Fair value and decentralized governance of data

Roger Chen, Bharath Ramsundar, Rob Robbins
computable.io

Version 1.0
June 10, 2019

Contents

Introduction

Everyday our interactions with technology leave behind digital footprints of how the world works. Collectively that data creates a machine learnable representation of the world that is crucial for training and advancing artificial intelligence (AI). However, it presently remains a privileged playground for technology giants and political rulers. This kind of digital rent seeking stifles innovation by making it hard to work on AI anywhere outside of monopoly technology companies. Furthermore, history has revealed that the absence of an open and transparent market for data leads directly to negligence and misappropriation by the powerful few [1,2]. Big businesses readily broker the personal data of their customers without informed consent, and governments unilaterally subject citizens to mass surveillance.

The only way to truly prevent abuse of power is to make it structurally impossible. Data regulations such as GDPR [3] and CCPA [4] are steps in the right direction, but the current conversation and laws mainly address privacy. There should also exist a way to establish *economic* ownership because the rights of property include the right to sell. Instead of being cut out of a corporate shadow economy, individuals and small businesses deserve to become knowing market participants. The pervasive thirst to monopolize data has deep roots in existing structural incentives resulting from how businesses and web technologies evolved together. Even in the absence of malice, companies have felt compelled to consolidate control of data to win modern economies built on attention and market intelligence. Growing public awareness and newly available technologies offer hope that the time has come to redefine a better way to share and use data, and it begins with finding new ways to establish data lineage on the Internet.

Given the ease of data replication and distribution, a key to establishing data ownership is achieving provenance. An ongoing historical record of authorized use can facilitate a new social contract that enforces economic ownership by proving whether data consumers ever received explicit permission for use. Ultimately this paradigm promotes data sharing and access. While data monopolies silo data and only share if they perceive strategic value, individual owners would be economically incentivized to sell data as widely as possible, by choice. A source of truth like this would have to be public and trustless so that it can be audited without fear of manipulation. It must also be algorithmic because paper records and manual process cannot possibly handle the scale and complexity of data transactions. If achieved, a new system of checks and balances can finally decentralize control of data and restore a healthier equilibrium of power to a digital world.

In this paper, we propose a protocol that gives users access control over their data and promotes fair payment for the value that their data creates. A usage-based valuation model actively measures the fair market value of data by recording its real-time demand. Integration with decentralized systems meanwhile enables trustless data lineage properties that can be combined with regulations like CCPA to powerfully establish and protect data ownership. Together these building blocks and the outlined protocol provide an economic framework for truly cooperative data networks native to the decentralized Internet.

The Fair Value of Data

Making contributors economic agents can financially incentivize their participation in crowdsourced data networks. Each network member should receive appropriate dues as buyers consume their data. Determining how to fairly distribute value among members first requires a fair valuation model for data. However, accurately pricing data faces some rather unique and fundamental challenges not seen by other kinds of assets. In summary:

i. Data is a practically unbounded and replicable resource (inflation)
ii. Not all data points have the same intrinsic value (non-fungibility)
iii. The value of a data point depends on the data it is combined with (interdependence)
iv. Data’s true value can only be ascertained retroactively after computation and analysis (speculation)

These complexities conspire against any ambitions for a grand unified theory on the price of information. Fortunately, we can construct empirical models that work because we can measure demand for data. Data that sees more usage (via queries or computations) merits a higher value. The rest of this section lays out how a simple yet powerful usage-based valuation model can overcome the challenges laid out above. We begin with some definitions before diving into the mechanics of the model.

Definitions

Data Market

A data market refers to a decentralized data network that is specifically implemented and constructed according to the protocol described in this paper. The protocol itself agnostically supports any kind of data.

Listing

A listing represents some quantity of data made available for sale on the market as defined by that market's owners and creators. The definition of a listing can vary from market to market as we will see in some examples below.

Maker

Makers supply data to data markets for sale in the form of listings.

Buyer

Buyers consume data by purchasing computational access.

Patron

Patrons offer capital to the market to financially incentivize makers to supply data. In some instances, they may receive a stake in the market in exchange that includes an economic share of future payments from buyers.

Rewards

Market rewards measure financial ownership and governance rights in a data market. Makers earn them for supplying data to the market and as that data gets consumed. Patrons can acquire rewards by depositing funds into a data market.

Datatrust

A datatrust is the off-chain storage and computing solution that integrates with an on-chain data market. Datatrust operators receive fees for executing computations on behalf of buyers while protecting data privacy on behalf of makers.

Data Market Basics

In this section, a high-level sequence of transactions outlines how agents participate in a decentralized data market according to our protocol. The rest of this paper focuses on the economic and governance structures that functionalize these markets. Future work will further detail technical specifications and reference implementations.

Supporting a Data Market

1. Patron adds capital to the data market to fund gathering of data.

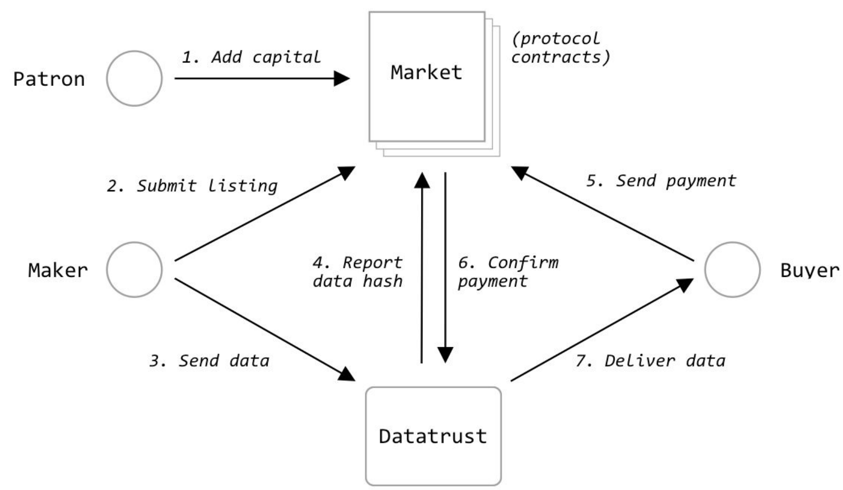
Listing Data on the Market

2. Maker submits a listing candidate to the market (protocol contracts).
3. Maker sends data associated with the listing to a datatrust supporting the market.

4. Datatrust verifies data and sends a hash of the data to the market, confirming receipt and linking the on-chain listing to the off-chain store. Maker earns listing rewards and a portion of the capital provided by patrons.

Buying Data Access

5. Buyer sends payment to the market.
6. Datatrust verifies appropriate payments have cleared via the market's records.
7. Datatrust delivers data to buyer and receives appropriate backend service fees. Datatrust notes which listings were accessed and reports to market contracts. Listing owners receive appropriate shares of buyer payment.



A sequence of basic market transactions

General Examples

A diversity of applications illustrates the generality of the concept and value behind a decentralized data market protocol. These examples also give us an opportunity to show the different ways that listings might be concretely defined.

Precision Medicine

Population-scale genomic and phenomic data is key to unlocking precision medicine. However, the sensitive nature of personal health information demands mechanisms to ensure that people stay in control of their data even as they share it. For this application, a listing might be a text file that combines genetic data in variant call format (VCF) with basic self-reported phenotypic data such as hair color, weight, and other attributes.

Weather Data
There exists valuable data about our ambient environment that arguably no one owns, yet we all deserve access to because it exists in the public domain. Weather information is an apt example. It should not belong to any one party, and in fact there should probably exist just one universal data network for everyone. A listing in this case might consist of a timestamp, location coordinates, and readings for temperature, pressure, and humidity. Given the time series nature of this data, a listing might be created for a fixed time period, e.g. every minute, at a fixed location.

High Definition Mapping
Crowdsourcing is a powerful way to aggregate data for building widely impactful applications. For example, OpenStreetMap powers numerous mapping products used by countless people everyday. Traditionally these projects have relied on the kindness of volunteers and enthusiasts. What if we could offer them more? By enabling financial ownership of the data they collect, perhaps more people and businesses will feel greater incentive to share more data. We are still in the early days of high definition mapping for autonomous vehicles, which may very well benefit from a more market-driven crowdsourcing model. Here a listing might consist of telematics data, video, and LIDAR results over the course of a fixed time period or a driven distance.

Cryptocurrency Market Data
While a wealth of data sources and terminals exist for traditional finance, it remains challenging to get a clear picture of activity in the cryptocurrency market. How can we find ways to aggregate data streams from exchanges, over-the-counter (OTC) desks, and other disparate sources while staying in line with the industry's ethos and culture of decentralized control? A decentralized data market for finance seems to be a natural fit. In this case, a listing could be a fixed volume of verified transactions that have been reported to the market.

Listing Rewards

Makers earn rewards as they add listings to a network. A maker receives total listing rewards in the amount of

$$R_{listing} = N_{listing} \times V_{listing} \ ,$$

where $N_{listing}$ is the maker's number of approved listings and $V_{listing}$ is a parameter for the number of rewards issued per successful listing.

Many data applications combine multiple data types for analysis. Consider the use case for genomics. While genetic data is required, combining it with phenotypic data unlocks vastly more value for researchers. Location data is another use case that often benefits from combining data. For example, some mapping applications may require an overlay of weather information on a points-of-interest dataset. This complicates how rewards should be distributed. After all, a temperature reading is a very different measurement from the GPS coordinates of a nearby fast food restaurant. They should therefore likely have different listing values.

One strategy is to initialize a data market with multiple $V_{listing}$ parameters, one for each data type supported. However, this presumes that a market creator has pre-existing knowledge of the relative value between two or more data types. We therefore expect that most practical market implementations will have to define a data point comprised of either one data type or multiple types with well-understood interdependency e.g. a GPS coordinate plus the name and address of a point of interest. Each one of these market economies can develop separately and converge on its own fair market value. Combining data from these different networks in complex ways can then come in the form of data market interoperability, which we will explore in future work.

Access Rewards

Non-fungibility of data is a factor even when working with a single type of data. For example, traffic data may have more use cases for dense cities than for a sparsely populated countryside. In this case, one way to measure the relative value of different data points is simply to measure how relatively often they are used. Presumably, city planners would want to query and buy far more urban traffic data than rural traffic data. Data points that see more usage signal more value and therefore merit more rewards for their contributors. Accordingly, each maker receives total access rewards in the amount of

$$R_{access} = N_{access} \times V_{access} \ ,$$

where N_{access} represents the amount of the maker’s data accessed and V_{access} is the number of rewards issued per unit of data access.

A usage-based reward mechanic has two key features. First, it proportionally rewards makers who supply the most valuable subsets of data based on real market feedback. Second, it has the natural effect of improving curation for a data market. Makers learn what data is more relevant to buyers and are more incentivized to supply it, i.e. city traffic data in this example, because it is more profitable. As a result, the aggregated dataset evolves into a resource that better matches real public demand.

Capitalization

Listing and access rewards get minted and issued as makers supply data and that data gets used, respectively. A maker's rewards total is then simply the sum of both parts $R_{listing}$ and R_{access}. A data market's overall reward supply is uncapped. As long as makers continue to contribute data or their data continues to see usage, minting and issuance will continue in stride.

Nothing precludes a single person or entity from being both a maker and a patron in a data market. Such a stakeholder would own an aggregate of market rewards

$$R = R_{listing} + R_{access} + R_{support} \; ,$$

where $R_{support}$ reflects holdings acquired through support as a patron. We define and detail how patronage works later in this paper.

For a data market, we define its capitalization as the ongoing accumulation of all holdings by all stakeholders:

$$\sum_i R_i \qquad \text{// Capitalization}$$

A stakeholder's market ownership level is then simply that user's share of total capitalization:

$$R_n \; / \; \sum_i R_i \qquad \text{// Ownership of the } n^{th} \text{ stakeholder}$$

Makers who grow their rewards at a faster rate than the overall market therefore garner a larger fraction of that market. As a result, the distribution of market ownership can be highly dynamic. It incentivizes healthy competition and naturally rebalances ownership in favor of the most productive makers. This is important because, as we will see, network economics and governance rights are directly proportional to stakeholders' ownership levels.

This model also aligns incentives between buyers and suppliers, lending well to a virtuous cycle. First, some makers earn a greater market share than others by supplying more and better curated data. Other makers follow suit to close the earnings gap. The data network then grows in both volume and quality, attracting more consumers and more spend. In turn, greater spend incentivizes even greater supply to come onto the network.

Defeating Rent Seekers

Questions about dilution may naturally arise. After all, makers do gain ownership at the cost of other stakeholders. For this protocol, this is a feature by design.

Consider a model where reward supply is capped so that ownership never dilutes. In this approach, issuance rates necessarily diminish as more users add more data to the network, or else the network would quickly run out of rewards to supply. As a result, later contributors would earn disproportionately lower ownership than early contributors who may have stopped providing data, even if later contributions are vastly larger and more valuable.

Our capitalization scheme defeats exactly this kind of economic rent seeking. Rather than reward makers for being first, this protocol rewards makers for being valuable. If makers stop supplying data or have made useless contributions, their ownership position would indeed dilute as others aggregate more valuable data points. Such rent seekers would eventually lose their influence and financial hold on a market, rightfully so.

An uncapped supply is also important because there is a need to perpetually incentivize data crowdsourcing. After all, any dataset must first reach critical density and quality before it truly becomes valuable. Furthermore, data enjoys its own sort of network effect in that the value of a piece of information increases when contextualized with more information. Establishing a *continuous* data supply can therefore create a multiplier effect for all members of a data network.

Said differently, data is a uniquely inflationary asset that benefits from inflationary economics. In this protocol, a unit of inflation directly corresponds to a unit of real economic growth via the supply of new data. Up to now, we have discussed fair distribution of network value in relative terms. A rewards system effectively indexes the relative value of network members based on their relative productivity. In the *Economics* section, we will show how to pin these relative values onto absolute market prices. In the next section on *Cooperative Governance*, we will see how network ownership also plays a prominent role in democratizing network control.

Cooperative Governance

The previous section describes a protocol for fairly distributing value amongst members of a data network. However, it requires a trustless accounting system to work in practice. If a centralized entity was responsible for tracking capitalization and ownership, market participants would have to place extreme trust in that organization to record transactions accurately and honestly. We have witnessed time after time how centralized oversight of data invariably leads to transgressions. When self-interest comes with singular control, even good intent can quickly turn into bad outcomes.

A potential solution has emerged with the rise of blockchain technology. Distributed consensus mechanisms enable public ledgers that are immutable and secure against tampering. By implementing a protocol with smart contracts, it can incorporate these features to become more trustless. Network members can then have full confidence in how their ownership is accounted for. They can also audit public logs and gain a level of transparency that centralized organizations would never provide. Furthermore, decentralized consensus technology guarantees users self-sovereignty. Anyone can join a blockchain network and independently elect whether or not to participate in a transaction without threat of censorship. This power of radical choice lends well to democratic rule and cooperative governance of data markets.

Voting

The durability of any network requires a governance model with enough flexibility to evolve and enough checks and balances to prevent fatal decisions. We propose to achieve this with a staked democratic voting structure for the data markets described in this paper.

In order to participate in a governance decision, a data market member must lock up at least STAKE rewards, where STAKE is a parameterized constant. A stakeholder who owns R rewards has to ability to execute R/STAKE votes simultaneously. Multiple votes can be allocated towards one governance decision or spread across several concurrent ones in progress.

Attaching an economic cost to voting has several virtues for decentralized networks. It honors the permissionless nature of these systems while also encouraging honest behavior. That is, anyone can participate in governance, but staking deters the abuse of that right by first requiring some minimum amount of financial commitment to earn it. It also creates a real penalty for bad actors. Anyone launching sybil or denial-of-service attacks would have to lock up vast economic resources to do so.

Because market ownership levels are always changing, the distribution of voting power in a data market is also highly dynamic. This defeats political rent seeking the same way it defeats economic rent seeking, as previously described. Stakeholders who do not continue to add value and earn rewards see their influence erode as more productive stakeholders accrue larger market shares and more voting power.

If STAKE is large, governance resembles the board of directors of a corporation. If STAKE is small, it looks more like common stock voting. Voting in favor of a proposal requires PLURALITY, which is a network parameter that defines a voting consensus threshold. By achieving PLURALITY, voters can execute on the following governance decisions:

- Listing candidacy
- Parameterization
- Storage and computing registration

The rest of this section describes these in detail.

Listing Candidacy

Listing Application

Every listing requires an application process. When a maker submits a listing, stakeholders can vote to accept it within a certain voting period. If PLURALITY is reached, the application successfully becomes a listing. Otherwise, the application fails.

Listing Challenge

Once a listing has been approved, it is still subject to removal. A market member may initiate a challenge by staking at least STAKE rewards to validate their interest in the market. If the challenge fails, the challenger loses the staked rewards to the listing owner. If the challenge succeeds, the listing is removed.

Parameterization

Ensuring that this protocol works in a trustless fashion requires implementing it with smart contracts. Anyone can interact with publicly deployed contracts and become a market participant. Because of their open source nature, smart contracts also allow anyone at anytime to instantiate a new data market of his or her own. When launching a market, its creator must first parameterize it. Our protocol includes the following parameters, the definition of which directly impacts the effective operating regime of the data network:

```
LIST_REWARD         // Listing reward amount
PLURALITY           // Voting consensus threshold
VOTE_BY             // Voting period
STAKE               // Challenge and voting stake amount
SPREAD              // Algorithmic market maker spread
PRICE_FLOOR         // Price floor for patron support
COST_PER_BYTE       // Cost to access one byte of data
BACKEND_PAYMENT     // Percentage of payments to datatrust operators
MAKER_PAYMENT       // Percentage of payments to makers
```

We expect different use cases will have different optimal settings. For example, a business consortium that wants to crowdsource fraud data may only want the largest contributors to gain governance rights. Accordingly they would set the STAKE parameter very high. On the other hand, a developer building a decentralized application for mass consumers may want to include a product feature that guarantees all users a democratic vote. STAKE would have to be a small (but nonzero) value in this case. In both cases and in general, it is likely that the initial parameterization is imperfect or that a data market has evolved in such a way that requires reparameterization. By reaching PLURALITY, voting stakeholders always have the power to tune the above parameters as appropriate.

Storage and Computing Registration

Protocols do not absolve us of the need for metal. Data and computations will still have to live on some physical object located somewhere. Ideally, storage and compute components are decentralized or else they re-introduce the problem of trust. Makers and consumers would have to believe that service providers are safely and appropriately handling data and computations. Fortunately, a number of exciting decentralized storage and computing solutions are on the horizon. However, they will need time for security hardening and to achieve scalability. As a result, integrating off-chain centralized computing components must be necessary for practical and functional protocols to extend from ideas to reality. By making decentralization a commitment, we can build a roadmap that steadily layers in features over time that eventually make the entire technology stack trustless.

The data market protocol allows for a “datatrust” to serve a data market. Such datatrusts are responsible for providing secure storage and computing capabilities for that market. Voting stakeholders take on the responsibility for authorizing a trusted party to run this datatrust via vote. Datatrust operators take on responsibility for storing data off-chain and for allowing secure computational access to it for users. In return for these services, operators receive a percentage of the access fee (BACKEND_PAYMENT) that users pay to perform computations.

At present, each data market is served by a single datatrust. Any interested party will be able to run their own datatrust using the open source datatrust implementation. Future iterations of the protocol will allow for multiple datatrusts to serve a given market which will help reduce the amount of trust required in any one datatrust provider. If one datatrust violates the trust of market stakeholders, they can always vote to de-register them from the market and use other datatrust operators instead. What keeps any network truly decentralized is not the absence of centralized participants, which is impossible, but rather the ability to prevent any centralized parties from seizing full control.

In addition, cryptographic verification schemes will allow users and voting members to verify that a given datatrust is behaving honestly at any given point in time. It will also be feasible to coordinate computations across multiple markets and multiple datatrusts in future versions of the protocol. This will allow for computations to run securely across many different datasets contained within many different data markets, unlocking new applications of machine learning while simultaneously respecting the rights of makers.

Privacy

Protecting the privacy of makers is a critical feature of the data market protocol. However, it’s important to recognize that privacy is a grand challenge, and there does not exist any one silver bullet that can prevent user information from leaking. We take a number of pragmatic approaches to ensure maker privacy in the current version of the protocol. To start, makers retain the right to delete their data from off-chain datatrusts at any time. The open source datatrust will implement this feature, and it is incumbent upon the market stakeholders to ensure that their datatrust provider is running a compliant version of the datatrust software.

While the capability to delete data provides powerful protections, it isn’t enough to protect maker privacy. What prevents an unscrupulous user from just scraping all data available in a data market? There are some potential options on the table. Advanced techniques such as differential privacy can provide considerable protection to users. However, it’s worth noting that these techniques are not universally applicable; there are classes of computations for which differential privacy provides no guarantees. In addition, software tooling for such privacy techniques is still a work in progress.

For this reason, the current version of the protocol provides *economic privacy protections*. That is, while there is not an absolute information-theoretic guarantee of privacy, the protocol can make scraping raw data from datatrusts prohibitively expensive. This is done by charging a COST_PER_BYTE for each byte of output from any computation that is performed on the data within a given datamarket. If this cost is chosen correctly, then market stakeholders can ensure that extracting large amounts of information from the data market will require considerable expense.

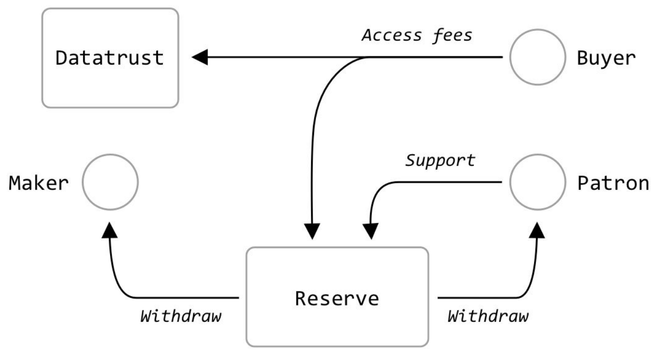
It's important to emphasize that as privacy algorithms mature, we fully anticipate that future versions of the protocol will have increasing algorithmic privacy protections to augment economic privacy protections. Until then, protecting individual privacy rights will lean primarily on explicit definition of consent through the protocol's recorded market transactions. Over time, the combination of all these techniques will allow for the strongest safeguards on maker privacy.

Provenance

Up to now, we have discussed on-chain governance mechanics, but on-chain provenance also enables practical off-chain governance. Implementing this protocol with blockchain technology produces a transparent and trustless audit log, which serves as a source of truth in case of disputes. When combined with user or license agreements, this ledger can define who has received explicit consent to use of data. If off-chain data usage is discovered to be in conflict with the on-chain permissions record, compromised data market members can then take legal recourse with proof in hand. As privacy laws like GDPR and CCPA take effect, anyone who uses data without such explicit consent faces potential regulatory action. Combined with growing public awareness around privacy, this compliance risk can help establish a new social contract around the fair use of data. Contrast this to the status quo. Today most companies operate with extreme opacity around whether or not they have violated even their own policies and user agreements.

Economics

A healthy decentralized market should offer incentives that align its stakeholders behind a common goal. For a data market, it is continuously creating, curating, and making available powerful data resources. Suppliers benefit from the profit potential of such a marketplace. Buyers benefit by gaining novel access to datasets that are relevant and valuable to their businesses. Patrons can meanwhile benefit in both of those ways. In this section, we detail the agents and payment mechanisms driving a data market's economic activity. At its heart lie two core concepts: the *Market Reserve* and *Algorithmic Liquidity*.



Reserve currency payment flows

Market Reserve

Each data market includes a reserve that functions like a bank account, accumulating currency spent by buyers and patrons. It represents wealth creation for a market and all its owners because each owner is entitled to a pro rata share of reserve holdings. The reserve model ensures patrons that their purchased market ownership is backed by real value and not just speculation. Patron rights include not only a share of the reserve's present holdings, but also the future payments it will receive. If a data market is in its growth phase, i.e. adding more supply increases purchases, the reserve will tend to function like a sink. Patrons observing the continued growth in access fees may want to buy a piece of that fee stream, adding even more currency to the reserve. If public markets believe that a data market's reserve is overcapitalized, both makers and patrons can choose to withdraw their holdings, and the reserve size decreases.

Up to now, we have discussed ownership and payments in terms of arbitrary rewards and currencies. That is because we believe the protocol design concepts presented in this paper are fairly generalizable to a variety of payments systems. However, powering the protocol with cryptocurrencies provides advantages for applications where no one party can be trusted to manage all of a market's data or transactions. In this case, tokens offer a technically elegant way to represent market rewards, and the reserve can simply be a smart contract address. Meanwhile, cryptocurrencies like Ethereum and Bitcoin can provide means of making payments and offering patronage. While initial support will be provided for Ethereum, we believe it will be important to make payments interoperable across a number of cryptocurrencies in the future.

Fee Splitting

Whenever buyers purchase data access, the protocol splits access fees between the makers who supplied the data, the datatrust operator that delivers computational access, and all other stakeholders. Makers and datatrust operators are entitled to MAKER_PAYMENT and BACKEND_PAYMENT percentages of access fees, respectively. Datatrust operators receive their portion of payments through direct transfers i.e. between wallet addresses. The rest of access fees are deposited into the reserve. Makers receive access rewards that give them a claim on the updated reserve that is equivalent to a MAKER_PAYMENT percentage of access fees. Note that though access rewards are newly minted, they are not inflationary because they have been appropriately backed by the deposit of new funds into the reserve. The rest of the access fees deposited into the reserve (1-MAKER_PAYMENT-BACKEND_PAYMENT) then get distributed amongst all market stakeholders pro rata to their ownership levels. As data markets attract real and paid usage, everyone profits.

Withdrawal

Both makers and patrons receive payments from the reserve by withdrawing their holdings. The amount of payment received by the n^{th} stakeholder is:

$$M \times R_n \,/\, \sum_i R_i$$

Here M represents the total currency held in the reserve. When executing the withdraw function, R_n rewards (or tokens for cryptocurrency implementations) are burned and removed from market capitalization.

Creator Block

To incentivize data market creation, creators can deploy markets with a creator block. A creator block is simply an initial capitalization of rewards assigned to the creator’s address upon first launching market smart contracts. This gives creators economic ownership in the market from the start. As makers and patrons join the market, the creator’s ownership dilutes, but this is counterbalanced by the increasing value of the reserve as makers contribute data and buyers and patrons contribute funds.

Algorithmic Liquidity

Like any market, a decentralized data market requires liquidity in order to function and optimize price discovery. Facilitating market transactions via smart contracts presents a novel approach to this. Rather than relying on a counterparty, patrons can buy or sell ownership directly from and to smart contracts. In effect, protocol smart contracts serve as an algorithmic market maker or liquidity provider. Given the reliance of cryptocurrency networks on liquidity and the cold start problem that any new marketplace faces, we expect that many practical future networks will benefit from incorporating algorithmic liquidity. The rest of this section discusses how we approach this for a data market protocol.

Support

Anyone can purchase ownership in a data market. To acquire R_n newly minted tokens, a patron would pay the following amount to the market reserve:

$$\text{PRICE\_FLOOR} + \text{SPREAD} \times M \times R_n \,/\, \sum_i R_i \qquad \text{// Support price}$$

Note how this purchase amount resembles the withdrawal calculation described above. Here the constant SPREAD parameterizes the spread between the support and withdrawal prices i.e. a SPREAD of 1.1 indicates a 10% price difference. PRICE_FLOOR meanwhile allows market creators to parameterize a minimum support amount. In effect, market making smart contracts can charge patrons a fee for providing liquidity. All market stakeholders then receive a claim on those fees proportional to their existing rewards holdings.

Setting SPREAD > 1 activates an additional profit stream on top of the access fees that flow in from buyers. Subsequent patron support dilutes the ownership of earlier stakeholders, but they still profit because the reserve balance increases super-proportionally if S > 1. Patrons only ever lose money if the reserve size is reduced below what it was when patrons first bought in. This can happen if market sentiment sours and enough makers and patrons withdraw their holdings to deplete the reserve. Payments into the reserve made by buyers are meanwhile non-dilutive. They fairly split among all market participants according to their ownership levels in the market.

Note that SPREAD and PRICE_FLOOR also offer a mitigation against denial-of-service attacks. If an attacker wants to disrupt services by recursively calling the smart contract’s support and withdraw functions, they would lose PRICE_FLOOR and some multiple of SPREAD with each call until their funds eventually erode away.

Patrons are also welcome to acquire ownership through secondary markets if they can find a better price, but many if not most will benefit from the ability to reliably and programmatically trade directly with the contracts. Any additional price that a patron pays for supporting a data market via the contracts can be considered a liquidity premium of sorts.

Withdrawal

As described previously, patrons and makers utilize the same withdraw function to exit a market with their entitled portion of the reserve. Namely, the amount of payment received by the n^{th} patron is:

$$M \times R_n \;/\; \sum_i R_i \qquad \text{// Withdrawal price}$$

As part of withdrawal, R_n rewards (or tokens for cryptocurrency implementations) are burned and removed from market capitalization.

Value to Patrons

There are times when a patron might be driven purely by philanthropic reasons e.g. supporting an open source repository of climate data. There are other times when patrons might be motivated by market forces and seek to make a return on the capital they provide. The protocol itself allows for both philanthropists and capitalists to participate as patrons. It remains agnostic to personal motives. For some data markets, it may even be structurally impossible for patrons to generate returns i.e. MAKER_PAYMENT+BACKEND_PAYMENT=1.

For patrons with a capitalist orientation, the economic value of ownership in a data market centers on real fundamentals with plenty of historical precedent. A reserve banks each data market by accumulating currency from both patrons and data buyers. By owning a percentage of the market, a patron owns a pro rata share of that reserve. If the reserve balance increases because of future patrons and buyers, an existing patron’s share of the reserve will increase in value. In effect, patrons can lend capital to the market to generate interest rate returns financed by future currency flows into the reserve.

It is worth reiterating that the algorithmic liquidity provided by the protocol lets patrons recollect their capital (with returns or losses) at any time. They can buy or sell market rewards from and to smart contracts even when human counterparties are not available. This level of liquidity presents a compelling case for potential patrons. Today the vast majority of cryptocurrencies passively remain on the sideline. That’s a waste. Instead, cryptocurrency holders can park their holdings in a market reserve, which would entitle them to payments while minimizing risk of loss because of the capitalization structure of our markets. They can always claim back their principal whenever they need it. They only ever lose funds incrementally through dilution as makers contribute listings, but that very activity is what then attracts access fees and drives returns for patrons.

Value of Patrons

This protocol measures the relative and fair value of makers by comparing data usage and distributing capital appropriately. Without capital in the reserve, makers cannot withdraw their rewards in exchange for real currency. In other words, the absolute asset value of data ownership depends on the reserve balance. Patrons add to this balance if they believe they can generate future returns or if they simply want to support the creation of a new public data resource. Before buyers emerge in a market, patron capital therefore provides a critical path towards bootstrapping the network economy. Even after consumers are actively paying makers to access their data, patrons can grow (or reduce) ownership value further by adding (or subtracting) capital. At scale and in concert with both makers and buyers, patrons therefore ultimately play a key role in market price discovery.

Patron Governance Rights

Like makers, patrons can vote and participate in network governance provided that they have at least STAKE rewards to stake their vote. They receive the same voting rights as any other market stakeholder.

Economic Growth

Listing rewards introduce an inflationary component that accommodates the inflationary nature of data. In fact, the LIST_REWARD parameter can be viewed as an inflation rate of sorts. However, the overall data market economy can actually be inflationary or deflationary depending on its current market dynamics. We introduce this concept here, but reserve deeper analysis and results for future publication. For clarity, we define inflation for a *local* data market economy as the condition under which a unit of market rewards can buy less reserve currency than before through the withdrawal process. That is, the price of goods (reserve currency) has increased as denominated by the local economy’s native medium of exchange (market rewards). Data supplied into a data market certainly also represents goods, but their price levels depend on a global economy beyond the scope of this particular analysis.

Inflationary Growth

A simple derivation reveals that the local economy experiences inflation over some time period ΔT when the following condition holds true:

$$\Delta C/C_o > \Delta M/M_o$$

Here ΔC and ΔM refer to the change in capitalization and total reserve currency, respectively. C_o and M_o refer to their values at the start of time period ΔT. Physically, this occurs when the normalized rate of new listings outstrips that of new payments coming into a market. In practice, every data market will likely start under these conditions because some amount of data supply must exist before buyers will purchase access.

During an inflationary period, economic growth is driven by the ongoing supply of new data made available for consumption. As data floods in, markets can experience data saturation effects. That is, new data may no longer add new value, which then commoditizes the value of all data in the market. However, the protocol naturally addresses this. Less profit opportunity diminishes new supply, which then diminishes inflation to more appropriate levels. Under certain conditions, a data market may even transition into a deflationary state.

Deflationary Growth

A data market's local economy becomes deflationary under the following conditions:

$$\Delta C/C_o < \Delta M/M_o$$

In this case, market rewards can buy back increasing amounts of reserve currency (through withdrawal). This can happen when new buyers join a market within a short period of time or if existing buyers suddenly increase their spend, perhaps because they identified a new business case or discovered new commercial value in the available data supply.

During deflationary periods, market stakeholders have extra incentive to protect their saved rewards and ownership levels. In many economies, this kind of behavior often leads to slowdowns because of decreased market activity. However, listing rewards addresses this by incentivizing productivity algorithmically. As more wealth enters a market, new and old makers alike can pursue their share of it simply by supplying more data to earn more listing rewards from the protocol. No one can stop this gold rush, and savers would also have to accelerate contributions if they want to protect their ownership positions.

Fair Market Value

As data markets oscillate between inflationary and deflationary states, they should trend towards some rational value. Recall that real fundamentals back the financial value of ownership in a data market. For example, owning five percent of capitalization entitles the owner to five percent of access fees after appropriately distributing MAKER_PAYMENT and BACKEND_PAYMENT to makers and datatrust operators. This opens an opportunity for analysts to apply familiar concepts like revenue multiples in valuing a data market. An enterprising patron who perceives the reserve total to be an undervalued multiple of incoming access fees may want to increase their stake in that market. As markets mature, other sophisticated patrons will even be able to start forecasting access fee streams and take a thesis-driven approach to supporting data markets.

Data Resale Market

Data leakage can lead to a data resale market the same way knockoffs generally emerge for any goods with demand. When embraced with the right business model in place, this effect is not only manageable, but can also be strategically advantageous. Nike profits handsomely from the marketing power of a secondary market for its sneakers. In disrupting the music industry, the Internet may have destroyed sales in earlier years, but since then has become a major revenue driver for the industry. As long as the right products and services are built around replicable goods, their creators still stand to benefit tremendously, especially if regulatory tailwinds are in their favor. In the case of data, value involves much more than just the dataset itself. That data must be clean and usable. For time-based applications, data must also be continuously updated, perhaps even in real time. It has to be hosted somewhere that facilitates computational analysis to unlock business value. Finally, as businesses feel the effect of CCPA and GDPR requirements, they will demand data sources that are provably compliant. By building an offering around data markets that fulfill all these features, a primary source offering provides vastly more value than secondary market options fraught with friction and risk.

Conclusion

In this letter, we have proposed a protocol for fairly valuing and governing data. We presented a valuation model that accounts for the peculiarities of data (inflation, non-fungibility, interdependence, and speculation) and discussed how this model can be combined with a decentralized system to enable user control and governance. We believe that this protocol provides a framework for building cooperative data networks that will power the next stage of a decentralized internet.

But even more fundamentally, the data governance protocol provides a first step towards decentralizing control of critical data resources by allowing data creators and owners to share fairly in the economic rewards that accrue to valuable datasets. Today’s world is set up to reward data aggregators such as big tech companies and governments, powerful centralized entities that track everyone else to capture and control valuable data feeds in perpetuity. The time has come to shift from a world of surveillance capitalism to one of cooperative capitalism. By allowing data creators to participate fairly in the new economic calculus of data, our protocol takes a step towards cracking the hegemony of these data fiefdoms and enabling a more democratic future for consumers and citizens alike.

Acknowledgements

This work benefited from the kind contributions of numerous supporters. In particular, we would like to thank the following for their valuable feedback and suggestions: Alok Vasudev, Jordan Cooper, Spencer Bogart, Aleks Larsen, Joey Krug, Kartik Talwar, Reid Williams, Shilpi Kumar, Caroline Sofiatti, Ken Mitsumoto, Fattaneh Bayatbabolghani, and Will Button. We would also like to acknowledge the Gauntlet team led by Tarun Chitra, Rei Chiang, and John Morrow who helped us test and improve our economic design.

References

[1] Zuboff, Shoshana. The Age of Surveillance Capitalism. New York, NY: Public Affairs. January 15, 2019.

[2] Snowden Archive - Canadian Journalists for Free Expression. https://snowdenarchive.cjfe.org/. Accessed June 9, 2019.

[3] General Data Protection Regulation (GDPR). https://gdpr-info.eu/. April 27, 2016.

[4] Assembly Bill 375 - California Consumer Privacy Act. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375. June 28, 2018.