Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier XX.XXXX/ACCESS.XXXX.DOI

On the Design of Decentralised Data Markets

Aida Manzano Kharman[1], Christian Jursitzky, Quan Zhou[1], Pietro Ferraro[1], Jakub Marecek[2], Pierre Pinson[1], Robert Shorten[1]

[1]The authors are affiliated with the Dyson School of Design Engineering, Imperial College London

[2]The authors are affiliated with the department of Computer Science, Czech Technical University in Prague

Corresponding author: Aida Manzano Kharman (e-mail: aida.manzano-kharman17@imperial.ac.uk).

This work was partially supported by the IOTA Foundation.

ABSTRACT We present an architecture to implement a decentralised data market, whereby agents are incentivised to collaborate to crowd-source their data. The architecture is designed to reward data that furthers the market's collective goal, and distributes reward fairly to all those that contribute with their data. This is achieved leveraging the concept of Shapley's value from Game Theory. Furthermore, we introduce trust assumptions based on provable honesty, as opposed to wealth, or computational power, and we aim to reward agents that actively enable the functioning of the market. In order to evaluate the resilience of the architecture, we characterise its breakdown points for various adversarial threat models and we validate our analysis through extensive Monte Carlo simulations.

I. INTRODUCTION

A. PREAMBLE

In recent years there has been a shift in many industries towards data-driven business models [62]. 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 monetary compensation. 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 these business models continue to be questioned, with more and more stakeholders arguing that such platforms should recompense citizens in a more direct manner for data that they control [65] [39], [5], [6]. To give more context, 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. 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 [7]. 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 regards fair recompense to the data harvester by data-driven businesses. 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 the exchange of value is fair. The second one arises from the potential

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

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 scandal 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] [37]. Cambridge Analytica has a vast portfolio of elections they have worked to influence, with the most notorious one being the 2016 US presidential elections [71], [48].

It is important to understand that these cases are not anecdotal. Without the adequate infrastructure to track and trade ownership, cases like the ones outlined above, with for example, mass privacy breaches, having 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, we argue for the shift towards a more decentralised data ownership model, where this ownership can be publicly verified and audited. We 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 is not new and there have been numerous attempts to design marketplaces to enable the exchange of data for money [63]. This, however, is an extremely challenging 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 [2] [10]. These properties make marketplace models difficult to design and have been an emergent research area.

In what follows, we describe a first step in the design of a marketplace where data can be exchanged. Furthermore, this marketplace provides certain guarantees 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 a specific use case from the automotive industry that, we hope, might represent a first step towards more general architectures.

B. SPECIFIC MOTIVATION

We 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. 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 remains of 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, still captures many pertinent aspects of more general data-market design: for example, detection of fake data; certification of data-quality; resistance to adversarial attacks.

The context of automotive data collection is a ripe 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

[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

[46] *"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 smartphones; 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 [75]. 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 (IP), 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 [66] "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 rigorously justified before such marketplaces emerge. Given the lack of verifiability of a centralised authority, such as a car manufacturing company, we propose exploring decentralised or hybrid alternatives.

The 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, we 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.

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

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 [51] [36] [70]. 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 [11]. Indeed, centralised solutions to data markets already exist, such as [12], 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 [52] 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 [70] also utilise. However, introducing an auditor simply shifts the trust assumption to their honest behaviour and forgoes decentralisation.

In [2], 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 Intellectual Property, especially when it is the core of their business model. The work of [47] 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.

There is a need to have trust assumptions in components of the data market, whether it be centralised or decentralised. However, we 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 proposition. For users to trust the consensus mechanism, they must have a series of provable guarantees that it was executed correctly. It is not sufficient for the consensus mechanism to function correctly, it should also prove this to the users.

We 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

Consensus mechanisms are crucial in distributed ledgers to ensure agreement on the state of the ledger. For example, in the context of the branch of Computer Science known as *distributed systems*, consensus can be mapped to the fault-tolerant state-machine replication problem [53]. In such systems, the users in the network must come to an agreement as to what is the accepted state of the network. Furthermore, it is unknown which of these users are either faulty or malicious. This scenario is defined as a Byzantine environment, and the consensus mechanism used to address this issue must be Byzantine Fault Tolerant (BFT) [20].

In permissionless networks, probabilistic Byzantine consensus is achieved through the means of certain cryptographic primitives [73]. Commonly this is done by solving a computationally expensive puzzle.

In permissioned networks consensus is reached amongst a smaller subset of users in the network. This is done through BFT consensus mechanisms such as Practical BFT (PBFT) [19] and PAXOS [18]. Often the permissioned users are elected according to how much stake in the network they hold, following a proof of stake (PoS) method. This centralisation enables a higher throughput of transactions at the cost of higher messaging overhead, but ensures immediate consensus finality. They also require precise knowledge of the users' membership [28]. Meanwhile, in permissionless consensus protocols the guarantee of consensus is only probabilistic but does not require high node synchronicity or precise node memberships (ie: exact knowledge of which users are in the quorum), and are more robust [72] [73].

When considering consensus mechanisms for permissionless distributed ledgers, there exist a wide range of consensus mechanisms that are a hybrid combination of either PoS and PoW (eg: Snow White), or PoW-BFT (eg: PeerCensus) or PoS-BFT (eg: Tendermint). Each consensus mechanism places greater importance in achieving different properties. For example, Tendermint focuses on deterministic, secure consensus with accountability guarantees and fast throughput [17]. Snow White is a provably secure consensus mechanism that uses a reconfigurable PoS committee [13], and PeerCensus enables strong consistency in Bitcoin transactions, as opposed to eventual consistency [24].

There also exists a class of probabilistic consensus mechanisms, such as FPC [50], Optimal Algorithms for Byzantine Agreements [27], Randomised Byzantine Agreements [69] and Algorand [32]. We find this class of consensus mechanisms of particular interest for the context of a data market. Namely, the fact that they are probabilistic makes coercion of agents difficult for a malicious actor. To ensure malicious actors are selected by the consensus algorithm, they must know a priori the randomness used in the mechanism, or coerce a supra-majority of agents in the network. Furthermore, we argue that selecting users in a pseudo-random way treats users equally, and is closer to achieving fairness

than selecting users with the greatest wealth or greatest computational power. Another consideration of fairness is made in [23], where the mechanism does not rely on a correct leader to terminate, therefore decentralising the mechanism.

In some examples described above, such as in [32], agents with greater wealth in the ledger are more likely to be selected by a Verifiable Random Function to form part of the voting committee. However, we believe that for the context of the work here presented, voting power should be earned and not bought. Indeed, this right should be earned irrespective of wealth or computational power. This opens the question of, how then, should this power be allocated? The market should trust the actors that behave honestly and further the correct functioning of the market. Agents should prove their honesty and only then earn the right to be trusted. A collective goal for the data market can be defined, and agents who contribute to this should be adequately rewarded. This goal can be characterised mathematically, and each agent's marginal contribution can be evaluated. In this manner, rights and rewards can be granted proportionally.

Algorand wishes to retain the voting power amongst the agents with the most stake, 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 our data market. Owning more cars (i.e. more stake) 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 [55] and [68], 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 crowd-sourced applications, usually mounted through a Sybil attack.

Furthermore, Algorand uses majority rule and their consensus mechanism only has two possible outcomes: accept the transaction or not (timeout or temporary consensus) [32]. In the context of the data market, this would not suffice. The consensus mechanism in the work here presented 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. The disadvantages of majority rule systems such as First-Past-the-Post voting are known of and extensively documented [40], [14], [22]. A common critique of these voting systems is that they do not achieve proportional representation and retain power within a wealthy minority. Consequently, it could be argued that they are not an appropriate consensus mechanism for a context where we aim for decentralisation and fairness.

[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

D. STRUCTURE OF THE PAPER

Firstly we 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 IV, as well as describing how each functional component contributes to achieving the desired properties described in the preceding section. Then we 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 VII, we describe the set of attacks considered, and in section VII-A 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 data ownership models, we 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:

a. How to protect the market against fake data or faulty sensors?
b. Given an oversupply of data, how to ensure that everybody receives a fair amount of write access to the market?
c. How to enable verifiable exchange of data ownership?
d. How to select data points from all those available to add most value to the marketplace?
e. 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 placing trust on 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 are publicly verifiable.

- *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 [58] under the definition of Shapley Fairness, and described in V.8. These are the definitions of fairness that we 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.

- *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 centralisation assumption.

- *Resistant to spam attacks:* The data market should not be susceptible to spam attacks; that is, malicious actors should not have the ability to flood and congest the network with fake or poor quality data.

III. PRELIMINARIES

The architecture for our proposed data market is illustrated in Figure 2, and makes reference to several technology components that are now briefly described in the subsequent section.

A. DISTRIBUTED LEDGER TECHNOLOGY

A distributed ledger technology (DLT) will be used to record access (or any other given right) to a dataset, in the data market. 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 more resilient against failure points than a centralised infrastructure. 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, DLTs strive to satisfy the following properties: have only one version of the ledger state, are scalable, make double spending impossible, and have fast transaction times. One example of a DLT is the IOTA Tangle, shown in Figure 1. 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 [56].

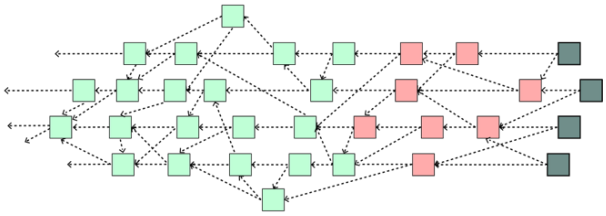


Figure 1. IOTA Tangle. credit: IOTA Foundation

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 [29].

C. CONSENSUS MECHANISMS

A consensus mechanism is a means for a collective to come to an agreement on a given statement. In section I-C3 some examples of consensus mechanisms are discussed that are appropriate for Byzantine environments. Some of these utilise a voting mechanism to enable said consensus, and we now discuss an alternative voting

mechanism that satisfies a different set of properties. It is important to note that there exist numerous methods of aggregating preferences, which are well studied in the field of social choice [57], and voting mechanisms provide different means to enable this aggregation [77].

The taxonomy of voting systems is diverse. They can be either be considered probabilistic or deterministic; proportional or plurality rule; or ordinal as opposed to cardinal. Depending on the set of practical constraints or preferred properties of the implementation context, we encourage selecting an appropriate voting mechanism that best satisfies the desired criteria for a given application. Subsequently, we discuss Maximum Entropy Voting, and why it has desirable properties for the context of this data market.

1) Maximum Entropy Voting

Within the classes of voting schemes, Maximum Entropy Voting (MEV) belongs to the family of probabilistic, proportional and ordinal systems. Arrow famously defined in [9] an impossibility theorem that applies to ordinal voting systems. In it, he states that no ordinal voting systems can satisfy all three of the following properties:

Definition III.1 (Non-Dictatorial). *There exists no single voter with power to determine the outcome of the voting scheme.*

Definition III.2 (Independence of Irrelevant Alternatives). *The output of the voting system for candidate A and candidate B should depend only on how the voters ordered candidate A and candidate B, and not on how they ordered other candidates.*

Definition III.3 (Pareto property). *If all voters prefer candidate A to candidate B, then the voting system should output candidate A over candidate B. Representative Probability states that the probability of the outcome of candidate A being placed above candidate B should be the same as the fraction of the voters preferring the one to the other.*

Whilst this impossibility theorem only applies to ordinal voting systems and not cardinal ones, it has been shown by Gibbard's theorem that every deterministic voting system (including the cardinal ones) is either dictatorial or susceptible to tactical voting [30]. Gibbard later then shows in [31] that the Random Dictator voting scheme satisfies a series of desirable properties, namely: voters are treated equally, it has strong strategy proofness and it is Pareto efficient. With this in mind, the reader can now understand the novelty that the work in [44] presents. Here, MEV is presented as a probabilistic system that first, determines the set of voting outcomes that proportionally represent the electorate's preference, whilst selecting the outcome within this set that minimises surprise. Lets proceed to elaborate: if one were to pick a voting system that is probabilistic and satisfies Arrow's properties to the greatest degree, the adequate system to choose would be Random Dictator. However, whilst computationally an inexpensive method to run, it suffers from a series of drawbacks. The one of greatest concern for the context of this work is the following: imagine a ballot is sampled that happens to contain a vote for an extreme candidate (or in this case, for a malicious actor). The entire choices of an individual that votes for extremes now dictate the entire electorate's leaders. In this scenario, a malicious agent would likely only vote for equally malicious agents, although the number of malicious agents is still assumed to be a minority. Could one reduce the amount of information taken from that sampled ballot? MEV proposes a way to sample ballots that while still representing the electorate's views, minimise the amount of information taken from their preferences. In essence, this is selecting a ballot that reflects the least surprising outcome for the electorate, whilst ensuring that it is still within the set of most representative choices. Furthermore, MEV still satisfies relaxed versions of the Independence of Irrelevant Alternatives and Pareto properties, whilst not being dictatorial. It also enjoys the benefits of proportional voting schemes as well as being less susceptible to tactical voting [44]. 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 [7] because it is costly to have high confidence of success. As a result, we believe MEV offers a suite of benefits and properties that are desirable for the context of the data market presented here.

IV. ARCHITECTURE OF THE DATA MARKET

As can be observed in figure 2, some of the functional 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.

[7] An attack wherein a malicious actor wishes to influence the outcome of an election with high certainty of success at a given instance in time.

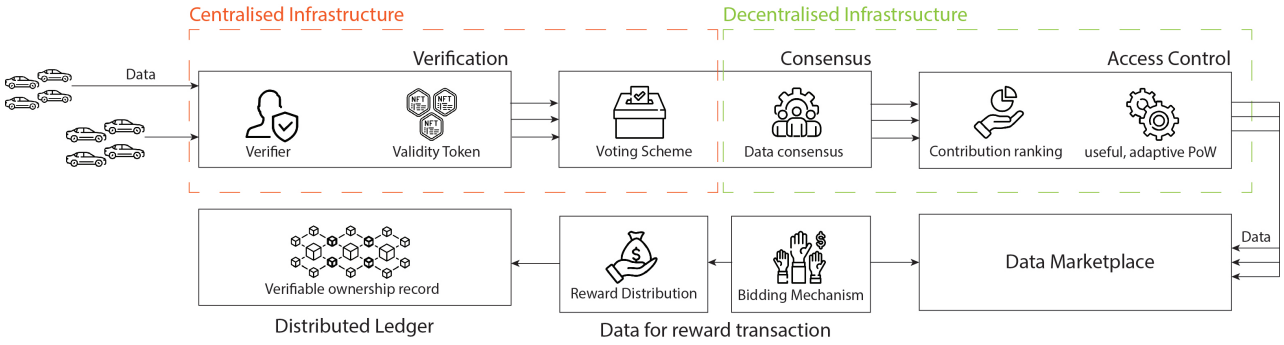


Figure 2. Data Market Architecture. Credit for the images is given in [9]

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 data market. The coalitions with the most valuable data perform the least amount of work.

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, and the reward of this sale is distributed amongst the sellers using the reward distribution function. Successful transactions are recorded on the distributed ledger, and the data corresponding to the successful transactions are removed from the data market.

In what follows, we describe the high level functioning of each of the components shown in figure 2, and how each contribute do achieving the desired properties described in section II.

A. VERIFICATION

Agents are verified by a centralised authority that ensures they provide a valid position, identity and dataset. This component ensures that spam attacks are expensive, as well as enabling verifiable centralisation. All agents in the market can verify the validity of a proof of position and identity because this information is public.

B. CONSENSUS

1) Voting Scheme

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. This is carried out through a voting scheme.

2) Data Consensus

Once a group of trusted agents is elected, they must then come to a consensus as to what the accepted data value of a location is. This is computed by the group following an algorithm that aggregates the coalition's data.

These components enable the property of *decentralised decision making*, allowing coalitions to govern themselves and dictate who to trust for the decision making process. Furthermore, they make uploading fake data to the market costly, as malicious agents must coerce sufficient agents in the voting system, to ensure enough coerced actors will be elected to compute the value of a dataset that they wish to upload.

C. ACCESS CONTROL

1) Contribution Ranking

Once datapoints are agreed upon for the given locations, it is necessary to determine which ones should receive priority when being sold. The priority is given to the data that increases the combined worth of the data market. This can be measured by using the Shapley value, defined in [58], that in this case is used to measure the marginal contribution of dataset towards increasing the value of the market with respect to a given objective function. A precise formalisation is presented in definition V.8. This component provides the property of generalised fairness of the market, and agents with more valuable data should do less work to sell their data.

[9] In order of appearance: Icons made by Freepik, Pixel perfect, juicy_fish, srip, Talha Dogar and Triangle Squad from www.flaticon.com

2) Useful Adaptive Proof of Work

Coalitions must perform a proof of work that is proportional to how valuable to the market their data is deemed. The work performed is adaptive, and furthermore, it is useful to the functioning of the market. This is because the work performed is in fact, calculating the worth of the new incoming data into the market. This feature ensures that spam attacks are costly and that the market is resistant to duplication of profit by simply duplicating data. This is because for every dataset a coalition wishes to sell, they have to complete a PoW.

D. DATA MARKETPLACE

This is where the collected and agreed upon data of specific locations is posted to be sold. The datasets themselves are not public, but rather a metadata label of the dataset, who it is owned by (the spatial coalition that crowd-sourced it) and the location it is associated with. Sellers can access and browse the market and place bids for specific datasets in exchange for monetary compensation. Sellers may wish to purchase access to the entire dataset, to a specific insight or to other defined rights, such as the rights to re distribute or perform further analytics on said dataset. Each right has a corresponding price.

E. BIDDING MECHANISM

The mechanism matches buyers to the sellers of data. This component determines the price-per-right. At this stage, a spatial coalition formed of multiple agents is considered to be one seller. Successful sales will be recorded in an immutable ownership record that is public, such that all participants of the market can see which agents have rightful access to a dataset.

F. REWARD DISTRIBUTION

Once a bid is successful, then the reward of the sale is distributed amongst the participants of the spatial coalition that generated the sold dataset. This is to ensure that all agents participating in the crowd-sourcing of the dataset receive adequate compensation for it.

G. DISTRIBUTED LEDGER

Successful transactions are recorded on a distributed ledger to provide a decentralised, immutable record of ownership. This ledger will represent which agents have access to who's data, and what access rights they are allowed.

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 is calculated by aggregating averages of different pollutant concentrations [10]. To illustrate the function of the proposed data market, we divide the city into a grid with constant sized quadrants. As agents drive across the city they measure pollution concentration values of varying contaminants at different quadrants of the city. Only agents with a valid license plate are granted access to collect and sell data on the marketplace.

B. ASSUMPTIONS

1) For each potential data point that can be made available in the marketplace, there is an oversupply 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 [43], [76] [16] [74] [15].
5) Following from 4 a Proof of Position algorithm is defined in V.14. 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.14.
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.

C. DEFINITIONS

Definition V.1 (Datapoint). *A datapoint is defined as $x_i \in X$ where x_i denotes the data point of agent i and X is the space of all possible measurements.*

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

[10] https://app.cpcbccr.com/ccr_docs/How_AQI_Calculated.pdf

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 $s \in \{1, 2\}$. In stage 1, agents are in the contribution ranking stage, where the value of their data is ranked according to their Shapley value. In stage 2, the must complete a useful, adaptive PoW before they can pass the access control mechanism and enter the data marketplace. For example, agent $a_{5,2}$ is the agent number 5, currently in stage 2 of the access control mechanism. For brevity, in sections where an agent is not in the access control mechanism, we omit the use of the second index.

Definition V.5 (Spatial Coalition)**.** *A spatial coalition is defined as a group of agents in the same location quadrant q. The coalition is denoted as C_q.*

Definition V.6 (Crowdsourced Dataset)**.** *Agents in a spatial coalition C_q aggregate their individual data to provide an agreed upon dataset D_q, of their location quadrant q.*

Definition V.7 (Value Function)**.** *The value function maps the aggregate of datapoints provided by a spatial coaltion to utility for a buyer. For the purpose of this case study, the data provided will be valued with respect to a linear regression model to predict Air Quality Index of a city. The function is denoted as $v(S) = y$ where y is the utility allocated to a dataset and S is a coalition of agents with corresponding datapoints.*

Definition V.8 (Shapley Value)**.** *The Shapley Value is defined in [58] 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 [58].

Definition V.9 (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 allows for the automation of certain tasks [21] [67]. In our context, a smart contract will be executed by agent $a_{i,2}$ to compute the Shapley value of agent $a_{j,1}$'s dataset. The outputs will be the Shapley value of agent $a_{j,1}$'s dataset and a new smart contract for agent $a_{i,2}$. Calculating the new smart contract generated serves as the proof of agent $a_{j,1}$'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.10 (Bidding Mechanism)**.** *Following from the assumption 6, there is a set of buyers M_q for each $q \in \mathcal{L}$ wishing to purchase a dataset D_q from that quadrant. A Bidding Mechanism is defined, BM, as a function that returns a buyer m that will purchase D_q, such that $m \in M$. Consequently, for all $q \in \mathcal{L}$: $m \leftarrow BM(M, D_q)$.*

Definition V.11 (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.12 (Commitment)**.** *An agent commits to their datapoint by generating a commitment that is binding, such that the datapoint cannot be changed once the commitment is provided. A commitment to a datapoint d_i, location quadrant q and ID i of an agent a_i is defined as $c \leftarrow \text{Commitment}(a_i, d_i, q, t)$*

Definition V.13 (Proof of ID)**.** *Let the Proof of Id be an algorithm, PoID, that verifies the valid identity of an agent a_i, 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.14 (Proof of Position). *Let Proof of Position be an algorithm,* PoP*, that is called by an agent* a_i*, 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:*
$\beta \leftarrow \mathrm{a_i^{PoP}}(q, c)$
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.15 (TimeCheck). *The function* TimeCheck *takes in three arguments, the timestamp,* t*, of a datapoint, 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.*
$\gamma \leftarrow \mathrm{TimeCheck}(t, timeNow, r)$

Definition V.16 (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.*
$Token \leftarrow \mathrm{Verify}(\alpha, \beta, \gamma)$

Definition V.17 (Reputation). *An agent* a_i *assigns a score of trustworthiness to an agent* a_j*. This score is denoted as* $r_{i \to j}$*. This reputation is given by one agent to another in a rational, efficient and proper manner, and is an assessment of honesty.*

Definition V.18 (Election Scheme). *We use a generalised definition for voting schemes, following from the work in [45] and [59].An* Election Scheme *is a tuple of probabilistic polynomial-time algorithms* $(\mathrm{Setup}, \mathrm{Vote}, \mathrm{Partial-Tally}, \mathrm{Recover})$ *such that:*

Setup *denoted* $(pk, sk) \leftarrow \mathrm{Setup}(k)$ *is run by the administrator. The algorithm takes security parameter* k *as an input, and returns public key* pk *and private key* sk*.*
Vote *denoted* $b \leftarrow \mathrm{Vote}(pk, v, k)$ *is run by the voters. The algorithm takes public key* pk*, the voter's vote* v *and security parameter* k *as inputs and returns a ballot* b*, or an error* ($\perp$)*.*
Partial − Tally *denoted* $e \leftarrow \mathrm{Partial-Tally}(sk, \mathfrak{bb}, k)$ *is run by the administrator. The algorithm takes secret key* sk*, bulletin board containing the list of votes* $\mathfrak{bb}$*, and security parameter* k *as inputs and returns evidence* e *of a computed partial tally.*
Recover *denoted* $\mathfrak{v} \leftarrow \mathrm{Recover}(\mathfrak{bb}, e, pk)$ *is run by the administrator. The algorithm takes bulletin board* $\mathfrak{bb}$*, evidence* e *and public key* pk *as inputs and returns the election outcome* $\mathfrak{v}$*.*

Definition V.19 (Ballot Secrecy). *Ballot secrecy can be understood as the property of voters having a secret vote; namely, no third party can deduce how a voter voted. We utilise the definition for* Ballot Secrecy *presented in the work [59]. This definition accounts for an adversary that can control ballot collection. The adversary must be able to meaningfully deduce which set of votes they have constructed a bulletin board for. This definition is formalised as a game where, if the adversary wins with a significant probability, the property of* Ballot Secrecy *does not hold. An* Election Scheme *is said to satisfy* Ballot Secrecy *if for a probabilistic polynomial-time adversary, their probability of success is negligible.*

VI. THE DATA MARKET

A. THE VERIFICATION ALGORITHM

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

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

The validity of the data submission must be verified before the data reaches the data marketplace, to avoid retroactive correction of poor quality data. This is done through the VerifyingAlgorithm. Firstly, an agent provides an immutable commitment of their datapoint, location quadrant, timestamp and unique identifier (Line 1). Next, the agent submits their unique identifier to a centralised authority that verifies that this is a valid and real identity (Line 2). In practise, for this context, this identifier will be the agent's vehicle license plate. Subsequently, the agent generates a valid proof of position (Line 3). Following from assumption 2, an agent can only provide one valid outcome from algorithm V.14 at a given time instance t. Then, the datapoint is checked to ensure it is not obsolete through TimeCheck (Line 4). Finally, the outputs of all previous functions are verified to ensure the agent has produced a valid proof (Line 5). 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 (Line 6).

B. VOTING SCHEME: REPUTATION-BASED MAXIMUM ENTROPY VOTING

In what follows we present an adaptation of the Maximum Entropy Voting scheme that takes into consideration the reputation of agents in the system. Both components are introduced and will work as a single functional building block in the the data market design.

1) Reputation

In the general sense, reputation can be seen as a performance or trustworthiness metric that is assigned to an entity or group. For the purpose of this work, reputation should be earned through proof of honesty and good behaviour. In this case, agents that can demonstrate they have produced an honest computation should receive adequate recompense.

In our context, agents can be administrators by running an election. They must prove that an election outcome was correctly computed to earn reputation.

In the case of Maximum Entropy Voting, the administrator running the election must prove that: the voting outcome was correctly computed, and that it does indeed satisfy the optimisation formulation defined in equations 5. To provide guarantees of correctness to the voters, we propose using an end-to-end (E2E) verifiable voting scheme. E2E voting schemes require that all voters can verify the following three properties: their vote was *cast as intended*, *recorded as cast* and *tallied as cast* [3]. An example of an E2E voting scheme is Helios [1]. This scheme uses homomorphic encryption, enabling the administrator to demonstrate the correctness of the aggregation of votes operation. The operations required in the aggregation of votes for MEV can be done under homomorphic encryption, and an E2E voting scheme such as Helios could be used to carry out this step. This aggregation is then used to solve the optimisation problem and yield a final vote outcome. Once the optimisation is solved, the administrator can release the aggregation of votes and prove that the aggregation operation is correct and that the solution of the optimisation problem satisfies the KKT conditions. Upon presenting the verifiable proofs mentioned above, agents behaving as administrators should receive reputation from other agents in the network.

Remark: We note that the Helios voting scheme has been proven not to satisfy Ballot Secrecy in [59] and [45], although a variant of Helios that does satisfy Ballot Secrecy is proposed in [60]. Proposing and testing an E2E verifiable voting scheme that satisfies definitions of Ballot Secrecy, receipt-freeness and coercion resistance is beyond the scope of this work, although of interest for future work.

2) Reputation-based Maximum Entropy Voting

Definition VI.1 (Vote). *The vote of agent $a_i \in A$, is defined as a pairwise preference matrix in $S(i)\mathbb{R}^{N\times N}$. Each entry is indexed by any two agents in A and its value is derived from datapoint x_i V.1 and reputation $r_{i\to j}$ V.17. An example of a pairwise preference matrix for three agents is shown in equation (2).*

Definition VI.2 (Administrator). *An agent that carries out the vote aggregation and the computation of the election outcome is defined as an administrator, $\mathbb{A}$.*

Definition VI.3 (Aggregation of Votes). *The aggregation of all agents' votes $S(A)$, is defined as the average of $S(i), i \in A$, as follows:*

$$S(A) := \frac{1}{N}\sum_{a_i\in A} S(i). \tag{1}$$

Definition VI.4 (Agent Ordering). *An agent ordering, denoted as t, is defined as a permutation of agents in [44], i.e., arranging all agents in order. Further, concerning computation complexity, we suggest t being a combination of agents, i.e., selecting a subset of agents as the preferred group, such that the order of agents does not matter.*

Definition VI.5 (Ordering Set). *The ordering set $\mathcal{T}$ is the set of all possible agent orderings, such that t is an element of $\mathcal{T}$. See Figure 3 for the example of an ordering set of combinations with 3 agents.*

Definition VI.6 (Probability Measure of Ordering Set). *The (discrete) probability measure, $\pi : \mathcal{T} \to \mathbb{R}_{\geq 0}$ gives a probability of each ordering $t \in \mathcal{T}$ being selected as the outcome ordering t^*. The measure π of maximal entropy whilst adhering to Representative Probability, described in definition III.3, i.e., the optimal solution of the optimisation problem defined in equations 5 is denoted as π^*.*

Given a set of agents of cardinality $|A| = N$, each agent a_i has a data point $x_i \in X$ and a reputation $r_{i\to j}$ for all agents $a_k \in A$. The data point x_i in this context is defined as measurements of pollutants of an agent which they want to submit and sell. The reputation $r_{i\to j} \in \mathbb{R}^+$ is a non-negative value that represents the individualised reputation of agent a_j from the perspective of agent a_i.

To combine maximum entropy voting and reputation, a key step is to move from reputation $r_{i\to j}$ to a pairwise preference matrix $S(i) \in \mathbb{R}^{N\times N}$. The entry of a pairwise preference matrix is indexed by every two agents of A, and its values is defined as follows:

$$S(i)_{j,k} = \begin{cases} 1 & \text{if } a_i \text{ prefers } a_j \text{ and } j \neq k \\ 0.5 & \text{if } a_i \text{ prefers both equally and } j \neq k \\ 0 & \text{if } a_i \text{ prefers } a_k \text{ or } j = k \end{cases}, \tag{2}$$

for $a_j, a_k \in A$ and a_j is preferred to a_k if and only if $\frac{1+|x_i|\cdot r_{i\to j}}{1+|x_i - x_j|} > \frac{1+|x_i|\cdot r_{i\to k}}{1+|x_i - x_k|}$ and both agents are equally preferred if the two values are equalised, such that the reputation is scaled by their absolute differences from agent a_i. Likewise, we could find a pairwise preference matrix $S(i)$ for each agent a_i. The average of pairwise preference matrices over all agents are denoted as the preference matrix $S(A)$, as in 1. $S(A)$ represents the pairwise preference of all agents in A, whose entries

$S(A)_{j,k}$, displays the proportion of agents that prefers agent a_j over agent a_k.

The original MEV [44] runs an optimisation over all candidate orderings, which strongly defines the computational complexity of the problem because the number of orderings is the factorial of the number of candidates. As a variant of MEV, we consider agent combinations, instead of permutations for the ordering set $\mathcal{T}$, such that A is divided into a preferred group $\mathcal{P}$ of cardinality M and non-preferred group $\mathcal{NP}$, where M is the number of winners needed. Hence, the cardinality of the ordering set decreases from $N!$ to $\frac{M!}{M!(N-M)!}$. For small M, this leads to to a dramatic reduction of the computational complexity.

For each ordering $t \in \mathcal{T}$, we could define its pairwise preference matrix $S(t)$, whose entry is defined in equation (3), and likewise in equation (2):

$$S(t)_{j,k} = \begin{cases} 1 & \text{if } a_j \text{ is placed over } a_k \\ 0.5 & \text{if both are in the same group and } j \neq k \\ 0 & \text{if } a_k \text{ is placed over } a_j \text{ or } j = k \end{cases}, \tag{3}$$

for $a_j, a_k \in A$. Let us define an **unknown** probability measure $\pi : \mathcal{T} \to \mathbb{R}_{\geq 0}$. $\pi(t), t \in \mathcal{T}$ gives the probability of t being chosen as the outcome ordering. Then, we construct a theoretical preference matrix $S(\pi)$ as follows:

$$S(\pi) := \sum_{t \in \mathcal{T}} \pi(t) \cdot S(t). \tag{4}$$

The entry $S(\pi)_{j,k}$ states that under probability measure π, the probability of the outcome ordering placing a_j over a_k. Recall the definition of Representative Probability in Section III.3 or [44], it simply requests $S(\pi) = S(A)$.

The entropy of π measures the uncertainty of choosing elements in $\mathcal{T}$. The uniform distribution has the maximum amount of entropy. Associated with π, the entropy is defined as $-\sum_{t \in \mathcal{T}} \pi(t) \log \pi(t)$ [33]. Hence, the original formulation of maximum entropy voting adhere to Representative Probability is as (5). In this formulation, when maximising the entropy, we ensure the solution π^* to be the most moderate probability measure with obeying Representative Probability in III.3.

$$\begin{aligned} \pi^* = \max_{\pi} & -\sum_{t \in \mathcal{T}} \pi(t) \log \pi(t) \\ \text{s.t.} & \sum_{t \in \mathcal{T}} \pi(t) \cdot S(t) = S(A) \\ & \sum_{t \in \mathcal{T}} \pi(t) = 1 \\ & \pi(t) \geq 0 \quad \forall t \in \mathcal{T} \end{aligned} \tag{5}$$

3) A Motivating Example

Consider $A = \{a_i, a_j, a_k\}$ and only one winner is needed ($M = 1$), all possible combinations are in shown in Figure 3, while the number of permutations would be $3!$.

$\mathcal{T}$	t_1	t_2	t_3
Preferred $\mathcal{P}$	(a_i)	(a_j)	(a_k)
Non-Preferred $\mathcal{NP}$	(a_j, a_k)	(a_i, a_k)	(a_i, a_j)

Figure 3. The lower-carnality ordering set when $A = \{a_i, a_j, a_k\}$ and $M = 1$. Agents in the same brackets are given the same rank in an ordering.

As an example, the pairwise preference matrix $S(t_1)$ is displayed in (6), following the definition in (3).

$$S(t_1) = \begin{array}{c|ccc} & a_i & a_j & a_k \\ \hline a_i & 0 & 1 & 1 \\ a_j & 0 & 0 & 1/2 \\ a_k & 0 & 1/2 & 0 \end{array} \tag{6}$$

Suppose an optimal measure π^* is extracted from the optimisation problem in equations 5. Assuming $\pi^*(t_1) = 0.3$, $\pi^*(t_2) = 0.4$ and $\pi^*(t_3) = 0.3$, to sample an outcome ordering t^* from π^*, consider a prize wheel as in Figure 4. The wheel includes $|\mathcal{T}|$ wedges where each wedge represents one ordering t and takes the share of $\pi^*(t)$. To obtain an outcome ordering, simply spin the wheel and t^* is the wedge where the red arrow stops, i.e., t_1 in Figure 4.

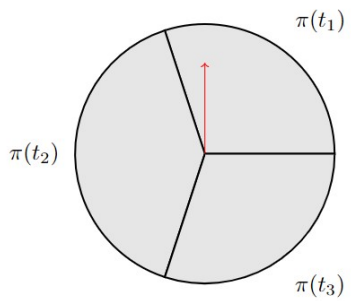


Figure 4. A prize wheel for sampling an outcome ordering $t \in \mathcal{T}$ from a probability measure π.

Maximum Entropy Voting is summarised in the following steps: first, each agent a_i constructs their vote, the pairwise preference matrix $S(i)$, from the data point x_i and reputation $r_{i \to j}$. Then, an average of all agents' pairwise preference matrix $S(A)$ is calculated by the administrator $\mathbb{A}$, which is seen as the aggregation of all agents' votes. Then, a low-cardinality ordering set of agents $\mathcal{T}$ is constructed from M, the number of necessary winners needed. For every possible ordering of candidates t, a theoretical pairwise preference matrix $S(t)$ is constructed. These two steps can be computed by any agent in the election, or the administrator. Then, the administrator solves the optimisation problem to maximise entropy as defined in equation 19 to find a probability measure π^* of a given ordering. This probability measure also adheres to the Representative Probability property III.3. Finally, the administrator samples

an outcome ordering t^* from π^*, independently and at random, using a "prize-wheel" sampling, as shown in Figure 4. This ordering is the final election outcome.

C. DATA CONSENSUS

In an oversubscribed environment, crowd-sourcing can be used to estimate an agreed upon measurement which should reflect the ground truth as closely 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 are: 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.

1) K-Anonymity

One way to define k-anonymity is that data sourced from multiple agents satisfies k-anonymity if any individual data point cannot be related to less than k agents, where k is a positive integer. [54][11] In other words, if a agreed upon dataset includes multiple measurements and is assigned a k-anonymity with $k = 2$, it is not possible to identify a single measurement without a second agent revealing their measurement.

2) Breakdown Point

In general, the breakdown point characterises the robustness of an estimator and is usually dependent on the sample size n. For this work the definition given below is used to characterise the *theoretical breakdown point*. In such a way the theoretical breakdown point BP_{th} characterises the minimum share of malevolent agents needed to break the system and alter the the agreed upon dataset arbitrarily, given the worst case configuration of the system. [35]. Complementary to that we define the *practical breakdown point* in definition VI.7

Definition VI.7 (Practical Breakdown Point)**.** *The practical breakdown point BP_{pr} is the average share of malevolent agents at which the agreed upon dataset is arbitrarily altered, given naturally occurring configurations of the system.*

3) Mean

The mean $\bar{x}$ in its simplest form is defined in equation (7).

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{7}$$

[11] latanyasweeney.org; k-anonymity

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

The mean can be calculated in a decentralised and privacy preserving manner [49]. The k-anonymity of the mean results then in $k = n - 1$. The theoretical 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. However, even with this mitigation in place, a larger coalition of malevolent agents is still able to influence the mean significantly. This, in combination with the fact that the presence of malevolent agents can be expected in a data market, may suggest that the mean is not sufficiently robust to compute an agreed upon dataset for most use cases.

4) Median

The median 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:

$$\mathrm{median}(x) = \begin{cases} x_{(n+1)/2} & \text{if n is odd} \\ \frac{1}{2} \cdot (x_{(n/2)} + x_{(n/2)+1}) & \text{if n is even.} \end{cases} \tag{8}$$

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 a selection algorithm that finds the kth smallest element(s) as long as the data of the agents can be shared with other agents in their coalition. If data cannot be shared with other members, calculating the median in a privacy-preserving way demands a more complex scheme [26] and is not trivial. The theoretical breakdown point of the median characterises it as one of the most robust estimators and is for the worst case given with $\frac{1}{2}$.

5) Mean Median Algorithm

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 an agreed upon dataset in a robust and privacy-preserving way. It must be said that it is a compromise and this algorithm is not as robust as the median and not as privacy-preserving as the mean, when compared individually.

Explaining the algorithm, the first step is to randomly assign every agent to a group in such a way that there are g groups with at least s agents each. The way the parameters g and s are chosen determine the anonymity and robustness properties of the algorithm and will

be discussed in the simulation section VII-C. 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 g groups, there are g ways in which the median is chosen. This gives a breakdown point given in equation 9.

$$\text{Breakdown point of } \text{meanmedian}(x) = \frac{g}{2n} \tag{9}$$

The relationship between s, g and the number of agents n, is given with the inequality (10).

$$n \geq s \cdot g \tag{10}$$

D. STAGE 4: ACCESS CONTROL MECHANISM TO THE DATA MARKET

The previous stages of verification, voting and data consensus run simultaneously in numerous rounds, as vehicles sense data and form coalitions to provide an agreed up value for a given location quadrant. By the time they reach the access control, 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) Contribution ranking: 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 manner, the Shapley value is used as a metric to rank the most valuable datapoints with respect to a value function.

2) Useful, adaptive 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, $\mathbf{D_{t+1}}$.

3) A contextual example

In the context of agents measuring different levels of pollution, we illustrate an example of how the Shapley value would be used to rank the datapoints in terms of value, and allocate a proportional proof of work correspondingly.

We use the data on pollution levels of a range of different contaminants, taken from a number of cities in India. The data has been made publicly available by the Central Pollution Control Board: [12] which is the official portal of Government of India. The cleaned and processed data was accessed from [13]. We illustrate an example wherein a public authority is interested in purchasing data on pollution levels of different contaminants in order to predict the Air Quality Index (AQI) of a given location. We generate a linear regression model to predict AQI, which has been previously done in [61], although other options for models to predict AQI have been explored in alternative works such as [4] and [38]. We note that it is up to the buyer to select a model that best defines the objective they wish to achieve.

A description of how AQI is calculated can be found in [14]. Following from this calculation, it is reasonable to observe how the variables PM2.5 (Particulate Matter 2.5-micrometer in $\mu g/m^3$) and PM10 (Particulate Matter 10-micrometer in $\mu g/m^3$) are highly correlated with AQI. It can be seen from Figure 5 that these are the two variables with the highest correlation to AQI. We include them as well as NO, NO2, NOx, NH3, CO, SO2 and O3 as training features for the linear regression model.

Agents collecting measurements of different pollutants have their dataset evaluated by a preceding set of agents that must calculate some proof of work. They receive the seller's objective value function, which in this case is the linear regression model, and access to another agent's dataset. We show the results of calculating the Shapley value of individual datapoints within a given dataset in Figure 6. We simulate this using the SHAP library, presented in [42]. Following from the SHAP documentation: "Features pushing the prediction

[12] https://cpcb.nic.in/
[13] https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india
[14] https://app.cpcbccr.com/ccr_docs/How_AQI_Calculated.pdf

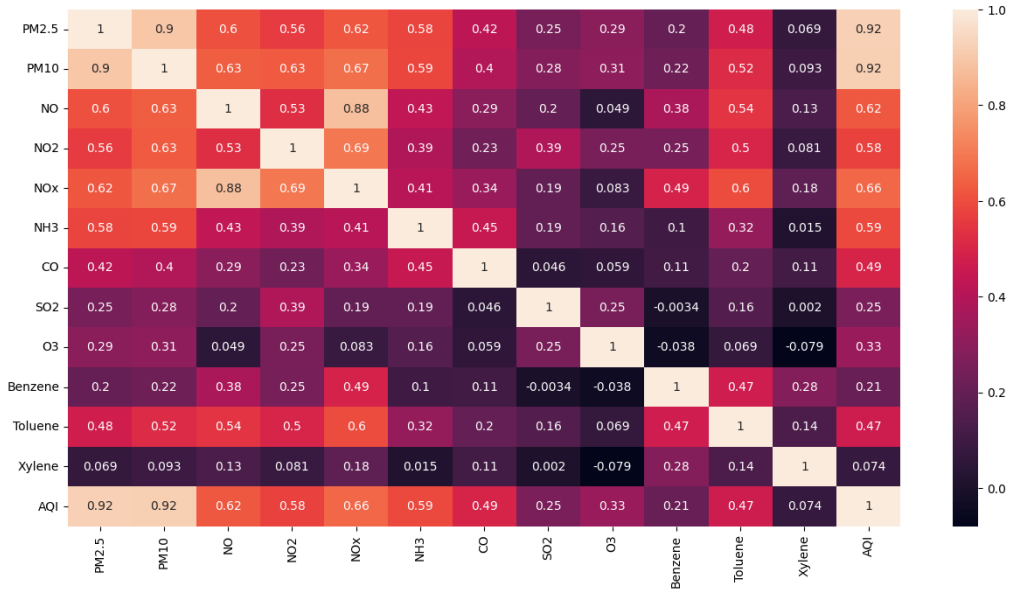


Figure 5. Correlation Matrix of the different pollutants measured with AQI

higher are shown in red, those pushing the prediction lower are in blue. The plot below sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red high, blue low)" [41]. This reveals that high PM2.5 and PM10 concentration increases the predicted AQI.

We also show the mean absolute value of the SHAP values for each pollution contaminant in Figure 7. From this we can deduce that any measurement belonging to the highest SHAP value classes will be deemed more worthy and thus the agent submitting it will have to perform less proof of work to sell it. Every spatial coalition C_q would have their own total Shapley value, which is the aggregate of the Shapley values shown in Figure 7.

4) Privacy Concerns

The reader may rightly question the privacy risks of an agent accessing another one's dataset to compute the Shapley value. What is to incentive them to compute the Shapley value honestly, and what is to prevent them from stealing or duplicating another agent's data if they realise it has a high Shapley value?

To address the first concern, a Shapley value calculation is only accepted and considered complete once enough agents have agreed on the same outcome. With the assumption that the system is Byzantine, we assume that at least $2/3$ of agents are honest, and thus once a consensus is agreed on the value, it must be true.

Secondly, in the market there is no protection against agents duplicating data, but they cannot monetise this copied data unless they go through the verification, consensus and then access control stages again. Because we are in an environment with an oversupply of data and that is crowd-sourced, the data is unlikely to be highly sensitive and thus the incentive to go through these steps is very small.

Finally, we address the concern of having a public value function. We note that the value function is not the same as the buyer's predictive task, but rather the mathematical representation of the market's utility function. This information should be public, as it is the way the market agrees to assign value to data. Given that we propose this work in the context of a collective environment where the value function should dictate the entire market's objective, this should be public knowledge and not sensitive Intellectual Property. Furthermore, malicious buyers could attempt to propose an objective function that penalise data they are interested in, but that would not ensure that they would pay a lower price for their desired data, it would only delay the access of said data into the market. The price would be dictated by the bidding mechanism, which is something that can be agreed upon by the entire collective to prevent issues like the aforementioned one.

VII. ADVERSARIAL ATTACKS

In an environment like decentralised networks or data markets one must take into account the possibility of attacks on the system. we proceed to describe their

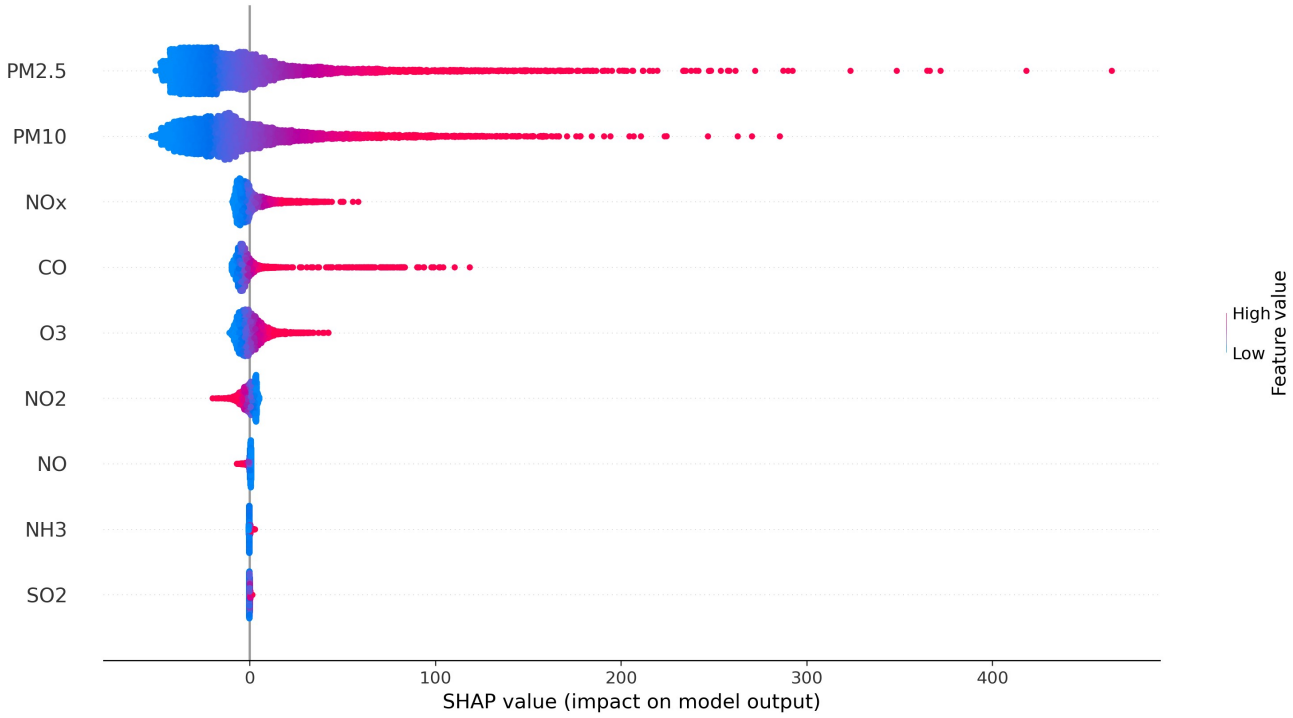


Figure 6. SHAP Values of the samples of each feature

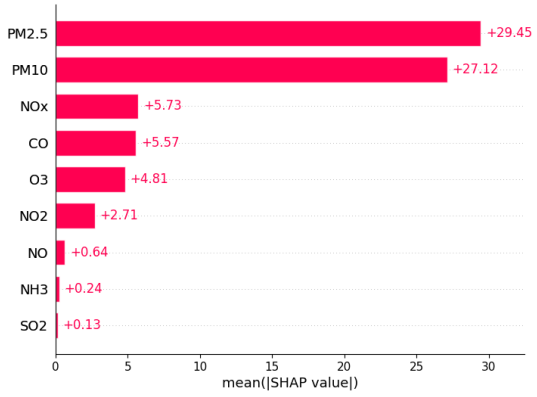


Figure 7. Feature ranking

nature and how these are mitigated by the functional components of the data market architecture.

Definition VII.1 (Sybil Attack)**.** *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 [25]*

Sybil attacks are mitigated in the verification stage, as agents must present a valid proof of identity. This proof is granted to them through a centralised authority but all other agents can verify that it exists and therefore that it must be valid. Generating multiple identities is made expensive in this proposed architecture, because agents must provide a valid license plate to enter the market and collect data. Unless the attacker purchases a real vehicle with a valid license plate, they cannot succeed in creating another identity, and therefore sell data in the market.

Definition VII.2 (Wormhole Attack)**.** *A Wormhole Attack involves a user maliciously reporting they are in a location that is not the one they are truly in [34].*

An attack can be mounted by a series of malicious actors claiming to measure data from a location they are not truly in, and wishing to monetise this fraudulent data. To mitigate against this attack, agents must present a valid proof of position in the verification stage (defined in V.14). This proof is assumed to be correct and sound, and by definition, agents are only able to present one valid proof.

Definition VII.3 (Data Poisoning)**.** *Data Poisoning is an attack where malicious agents collude to report fake data in order to influence the agreed upon state of a system [64].*

Malicious agents wishing to report fake data must influence enough agents in their spacial coalition to ensure that sufficient agents in the data consensus stage

will compute a fake data point. Probabilistic voting schemes make the cost of this coercion significantly high. Furthermore, to sell the uploaded data point, the agent must perform a useful proof of work that is proportional to how valuable the data point is deemed. The more useful the data point the less work the agent must carry out to sell it. Selling spam data will therefore be very time consuming for an attacker.

A. EVALUATION

In this chapter the data market as well as individual parts of it are analysed to assess the robustness against the earlier introduced attacks. Simulations of the trust and truth consensus mechanism are carried out to gain deeper insights.

As previously discussed in section VII, there are four types of unwanted instances that this work focus on, namely Sybil Attacks, Wormhole Attacks, Data Poisoning, and Faulty Data.

Data poisoning and faulty data are both similar in the sense that untrue measurements are submitted for different reasons. In the case of Faulty data this happens unintentionally while data poisoning is intentional. Further, due to the (assumed) random nature of faulty data, where untrue measurements happen to be on all sides of the ground truth, it can be said that it rather cancels each other out, when the agreed upon data is estimated. In contrast, when multiple agents build a malevolent coalition to influence the vote by submitting the same untrue measurement, their influence is greater. Therefore, data poisoning can be seen as the worst case scenario among the two and by investigating it, a bound for both can be found. Note that the case of Faulty data with the same systematic error on multiple sensors results in the same outcome as data poisoning, with the difference that the bias is chosen randomly. To further investigate data poisoning, simulations have been carried out.

B. SIMULATION SETUP

A class of agents was created and a ground truth established from which the honest agents measure their data point. To account for imperfect sensors and other sources of errors, the process of taking a measurement is represented by sampling from a Gaussian distribution with μ and σ. Additionally, a set of agents was created which have the same untrue measurement μ_{adv} to represent the group of dishonest agents forming a malevolent coalition to mount a *data poisoning* attack.

Further, a base reputation of 1 is assigned to all agents, and in a second step, every honest agent has a probability to have a high reputation assigned. This is modeled by a weighted coin toss deciding if the agent is assigned a high reputation, and if yes, the reputation is sampled from a Gaussian with μ_{rep} and σ_{rep}.

To simulate the governance and consensus mechanisms, models of the different data consensus algorithms and the voting mechanism were built and applied to the created agents. It is important to note that the mean-median algorithm was implemented twice with different parameters, namely triplets and squareroot. The former means that the minimum number of agents per group is $s = 3$, while for the squareroot implementation it is chosen dynamically depending on the number of agents N, with $\sqrt{N} = s$. In return, the triplets algorithm has a higher number of groups g than the squareroots implementation, and therefore a higher robustness can be expected, as discussed in section VI-C.

The simulations have been done with S number of samples using Monte Carlo methods, varying numbers of agents N, and differently sized malevolent coalitions. Note that this setup assumes the honest actors to be independently, identically distributed, which implies that measurement errors are not systematical or correlated. It can be translated to a world where every agent takes their measurements independently with the same unbiased sensor system and spatio-temporal effects do not occur (within a spatial coalition).

Remark: We note that the purpose of the simulations in Figures 8 and 9 is to characterise the robustness of the data consensus algorithms against *data poisoning* attacks. Therefore, the absolute value of the data that each agent submits is not of importance. Rather, our objective is to understand how resilient the algorithm is when a series of malevolent agents collude to send the same, malicious value, that differs highly from the ground truth.

C. EVALUATION OF RESULTS

Figure 8 shows a simulation of the different data consensus algorithms which can be used to find the agreed upon dataset. The simulation has been repeated $S = 15$ times with $N = 1000$ agents. The behaviour of the system with a high number of agents can be considered to reflect the upper limit scenario of the system, and scaling effects can be observed when the number of agents are lower. It is important to examine this scaling effect because in practise, varying numbers of agents will occur. To do so, the behaviour in the limit is examined to establish the baseline.

Generally, it can be seen that the more adversaries are present, the higher is the deviation of the agreed upon dataset from the ground truth.

For the median algorithm, in green, and the triplets algorithm, in blue, discrete steps can be seen, where at defined percentages of adversaries, the deviation of the agreed upon dataset changes drastically. For the median algorithm, this happens once at 50% which is the theoretical breakdown point. For the triplets algorithm this happens three times, which reflects the fact that there are three agents per group. It can be

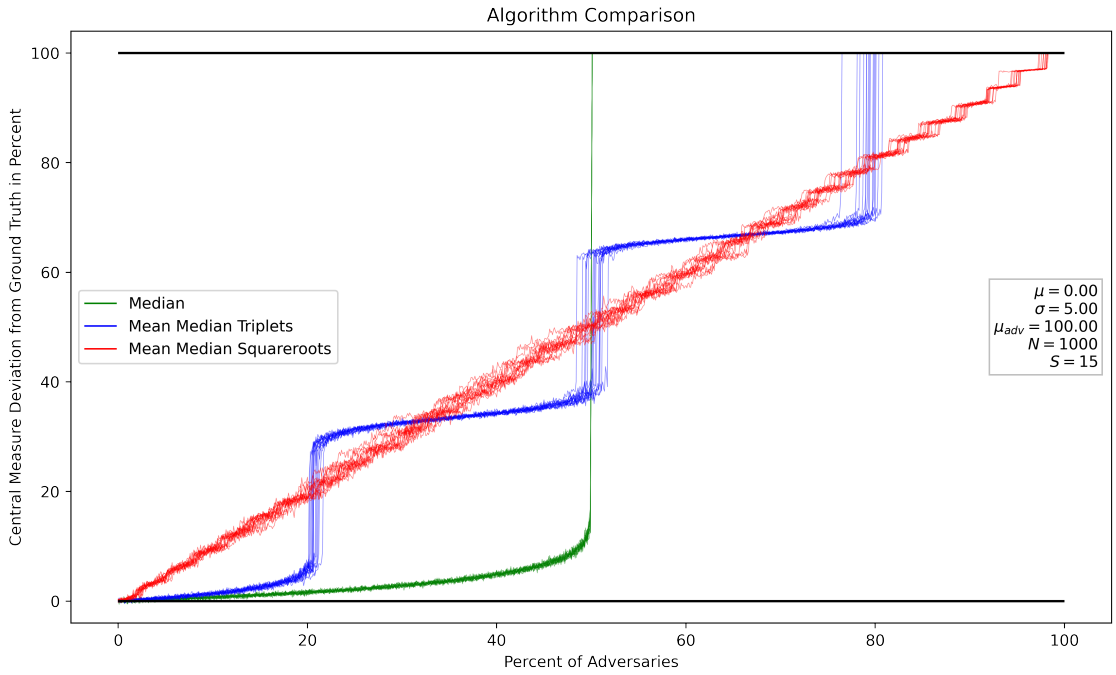


Figure 8. Characterisation of data consensus algorithms' behaviour under different degrees of coordinated data poisoning attacks

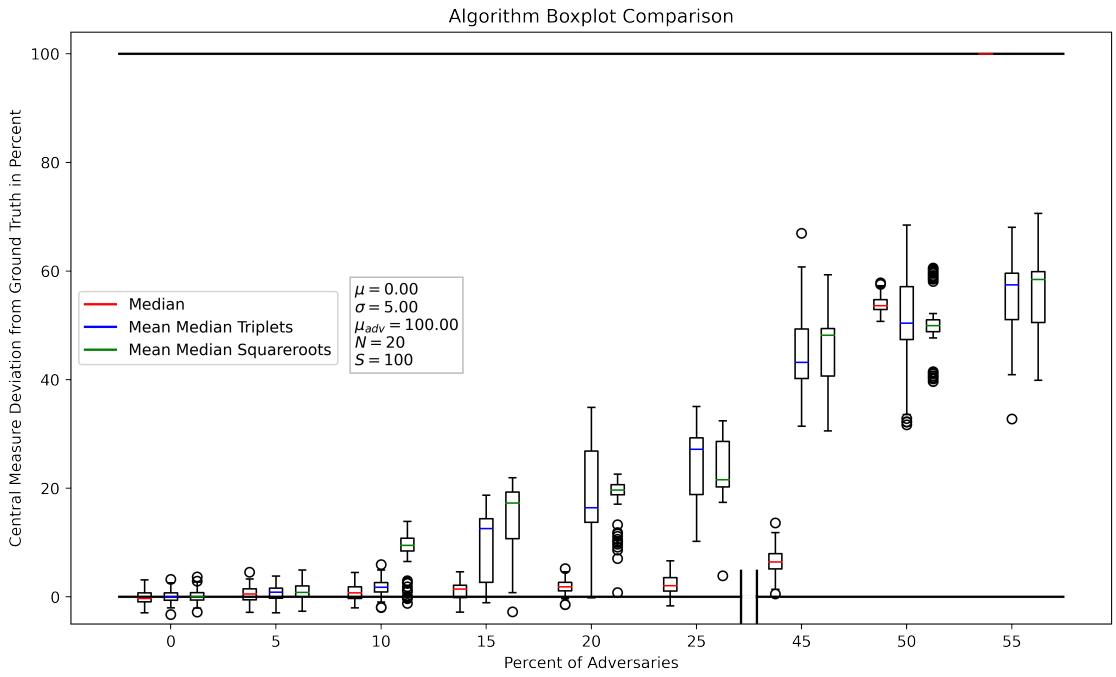


Figure 9. Breakdown analysis of data consensus algorithms, with a coordinated data poisoning attack

interpreted that for up to 20% of adversaries, there are high chances of all three agents in the median group being honest. Further, between 20% and 50% it is likely that one out of three is malevolent. This continues up to where from 80 % onwards, there are high chances of all agents in the median group being malevolent.

Between the jumps it can be observed that there is an upwards trend. This can be explained with the fact that the honest actors sample their data from a Gaussian distribution with non zero σ. Due to the fact that more and more adversaries are distorting the data, honest agents further away from the μ are selected.

Lastly, the squareroots algorithm shows a fairly continuous, almost linear, reaction to adversaries. This is a result of the bigger groups, with $s = \sqrt{N}$ and $N = 1000$, which allows finer blends of honest and malevolent agents in the median group. In theory, there should be also a discrete step-wise increase similar to the triplets algorithm because of the discrete number of agents s underpinning the process. However, due to the smaller step size and the inherent randomness to the simulation, this can only be observed in the region between 80% and 100% of adversaries.

To conclude, it can be said that at the baseline, the median algorithm is the most robust with a practical breakdown point at 50%, after which comes the mean median triplets with 20%, and the mean median squareroot with a breakdown point of about 2%. When using equation (9) to calculate the theoretical breakdown points for the mean median algorithm of 16.65% and 1.55%, it is easy to see, at least for the triplets implementation, that in the practical breakdown points are higher.

To investigate the scaling effect and behaviour of the algorithm when smaller numbers of agents are present, the same simulation was carried out but with $N = 20$ agents, see Figure 9, and presented as boxplot instead. This results in bigger steps in which the percent of adversaries is increased. For $N = 20$, adding one adversary translates to an increase in 5%. The number of repeated sampling $S = 100$ was increased to compensate the increased uncertainty due to the lower agent count. The boxplot shows the three algorithms for 9 different shares of adversaries, namely 0, 5, 10, 20, 25, 45, 50, and 55.

It can be seen that in general, the algorithms act similarly as in Figure 8 with two major differences. First, the boxplot clearly shows that given a share of adversaries, the spread of the deviation is higher. This is a direct result of the lower number of agents, where chances of fluctuations are higher. Secondly, the practical breakdown point is shifted to 10% and 15% for squareroot and triplets, respectively. Given equation (9), the theoretical breakdown points for the mean median implementation are 10% and 15% which confirms the observation.

For the squareroot implementation, this is a great improvement which can be explained by the low number of agents. This results in four groups $g = 4$ which is in proportion to the number of agents $N = 20$ higher than for $N = 1000$ with $g = 31$ groups. This shifts the theoretical breakdown point and therefore also the practical one. The explanation behind the shifting of the practical breakdown point of the triplets implementation lies less in a shift of the theoretical breakdown point (although there is a slight one) and more in the random nature of the allocation of the agents to the groups. Specifically, it is more likely to end up with two or more malevolent agents per group with higher numbers of agents. Equation (11) gives the probability that the breakdown occurs at the theoretical breakdown point, given the share of adversaries of the theoretical breakdown point. The chances are higher for the case of triplets with $N = 20$, $r = 6$ and $a = 3$, than for triplets with $N = 1000$, $r = 333$ and $a = 167$.

$$\mathrm{BP}_{th} = \frac{(g-1)!}{g^{a-1}(g-a)!} \tag{11}$$

where BP_{th} is the probability of the breakdown of the Mean Median Algorithm at the theoretical breakdown point, g is the number of groups, and a is the number of adversaries at the theoretical breakdown point.

Figure 10 represents the entire consensus mechanism, whereby the main purpose is to demonstrate that the reputation system can increase the robustness, given a functional reputation system exists. The way this simulation works it that the voting scheme (MEV) outputs a set of agents, the committee, which then compute the agreed upon dataset of a given location. Given the context of this data market, where anonymity is not demanded, the use of the median algorithm as a data aggregation algorithm can be justified. The size of the committee was chosen with $K = 3$ to be able to make use of the proposed reduction in computational complexity, described in section VI-B.

Figure 10 shows two simulations plotted which present the share of adversaries at which the practical breakdown point occurs (y-axis) as function of the share of highly reputational agents among the honest ones (x-axis). The latter resembles the weight of the weighted coinflip mentioned above. The plot in blue has $N = 15$ and $S = 10$, and the one in orange, $N = 10$ and $S = 15$. For both simulations the same Gaussian was used to sample the reputation for the high-reputation honest actors, $\mu_{adv} = 100$ and $\sigma_{adv} = 30$.

For both simulations, a clear trend can be observed that with higher shares of reputational agents among the honest ones, the practical breakdown point is at higher shares of adversaries, with some outliers having a breakdown point of more than 80%. Precisely, this means that two out of the three committee members are honest. At the same time it is visible that a great

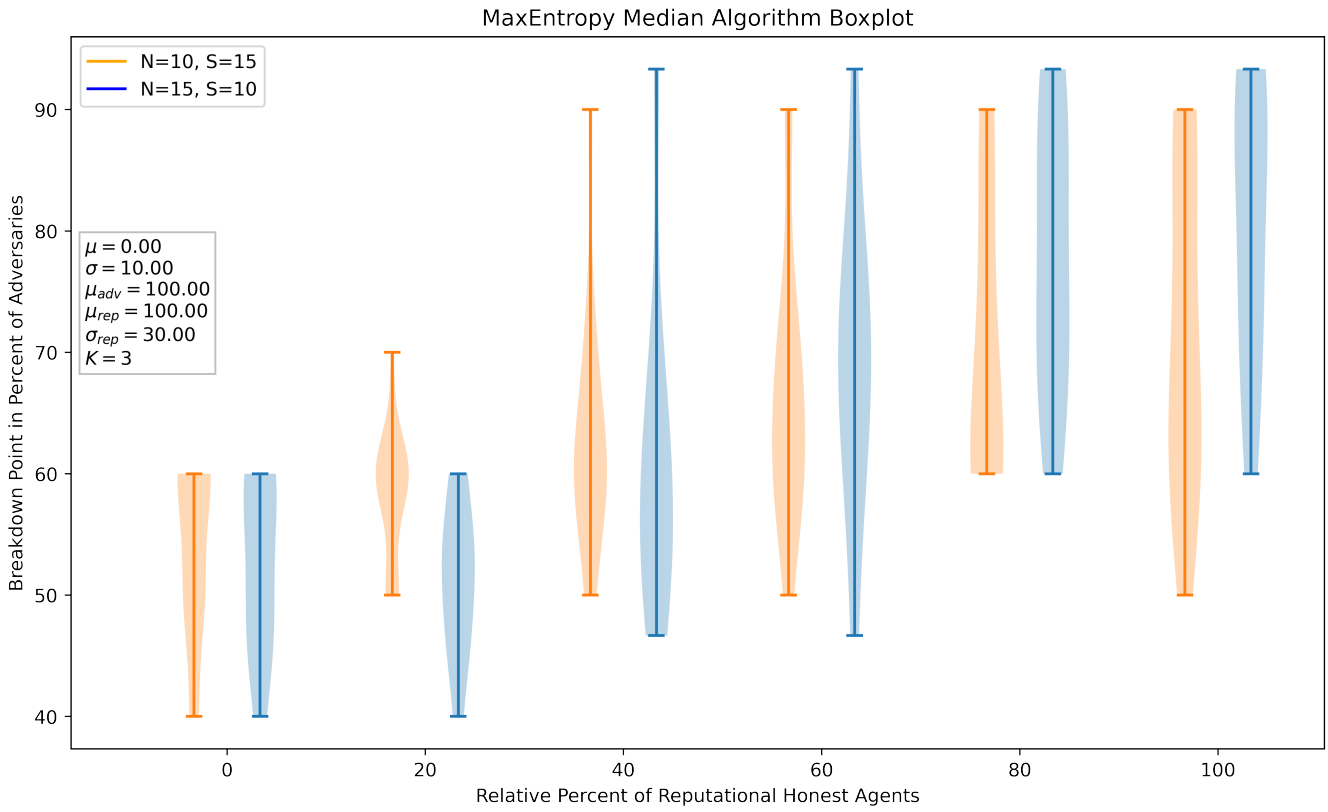


Figure 10. Characterisation of breakdown of MEV combined with Median Algorithm

spread is introduced by adding the voting consensus mechanism. This is due to the fact that the voting mechanism is probabilistic. This makes it very expensive for an attacker to have certainty that their attack will be successful.

To conclude the analysis on the robustness against *data poisoning*, it can be said that with both, the Maximum Entropy Voting and the median in place, the system offers protection against *data poisoning* by introducing a spread in the breakdown point of the system, as well as increasing the number of adversaries needed on average to reach the breakdown point. In order to be have a probability to succeed in subverting the network, the malevolent coalition has to be in control from 40% to 80% of the network, depending on the reputation system and the honesty of the other agents. Under the security assumptions used in Byzantine environments, where it is assumed to have 2/3 honest actors, on average the breakdown point in percent of adversaries increases to 60% to 70%.

D. CONCLUSION

Fairness, decentralisation and verifiability are fundamental to the body of this work. The novelties of this work include: ranking data in terms of how valuable it is to the market using the Shapley value, and proportionally adapting the proof of work to it. Furthermore, the proof of work is itself useful and necessary for the functioning of the market, and thus not wasteful. We also propose consensus through a voting scheme that satisfies desirable properties of fairness, and introduce an optimisation to make its computational complexity significantly lower for the context of this work. Most importantly, this voting scheme favours agents that can prove their honesty, as this is how reputation is earned in the system.

Indeed, at time of writing, Algorand just announced that they are aware of the critique towards their voting algorithm. Passive agents with more wealth have more voting power than the active agents that are enabling the functioning of the network. Algorand have stated they agree with this critique and will be rolling out changes in June 2022 to reward active network users [15]. We hope the work here presented can be a first step in enabling the shift towards this direction.

For future work, we wish to explore how our voting scheme can be implemented in an End-to-end verifiable manner, and how the computation of the Shapley value for each agent's dataset can be done in a privacy preserving way. Achieving the latter may enable us to relax security assumptions of the honesty of agents

[15] https://algorand.foundation/news/governance-voting-update-g3

computing the Shapley value.

References

[1] Adida, B. Helios: Web-based open-audit voting. In USENIX security symposium (2008), vol. 17, pp. 335–348.

[2] 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.

[3] Ali, S. T., and Murray, J. An overview of end-to-end verifiable voting systems. Real-World Electronic Voting (2016), 189–234.

[4] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., and Asghar, M. N. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access 7 (2019), 128325–128338.

[5] Andrews, L. Facebook is using you. New York Times.

[6] Aperjis, C., and Huberman, B. A. A market for unbiased private data: Paying individuals according to their privacy attitudes. Available at SSRN 2046861 (2012).

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

[8] ApS, M. MOSEK Optimizer API for Python 9.3.20, 2019.

[9] Arrow, K. J. Social choice and individual values. In Social Choice and Individual Values. Yale university press, 2012.

[10] Arrow, K. J. Economic welfare and the allocation of resources for invention. Princeton University Press, 2015.

[11] 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.

[12] 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.

[13] Bentov, I., Pass, R., and Shi, E. Snow white: Provably secure proofs of stake. IACR Cryptol. ePrint Arch. 2016, 919 (2016).

[14] Blais, A., and Massicotte, L. Electoral systems. Comparing democracies 2 (1996), 40–69.

[15] Boeira, F., Asplund, M., and Barcellos, M. Decentralized proof of location in vehicular ad hoc networks. Computer Communications 147 (2019), 98–110.

[16] 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.

[17] Buchman, E. Tendermint: Byzantine fault tolerance in the age of blockchains. PhD thesis, University of Guelph, 2016.

[18] Cachin, C. Yet another visit to paxos. IBM Research, Zurich, Switzerland, Tech. Rep. RZ3754 (2009).

[19] Castro, M., and Liskov, B. Practical byzantine fault tolerance and proactive recovery. ACM Transactions on Computer Systems (TOCS) 20, 4 (2002), 398–461.

[20] Castro, M., Liskov, B., et al. Practical byzantine fault tolerance. In OsDI (1999), vol. 99, pp. 173–186.

[21] Christidis, K., and Devetsikiotis, M. Blockchains and smart contracts for the internet of things. IEEE Access 4 (2016), 2292–2303.

[22] Courtney, J. C. Plurality-majority electoral systems: A review. Electoral Insight 1, 1 (1999), 7–11.

[23] Crain, T., Gramoli, V., Larrea, M., and Raynal, M. Dbft: Efficient leaderless byzantine consensus and its application to blockchains. In 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA) (2018), pp. 1–8.

[24] Decker, C., Seidel, J., and Wattenhofer, R. Bitcoin meets strong consistency. In Proceedings of the 17th International Conference on Distributed Computing and Networking (2016), pp. 1–10.

[25] Douceur, J. R. The sybil attack. In International workshop on peer-to-peer systems (2002), Springer, pp. 251–260.

[26] Duguépéroux, J., and Allard, T. From task tuning to task assignment in privacy-preserving crowdsourcing platforms, 07 2020.

[27] Feldman, P., and Micali, S. Optimal algorithms for byzantine agreement. In Proceedings of the twentieth annual ACM symposium on Theory of computing (1988), pp. 148–161.

[28] Foundation, I. Consensus in the iota tangle — fpc, 2019.

[29] 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).

[30] Gibbard, A. Manipulation of voting schemes: a general result. Econometrica: journal of the Econometric Society (1973), 587–601.

[31] Gibbard, A. Manipulation of schemes that mix voting with chance. Econometrica: Journal of the Econometric Society (1977), 665–681.

[32] Gilad, Y., Hemo, R., Micali, S., Vlachos, G., and Zeldovich, N. Algorand: Scaling byzantine agreements for cryptocurrencies. In Proceedings of the 26th symposium on operating systems principles (2017), pp. 51–68.

[33] Gray, R. M. Entropy and information theory. Springer Science & Business Media, 2011.

[34] Hu, Y.-C., Perrig, A., and Johnson, D. B. Packet leashes: a defense against wormhole attacks in wireless networks. In IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428) (2003), vol. 3, IEEE, pp. 1976–1986.

[35] Huber, P. J., and Ronchetti, E. M. Robust Statistics. Wiley-Blackwell, 2nd ed., 2009.

[36] 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.

[37] Isaak, J., and Hanna, M. J. User data privacy: Facebook, cambridge analytica, and privacy protection. Computer 51, 8 (2018), 56–59.

[38] Kumar, A., and Goyal, P. Forecasting of daily air quality index in delhi. Science of the Total Environment 409, 24 (2011), 5517–5523.

[39] 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.

[40] Laslier, J.-F. And the loser is... plurality voting. In Electoral systems. Springer, 2012, pp. 327–351.

[41] Lundberg, S. Shap documentation, 2018.

[42] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[43] 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.

[44] MacKay;, R. S. D., and McLean, I. Probabilistic electoral methods, representative probability, and maximum entropy. Voting matters (2009).

[45] Manzano Kharman, A. M., and Smyth, B. Is your vote truly secret? ballot secrecy iff ballot independence: Proving necessary conditions and analysing case studies, 2021.

[46] McKinzie, and Company. Monetizing car data: New service business opportunities to create new customer benefits, 2016.

[47] 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.

[48] OpenSecrets. Expenditures breakdown, donald trump, 2016 cycle, 2016.

[49] 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).

[50] Popov, S., and Buchanan, W. J. Fpc-bi: Fast probabilistic consensus within byzantine infrastructures. Journal of Parallel and Distributed Computing 147 (2021), 77–86.

[51] 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.

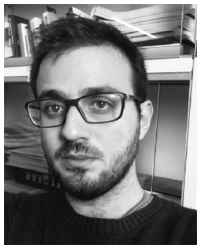
[52] Rasouli, M., and Jordan, M. I. Data sharing markets. arXiv preprint arXiv:2107.08630 (2021).

[53] Raynal, M. Communication and agreement abstractions for fault-tolerant asynchronous distributed systems. Synthesis Lectures on Distributed Computing Theory 1, 1 (2010), 1–273.
[54] Samarati, P. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13, 6 (2001), 1010–1027.
[55] 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.
[56] Schueffel, P. Alternative Distributed Ledger Technologies Blockchain vs. Tangle vs. Hashgraph - A High-Level Overview and Comparison -. SSRN Electronic Journal (2018), 1–8.
[57] Sen, A. Social choice. the new palgrave dictionary of economics, abstract & toc, 2008.
[58] Shapley, L. Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, 1953.
[59] Smyth, B. Ballot secrecy: Security definition, sufficient conditions, and analysis of Helios. Journal of Computer Security 29, 6 (2021), 551–611.
[60] Smyth, B., Frink, S., and Clarkson, M. R. Election verifiability: cryptographic definitions and an analysis of helios, helios-c, and jcj. Cryptology ePrint Archive (2015).
[61] Sonu, S. B., and Suyampulingam, A. Linear regression based air quality data analysis and prediction using python. In 2021 IEEE Madras Section Conference (MASCON) (2021), pp. 1–7.
[62] Stahl, F., Schomm, F., and Vossen, G. The data marketplace survey revisited. Tech. rep., ERCIS Working Paper, 2014.
[63] Stahl, F., Schomm, F., and Vossen, G. Data marketplaces: An emerging species. In DB&IS (2014), pp. 145–158.
[64] Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. Advances in neural information processing systems 30 (2017).
[65] Stucke, M. E. Should we be concerned about data-opolies? Geo. L. Tech. Rev. 2 (2017), 275.
[66] Störing, M. What eu legislation says about car data legal memorandum on connected vehicles and data, 2017.
[67] 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.
[68] 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.
[69] Toueg, S. Randomized byzantine agreements. In Proceedings of the third annual ACM symposium on Principles of distributed computing (1984), pp. 163–178.
[70] Travizano, M., Sarraute, C., Ajzenman, G., and Minnoni, M. Wibson: A decentralized data marketplace. arXiv preprint arXiv:1812.09966 (2018).
[71] ur Rehman, I. Facebook-cambridge analytica data harvesting: What you need to know. Library Philosophy and Practice (2019), 1–11.
[72] Vukolić, M. The quest for scalable blockchain fabric: Proof-of-work vs. bft replication. In International workshop on open problems in network security (2015), Springer, pp. 112–125.
[73] Wang, W., Hoang, D. T., Hu, P., Xiong, Z., Niyato, D., Wang, P., Wen, Y., and Kim, D. I. A survey on consensus mechanisms and mining strategy management in blockchain networks. IEEE Access 7 (2019), 22328–22370.
[74] 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.
[75] Zhang, S. Who owns the data generated by your smart car. Harv. JL & Tech. 32 (2018), 299.
[76] 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.
[77] Zwicker W. S., M., Herve (2016), B., Felix; Conitzer, V. E., and Ulle; Lang, J. Introduction to the theory of voting, 2016.

AIDA MANZANO KHARMAN is a first year Ph.D. student at Imperial College London, working under the supervision of Professor Robert Shorten and Dr Pietro Ferraro. She is working on the topic of data ownership and data privacy using distributed ledger technologies. Her work is funded by IOTA Foundation. She received her undergraduate degree at Imperial College London of Design Engineering in 2022. Her research interests include fairness in decentralised architectures, data centric applications of distributed ledgers, homomorphic encryption and zero knowledge proofs.

CHRISTIAN JURSITZKY received a M.Sc. in Sustainable Energy – Energy Systems Analysis from the Technical University of Denmark (DTU) and a B.Sc. in Materials Science from the University of Leoben, Austria. He was a visiting student at the Imperial College London, UK, while writing the Master's thesis under the supervision of Pierre Pinson and Robert Shorten.

QUAN ZHOU is working towards her Ph.D. at the Imperial College London, London, U.K. She received her undergraduate degree from the College of Finance and Statistics, Hunan University, and her MSc in Operational Research with Risk from the School of Mathematics, University of Edinburgh, in 2018 and 2019, respectively. Her current research interests include fairness in artificial intelligence.



PIETRO FERRARO received the Ph.D. degree in control and electrical engineering from the University of Pisa, Pisa, Italy, in 2018. He is currently a Research Associate with the Dyson School of Design Engineering, Imperial College London, London, U.K., and a Research Scientist with IOTA Foundation, Berlin, Germany. He works closely with a number of research teams within IOTA, primarily on topics related to networking and compliance. His research interests include control theory, distributed ledger technologies, and the sharing economy.

JAKUB MAREČEK received his Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 2012. He is currently a faculty member at the Czech Technical University in Prague, the Czech Republic. He designs and analyses algorithms for optimisation and control problems across a range of application domains.

PIERRE PINSON received his Ph.D. degree from Ecole des Mines de Paris, France, in 2006. He is currently the chair of data-centric design engineering at Imperial College London, Dyson School of Design Engineering, as well as chief scientist at Halfspace, Copenhagen (Denmark). He is the editor-in-chief of the International Journal of Forecasting.

ROBERT N. SHORTEN received his Ph.D. degree from University College Dublin, Ireland in 1996. He currently holds appointments at Imperial College London and University College Dublin, where he is a Professor Cyber-physical Systems and Professor of Control Engineering and Decision Science, respectively. He has been active in computer networking, automotive research, collaborative mobility, control theory, and linear algebra.

.

A. CONIC OPTIMISATION AND LAGRANGIAN RELAXATIONS

Relative entropy programs (REPs) and second-order cone programs (SOCPs) are conic optimisation problems in the relative entropy cones and second-order cones, possibly subject to other linear constraints. They could be solved via interior-point methods.

Let $\pi, \delta, \mathbf{1}$ be $|\mathcal{T}|$-dimensional vectors. The elements of π are $\pi(t), t \in \mathcal{T}$, and $\mathbf{1}$ is an all-ones vector $\mathbf{1}$ of compatible size. A relative entropy cone $(\pi, \mathbf{1}, \delta) \in \mathcal{RE}$ is defined as:

$$\mathcal{RE} := \left\{ (\pi, \mathbf{1}, \delta) \in \mathbb{R}_{\geq 0}^{|\mathcal{T}|} \times \mathbb{R}_{\geq 0}^{|\mathcal{T}|} \times \mathbb{R}^{|\mathcal{T}|} \middle| \quad \pi(t) \log(\pi(t)/1) \leq \delta_t, \forall t \in \mathcal{T} \right\}, \tag{12}$$

The objective function in (5) can be reformatted into (12). The relative entropy cone $(\pi, \mathbf{1}, \delta) \in \mathcal{RE}$ induces that $-\sum_{t\in\mathcal{T}} \pi(t) \log \pi(t) \geq -\sum_{t\in\mathcal{T}} \delta_t$ and we can just minimise $\sum_{t\in\mathcal{T}} \delta_t$ to obtain a maximum entropy solution. Hence, the Problem 5 is re-formulated as

$$\begin{aligned} &\max_{\pi,\delta} \sum_{t\in\mathcal{T}} \delta_t \\ &\text{s.t.} \sum_{t\in\mathcal{T}} \pi(t) \cdot S(t) = S(A), \quad \sum_{t\in\mathcal{T}} \pi(t) = 1 \\ &\pi(t) \geq 0 \ \forall t \in \mathcal{T}, \quad (\pi, \mathbf{1}, \delta) \in \mathcal{RE}. \end{aligned} \tag{13}$$

If $\mathcal{T}$ is the set of combinations, the constraint $\sum_{t\in\mathcal{T}} \pi(t) \cdot S(t) = S(A)$ in Problem 5 or 13 cannot always be satisfied. Correspondingly, we lift up this constraint to the objective function, with a multiplier $\lambda > 0$. Let

$$S^{\text{diff}} := \sum_{t\in\mathcal{T}} \pi(t) \cdot S(t) - S(A) \tag{14}$$

According to the definitions of $S(A), S(t)$, S^{diff} is an $N \times N$ symmetric matrix, with its diagonal being all zeros. On the other hand, S^{diff} implies the distortion of solution π from Representative Probability property III.3. Further, a second-order cone $(S^{\text{diff}}, \eta) \in \mathcal{SO}$ is defined as (15).

$$\mathcal{SO} := \left\{ (S^{\text{diff}}, \eta) \in \mathbb{R}_{\geq 0}^{N\times N} \times \mathbb{R}_{\geq 0} \middle| \sqrt{\sum_{i,j\in A, i<j} 2 \left(S_{i,j}^{\text{diff}}\right)^2} \leq \eta \right\}, \tag{15}$$

where $S_{i,j}^{\text{diff}}$ denotes the element of S^{diff} in row i and column j. The Lagrangian relaxation of Problem 13, using second-order cone, reads

$$\begin{aligned} &\max_{\pi,\delta,\eta} \sum_{t\in\mathcal{T}} \delta_t + \lambda\, \eta \\ &\text{s.t.} \sum_{t\in\mathcal{T}} \pi(t) \cdot S(t) - S(A) = S^{\text{diff}}, \quad \sum_{t\in\mathcal{T}} \pi(t) = 1 \\ &\pi(t) \geq 0 \ \forall t \in \mathcal{T}, \quad (\pi, \mathbf{1}, \delta) \in \mathcal{RE}, \quad (S^{\text{diff}}, \eta) \in \mathcal{SO}. \end{aligned} \tag{16}$$

B. DATA GENERATION

Algorithm ?? indicates that the input $S(A)$ for Problems 5, 13 and 16, is obtained from $x_i, r_{i\to j}$. The generation of measurements x_i and reputation $r_{i\to j}$, could be divided into the honest-agent and the adversarial-agent cases. For the former case, the measurement is sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ while an untrue measurement μ_{adv} is assigned to all adversarial agents directly, as in (17).

$$\begin{cases} x_i \sim \mathcal{N}(\mu, \sigma) & \text{if } a_i \text{ is honest,} \\ x_i = \mu_{adv} & \text{if } a_i \text{ is adversarial.} \end{cases} \tag{17}$$

For simplicity, we assume the reputation $r_{i\to j} = r_i$, for all $a_j \in A$. For generating $r_i, a_i \in A$, a base reputation of 1 is assigned to all agents. Further, a Binomial distribution variable $r_i^{\mathcal{B}} \sim \mathcal{B}(1, p)$ is used to determine if an honest-agent is respected: a_i is honest and respected if $r_i^{\mathcal{B}} = 1$. Further, an honest-and respected-agent would be added extra reputation $r_i^{\mathcal{N}}$ sampled from a Gaussian distribution $\mathcal{N}(\mu_{rep}, \sigma_{rep})$. The procedure is displayed in (18).

$$r_i = \begin{cases} 1 + r_i^{\mathcal{B}} \cdot r_i^{\mathcal{N}} & \text{if } a_i \text{ is honest} \\ 1 & \text{if } a_i \text{ is adversarial} \end{cases} \tag{18}$$

Combination Set $\mathcal{T}_{\text{com}}$		Permutation Set $\mathcal{T}_{\text{per}}$	
t_1	$a_i > (a_j, a_k)$	τ_1	$a_i > a_j > a_k$
		τ_2	$a_i > a_k > a_j$
t_2	$a_j > (a_i, a_k)$	τ_3	$a_j > a_i > a_k$
		τ_4	$a_j > a_k > a_i$
t_3	$a_k > (a_i, a_j)$	τ_5	$a_k > a_i > a_j$
		τ_6	$a_k > a_j > a_i$

Figure 11. This table displays two ordering sets, i.e., combination set and permutation set, when $A = \{a_i, a_j, a_k\}$ and $M = 1$. All orderings, including combinations and permutations, in the same row are equivalent, in terms of election results.

C. MEASURING ENTROPY

Given a set of agents A and the number of winners needed M, we can build two orderings sets: one of combinations $\mathcal{T}_{\text{com}}$ and the other one of permutations $\mathcal{T}_{\text{per}}$. Suppose an optimal probability measure is obtained from Problem 5 for each ordering set, denoted as π^*_{com} for $\mathcal{T}_{\text{com}}$ and π^*_{per} for $\mathcal{T}_{\text{per}}$, with the same input $S(A)$. See Figure 11 for an example when $A = \{a_i, a_j, a_k\}$ and $M = 1$.

Notice that for each element $t \in \mathcal{T}_{\text{com}}$, we can find $M!(N-M)!$ elements in $\mathcal{T}_{\text{per}}$ that are equivalent to t, in terms of the election results. We use $\sim$ to denote this equivalence relation. For instance, each row of Figure 11 displays a equivalent tuple of $t \in \mathcal{T}_{\text{com}}$ and $\tau \in \mathcal{T}_{\text{per}}$. Specifically, $t_1 \in \mathcal{T}_{\text{com}}$ is equivalent to $\tau_1, \tau_2 \in \mathcal{T}_{\text{per}}$, because their election results are the same, i.e., only agent a_i gets elected. Then, we have $t_1 \sim \tau_1 \sim \tau_2$.

To compare the entropy of π^*_{com} and π^*_{per}, we suggest

$$\begin{aligned} \text{Entropy}(\pi^*_{\text{com}}) &:= \sum_{t \in \mathcal{T}_{\text{com}}} \pi^*_{\text{com}}(t) \log \pi^*_{\text{com}}(t) \\ \text{Entropy}(\pi^*_{\text{per}}) &:= \sum_{t \in \mathcal{T}_{\text{com}}} \left(\sum_{\tau \in \mathcal{T}_{\text{per}}, \tau \sim t} \pi^*_{\text{per}}(\tau) \right) \log \left(\sum_{\tau \in \mathcal{T}_{\text{per}}, \tau \sim t} \pi^*_{\text{per}}(\tau) \right) \end{aligned} \tag{19}$$

D. NUMERIC ILLUSTRATIONS

With $S(A)$ extracted from data generated in -B, we have the following implementations:

- "Permutation": solving Problem 13 with input $\mathcal{T} = \mathcal{T}_{\text{per}}, S(t), S(A)$, and optimal solutions π^*_{per}.
- "Combination_Lag": solving Problem 16 with input $\mathcal{T} = \mathcal{T}_{\text{com}}, \lambda = 2, S(t), S(A)$, and optimal solutions π^*_{com}.

Both are solved by MOSEK Optimizer API for Python 9.3.20 [8]. Figure 12 displays the results of runtime, entropy in (19) and RP distortion S^{diff} in (14), of optimal solutions π^*_{com} and π^*_{per}, when the number of agents N are $6, \ldots, 15$ for "Combination_Lag" and $6, \ldots, 9$ for "Permutation". Note that "Permutation" with larger N is not implemented due to its spike in runtime. Under each N, both implementations are conducted 6 times (6×2 runs in total), with a new $S(A)$ generated every time. The average entropy, runtime and RP distortion of 6 runs are presented as solid curves for "Permutation" and dashed curves for "Combination_Lag".

...

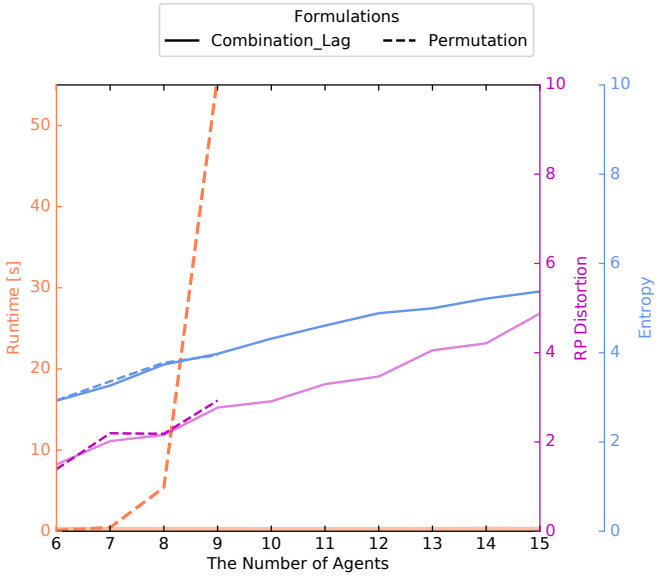


Figure 12. The average results of entropy, runtime and RP distortion, of implementing "Combination_Lag" and "Permutation" for 6 times, with the number of agents N being $6, \ldots, 15$ and $6, \ldots, 9$ respectively.