalition owning it should complete, to enter the market. This assigned proof of work, in fact, is computing the Shapley Value of the next set of datapoints, $\mathbf{D_{t+1}}$.

X. Adversarial Attacks

1) Sybil Attacks:

Definition X.1. *Sybil Attacks are a type of attack in which an attacker creates a large number of pseudonymous identities which they use to exert power in and influence the network.*

2) Wormhole Attacks:

3) Collaborative Lying:

4) Faulty Data Collection:

XI. Adversarial Attacks

In an environment like decentralised networks or data markets one must take into account the possibility of attacks on the system. In the following different types of attacks will be

- Sybil Attacks are a type of attack in which an attacker creates a large number of pseudonymous identities which they use to exert power in and influence the network.
- Wormhole Attacks are attacks in which an attacker fakes their location in order to gain an advantage. This could involve jumping to different locations to be able to submit multiple measurements.
- Collaborative Lying involves two or more malevolent actors who act in a coordinated fashion to influence the verification of data, i.e. submit the collaboratively a wrong measurement during the selection of a central measure with the aim to corrupt the outcome.
- Faulty Data is not necessary an attack but rather resembles the fact that during the sensing process errors can occur and thus invalid measurements are submitted.

XII. Evaluation

A. Simulation Setup

1) Median Algorithm:

- Decentralised

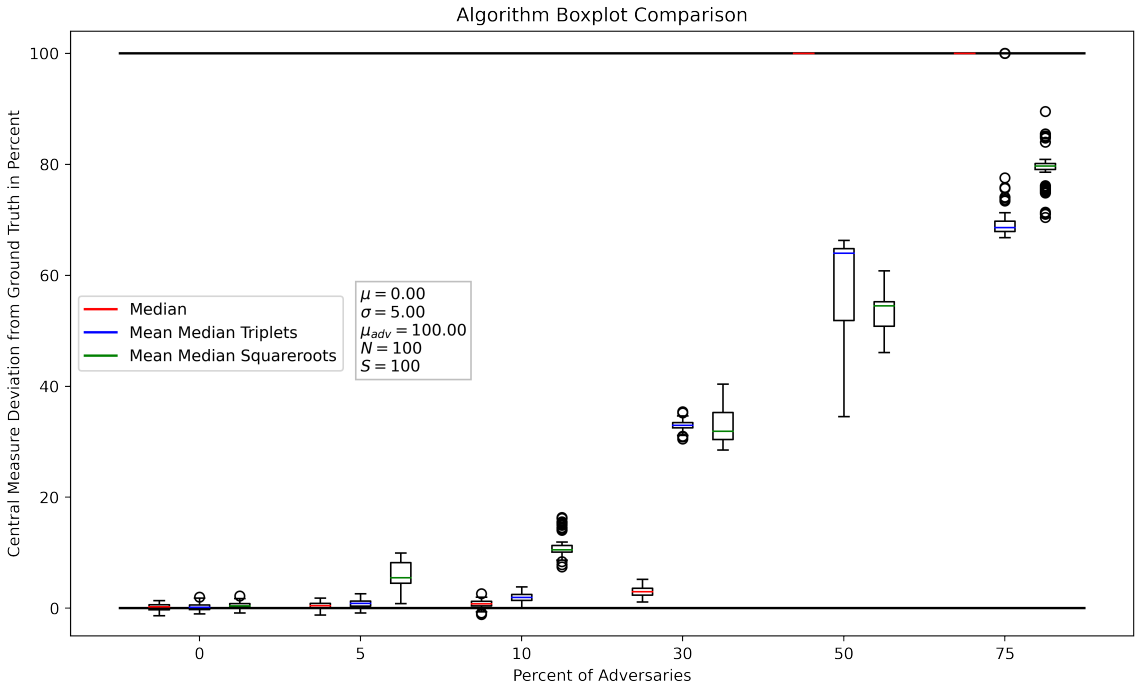

Figure 4. Breakdown analysis of data consensus algorithms, with a coordinated data poisoning attack

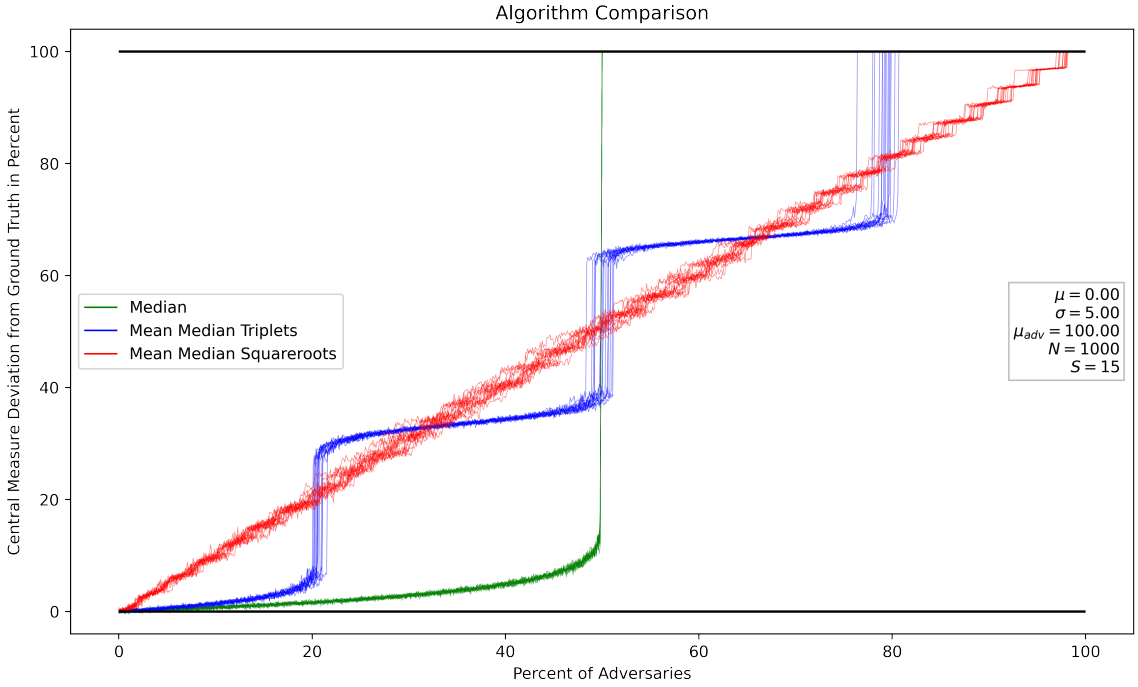

Figure 5. Characterisation of data consensus algorithms' behaviour under different degrees of coordinated data poisoning attacks

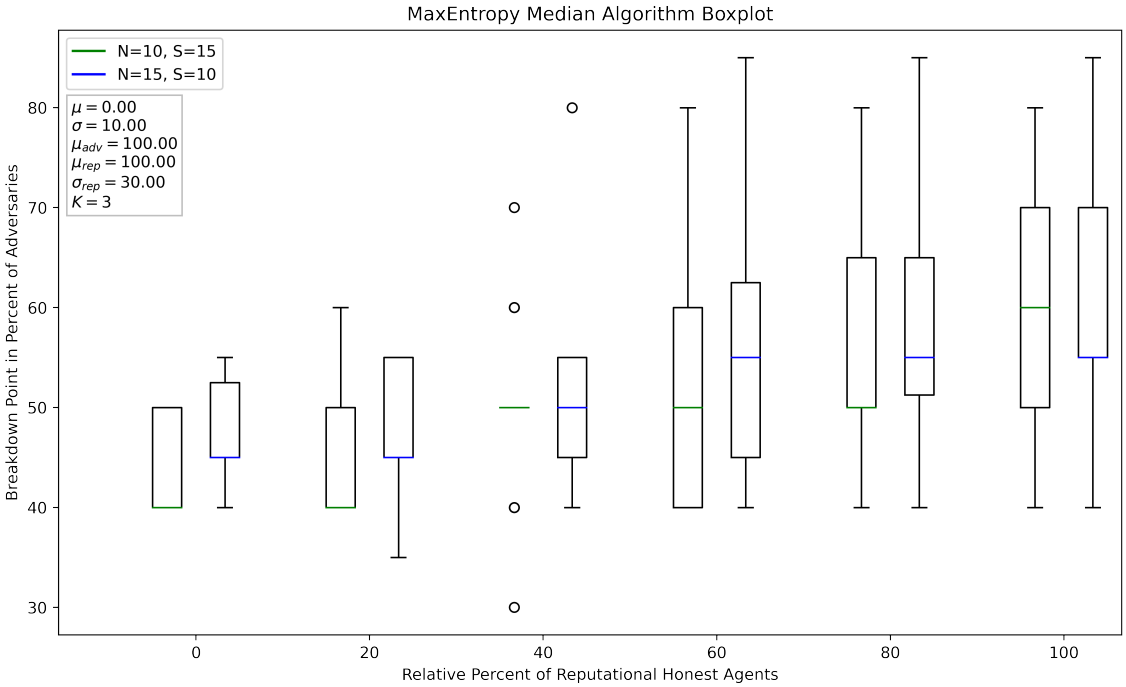

Figure 6. Characterisation of breakdown of MEV combined with Median Algorithm

- Robust up to 50%

The Median

2) Mean Median Algorithm:

- Decentralised
- Privacy Preserving
- Robust up to 50%

B. Max Entropy Voting

- centralised
- adapted to our problem

The authors have already introduced Max Entropy Voting in the section IV-D2 and will continue here to adapt the technique for the data market. This is necessary because of the computational complexity the original problem poses with $N!$ where N is the number of candidates that are voted on. For the case of data markets this is computationally infeasible because it is necessary to process on a high frequency double digits numbers of candidates.

One way to reduce the complexity of the problem is to consider a different set of combinations in which the candidates are ranked instead of all possible permutations like in the original problem. The key idea is to look at it from perspective of sampling with replacement, i.e. the set of candidates is seen as pool from which a subset of preferred agents is drawn without replacement. By considering only the combinations in which a subset can be drawn without demanding a ranking within the groups reduces the complexity to $\frac{N!}{K!(N-K)!}$, where N is the number of candidates and K the number of drawn agents. This adapted way allows the application of the max entropy voting strategy to a reasonable large set of candidates as long as the number of drawn agents is relatively small.

A practical challenge that poses this change is that the constraint (**??**) cannot be satisfied anymore and therefore the whole linear problem becomes infeasible. To solve this problem an amendment of the objective function (**??**) and the aforementioned constraint as follows.

$$\max_{\pi, S_{free}} -\alpha ||S_{free}||_2 - \sum_t \pi_t \log \pi_t \tag{11}$$

$$\sum_t^T \pi_t S(t) + S_{free} = Z_{j,k} \tag{12}$$

where S_{free} is a matrix of free, unbound variables, $||.||_2$ denotes the Frobenius norm, and α a constant which in practice is assigned a high value to force the linear program to the solution closest to the one of the original problem.

C. Conclusion

1) Ownership Traceability: The purpose of using a decentralised ledger is to ensure any agent can verifiably trace who has access to their data. Centralisation forgoes verifiability, so the DLT's access control mechanism must ensure that the system does not converge to centralisation over time (goal C). To achieve ownership traceability, we will design a data market inspired in the Directed Acyclic Graph architecture, (given Blockchain's aforementioned issues). Each transaction, instead of representing an exchange of wealth, will represent an exchange in wealth for access to a dataset. Thus, the ledger will be a graphical representation of which users have access to what datasets. Our data market can include a suite of services available on top of access to data as well. For example, agents can pay for the right to also carry out analytics on the dataset, or for the right to re-sell to third parties. This will be managed through the use of smart contracts, that will record which users have been granted what rights. The tangle will thus reflect these interactions, in an immutable, publicly verifiable manner.

REFERENCES

[1] AGARWAL, A., DAHLEH, M., AND SARKAR, T. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation* (2019), pp. 701–726.

[2] ANDREWS, L. Facebook is using you. *New York Times*.

[3] APERJIS, C., AND HUBERMAN, B. A. A market for unbiased private data: Paying individuals according to their privacy attitudes. *Available at SSRN 2046861* (2012).

[4] APPSFLYEA. The impact of ios 14+ & att on the mobile app economy.

[5] ARROW, K. J. *Economic welfare and the allocation of resources for invention*. Princeton University Press, 2015.

[6] BEIKVERDI, A., AND SONG, J. Trend of centralization in bitcoin's distributed network. In *2015 IEEE/ACIS 16th international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)* (2015), IEEE, pp. 1–6.

[7] BELL, F., CHIRUMAMILLA, R., JOSHI, B. B., LINDSTROM, B., SONI, R., AND VIDEKAR, S. Data sharing, data exchanges, and the snowflake data marketplace. In *Snowflake Essentials*. Springer, 2022, pp. 299–328.

[8] BOEIRA, F., ASPLUND, M., AND BARCELLOS, M. Decentralized proof of location in vehicular ad hoc networks. *Computer Communications 147* (2019), 98–110.

[9] BORNHOLDT, L., REHER, J., AND SKWAREK, V. Proof-of-location: A method for securing sensor-data-communication in a byzantine fault tolerant way. In *Mobile Communication - Technologies and Applications; 24. ITG-Symposium* (2019), pp. 1–6.

[10] BORYCZKA, J. M. Democracy, 2008.

[11] CHRISTIDIS, K., AND DEVETSIKIOTIS, M. Blockchains and smart contracts for the internet of things. *IEEE Access 4* (2016), 2292–2303.

[12] FOUNDATION, I. The tangle: an illustrated introduction. https://blog.iota.org/the-tangle-an-illustrated-introduction-4d5eae6fe8d4/.

[13] GHAFFARI, F., BERTIN, E., HATIN, J., AND CRESPI, N. Authentication and access control based on distributed ledger technology: A survey. *2020 2nd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)* (2020).

[14] HYNES, N., DAO, D., YAN, D., CHENG, R., AND SONG, D. A demonstration of sterling: a privacy-preserving data marketplace. *Proceedings of the VLDB Endowment 11*, 12 (2018), 2086–2089.

[15] ISAAK, J., AND HANNA, M. J. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer 51*, 8 (2018), 56–59.

[16] JIA, R., DAO, D., WANG, B., HUBIS, F. A., GUREL, N. M., LI, B., ZHANG, C., SPANOS, C., AND SONG, D. Efficient task specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment 12*, 11 (2018), 1610–1623.

[17] LAOUTARIS, N. Why online services should pay you for your data? the arguments for a human-centric data economy. *IEEE Internet Computing 23*, 5 (2019), 29–35.

[18] LUO, W., AND HENGARTNER, U. Veriplace: A privacy-aware location proof architecture. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2010), GIS '10, Association for Computing Machinery, p. 23–32.

[19] MACKAY;, R. S. D., AND MCLEAN, I. Probabilistic electoral methods, representative probability, and maximum entropy. *Voting matters* (2009).

[20] MCKINZIE, AND COMPANY. Monetizing car data: New service business opportunities to create new customer benefits, 2016.

[21] NARULA, N., VASQUEZ, W., AND VIRZA, M. zkledger: Privacy-preserving auditing for distributed ledgers. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)* (2018), pp. 65–80.

[22] OPENSECRETS. Expenditures breakdown, donald trump, 2016 cycle, 2016.

[23] OVERKO, R., ORDOPOEZ-HURTADO, R., ZHUK, S., FERRARO, P., CULLEN, A., AND SHORTEN, R. Spatial positioning token (SPToken) for smart mobility. *2019 8th IEEE International Conference on Connected Vehicles and Expo, ICCVE 2019 - Proceedings* (2019).

[24] RAMACHANDRAN, G. S., RADHAKRISHNAN, R., AND KRISHNAMACHARI, B. Towards a decentralized data marketplace for smart cities. In *2018 IEEE International Smart Cities Conference (ISC2)* (2018), IEEE, pp. 1–8.

[25] RASOULI, M., AND JORDAN, M. I. Data sharing markets. *arXiv preprint arXiv:2107.08630* (2021).

[26] RAYNAL, M. Communication and agreement abstractions for fault-tolerant asynchronous distributed systems. *Synthesis Lectures on Distributed Computing Theory 1*, 1 (2010), 1–273.

[27] SANCHEZ, L., ROSAS, E., AND HIDALGO, N. Crowdsourcing under attack: Detecting malicious behaviors in waze. In *IFIP International Conference on Trust Management* (2018), Springer, pp. 91–106.

[28] SCHUEFFEL, P. Alternative Distributed Ledger Technologies Blockchain vs. Tangle vs. Hashgraph - A High-Level Overview and Comparison -. *SSRN Electronic Journal* (2018), 1–8.

[29] SEN, A. Social choice. the new palgrave dictionary of economics, abstract & toc, 2008.

[30] SHAPLEY, L. *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, 1953.

[31] STAHL, F., SCHOMM, F., AND VOSSEN, G. The data marketplace survey revisited. Tech. rep., ERCIS Working Paper, 2014.

[32] STAHL, F., SCHOMM, F., AND VOSSEN, G. Data marketplaces: An emerging species. In *DB&IS* (2014), pp. 145–158.

[33] STUCKE, M. E. Should we be concerned about data-opolies? *Geo. L. Tech. Rev. 2* (2017), 275.

[34] STÖRING, M. What eu legislation says about car data legal memorandum on connected vehicles and data, 2017.

[35] SZABO, N. The idea of smart contracts. https://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_idea.html, 1997.

[36] TAHMASEBIAN, F., XIONG, L., SOTOODEH, M., AND SUNDERAM, V. Crowdsourcing under data poisoning attacks: A comparative study. In *IFIP Annual Conference on Data and Applications Security and Privacy* (2020), Springer, pp. 310–332.

[37] TRAVIZANO, M., SARRAUTE, C., AJZENMAN, G., AND MINNONI, M. Wibson: A decentralized data marketplace. *arXiv preprint arXiv:1812.09966* (2018).

[38] UR REHMAN, I. Facebook-cambridge analytica data harvesting: What you need to know. *Library Philosophy and Practice* (2019), 1–11.

[39] WU, W., LIU, E., GONG, X., AND WANG, R. Blockchain based zero-knowledge proof of location in iot. In *ICC 2020 - 2020 IEEE*

International Conference on Communications (ICC) (2020), pp. 1–7.

[40] ZHANG, S. Who owns the data generated by your smart car. *Harv. JL & Tech. 32* (2018), 299.

[41] ZHAO, L., VIGNERI, L., CULLEN, A., SANDERS, W., FERRARO, P., AND SHORTEN, R. Secure Access Control for DAG-based Distributed Ledgers. *IEEE Internet of Things Journal* (2021), 1–15.

[42] ZHU, Z., AND CAO, G. Toward privacy preserving and collusion resistance in a location proof updating system. *IEEE Transactions on Mobile Computing 12*, 1 (2013), 51–64.

[43] ZWICKER W. S., M., HERVE (2016), B., FELIX; CONITZER, V. E., AND ULLE; LANG, J. Introduction to the theory of voting, 2016.