The Liberalization of Data: A Welfare-Enhancing Information System

Preprint · December 2018

CITATIONS
0

READS
161

2 authors:

Jose Parra-Moyano
Copenhagen Business School
16 PUBLICATIONS **103** CITATIONS
SEE PROFILE

Karl Schmedders
University of Zurich
103 PUBLICATIONS **889** CITATIONS
SEE PROFILE

The Liberalization of Data: A Welfare-Enhancing Information System

José Parra Moyano
Chair of Quantitative Business Administration
University of Zurich
Zurich, Switzerland
jose.parramoyano@business.uzh.ch

Karl Schmedders
Chair of Quantitative Business Administration
University of Zurich
Zurich, Switzerland
karl.schmedders@business.uzh.ch

***Abstract*—Users' data has become a crucial production factor for companies and a necessary asset if they are to compete in the digital ecosystem. However, users' data is a production factor that is not mobile across companies, since a company can only use the data that its customers—its "users"— generate within its own environment and not the data that its users produce outside of it. This represents a market friction that hinders competition, leads to monopolies, and raises the entry barriers for new companies. Additionally, the users generating and owning the data stored in a company have no control over or overview of their data and cannot monetize it. We model the users' data as a production factor in the value generation function of companies and introduce the concept of *data elasticity of value*. Further, and in light of advances in distributed database management, blockchain technology, and data protection regulation, we propose an information system that allows users to sell their data freely to companies other than those within which the data was generated, receiving a self-generated, market-driven basic income. A consequence of this system is that data becomes a mobile production factor, since any company can work with the data that its users generate outside of that company's own environments. Moreover, our system solves some of the data-ownership problems of the current Internet business model, lowers the entry barriers for new data-intensive companies, and enables new income streams for data-intensive companies, which in the case of online platforms allows them to avoid a dependence on online advertisement to finance their operations. We propose this ecosystem at a conceptual level and simulate the impact of companies having access to higher fractions of their users' data under different data elasticities of value. We show that the introduction of such a system could theoretically, and under the taken assumptions, more than double the aggregated value of data-intensive companies.**

Keywords—*Data sharing, blockchain, distributed databases, GDPR, universal basic income, data as labor, data elasticity, information system*

I. INTRODUCTION

The existing Internet business model is dominated by providers of free content and services. The free content and services are offered through platforms on which users spend time and to which they devote attention. The users' interaction with the platform generates data that platform owners can use to better understand their users' behavior and needs. Examples of such platforms include the Google and Yahoo browsers, Google's map service, the video streaming portal YouTube, the music platform Spotify, social and professional networks such as Facebook, LinkedIn, Twitter, or Instagram, online retailers including Amazon, Alibaba, or e-bay, and online newspapers[1].

In order to finance their operations, companies owning platforms sell advertising space to companies interested in reaching those platforms' users. Online platforms are therefore advertising spaces that compete with each other in terms of the type and number of active users they have and of their use of the users' data, which they gather and analyze. Obviously, the higher the number of active users of a platform, the higher the potential reach of the advertisement displayed on the platform. However, the knowledge generated by the platform's owners based on the users' data represents a competitive advantage for the companies owning the platforms, since it allows them to identify which users are more susceptible to specific types of advertising. The better the advertising optimization that companies carry out for their platforms, the higher their competitive advantage, due to the higher return on marketing investment (ROMI) that they can offer to the companies advertising on their platforms. The knowledge generated by analyzing the users' data is so valuable that some platforms with a high number of users but relative low fractions of data per user rent their ad space to platforms with relatively higher fractions of the user's data, such that the latter platforms can directly display optimized advertisement to the users of the former platforms. Examples of these service include AdSense (provided by Google), Amazon Associates (provided by Amazon), and Media.net (provided by Yahoo).

This system has enabled the creation of big content platforms, free services, and markets from which companies and users benefit. Companies providing free content and services have reached the top ranks of company valuations in recent years and online advertising had an estimated global turnover of more than 200 billion dollars in the year 2017 [1]. According to [2], four of the top ten companies in the world by market value in 2017 provided free services to their final users: Amazon.com, Alphabet, Facebook, and Alibaba. Further, six of the top ten companies by market value in 2018 base a significant part of their competitive advantage on the knowledge that they generate from their users' data, as illustrated in Table I. Therefore, users' data has a value.

A study published in 2013 by [3] estimates the net income per individual record for certain datasets, concluding that the "implied market capitalization per Facebook user has fluctuated between USD 40 and USD 300 at different times between 2006 and 2012". Further, the study states that in 2013 "prices in the United States for personal data ranged from USD 0.50 for a street address to USD 2 for a date of

[1] Some of these platforms, including Twitter, LinkedIn, Facebook, and YouTube, do not provide the content themselves. Instead, their users provide the content.

birth, USD 8 for a social security number, USD 3 for a driver's license number, and USD 35 for a military record." These numbers reinforce the idea that users' data has a significant value for companies.

TABLE I. TOP TEN COMPANIES BY MARKET VALUATION IN 2018

Company	Market Value in billion USD
Apple	926.9
Amazon.com	777.8
Alphabet	766.4
Microsoft	750.6
Facebook	541.5
Alibaba	499.4
Berkshire Hathaway	491.9
Tencent Holdings	491.3
JPMorgan Chase	387.7
ExxonMobil	344.1

However, it is worth noting that companies generating data-based knowledge are only using the data that their users generate in their own company environments (i.e., a fraction of their users' total available data). Since typical users might interact not only with one platform or company but with many during their lives, there is a lot of valuable data related to each user that companies are not analyzing. This situation represents a market friction since the data that can be used to produce better products and services is not free to move across companies like capital and labor are. Additionally, this situation is yielding monopolies that own vast amounts of users' data, which is hindering both competition among companies and the entrance of new competitors into the market. We conjecture that if companies could have the data that their users generate outside of their own company systems, they could generate additional economic value.

The reason why companies only analyze the data that their users generate in their own company ecosystems is twofold. First, companies do not have the right to sell their users' data. The data of a user belongs to the user and that user is the only one who can allow the portability of data between companies. Since there is nowadays no easy mechanism, nor incentive, for the users to securely share their data with other instances that are able to monetize those data, users do not allow companies to sell their data. Second, companies have no incentive to sell their users' data to other companies (competitors or not) even if they get the users' permission to do so since they base part of their competitive advantage on the knowledge that they generate from this data.

Nevertheless, thanks to advances in distributed database management and blockchain technology, as well as to the newest regulation on data protection, it is now technically and legally possible to establish a marketplace for data in which users can have full ownership of all their data and in which they are encouraged to monetize it. Additionally, and thanks to these advances, it is possible to define incentives for the companies participating in this marketplace to buy and facilitate the selling of users' data whenever the users grant their permission to do so. Such a marketplace would allow users to share or sell their data with companies other than those where these data were generated, turning data into a mobile production factor for existing companies and new competitors, and therefore solving the abovementioned market friction. Further, users could monetize this data in ways that generate value for themselves and for the data purchasers.

As already mentioned, such a marketplace is only possible thanks to the introduction of one legal and two technical innovations. First, securely sharing private information across distributed databases is possible under specific database architecture, as suggested by [4]. While the system proposed by [4] relies on a central party to verify the right of the network's participants to read and write in the distributed databases, thanks to blockchain technology it is possible today to leave this task to a blockchain protocol and a combination of smart contracts, eliminating the need for all the system's users to trust any third party, converting the system into a purely P2P, decentralized one. Second, according to the General Data Protection Regulation (GDPR) introduced in 2018 by the European Commission [5], companies need to "provide a copy of the personal data undergoing processing" in digital form and free of charge to the data subject (i.e., to the platform's user in our context) and ensure portability of the data between data controllers. Although this regulation only applies to citizens of the European Union, international companies have adapted their data protection standards and practices for all their users to meet the requirements of the GDPR and to avoid parallel standards when it comes to data handling. These three innovations make the system that we introduce in this paper possible.

Part II describes the current business model of, and the ecosystem around, the companies offering free services through online platforms. Part III presents the advances in distributed database architecture that allow the secure and private sharing of information as well as the blockchain technology required to prevent third parties from operating in the system, and describes the relevant points of the GDPR. Part IV introduces the new ecosystem, focusing on its architecture and on the emerging business model that it enables. Part V simulates the aggregated value of the system, the income for the data owners, and the increase in aggregated company values, using on different metrics. Part VI concludes.

II. CURRENT ECOSYSTEM

In this section we briefly describe the existing Internet business model and define schematic production functions in order to present the role of data in the value function of online platforms.

A. Current Business Model

The Internet ecosystem comprises different business models and data and financial flows between its participants. We are particularly interested in the interaction between three types of agents of the Internet economy:

- Companies offering free or "freemium" services through online platforms such as web browsers, social networks, search engines, and music and video streaming platforms, or online publishers.
- Companies making online advertising that is displayed on online platforms.
- Final users or consumers that spend time using the services provided by the online platforms.

We present a simplified model of data, services, and money flow between these three agents, inspired by the map of dollar flows presented by [6]. While other agents and interaction possibilities do also exist in the Internet business model, we use this simplified model as a baseline to illustrate the current flow of money, data, and services, and to be able to compare this model with the system that we later introduce. In this model, represented in Figure 1, online platforms provide free content and services to the final users[2]. The content and services offered by online platforms might be very heterogeneous. What companies owning online platforms have in common is that they sell space on their platforms for other companies (to which we refer here as "online advertisers") to broadcast advertising to the platforms' users. Online advertisers are entities interested in getting the users' attention for purposes as diverse as carrying out online marketing, increasing brand awareness, spreading political ideas, or gaining members or affiliates, etc. Users benefiting from the free content and services "pay" the platform owners with data and attention. The attention that the users devote to enjoying and benefiting from the platforms' services and content is necessary if the platforms are to display the advertisements of the online advertisers. The data that users generate when interacting with the platform is analyzed by the company owning the platform both to identify which users are more susceptible to specific types of advertisement and to profile and segment the users, such that they can optimize the target population of their clients' advertisement.

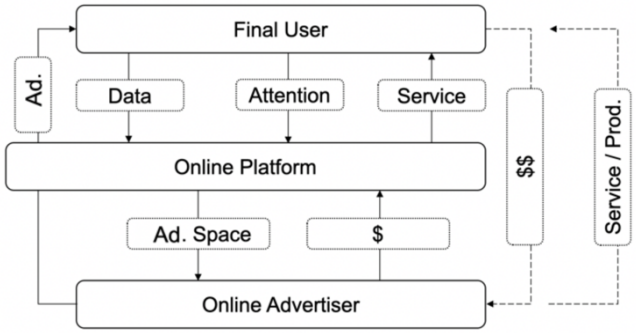

Figure 1: Simplified model of data, revenue, advertising, products, and services flow in the Internet business model, derived from [6].

Different models are used to measure the pricing of online advertisement. The model described by [7] explains how the choice of a pricing model for advertising has become a critical issue for firms like Google that base their value on online advertisement. Specifically, the authors present a comparison of pricing models in online advertising using the principal–agent framework to study the two most popular pricing models: input-based cost per thousand impressions (CPM) and performance-based cost per click-through (CPC). They state that advertisers often classify the consumer population into a target market segment and a non-target market segment depending on how likely they are to purchase their product [8], and that consequently advertising to consumers in the target segment yields a higher return relative to advertising to consumers in the non-target segment. Further, they identify the cost of mis-targeting advertisement as one of the four factors with the highest impact on the revenue generated by the CPM and CPC models. The classification of consumers into a target market segment and a non-target market segment, as well as the quota of mis-targeting advertisement is influenced by the knowledge generated using the users' data. Therefore, the data and the knowledge derived from it are crucial for the platforms that we are describing here.

B. Economic Value Functions

Continuing with the simplified model that we are describing, we introduce a schematic representation of the online platforms' value functions to illustrate the value-generating nature of data in the Internet business model. We illustrate the value function of online platforms as a Cobb–Douglas production function with three input factors: labor (L), capital (K), and data (D), with parameters α, β, and γ as output elasticities of the respective factors:

$$Y = K^{\alpha} L^{\beta} D^{\gamma}. \quad (1)$$

Using this equation, which is the equation the most widely used in economics to model firms' economic output, we can define the *data elasticity of value* as the responsiveness of the value of the platform to a change in the amount of data that the platform has at its disposal regarding its users. This elasticity is defined by γ. Equation (1) only holds if the production factors are free to move between companies.

Given the current nature of the economy, both labor and capital are mobile production factors that online platforms can use and for which they can compete. Nothing prevents labor—"workers"—from switching between platforms and nothing prevents investors from moving their capital from one platform to another. However, it is currently impossible for a platform to use the data of another platform. Data is therefore not mobile, which results in a market friction. Additionally, this friction underpins the monopolistic power of online platforms, since it makes it very difficult for new competitors to work with as much data as established platforms.

III. Technology and Legal Framework

In this section we describe the three innovations that allow the design of the system that we propose. Namely, the private sharing of information across distributed databases, blockchain technology, and the GDPR.

A. Private Information Sharing across Databases

References [4] and [9] present a database architecture that allows the electronic sharing of privacy-sensitive data across distinct organizations while enabling the organizations to keep their legacy databases and maintain ownership of the data that they generate and store. Specifically, [9] present a system in which health care institutions (which they refer to as data producers) generate data about patients. We will refer to the patients as data subjects. In [9]'s system, the data subjects are the subjects whose data is produced and treated by the health care institutions (i.e., by the data producers). Further, many data producers generate data about each data subject. This data is privately stored in the internal database

[2] Although some of these services might follow a so-called freemium model in which basic functionalities and content are offered free of charge and extended content or service is provided for a fee, we take a simplified version of the model to enable a structured comparison between the currently existing model and the model that we introduce in later chapters.

of each data producer. Therefore, the data set of any one data subject may reside distributed across the databases of the data producers. It is important to note that data producers only store and see the data that they produce for each subject (i.e., a fraction of the total data that exist regarding each user) such that the full dataset of a data subject might be split among different data producers and it may be that no data producer has a full copy of a subject's data. For the system to work, the databases' schema is irrelevant; it is sufficient for the data producers to provide some read-only views into this data, conforming to standardized schemas. References [4] and [9] provide three privacy features that their system needs to fulfill:

1) *Data privacy:* The querier learns only the answer to the query and not any of the data used to compute that answer.

2) *Query privacy:* The data owner does not learn the query, only that a query was performed against a particular user's information.

3) *Anonymous communication*: Queriers and data owners do not know who the opposite party is.

In their system, each view consists of tuples of the form *(id, attr1, attr2, ...)*, where *id* is a locally unique identifier for each user. Further, they suggest a network model in which data providers have universal connectivity and asynchronous messaging such as a web service layer built on HTTP/SSL. The security model suggested by [9] is standard. Specifically, [9] suggest that data providers in the system have public keys, *K*, and private keys, *k*. With such a security model, a provider A communicating with a provider B is able to sign messages using a function *S(kA, msg)*. Provider B can verify the signature on messages that it receives using *V(KB, msg)*. Providers in this system are also allowed to encrypt and decrypt messages using functions *E(KB, msg)* and *D(kA, msg)*, respectively. For the system to work, providers must be able to verify that any other provider with which they communicate is authorized to participate. Reference [9] accomplish this by having the public keys signed by a trusted accreditation agency. The ultimate goal of this system is to enable participants to run queries against the stored, distributed data according to their rights to do so. For this purpose, the system participants use a known, unique identifier such as a social security number or a combination of name and date of birth to indicate whose data the asker is interested in. Queries are written in a relational algebraic language similar to SQL, which makes it possible to perform standard select, project, and join operations against tables that are fragmented across various nodes.

This system is sustained by an architecture that supports queries that reveal enough information for the organizations to run their businesses, but no more. The querying process designed by [9] is split into two phases. Phase 1 performs a global search for records pertaining to the data subject in question and returns a set of data handles, each of which indicates the presence of a record somewhere in the system but does not reveal where that record resides. Phase 2 uses the data handles to execute a relational algebraic query that keeps the original data hidden from the asker and keeps the query hidden from the data owners. In Phase 1, a message, *m*, from provider A to provider B is encrypted first with the public key of B, *KB*, generating the tuple *(B, E(KB, m))*. B uses its private key, *kB*, to recover the original message, *m*.

Once the data handles pertaining to the desired query have been discovered, the actual query (Phase 2) is executed. Reference [9] make use of a blind comparer owned by a third party (the comparer cannot see where either the data or the query came from) and two onion skin roots to perform the query such that the query maker does not learn from which node it is retrieving the information.

This system makes it possible to share data between nodes in a secure and encrypted manner, making use of standard cryptography. The work of [9] shows the feasibility of such an architecture for data as sensitive as health-related data. However, this system requires a trusted third party to grant the participant nodes in the system the right to make queries in other nodes' databases. In order to establish a decentralized permission key system that allows nodes to decide which nodes have the right to make queries against their data without the need of a trusted third party, and to establish contracts between nodes, we make use of blockchain and smart contract technology.

B. Blockchain and Smart Contract Technology

The idea of using a proof-of-work (PoW) system to implement a distributed timestamp server that prevents the double spending problem in peer-to-peer networks was introduced by [10]. In the system described by [10] a server takes a hash of a block of items (transactions) that is timestamped and broadcasts this new hash to a network of nodes. Each timestamp includes the previous timestamp in its hash, which forms a chain with each additional timestamp reinforcing the ones before it. The timestamp proves that the data broadcasted must have existed at the time of broadcasting, and this is easily verifiable for the nodes in the network. The chain of blocks that emerges is the so-called blockchain. A block is created by the instance (called a *miner*) that finds the next block's hash with the required amount of zero bits. Reference [10] notes that the average work required to generate a hash is exponential in the number of zero bits of the hash but that it can be verified by executing a single hash. Additionally, [10] states that once the computational effort required for it to satisfy the proof-of-work protocol has been expended, the block cannot be changed without repeating the work. As later blocks are chained after it, the work to change the block would include changing all the blocks after it. This makes the blockchain immutable after even a low number of blocks.

Reference [10] defines an incentive structure that rewards "miners" for verifying the transactions in the system and creating new blocks. This is accomplished by making the first transaction in a block a special transaction that generates a new coin (or coins as it was later implemented) that is owned by the creator of the block. This gives incentives for nodes to support the network and, of equal importance, provides a way to initially distribute coins without the need for a central authority to issue them. Additionally, [10] defines a system of transaction fees paid to the miner. This incentive structure encourages nodes to stay honest since an attacker that assembles more hash power than all the other nodes has to choose between using this hash power to steal back its own payments and using it to generate new coins. The contribution of [10] is twofold. First, it introduces the idea of the "bitcoin", a decentralized peer-to-peer online currency that does not need a central issuer to be fairly minted. Second, it introduces the concept of a proof-of-work-based blockchain that allows the public agreement and consensus that avoids double spending, without the need for a third party.

Since the publication of [10], further contributions have been made to the field. Reference [11] introduced the idea of blockchain-based smart contracts, systems that automatically move digital assets according to arbitrary pre-specified rules. In his work, [11] claims that the logical extension of this idea is decentralized autonomous organizations (DAOs), which are smart contracts that contain the assets and encode the bylaws of entire organizations. The work of [11] gave rise to Ethereum. Ethereum is a blockchain with a built-in Turing-complete programming language that allows anyone to write smart contracts and decentralized applications following their own arbitrary rules for ownership, transaction formats, and state transition functions. Smart contracts can be seen as cryptographic "boxes" that contain value and only unlock it if certain conditions are met. Smart contracts written and verified in Ethereum can interact with off-chain systems and trigger transactions outside of a blockchain, for example in a decentralized database.

The idea of [10] was to have a decentralized peer-to-peer currency that would be mined by CPUs. In fact, [10] does not use the term "hash power" but rather writes "CPU power" to refer to the capability of miners to find the nonce that yields the hash of the next block. Since the introduction of Bitcoin, specialized hardware such as GPU, FPGA, and ASIC miners have been developed to increase the efficiency of the mining process. Nevertheless, the use of this specialized hardware has led to the centralization of resources and has raised the entry barriers for miners. In order to increase decentralization and allow mass mining adoption, protocols such as Monero [12] and others have limited the type of hardware that can mine their currencies to GPU-type hardware only. Although all miners, regardless of their hardware and coin, can connect to a mining pool to secure a smooth mining income across time, such a limitation of the type of usable hardware lowers the barriers to participating in the mining process.

C. General Data Protection Regulation

The GDPR is a regulation introduced by [5] in which the European Union "lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data. This regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data. The free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data."

Among other matters, the GDPR states that "the data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller ["controller" understood as the instance within a company that is responsible for curating the subject's data] in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided". Further, the GDPR states that "the data subject shall have the right to have the personal data transmitted directly from one controller to another, where technically feasible".

IV. New Ecosystem

Both [10] and [11] allow us to build a blockchain-based architecture in which the architecture described by [4] and [9] does not require a trusted third party to define which data producers have the right to read which other producers' data and can instead rely on a blockchain architecture to enforce the arbitrary permissions they desire. Further, [10] allows us to introduce and distribute a Bitcoin-like currency to incentivize miners of a public blockchain to maintain the system, which in turn allows data subjects to monetize their data. Setting limitations that allow GPU-type hardware only lowers the entry barriers for miners and allows a greater number of subjects to participate and "co-own" the system. The GDPR introduced by [5] gives the adequate regulatory context for the system we propose. Specifically, we present here a new infrastructure that allows the development of the ecosystem described in Part II.

A. Ecosystem's Architecture

The system that we propose is defined by two layers. Layer 1 is composed of a distributed database with the same architecture as the one proposed by [4] and [9]. In this architecture, platforms play the role of the data producer in [4] and [9]. Note that platforms' owners already store users' data in databases. In order to define which platform owners have the right to conduct queries against other databases, a blockchain is used. This blockchain is Layer 2 of the system. In this layer, users can publicly state which platform owners have the right to read their data stored in specific databases. We suggest the use of an open blockchain that uses a PoW protocol to allows miners to provide hash power to the network and validate transactions. This blockchain mints a fixed and constant number of coins with each new block, whose difficulty is adapted after a periodic number of blocks to ensure an average constant block-generation time like in [10]. While this is open to future research, we foresee neither a limitation nor a reduction in the number of coins generated per block, this to avoid a deflationary currency (in which a unit of the currency is today worth less than tomorrow), such that the use of the coin to make transactions is not penalized by the coin being more valuable tomorrow than it is today. Additionally, we suggest limiting the type of hardware that can mine the currencies to GPU-type hardware only, like in [12]. By doing this, we intend to reduce the entry barriers for miners. Similar to the Ethereum blockchain, this blockchain allows for the execution of smart contracts and interaction with off-chain applications. In Layer 2, all participants in the system (online platforms and users) hold a node. This node allows them to make transactions in the system, hold currency, make payments, as well as grant and receive reading rights for the databases of the online platforms in Layer 1. Due to the portability that is legally required by [5], online platforms are obliged to provide a copy of all processed user data in an online format to the data controllers that the data owners decide on.

There are three types of transactions possible in Layer 2: (1) monetary transactions between nodes, (2) permission transactions, and (3) smart contract transactions. The monetary transactions are those that occur to move coins between nodes' wallets. They work like Bitcoin transactions. Permission transactions are data entries made by users to write in the blockchain the public key of the node that has the right to conduct queries against the database of another particular node. They serve to ensure that only participants who have been "white-listed" by the users can read the users' data on platforms other than their own. These transactions allow us to substitute the trusted third party required in [4] and [9] by a blockchain. Smart contract transactions are conducted to store and call smart contracts in the system. The purpose of the smart contracts in this system is to allow the

automatic and secure exchange of data and coins between data sellers and data buyers.

At this stage, it is useful to differentiate between "purchasing platforms", which purchase users' data stored on other platforms, and "storing platforms", which store users' data and allow reading access to purchasing platforms. In order for the purchasing platforms to decide which user's data might be interesting for them, and also to price this data, smart contract transactions can be made. Users' public keys need to be linked with the storing platforms in which they have data such that a purchasing platform can make queries against a storing platform's database to learn which data about a user is stored in databases other than their own without being able to read the data. The purpose of this query, which resembles Phase 1 in [4] and [9], is to allow purchasing platforms to price the data that they might eventually buy. By doing this, every purchasing platform could generate a "price tag" for each user's data on each storing platform and purchase the data individually, for each user, at an individually set price. Since storing platforms' owners know the architecture of their databases better than any other instance, they can offer the best smart contracts themselves, guaranteeing the quality, structure, correctness, and standards of the data. Since storing platforms themselves can design and offer these smart contracts, they can receive a fee whenever a purchasing platform uses their smart contracts to read from their databases and therefore benefit when purchasing platforms intend to purchase the data that they ward and helped to generate. This opens a new revenue stream for storing platforms, which become a sort of "identity" custodian for users and are compensated for this.

Resembling Phase 2 in [4] and [9], once the price is agreed between the data purchaser and the user, a query, executed by another smart contract, would be carried out in order to allow the purchasing platform to read the data from the storing platform and to conduct the previously agreed monetary transaction with the user.

It is important to note that a user can only generate data in his or her interaction with a platform and can therefore not just "freely" create useless data to sell to potential purchasers.

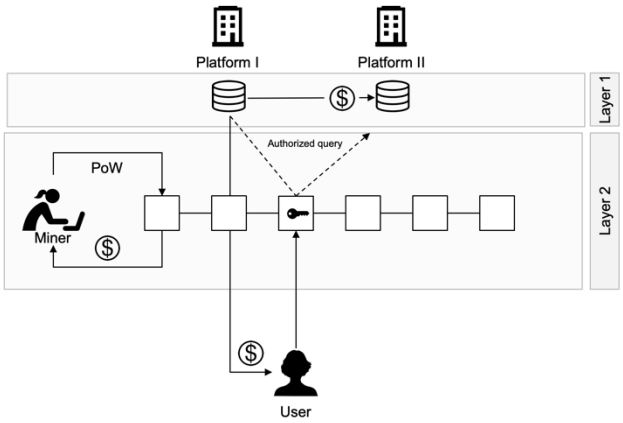

Fig. 2. The ecosystem's architecture

An exemplary transaction between a user and two platforms with which he or she interacts would work as follows. Platform I (the purchasing platform) would conduct a query against the database of Platform II (the storing platform) to see which type of data the user has stored there. In order to carry out this query, a smart contract written by Platform II would be used. In this operation, Platform II is already compensated for offering a standard, correct reading of its database. This query is similar to Phase I in [4] and [9], and would allow Platform I to price the user's data stored on Platform II without actually seeing the data. If the user accepts the price offered by Platform I for reading her data stored on Platform II, she grants Platform I the right to read from Platform II by writing into the blockchain a smart contract containing the terms of the purchase and the public key of both platforms. Then, whenever Platform I attempts to read the user's information in the databases of Platform II, it pays the corresponding price to the smart contract and the contract distributes the payment to the user for her data and triggers the query against Platform's II database outside of the blockchain. Platform I will have then securely purchased the user's data stored on Platform II, Platform II will have received revenue it return for ensuring the data's quality, and the user will have received income for the data she generated. By executing these operations, the data becomes a mobile production function such that each platform can use the data produced on the other platform at a market-driven price. Of course, interaction layers are required to make these operations feasible for all participants.

B. *Economic Implications*

By introducing this new construction, we allow the Internet business model to evolve since now online platforms and companies of any kind can purchase data directly from users. This fosters competition between current online platforms, makes data a mobile production factor, contributes to lowering the entry barriers for new competitors since capital can now acquire data in a legal way, and empowers users to monetize their online activity.

The proposed ecosystem transforms data from a platform-specific asset into a mobile production factor that can be used, after a payment to the user and to the owner of the platform on which the data was generated, by any participant in the system. This turns the data in (1) into a mobile production function and contributes to having a frictionless market.

Users in this system would receive income from the data that they generate online. Should technology in the future make the work and abilities of a part of the labor force irrelevant, an income like the one that would emerge from our system could become very relevant.

V. SIMULATION

The system that we propose would result in platforms having access to a higher fraction of the data of their current user bases. In order to measure the impact of our system on the aggregated value generated by platforms, we simulate an economy with p platforms and U users. Each platform in this economy has a fixed user base equal to u_p, which is a fraction of the total users', U, in the economy. The user bases of different platforms might overlap with each other. Each platform owns, on average, a fraction φ_p of the total data of its users. Since a user might participate in many platforms, the platforms might have different fractions of the total data of a user. Platforms generate value from the data that they own, following (1). We are interested in estimating the theoretic impact of increases in φ_p, under different values for data elasticity, γ.

In order to create values for a baseline economy that serves to compare different scenarios, we generate random

values for each company's φ to fit a truncated Chi-squared distribution with mean equal to 0.1, being φ bounded between 0 and 1. Such a distribution allows us to generate few platforms owning a high share of their users' data and many platforms owning very low fractions of their users' data:

$$\varphi \sim X^2_{0.1} \in [0,1]. \quad (2)$$

Similarly, we generate random values for each company's u to fit a truncated Chi-squared distribution with mean equal to 0.1, being u bounded between 0 and 1. Such a distribution allows us to generate few platforms owning a high share of U and many platforms owning very low fractions of U:

$$u \sim X^2_{0.1} \in [0,1]. \quad (3)$$

Further, we define the total volume of data, D, available for production in the economy[3] as the sum of the data that each company owns for the total population:

$$D = \sum_{p=1}^{P} U * u_p * \varphi_p \quad (4)$$

Defining the economy in this way allows us to simulate the total value of this economy, which using (1) and fixing capital and labor to 1 is given by:

$$Y = D^{\gamma}. \quad (5)$$

We simulate the aggregated value, Y, of an economy with 1'000 platforms, 100'000 users, and a value of $\varphi = 0.1$. We use the result of this simulation as a baseline model to compare with higher mean values of φ under scenarios with different data elasticities, γ.

Table II presents the result of simulating this economy 100 times for different values of φ and different elasticities. This simulation serves the sole purpose of comparing a baseline economy, in which platforms do not share users' data (baseline $\varphi = 0.1$), with an economy in which the sharing of users' data is possible and yields higher values of φ. We compare the aggregated value of the economy in the baseline model with the aggregated value of economies with higher levels of φ.

In Table III we present the results of the simulation in relative terms, in which the economy with $\varphi = 0.1$ represents the baseline economy. We observe the impact of increasing the average share of users' data available to the platforms under four levels of elasticity. The resulting aggregated value of the platforms would increase by a factor ranging between 1.51 and 2.25 times, depending on the increase in φ and the level of elasticity, γ, in the economy.

TABLE II.
SIMULATION OF THE OUTPUT OF DIFFERENT ECONOMIES

Y_p	$\varphi = 0.1$	$\varphi = 0.2$	$\varphi = 0.3$
$\gamma = 0.1$	320.260	482.332	615.201
$\gamma = 0.2$	355.990	573.514	770.237
$\gamma = 0.3$	530.957	1040.61	1195.22

TABLE III.
SIMULATION OF THE RELATIVE OUTPUT OF DIFFERENT ECONOMIES

$Y\%$	$\varphi = 0.1$	$\varphi = 0.2$	$\varphi = 0.3$
$\gamma = 0.1$	1	1.51	1.92
$\gamma = 0.2$	1	1.61	2.16
$\gamma = 0.3$	1	1.96	2.25

VI. Conclusion

In this paper we have modeled the impact of data in the value generation process of companies and identified that data, in the current ecosystem, is not a mobile production function but a company-specific production function. We consider this a market inefficiency subject to be solved. In order to solve it, we propose an information system that allows the free movement of data across platforms, incentivizing both companies and users to enable this exchange of data. We estimate that such an ecosystem could more than double the aggregated value of data-intensive companies.

Additionally, an ecosystem like the one described in this paper would result in a series of changes for the agents participating in it. First, it would generate a data-based, market-driven basic income for users. In such an ecosystem data would start substituting or complementing labor to generate income for users. Second, it would lower the entry barriers for new data-intensive companies and increase the level of competition in the economy, which could result in better products and services for users and reduce the risk of monopolies. Third, data piracy would be less incentivized since companies could legally acquire structured and up-to-date user data, which would help to enforce data protection regulations such as the GDPR.

While the system that we propose still requires a lot of effort to precisely define the database architecture, smart contracts, and processes within the system, as well as to create good usability for the users, we consider it a good basis upon to which further research in information systems can build. We find the system particularly interesting because it builds incentives for competitors to collaborate in certain areas of their business—namely, in allowing data sharing, fostering economic growth, and increasing their market valuations. Further research would also be required to identify the data elasticity of value in the economy and to compute a more precise simulation.

Acknowledgments

We would like to thank the participants of the seminar "Current Research in Management Science" held during the fall of 2018 at the University of Zurich for their constructive criticism, and in particular Gregor Reich for suggesting specific improvements in the database architecture of the system. Further, we would like to thank Dave Brooks from ELCS for his valuable support and comments on the paper.

[3] Note that D in this case is the same D as the one we defined in (1).

REFERENCES

[1] Bundeskartellamt, "Competition and Consumer Protection in the Digital Economy," Working Paper Series, 2018.

[2] NCES. "The 100 largest companies in the world by market value in 2018 (in billion U.S. dollars)" Statista - The Statistics Portal. Last accessed on December 6, 2018.

[3] C. Reimsbach-Kounatze, T. Reynolds, and P. Stryszowski, "Exploring the economics of personal data: a survey of methodologies for measuring monetary value," OECD, April 2013.

[4] M. Siegenthaler and K. Birman, "Privacy enforcement for distributed healthcare queries," in *Pervasive Health,* 2009

[5] European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," 2016

[6] S. O'Donnell, "An economic map of the internet" TPRC 30th Research Conference on Communication, Information, and Internet Policy, Alexandria, VA, 2002.

[7] V. Jacob, N. Kumar, and K. Asdemir, "Pricing Models for Online Advertising: CPM vs. CPC," nformation Systems Research, Vol. 23, No. 3, Part 1 of 2, September 2012, pp. 804–822

[8] H. Assel, and H.M. Cannon, "Do demographics help in media selection?", Journal of Demographic Research, 19 (6), 7–11, 1979

[9] M. Siegenthaler and K. Birman, "Sharing private information across distributed databases," 2009 Eighth IEEE International Symposium on Network Computing and Applications.

[10] S. Nakamoto, "Bitcoin a peer-to-peer electronic cash system," https://bitcoin.org/bitcoin.pdf, 2008

[11] V. Buterin, "Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform," http://ethereum.org/ethereum.html, 2013.

[12] N. van Saberhagen, "Cryptonote V 2.0," https://whitepaperdatabase.com/monero-xmr-whitepaper/, 2013