ETLA Working Papers

No. 53

29 September 2017

Pantelis Koutroumpis
Aija Leiponen
Llewellyn D W Thomas

The (Unfulfilled) Potential of Data Marketplaces

Suggested citation: Koutroumpis, Pantelis & Leiponen, Aija & Thomas, Llewellyn D W (29.9.2017). "The (Unfulfilled) Potential of Data Marketplaces". ETLA Working Papers No 53. http://pub.etla.fi/ETLA-Working-Papers-53.pdf

ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

ISSN-L 2323–2420
ISSN 2323–2420 (print)
ISSN 2323–2439 (online)

DATAMARKKINOIDEN LUNASTAMATTOMAT LUPAUKSET

Pantelis Koutroumpis
Imperial College Business School

Aija Leiponen
Cornell University and ETLA – The Research Institute of the Finnish Economy

Llewellyn D W Thomas
Imperial College Business School

TIIVISTELMÄ

Vaikka teollisia massadata-aineistoja kertyy valtavaa vauhtia, niitä harvoin jaetaan laajalti tai myydään avoimilla markkinapaikoilla. Tässä tutkimuksessa arvioidaan datamarkkinoiden erityispiirteitä ja vertaillaan olemassa olevia markkinapaikkoja markkinoiden suunnittelun teorioiden valossa. Kahdenvälisen datavaihdannan lisäksi kuvaillaan keskitettyjä ja hajautettuja monenvälisiä markkinamekanismeja. Tutkimuksessa osoitetaan, että datan *provenienssi* eli luotettava tieto datan alkuperästä on olennaista säilyttää, jotta dataa voidaan suojella ja sen laatua arvioida. Monenvälisillä markkinoilla alkuperää koskeva tietoa on mahdollista luotettavasti ylläpitää joko suljetuissa datakonsortioissa tai hajautetuilla data-alustoilla jos otetaan käyttöön lohkoketjun tapaisia teknologioita liiketoimien todentamiseksi.

Asiasanat: Datamarkkinat, datan kauppa, markkinoiden suunnittelu

JEL luokat: D47, D82, K12, L14, O34

THE (UNFULFILLED) POTENTIAL OF DATA MARKETPLACES

Pantelis Koutroumpis
Imperial College Business School

Aija Leiponen
Cornell University and ETLA – The Research Institute of the Finnish Economy

Llewellyn D W Thomas
Imperial College Business School

ABSTRACT

Although industrial datasets are abundant and growing daily, they are not being shared or traded openly and transparently on a large scale. We investigate the nature of data trading with a conceptual market design approach and demonstrate the importance of provenance to overcome protection and quality concerns. We consider the requirements for data marketplaces, comparing existing data marketplaces against standard market design metrics and outline both centralized and decentralized multilateral designs. We assess the benefits and potential operational features of emerging multilateral designs. We conclude with future research directions.

Keywords: Data marketplaces, data trading, market design

JEL codes: D47, D82, K12, L14, O34

Acknowledgment: This paper was completed while Leiponen was on a sabbatical leave at ETLA. She gratefully acknowledges the support of BRIE-ETLA (Tekes project 2666/31/2015) and SWiPE (Strategic Research Council at the Academy of Finland project no 303667).

INTRODUCTION

Data have long been shared and traded: for example, academics share research data and businesses share household credit data. In recent years, much of the data being traded are "exhaust data", created as a by-product of other activities such as online shopping or socializing, rather than specifically created for an analytical purpose (Manyika et al. 2011; Mayer-Schonberger and Cukier 2013). Data concerning the purchasing habits of consumers have become the first data industry segment that now sees significant trading. Concerns regarding the associated trading practices lead to the Data Brokers report by the US Federal Trade Commission (Ramirez et al. 2014). Technology firms such as Amazon and Google also are data vendors, collecting and aggregating data from online sources and sharing with or selling to third parties. Additionally, there have always been thriving marketplaces for stolen data (Holt and Lampke 2010), such as credit card numbers or user profile data (Shulman 2010). However, the trading of data has barely been studied.

New types of data platforms have been envisioned that would either have data trading as their core activity, or that trade data that arise from their core operations (Parmar et al. 2014; Thomas and Leiponen 2016). Such data marketplaces would allow other parties to upload and maintain datasets, with access, manipulation and use of the data by other parties regulated through varying licensing models (Schomm et al. 2013). Conceptually, data marketplaces are multi-sided platforms, where a digital intermediary connects data providers, data purchasers, and other complementary technology providers (Eisenmann et al. 2006; Parker and Van Alstyne 2005). These platforms would, in principle, generate value for both data buyers and sellers through enhanced market efficiency, resource allocation efficiency, and an improved match between supply and demand (Bakos 1991; Soh et al. 2006). The initial entrants appear to be large cloud computing incumbents, such as Microsoft Azure Marketplace and Amazon Data Marketplace. Other data marketplaces, such as those established by insurers,[1] are based around industry verticals in which the pooling of data enhances the performance of the whole industry. These within-industry marketplaces tend to focus on using data to address a specific shared risk—regulatory, commercial, or technical—while still allowing participating organizations to normally compete for business from customers (Gopalkrishnan et al. 2013).

However, despite these examples, in practice data are rarely shared or traded on a large scale through multilateral platforms (Borgman 2012). There are surprisingly few data "stores" or platforms for either individual or industrial data, and of those that exist, most are either not commercial (e.g., the London Data Store does not actually sell data), or sell via bilateral and negotiated contractual relationships (e.g., consumer data sold by data brokers such as Acxiom). Moreover, there are abundant examples of failed data platforms (Carnelley et al. 2016; Markl 2014). It thus appears challenging to set up systems to trade data through open markets in the same way we trade many other goods, including many intangible goods such as content and inventions.

This paper develops a conceptual market design framework to examine possible forms of governance for trading data, including their benefits and deficiencies. We make several contributions to the emerging literature on data markets. First, we demonstrate that markets for data require the establishment of rigorous provenance, as expressed through verifiable metadata, for the data goods being sold, as the protection, or control rights, regime is not sufficiently robust. Intellectual property rights do not appear to facilitate controlling the use and dissemination of data products (Duch-Brown et al. 2017; Mattioli 2014; Wald 2002), and, therefore, data products are highly likely to be associated with significant unintended knowledge spillovers. Furthermore, traditional mechanisms ensuring quality in multilateral

[1] http://www.out-law.com/en/articles/2016/january/more-data-sharing-at-heart-of-taskforce-plans-to-tackle-insurance-fraud/; retrieved 07/04/2016.

markets such as reputation systems (Dellarocas 2005; Moreno and Terwiesch 2014; Pavlou and Gefen 2004) are insufficient when the sellers themselves may not be aware of the (lack of) quality of their goods. For example, personal and some organizational data may be associated with laws and regulations about their use, and the landscape around privacy and secrecy is evolving rapidly around the world, creating substantial uncertainty about the value and usability of certain types of data, particularly those compiled from multiple sources. As the value of data critically depends on the appropriateness of the procedures associated with collecting, organizing, and curating them, knowledge about the origins of data is usually critical to discern their quality and protection status.

Second, building upon the market design literature of Roth (2002; 2008), we characterize the main data market matching mechanisms and present illustrative examples of actual data marketplaces utilizing these designs. One-to-one matching is a bilateral relationship that involves two parties, involving negotiated terms of exchange, as exemplified by Google and Acxiom. One-to-many matching is a dispersal marketplace, characterized by standardized terms of exchange, such as data distributed through APIs. A many-to-one marketplace is characterized by the harvesting of data, where many sellers provide their data to a single buyer in exchange for services under terms of exchange that typically resemble barter. An example of this is Google Search where users provide search terms (data) in return for the search results. Multilateral or many-to-many marketplaces are trading platforms upon which anybody (or at least a large number of registered users) can upload and maintain datasets, and where access to and use of the data is regulated through varying licensing models. Roth’s (ibid.) market design framework allows us to assess the benefits and shortcomings of each matching mechanism and thereby derive conclusions about the types of data and trades that can be exchanged via each matching mechanism.

Third, we view data as a common-pool resource due to the poor legal and technical control associated with it (Koutroumpis and Leiponen 2013). Integrating Ostrom (1990) with Roth (2002; 2008), we argue there are three potential multilateral market designs, but only two of which may be feasible in the future. We find that multilateral marketplaces are characterized by a trade-off between provenance (representing control and quality) and transaction costs. In a centralized model, the market intermediary trades off quality (provenance control) for lower transaction costs, while in a decentralized model based on a distributed ledger technology (Catalini and Gans 2016), the market intermediary trades off transaction costs for increased provenance control. We argue that a decentralized market design based on distributed ledgers reduces the impediment to trading posed by data quality (including legitimacy) as trades can be found in the public ledger and enforced via authentication of each transaction, and it clearly differentiates between the holder of a data asset and the original entity that created the data assets. While the centralized multilateral platform is technically entirely feasible, it provides poor control rights enforcement and therefore may not be commercially feasible.

A third emerging multilateral design trades off liquidity against enforcement of control rights. Per Ostrom (ibid.), contract enforcement is a critical feature of sustainable common-pool resource management. Collective governance of a small-scale and closed multilateral platform is likely to involve fewer trading partners but more well-defined boundaries and rules of exchange, and more stringent enforcement of relatively comprehensive contracts. Thus, in settings where distributed ledger technologies do not sufficiently address the problem of data provenance, we expect forms of collective multilateral governance to emerge. We discuss examples of the Industrial Internet of Things. We finish by identifying future research directions for Information Systems scholars.

Throughout our analyses, we focus on observational data that has not yet been significantly processed or manipulated into a content form, such as a business report. Such data corresponds to social, laboratory, and measurement data compiled by humans or machines (Uhlir and Cohen 2011), and business datasets used for analysis purposes – collections of data items that have grouping, content, relatedness and purpose (Borgman 2012). These data are rarely valuable alone and are generally input into analytics (such as a software program) to generate insights that can become expressed as content-based information good (such as a scientific report). Data are intermediate goods in that they are generally produced with the view of being combined and transformed to create other goods (Koutroumpis et al. 2016).

This paper is organized as follows. The following section reviews the specific trading challenges of data marketplaces, considering the protection regime and quality control, and highlights the importance of provenance in data marketplaces. The third section develops and compares possible data market designs. We then discuss the viability of existing models, focusing specifically on multilateral markets, discussing the benefits and drawbacks of centralized, decentralized, and collective data platforms. We then consider future research directions.

DATA TRADING CHALLENGES AND PROVENANCE

Marketplaces for data exhibit similar characteristics to those for ideas and patents. Ideas, patents, and data are non-rivalrous in use, in that a single idea or datum may be usable by many individuals and replicated at low marginal cost (Koutroumpis et al. 2016; Romer 1990). Furthermore, ideas and patents require matching with complementary ideas and assets for their commercialization (Bresnahan and Trajtenberg 1995; Gans and Stern 2010; Teece 1986). Similarly, data is an intermediate good and needs to be further processed and analyzed with complementary technologies and data in order for the consumer to gain utility (Chebli et al. 2015; Koutroumpis et al. 2016). In addition, like data, it can be difficult for sellers to appropriate the full value of an idea good in the absence of strong intellectual property rights (Arrow 1962; Teece 1986). Encouragingly, Gans and Stern (2010) suggest that effective markets for ideas may exist in settings where there is protection for the good being traded. In contrast, Hagiu and Yoffie (2013), considering the markets for patents, have argued that multilateral online marketplaces for patents are not viable due to the burdensome arrangements that would be required to ensure that high quality patents that are offered for sale. Indeed, when the quality of the good is imperfectly observable, markets tend to be flooded with low-quality goods (Akerlof 1970), and electronic markets for such goods may function particularly poorly (Overby and Jap 2009).

Taken together, scholarship into markets for ideas and patents suggests that specific governance mechanisms are needed for a multilateral data market to succeed. In particular, this literature indicates that for market participants to safely transact (Roth 2007; 2008), adequate protection and quality assurance of the traded goods are necessary.

Protection regime

The protection of intangible goods such as ideas and data is enforced through legal institutions that are normally outside of the market institutions within which they are traded. For instance, patent enforcement occurs in national or international justice systems, and the organizing principles are determined by wider societal institutions, such as legislation. In the market for ideas, legal instruments such as patents, copyrights and trademarks are often available to protect the idea, technology or innovation (Gans and Stern 2010).

In contrast, the legal instruments that are available to protect data are less conducive to an efficient marketplace. Although databases are theoretically protected under copyright, the strength and extent of the protection are limited and variable. For databases, copyright typically only protects an empty shell – the structure and organization of the database, not the individual data it contains (unless the data themselves are creative content), provided there is an original contribution in putting the dataset together. This weak protection is compounded by jurisdictional differences, with the US having no specific database rights, Australian copyright law protecting databases, and with the Canadian approach somewhere in the middle (Zhu and Madnick 2009). In the EU, the database directive of 1995 seeks to extend protection to the non-copyrightable aspects of databases, for example, when the data are provided in a different order or in a manipulated format, and even to substantial parts of the database, as long as there has been a substantial investment to compile it. In the US, despite some extensions of copyright to situations where the selection or arrangement of data required judgment,[2] it is difficult to prevent a competitor from taking substantial amounts of material from collections of data and using them in a competing product (Wald 2002). To remedy such legal challenges, law scholars have proposed limited datarights that would prevent unauthorized use of the data for a specified amount of time, but not distribution of data (Mattioli 2014). The goal of such datarights is a balance between protection and the encouragement of innovation in data practices.

However, designing data protections has proved difficult. The European database right appears to have had no measurable impact on database production.[3] When data is an observational record it can be particularly challenging to track and protect. Numerical data can be streamed or shared from a database, after which it may be impossible to detect where the data originated. The order of the individual observations or variables may be substantially altered, after which the data are no longer protected by copyright that essentially covers the "expression", i.e., the original structure of the database itself. The data may also be used to analyze a statistical question, and the results of the analyses are not subject to the original

[2] 945 F. 2d 509 - Key Publications Inc. v. Chinatown Today Publishing Enterprises, Inc. 1991.

[3] European Commission, 2005, "First evaluation of Directive 96/9/EC on the legal protection of databases"; Source: http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf

copyright, nor is it clear how datarights would apply to them. Moreover, barring legal access to audit the data management and analytical procedures, an outside party will not be able to prove that a particular data source was utilized for an analytical output. Such audit rights are regularly stipulated in bilateral data license agreements, but they would be impossible to enforce in a large-scale multilateral context.

Therefore, as data goods have a weak protection regime, protection is usually effected through contractual means. Data license agreements provide for the derivation, collection, reproduction, attribution, confidentiality, and commercial use of data. These licenses tend to be lengthy and complicated, as contract terms are contextual in that they are based upon laws, regulations, measurement units and values of a particular jurisdiction (Truong et al. 2012), and seek to define the admissible commercial utilization of data in explicit terms that depend on the market. As such, the legal implications of merging data sets that are governed by separate contracts can be problematic.

Quality control

Quality assurance within markets for intangible goods is normally addressed through verification processes offered by the market intermediary for a fee (Catalini and Gans 2016; Dushnitsky and Klueter 2011; Gefen and Pavlou 2012). Studies have shown that the reputation of the online marketplace itself can reduce the perceived risk of a market (Gefen and Pavlou 2012; Pavlou and Gefen 2004).

When goods being traded within the market are heterogenous in form and content, the intermediary offers verification services that are often focused on the seller, not the goods themselves. This can take the form of controlling the entry of sellers into the marketplace, as well as establishing reputation systems that rate the quality of the participants. Studies have shown how the reputation of the market participants themselves influences the efficiency of the market (Dellarocas 2005), such as through the publication of previous transactions

(Moreno and Terwiesch 2014) or through buyer feedback (Pavlou and Dimoka 2006). In contrast, when the goods have a homogenous legal form while heterogenous in content, such as patents, the intermediary can undertake verification processes that consider specifically the good itself. For instance, in the markets for patents, Dushnitsky and Klueter (2011), have shown that for effective operation multilateral markets have required thorough screening and disclosure requirements of the patents themselves to surmount the adverse selection problem in which only weak patents are offered.

Participant-level quality control verification by intermediaries in data markets undoubtedly will be required, as they are effective means of ensuring market safety when there are high levels of moral hazard (Dellarocas 2005; Pavlou and Gefen 2004). Goods-level verification by intermediaries such as screening and disclosure are much more problematic, given the vast heterogeneity in both the format and content of data goods.

However, the key challenge for quality control, independent of participant-level verification, is the challenge posed by privacy. The sellers themselves may not be aware of the legal status of their data goods and hence their quality. This is particularly true when the data includes personal (customer) information. Personal data raise privacy concerns because of the inalienability and inferability of data (Koutroumpis et al. 2016). Personal data such as health records or mobile phone records permanently point to a specific individual, and once several such data streams are integrated, the person in question can be identified despite anonymization (Ohm 2010). Computer scientists have convincingly demonstrated that they can often “reidentify” or “deanonymize” individuals hidden in anonymized data with ease (Sweeney 2000), highlighting that regulation of privacy is of crucial concern (athough there is increasing effort to ensure such anonymization processes are effective, see Menon and Sarkar 2016). Furthermore, privacy is enacted through regulations, and the regulatory environment is typified by complexity. National regulations tend to represent a patchwork of solutions for collecting and using data in support of different institutional and corporate aims (Schwab et al. 2011). The challenges of regulatory complexity within a jurisdiction are magnified by the limited coordination mechanisms between legal frameworks, policies and guidelines for different sources of data (Zuiderwijk and Janssen 2013). This regulatory complexity is further compounded by a lack of global interoperability with each jurisdiction creating its own privacy framework (Schwab et al. 2011), with different structures of enabling legislation, regulatory enforcement agencies, and jurisprudence (Perrin et al. 2013).

Taken together, this complex regulatory landscape suggests that the legality of data sales from proscribed industries, for particular functions, or across international borders may not be clear. Furthermore, when datasets have been combined into the hybrid data goods, consisting of data from a variety of industries, a variety of jurisdictions, differing originating contractual conditions, and for use in particular corporate functions, the legal privacy status of the hybrid product may not be clear. By not having certainty on the legal and regulatory status of any particular data good, the seller themselves may be (perhaps unwittingly) offering a lower quality product. In such cases, verification processes that consider the participant may only be partially effective. Furthermore, goods-level verification processes such as disclosure and screening may either be not feasible, due to opacity in the original process that combined the data or the sources of some of the constituent data, or so time consuming and costly as to result in extremely high transaction costs.

Data provenance

As a consequence of both the protection regime and quality control challenges, there is a strong need for data to have rigorous and comprehensive records of its origin, its characteristics, and history. The value of data significantly depends on this complementary "metadata" information about its provenance, in other words, its quality and legality. Data and metadata are strongly complementary in creating value, and the nondisclosure of the provenance and pedigree of data may prevent innovative applications of data analytics (Mattioli 2014). There may be significant additional incentives not to disclose the underlying metadata, concerning the associated data sources and practices. For instance, privacy regulations may prohibit the disclosure; relevant information may be strategically hidden, especially if it reveals the low quality of the data or helps de-anonymize an otherwise anonymous pool of individuals; and methods of data preparation themselves can be valuable trade secrets (Mattioli 2014).

Although personal observational data is possibly the most difficult data asset to assign provenance and to control, even the most generic data exhibits challenges. For example, generic data such as disaggregated weather-station readings can be used for a casual forecast but they can also be combined with demographic, health, economic and social information in different contexts. Weather may affect health outcomes (Maccini and Yang 2008), financial performance (Loughran and Shultz 2004), happiness (Levinson 2012), the chances to find a taxi (Faber 2014) or even violent outbreaks (Cohn 1993). Weather data may thus be analyzed and combined with other data in an infinite number of ways, completely disguising its provenance in the process. After data merge and analytics the various original data sources and observations have been perfectly "fused" and are thus indistinguishable (President's Council of Advisors on Science and Technology 2014). Although data may also be controlled to some degree by using proprietary formats and other technical standards, these can be worked around.

There have been few, if any, institutional responses to the necessity for proving provenance, although there have been some recent calls to action for the development of "sector-specific and trans-sector standards for metadata, calibration, accuracy and timeliness to provide a firm and trusted foundation for data capture, trading and re-use" (Royal Academy of Engineering 2015:p 5). Encouragingly, there are technical efforts to design improved provenance, such as trust management mechanisms for monitoring data consumers' contractual compliance (Moiso and Minerva 2012; Noorian et al. 2014; Schlegel et al. 2014). However, at present much of the provenance of data is shallow, in the sense that data sellers claim provenance, and then once the data leaves their control, provenance is lost.

MARKETPLACE DESIGN PRINCIPLES AND MATCHING MODELS

Market design principles

Marketplaces match buyers and sellers to exchange goods under agreed terms of exchange. At its most basic, a marketplace needs to provide a clear ongoing benefit for everyone to take part instead of bypassing it and trading directly (bilaterally) with the other participants. To do so they need to offer low transaction costs and effective trading arrangements (pricing, contracting, and fulfillment) that support the continued trading and engagement of participants. The marketplace also needs to reassure participants of the stability of its matching algorithm in the sense of Gale and Shapley (1962)—there is never a seller and a buyer who would have mutually preferred to be matched to each other rather than to their assigned matches.

The market design literature (Roth 2002; 2008) identifies several requirements that are associated with efficient market operation.[4] Firstly, an efficient market needs to provide "thickness" (liquidity) so that both buyers and sellers have opportunities to trade with a wide range of potential partners. Put differently, a market is "thick" when there is a sufficient pool of market participants willing to transact with one another. Without a critical mass of participants, positive externalities cannot have an impact and the market will not grow.

[4] The markets discussed in Roth (2008) cannot experience platform competition – schools, doctors' residencies, etc. cannot be managed by more than one clearinghouse. However, we do not feel this distinction influences our use of his model here, nor has this limitation restricted its application to other markets that can be managed by more than one clearinghouse, such as that for technology (Gans & Stern, 2010).

Secondly, while thickness is a necessary precondition for an efficient market, popularity can also create “congestion” by slowing down transaction time and thus limiting participants’ alternatives. As such, an efficient market requires rapid transactions to ensure market clearing, but not too rapid so that individuals, when considering an offer, do not have an opportunity to evaluate alternatives. Depending on the technological choices related to matching, payment, and fulfillment, congestion may be a nonissue or a significant source of friction.

Thirdly, the market needs to be perceived as “safe”. Safe markets are those where participants do not have incentives to misrepresent or undertake strategic action that might reduce efficiency. In the case of data marketplaces, the platform must be able to preclude behavior that influences the actions or preferences of other participants. For example, it would be important to exclude buyers from colluding or sharing data and sellers from making side contracts with buyers or other sellers, or trade outside the market altogether. Furthermore, the marketplace should provide provenance information: if a buyer is unable to assess the provenance of a data good, asymmetric information results as the seller will know more about the quality of the goods than the buyer. This information asymmetry may be complicated further when seller is not in full possession of all the provenance information.

Finally, the marketplace must not be “repugnant”, in that there are social norms or legal restrictions that limit the use of pricing to act as an allocation mechanism. Put differently, effective automated matching algorithms may be insufficient if rules, policies, norms and cultural expectations beyond those codified within the marketplace affect the attractiveness of the market itself (North 1990; Roth 2008). In the case of data marketplaces, the privacy implications of trading data potentially limit their development. Not only is there is increasing public interest in the societal impacts of big data, privacy and data trading,[5] in addition, there is increasing regulatory interest in the transparency and quantity of the personal data that has been amassed and is being traded (Ramirez et al. 2014).

Matching models for data markets

Conceptually, the matching of buyers to sellers for data is no different to any other type of markets. Gans and Stern (2010), applying the market design approach of Roth (2002; 2008) to markets of ideas or technology, suggested that effective market design might be possible for some innovation markets. However, they warn that the non-rivalry of the goods, requirement for complementary assets, and the weak intellectual property rights undermine the spontaneous and uncoordinated evolution of a market for ideas or technology. In particular, they note that when intellectual property rights are weak, the conditions for market thickness and market safety may not be met.

Table 1 classifies the main exchange matching mechanisms by the number of parties on each side and presents some examples of actual data marketplaces utilizing these designs. We next assess how each of these matching models addresses the market design issues reviewed above.

Table 1 – Types of marketplaces by matching mechanism

Matching	Marketplace design	Terms of Exchange	Examples
One-to-one	Bilateral	Negotiated	Data brokers; Acxiom
One-to-many	Dispersal	Standardized	Twitter API
Many-to-one	Harvest	Implicit Barter	Google Services
Many-to-many	Multilateral	Standardized or negotiated	Microsoft Azure Marketplace

[5] See for instance: Amnesty Global Insights, 27/02/17, "Why build a Muslim registry when you can buy it?"; www.medium.com/amnesty-insights/data-brokers-data-analytics-muslim-registries-human-rights-73cd5232ed19#.toi4vrsrm; accessed 04/03/17. Helbing et al, 2017, "Will Democracy Survive Big Data and Artificial Intelligence?", www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/; accessed 04/03/17.

To begin, one seller can trade simultaneously with one or more buyers. One-to-one matching is a bilateral relationship that involves two parties and is typically characterized by negotiated terms of exchange. Examples of bilateral traders include the personal or industrial data vendors and brokers, such as Acxiom. Markets based on bilateral trading relationships can be rather inefficient. Thickness (liquidity) can be a problem because it is difficult to locate trading partners when transactions are secretive, although this may also limit strategic behavior of participants (safety and clarity related to provenance and excludability due to more comprehensive contracts and their enforcement through monitoring). Congestion is unlikely to be a concern, but transaction costs will be high due to costs of search, negotiation, and ongoing relationship management. Furthermore, as bilateral markets are often opaque, there are potentially greater issues with repugnance, as typified by the recent interest by the US Federal Trade Commission (Ramirez et al. 2014).

When a single seller transacts with many buyers for the same data, using a one-to-many matching, standardized terms of exchange usually apply as it can be prohibitively costly to individually negotiate each exchange relationship. We call this a dispersal marketplace, and there are many examples of such markets, including most data distributed through APIs. Market thickness may require marketing and branding efforts, but fulfillment can be automated, reducing congestion and transaction costs. However, API-based automated trading without relationship monitoring (such as contract enforcement and auditing) is unlikely to address strategic behavior by buyers. Buyers may thus use the data in ways that reduce their value for other buyers. Automated standard contracts may also fail to comprehensively describe the sources and quality of the data, hence weakening provenance. Nevertheless, given the open nature of this marketplace, it is unlikely it will face repugnance concerns.

In a marketplace with more than one seller, transactions can be performed in a many-to-one or many-to-many fashion. This means that one or more sellers can trade simultaneously with one or more buyers. Many-to-one marketplaces are characterized by the harvesting of data, where users make their data available to a single service provider, under terms of exchange that resemble barter: The user receives access to a "free" service in exchange for their data; an example of this is Google Search where users provide search terms (data) in return for the search results. Google has effectively harvested the search behavior data that it uses for other purposes. Online social networks have similar harvesting arrangements.

The thickness of harvest markets depends on the popularity of the adjacent market for bartered "free" services. If the services are highly desirable, then there will be liquidity in the data market, too. However, the only types of data available are those related to the activities provided in the adjacent market. Meanwhile, congestion and transaction costs of harvesting can be very low, because there is no need for individual search, negotiation, or relationship management. Transaction costs may quickly balloon, however, if data harvesting runs afoul of repugnance concerns such as norms related to privacy. For example, the General Data Protection Directive of the EU gives users a "right to be forgotten", and should this become popular to exercise, it might be very costly to online service providers aiming to monetize user data. Moreover, strategic behavior can also be a concern, such as in the cases where users attempt to manipulate the search engine results by feeding biased data into the harvesting process. Furthermore, as the imposition of the "right to be forgotten" directive suggests, there are increasing levels of repugnance to these types of markets. The European data protection directive also stipulates a right for consumers to port their data from one service to another, thus increasing transaction costs and weakening excludability for service providers. Excludability is also weak for data providers (service users) because, with the allencompassing standard terms and conditions of most online services,[6] the user retains little control over subsequent utilization and commercialization of their data. Provenance and excludability are thus likely to be compromised in many-to-one markets.

Multilateral or many-to-many marketplaces are trading platforms upon which anybody (or at least a large number of registered users) can upload and maintain datasets, and where access to and use of the data are regulated through varying licensing models, either standardized or negotiated (Schomm et al. 2013). In principle, many-to-many marketplaces may enable the sourcing and integration of multiple disparate datasets, their taxonomic tagging for easy discovery, harmonization of formats, and subsequent aggregation and combination. To do so, many-to-many marketplaces require a regulatory environment, communication standards, data protocols, and procedures for data import, storage, transformation, aggregation, analysis, and delivery. Multilateral markets may provide several desirable features over other matching models, and, thus, they are technical systems that allow the participants to interact, potentially enabling economies of scale and scope, innovation, transaction and search (Thomas et al. 2014; Tiwana et al. 2010).

In its generic form, a many-to-many data marketplace is a two-sided market where sellers (data holders) and buyers exchange data products (Hagiu 2006; Parker and Van Alstyne 2005).[7] Unlike traditional market intermediaries, two-sided markets usually do not take ownership of the goods, instead alleviating (and profiting from) bottlenecks by facilitating transactions (Hagiu 2006; Hagiu and Yoffie 2009). Facilitation activities include services such as search/discovery, transaction validation, transaction history, and payment. A typical example is the eBay auction service, where sellers can list their goods for sale and

[6] E.g., see Facebook terms of service: https://www.facebook.com/terms

[7] There could be alternative types of transactions in cases where the traded dataset is auction based pricing or the trades are performed through machine learning (high frequency) trading. In the first case market participants will be strategic about their prices and in the second the price volatility of the commodity may create incentives for participants to bypass the clearinghouse (e.g. because of added delays in processing).

close the transaction through the eBay platform. Such digital platforms generate value for buyers and sellers through enhanced market efficiency, such as transaction volume, resource allocation efficiency, and an improved correspondence between suppliers and buyers (Eisenmann et al. 2006; Thomas et al. 2014).

Multisided platform theories (Bolt and Tieman 2008; Rochet and Tirole 2006; Weyl 2009) appear to have straightforward implications for the structure and pricing of multilateral data marketplaces. Data platform owners can utilize pricing strategies to optimize participation and achieve profitability by internalizing the bulk of the network externalities. In practice, however, achieving market thickness is in many cases very challenging (see Carnelley et al. 2016; Markl 2014 for some examples of failed data platforms). Furthermore, whereas digital technologies can mitigate direct transaction costs and achieve stable matching, strategic behavior may present market design and governance problems for multilateral data platforms (Tiwana et al. 2010; Wareham et al. 2014). In particular, suppliers of data may not reveal the origins and quality of the data and adverse selection issues may appear with poor quality data flooding the market (Holmstrom and Weiss 1985). This concern echoes the findings of Hagiu and Yoffie (2013) in the context of patent trading that the screening, listing fee and disclosure requirements to ensure that it is not only poor quality patents that are offered for sale reduces the efficiency of the market (Dushnitsky and Klueter 2011). On the other side, buyers of data may not respect the use and sharing restrictions and may degrade the value, privacy, and security of the data. Designing technical or contractual systems that incentivize and enforce appropriate behavior of the participants on a multilateral platform may be difficult if not impossible.

Table 2 Design principles for data markets

Matching	Marketplace design	Liquidity	Transaction costs	Provenance	Excludability
One-to-one	Bilateral	Low	High	Clear	Medium
One-to-many	Dispersal	High	Low	Unclear	Low

Many-to-one	Harvest	High	Low	Unclear	Low
Many-to-many	Multilateral Centralized	High	Low	Medium	Low

Table 2 summarizes the foregoing discussion of Roth's market designs principles for data. The bilateral design is likely to suffer from low liquidity but the other three designs are expected, in principle, to be able to achieve sufficient thickness. Bilateral markets also stand out in terms of their high transaction costs, but in return, they may provide greater provenance and excludability of undesirable trades, thus relatively limited strategic behavior of participants. In other words, with the currently available market mechanisms, it is possible to achieve large markets with little control or small markets with greater control.

MULTILATERAL MARKETPLACE DESIGN

Data as a common-pool resource

In this section, we focus on the capacity of emerging multilateral data marketplace models to address the critical need to limit strategic behavior with provenance and excludability presented in the previous section. Given the challenges of incomplete intellectual property protection, a complex and evolving regulatory environment around privacy and security, and the need for metadata, including provenance, as discussed in the second section, the operation, transaction, and dispute resolution of a multilateral data marketplace appears tricky. Here we build upon Ostrom's (1990) seminal contributions in the analysis of collective action in resolving the "tragedy of the commons" and assume that data can be viewed as a common-pool resource due to the limited control rights associated with it (Gil and Baldwin 2014; Hess and Ostrom 2003; Koutroumpis and Leiponen 2013). Using Ostrom's collective governance framework, and integrating this with Roth (ibid.), we shed additional light on the feasible design options for multilateral data markets.

The characteristics of a common-pool resource make it costly, but not impossible, to exclude potential beneficiaries from obtaining benefits from its use. According to Ostrom

(1990), collective action can overcome the tragedy of the commons and maintain a shared resource where a set of institutional requirements are met: boundaries, rules, metarules about changing the rules, and monitoring. Data can be conceptualized as a common-pool resource if excessive sharing depreciates its value for everyone. For example, valuable data can be shared within a community of users, but if some users extensively share the resource with outsiders, its distinctive market value may be diminished or lost, and thus the incentives to create and develop data in the first place are diminished. As a result, too little data are made available in the economy. In this section, we first examine Ostrom's original (1990) framework in the context of data marketplaces, and then assess how the requirements for managing data as a common-pool resource relate to the market design requirements and data market challenges.

Boundaries. Data marketplaces require clearly defined boundaries that enable the identification of a legitimate user. Marketplace boundaries are related to the thickness (liquidity) of the marketplace, as the stricter the boundary conditions, the fewer the number of participants within the marketplace. In principle, controlled access to a marketplace for both buyers and sellers should reassure participants regarding the origins of the traded goods (sellers) and the legal standing and reputation of prospective users (buyers). Clear boundaries are particularly important when considering the complexity of the regulatory environment, as being able to identify data users is often fundamental to data protection regimes. As long as every data transfer goes through the marketplace and the market participants are verified, it is up to the matching and enforcement mechanisms to steer the market towards equilibrium. However, in practice, it may be difficult to ensure all users only trade over the platform, rather than privately or even with unauthorized users outside of the platform.

Rules and their modification. Data marketplaces require rules that define how the resources are to be used (both on platform and off-platform) and the penalties for not doing so. However, these rules need to be such that they do not lead to congestion within the marketplace, reducing the efficiency of the market, as well as ensure safety and counter any repugnance concerns. Rules are vital to ensure the contractual or regulatory requirements for privacy, as data dissemination must be appropriate for the context and obey the governing norms of distribution within it (Nissenbaum 2004). These rules are for the benefit of both the data sellers and the platform itself. Moreover, there can be procedures for ex-post modifications of the rules in the data marketplace through "collective-choice arrangements" whereby users can take part and vote for changes.

Monitoring. Data marketplaces need effective monitoring by a group of core users or a neutral third party accountable to the core users (Ostrom 1990). The marketplace can also institute automated mechanisms that monitor the use of the data and flag any suspicious activity. Such monitoring mechanisms should (temporarily) be able to halt trading by suspected infringers and implement sanctions that limit the damage from illegal release of records. Misconduct can limit trust on the marketplace and curtail trading activities. Monitoring thus needs to be sufficient to enable the safe operation of the marketplace, but not so invasive as to cause congestion and reduce the operational efficiency of the market. Furthermore, the data protection and security mechanisms need to align with the nature of the traded assets. Indeed, privacy regulations such as those in Europe often require constant monitoring to ensure that data are not misused. Therefore, monitoring may be provided by a trusted intermediary who sponsors and operates the data platform (Eisenmann et al. 2009). For example, when fiat currency bank notes are exchanged, the value of the notes depends on the bank's signature; otherwise the notes are worthless. A similar approach has been used in crypto-currencies (like Bitcoin) where the "value" of an original quantum of the currency depends on the result of the hash function. Reputation-based systems could undertake this role (Dellarocas 2005; Moreno and Terwiesch 2014; Pavlou and Gefen 2004).

Data marketplace designs and the management of common-pool resources

Table 3 describes the different data marketplace designs with respect to their ability to address the governance problems related to common-pool resources. We now introduce three specific designs that have recently been considered for the multilateral marketplace: centralized platform, decentralized platform based on a Distributed Ledger Technology (DLT, see Catalini and Gans, 2016), and collective platform. The last column concludes with our assessment regarding the type of data that can feasibly be traded on each type of marketplace.

Table 3 – Marketplace and data typology

Matching	Marketplace design	Boundaries	Rules	Monitoring	Characteristics of data
One-to-one	Bilateral	Clear	Strong	Invasive	High value, High confidentiality
One-to-many	Dispersal	Unclear	Weak	Minimal	Low value, Low confidentiality
Many-to-one	Harvest	Unclear	Weak	Minimal	Low value, Low confidentiality
Many-to-many	Multilateral Centralized	Medium	Medium	Medium	Medium value, Medium confidentiality
Many-to-many	Multilateral Decentralized	Unnecessary	Strong	Effective	High value, High confidentiality
Many-to-many	Multilateral Collective	Strong	Strong	Effective	High value, High confidentiality, Small market

One-to-one marketplaces, or bilateral marketplace designs, have clear boundaries, as only parties to the contract are participants. There is thus high seller control over the rules for the use of the data, as well as substantial power through the contract for monitoring use. However, per Table 2, bilateral markets also have high transaction costs and limited liquidity, as much effort is expended to search for trading partners, design and agree on the final contract, and monitor its execution. It is for these reasons that bilateral trading can preserve the common-pool resource but is most likely used for high value and highly confidential data.

In contrast, one-to-many and many-to-one marketplaces, i.e., dispersal and harvest marketplaces, have weak boundaries, as the participants often give little identifying information to participate in the marketplace. As such, the rules that define data use are broad and general, and monitoring is minimal or, if required, not very effective, resulting in low control. However, per Table 2, transaction costs are also low and the transacting is almost fully automated. Therefore, the harvest and dispersal marketplace models are expected to be used for low value, low confidentiality data. This is not to say that it is not private data that is shared (particularly in the harvest models) but that the small quantities that are traded (or bartered) are perceived to be insignificant enough.

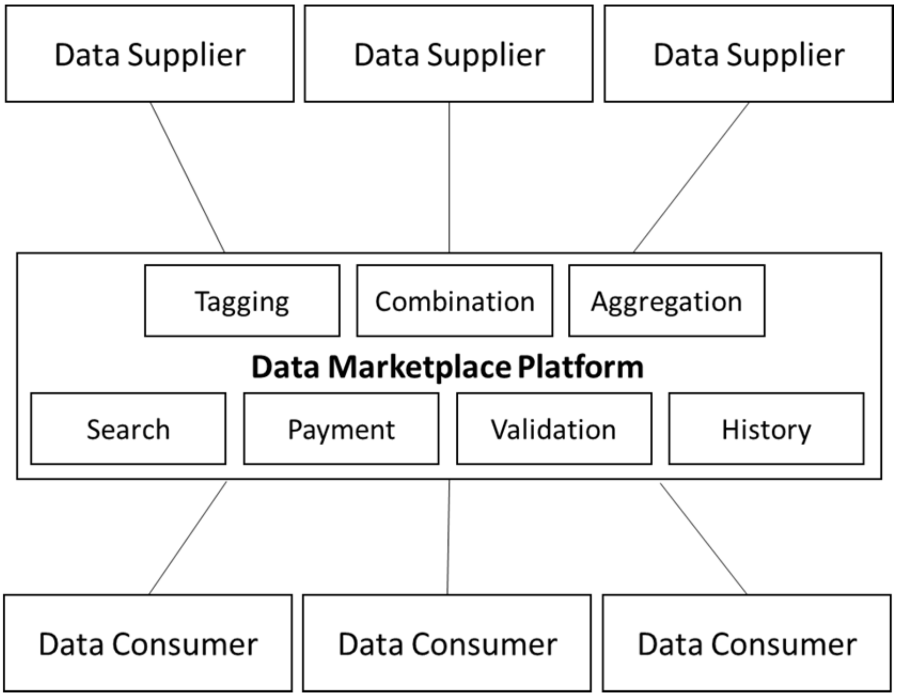
Regarding many-to-many marketplaces, we suggest that there are three distinct emerging data-trading frameworks: a centralized platform, a decentralized platform, and a consortium-like collective governance model.

Centralized multilateral marketplace

The generic multilateral marketplace design is the centralized marketplace, reflecting a "standard" multi-sided platform. In the centralized design the platform operator attracts participants from across the data ecosystem, including data creators, managers, analysts, service providers, and aggregators (Thomas and Leiponen 2016). These participants take advantage of the facilitation services—search/discovery, transaction validation, transaction history, and payment—delivered through the technological platform. In doing so the platform operator benefits from positive participation externalities among the participants, assuming that transaction costs prevent trading parties from trading outside of the platform. This means that the centralized design can only work as long as its facilitation services result in it being preferable to users compared to repeated bilateral exchanges in terms of cost, reach of suppliers, speed, and so on. This design is illustrated in Figure 1.

A centralized design favors larger data suppliers over smaller ones as the cost of exchanging and processing required to taxonomically tag and aggregate diminishes (per data quantum) with scale. The breadth of options and spectrum of uses that every possible raw data record may fit in makes the job of creating and maintaining the taxonomy (for search and purchasing purposes) challenging. Ideally, the marketplace should have some meta-information (ex-ante) for any possible data record that arrives in the marketplace to allow for future trades of the underlying information. Apart from the costly maintenance of this platform the distribution of exchange volumes (and profits) will be heavily skewed towards the upper end (high volumes of data for a fixed cost of meta-information). Furthermore, due to the scale economies and network effects, it is conceivable that there would be winner-take-all dynamics, meaning that only one or perhaps two data platforms would emerge for specific classes of data.

Figure 1 – Centralized marketplace design



The ability of a centralized platform to set formal entrance policies and fees is expected to result in a relatively strong boundary (see Table 3). However, the platform owner will need to balance the economies of scale against the risks of strategic behavior that grow with size. We further assess that the rules are medium strong as the platform owner may unilaterally define the terms and conditions, such as licensing terms, and then execute these.

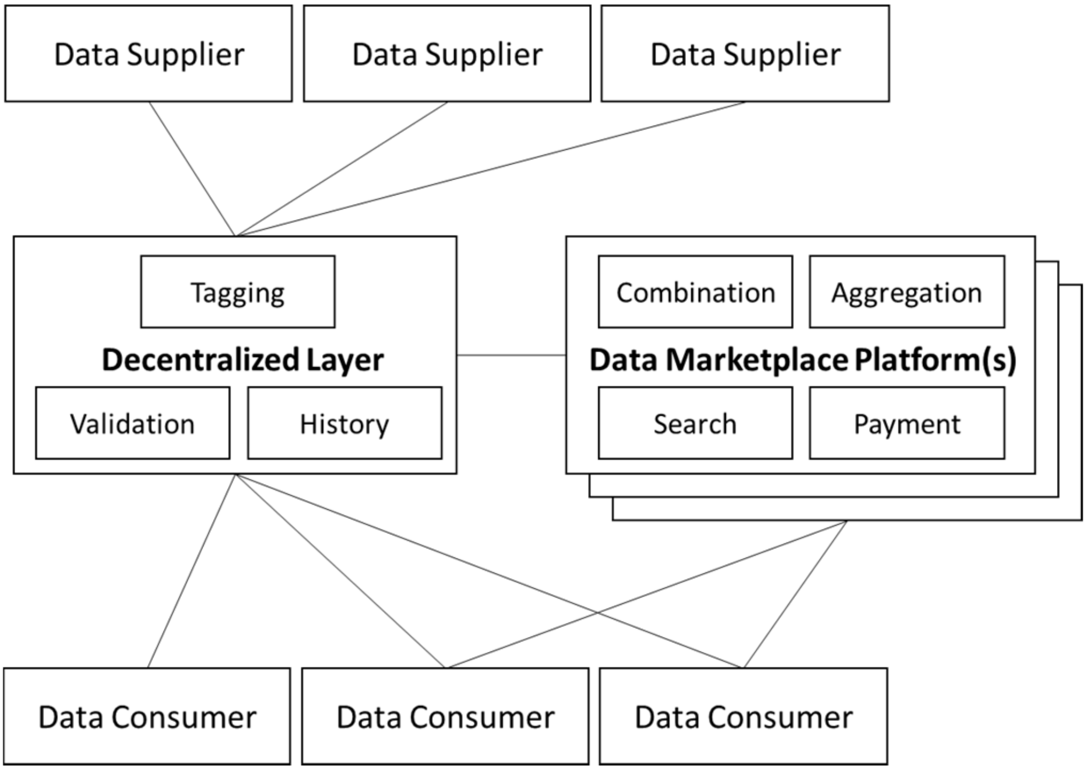
We expect limited recourse for platform users to renegotiate them. Finally, we expect a medium level of monitoring, as this can be quite strict within the platform environment but virtually nonexistent if the data can be removed from the platform. In summary, a centralized data platform appears only weakly able to manage data as a common-pool resource, because strong boundaries and effective monitoring/enforcement require a limited size and extensive restrictions on data use, which go against the fundamental economics of digital platforms and the utility from data integration and further analyses.

Without such use restrictions, enforcement, and strong boundaries, the origins of any downstream data transactions cannot be traced back to the original suppliers: outside the platform the provenance is lost. Hence, any future exchange will not consider the original content owners' preferences. A centralized design without boundaries and enforcement is therefore most feasible for non-private data, and potentially time-sensitive data. If the data are private, then in the future they might be used in ways that harm their originator after the exchange has taken place (identity theft, commercial profiling, etc.). If the data are not time sensitive, their value would drop rapidly as users share the data or the analyzed information thereof. This inability to address the privacy and value preservation requirements of personal data providers in the centralized design represents a market inefficiency: The unrestrictive centralized model sacrifices seller control – traceability – in return for lower transaction costs.

Decentralized multilateral data marketplace

Decentralized marketplace designs that use *distributed ledger technologies* (DLTs) have recently been proposed (Catalini and Gans 2016; Evans 2014). DLTs have initially been adopted in virtual currency markets, but they are subsequently being developed for a variety of digital markets. The decentralized design is described in Figure 2.

Figure 2 – Decentralized marketplace design



A decentralized marketplace facilitated by a DLT shares many basic attributes with the centralized marketplace but alleviates some of its limitations. Most importantly, DLTs may enable trades to be directly executed and verified by market participants. Furthermore, the (subsequent) proof of provenance is now decentralized and can be independently verified. In the decentralized design, the modes of exchange range from simple bilateral transactions to complex multilateral exchanges. Thus, what we previously called centralized marketplaces are now simple communication structures that facilitate the operation of the decentralized market. Information about the timing, quantity and value of transactions, the trustworthiness of a participant, and the type of data available for exchange are now decentralized and not executed by a central structure. The validation of each transaction rests upon a public ledger available to all users tracing back the history of all exchanges (Swan 2015; Walport 2016). Utilizing a DLT that reports all data transactions and a unique identification for every data block, the decentralized framework guarantees a complete and transparent history of every transaction. This process (resembling the blockchain paradigm) matches owners to their data points and precludes unauthorized transactions.

In this context, both the enforcement of the trading terms and the resolution of disputes can be transparently accommodated. The terms of exchange may now be extended to include features such as a timestamp for deletion (reminiscent of the "right to be forgotten"), the limitation of further trades (irrespective of the temporal restrictions), and mechanisms of data aggregation. Furthermore, given that the validation, history, and taxonomic tagging are independent of the central platform, we anticipate substantial competition between the trading platforms, with competition revolving around pricing, search/discovery, and aggregation capabilities.

The enhanced market clearing and transparency in the decentralized design means that the effort and cost of taxonomic tagging likely shifts from the centralized marketplace to each data supplier. The data provider must maintain a data taxonomy that annotates every record with a set of common standards that can be searched, aggregated, analyzed and resold (including the type of data, temporal granularity, accuracy, breadth of coverage, etc.). The data taxonomy helps characterize individual data blocks and allows for automated aggregation of complex data products. Taxonomies have been evolving over the past decade, e.g., under the umbrella of Open Linked Data project by W3C.[8,9] As the universal annotation mechanism allows for comparisons among products, the decentralized design tackles the other major challenge: traceability.

The decentralized market does not necessarily feature any specific boundaries as anyone with data can potentially register their data into the decentralized layer. Each hop of

[8] See for examples: http://www.w3.org/standards/semanticweb/data and http://linkeddata.org/

[9] In a recent pilot a universal vocabulary of Internet of Things has been created http://www.prnewswire.com/news-releases/rwe-holds-live-demo-of-lemonbeat-for-international-experts-555639401.html; retrieved 1st February, 2016.

the transaction path can be traced through the publicly available list. In this context, by "handshaking" each transaction to the individuals' data portfolio, the provenance of each data element can be monitored. These exchanges may or may not carry a cost but the process safeguards owners that their data will always be traceable. An authentication mechanism may also be introduced in each exchange that controls whether content owners accept the terms of each exchange. As long as all transactions are traceable and authenticated, content owners are assured that their data will not be used without them explicitly consenting.

In a decentralized platform, all data would be disseminated using the protocol whose standards and terms the receiving ends must accept. This means that each receiving end will agree to update the distributed ledger and notify about the steps in the data value chain. Some entities may choose to bypass this system – and there can be incentives to do so – and exchange data in the traditional (usually bilateral) fashion. However, as long as there are no specific sacrifices in service delivery, sellers (individuals) will be better off in the decentralized context and thus will help increase market thickness. The increase in transparency and control over the traded information are expected to raise content owners' welfare by controlling information reuse and thus improving individuals' privacy.

Put differently, the decentralized DLT-based marketplace would either address or defuse all conditions of common-pool resources: no boundaries would be required because of the distributed ledger; rules would be defined (and set in stone) in the shared dissemination protocol; and monitoring would be perfect because of the verifiable ledger. Arguably, a well-functioning DLT-based decentralized market would make data a private and excludable good. This would address the provenance issues as well as meet the market design principles of Roth. In principle, a DLT-based data marketplace would function as efficiently as any marketplace for rival and excludable goods, as long as the technology functions as expected and the participants find it beneficial to trade on the platform rather than circumvent it.

To summarize, the decentralized design based on DLT addresses three main issues. Firstly, it reduces the impediment to trading posed by the privacy of data as trades can be found in the public ledger and enforced via authentication of each transaction. Secondly, it differentiates between the holder of a data asset and the original entity that created the data assets. As part of its design it grants the original entity the right to control future access and use if they so desire. Lastly, it allows for everyone's participation in an open and transparent exchange as all data points can be tagged by the data taxonomy classes and searched based on their unique characteristics. In a sense, a decentralized market design with a verification technology privatizes the data assets. Data associated with a transparent chain of provenance and enforceable usage restrictions is no longer a common pool resource (Ostrom 1990), and its benefits can now be appropriated by individuals. These fundamental principles guarantee data reuse while respecting the dual—both intermediate and final—nature of record data (for a review of the economic characteristics of data goods, see Koutroumpis et al. 2016). However, the scalability of DLT systems remains unclear, and therefore the large-scale implementation of decentralized data marketplaces is uncertain.[10]

Collective multilateral platform

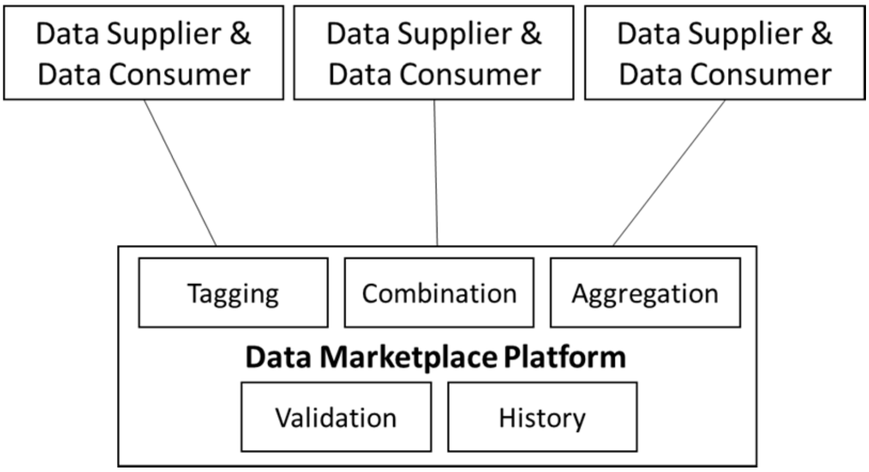
Although the benefits of a decentralized multilateral data market are undeniable, the technical solutions for DLT-enabled data markets are some years into the future. A more modest multilateral market platform can be set up using Ostrom's collective governance principles. A multilateral platform that adopts strong boundaries via complex contracts, clear rules such as bylaws and procedures to collectively change them, and effective monitoring and enforcement practices, could conceivably enable multilateral trading of data and overcome

[10] See e.g. http://hackingdistributed.com/2016/08/04/byzcoin/ and https://www.technologyreview.com/s/600781/technical-roadblock-might-shatter-bitcoin-dreams/

the common-pool resource challenges. Early examples of such data pools or data consortia can be identified.[11] See Figure 3 for a schematic overview.

However, a collective platform is expected to run into other market design issues, per Roth's framework. Whereas it may address some of the strategic behavior problems, a data consortium is expected to be associated with high transaction costs due to the need to vet partners well and write complex and comprehensive contracts. Effective monitoring is also costly, and hence the size of the collective may be limited. Therefore, market liquidity will be an issue. Thus, a small-scale data platform with many use restrictions is not likely to offer a strong value proposition or high efficiency. However, for an industry vertical or a stable consortium of innovation partners with pre-existing trust-based relationships and clear and significant shared interests might find such a small and restrictive multilateral platform a value-adding structure, even for high-value and highly confidential data.

Figure 3 – Collective multilateral design


Data Supplier & Data Consumer
Data Supplier & Data Consumer
Data Supplier & Data Consumer
Tagging
Combination
Aggregation
Data Marketplace Platform
Validation
History

[11] SSAB Smart Steel consortium: https://www.ssab.com/globaldata/news-center/2016/12/08/08/30/ssab-towards-the-internet-of-materials-with-ssab-smartsteel
ABB and Konecranes building the industrial internet campus: http://www.aalto.fi/en/about/for_media/press_releases/2016-04-07/
Jakamo solution to share data across the supply chain: http://jakamo.net/

RESEARCH IMPLICATIONS

We have demonstrated that markets for data require the establishment of rigorous provenance, as expressed through verifiable metadata, for the data goods being sold, because the protection, or control rights, regime is not sufficiently robust. We also characterized the main data market matching mechanisms and present illustrative examples of actual data marketplaces utilizing these designs. We have also argued there are three potential multilateral market designs, but only two of which, the decentralized and collective models, may be feasible in the future. The above analysis suggests two main avenues for future research for information systems scholars. The first research direction considers the conceptual and technical development of decentralized platforms for data. The second future research direction addresses the conceptual and technical development of data contract management.

Decentralized platform governance

An important direction for future research is the architecture of platforms where the provenance functionality is provided through a decentralized registry. As Figure 2 presents, an effective decentralized data marketplace consists of two layers – a decentralized layer that handles the tagging, validation and history functions required for provenance, and multiple platforms for data trading. This separation of functionality has not been considered by extant IS research into platform functionality and governance (see for instance Tiwana et al. 2010; Wareham et al. 2014). Future research could consider how multiple trading platforms would interface with the DLT and with each other. How would the costs of operation of the DLT layer be allocated? How transparent does the distributed registry need to be to both suppliers, consumers and data marketplaces? In the event of conflict, how would dispute resolution be handled? How would multiple DLTs be handled? What would the standardization of cross-platform data records and the automation of the data gathering process look like?

Beyond the governance challenges of integrating a distributed registry into platform governance, regulatory oversight of the distributed registry also needs to be considered. For instance, technical failures or congestion in the public ledger might lead to excessive times for transactions to be cleared. Even without unusual technical problems, the scalability of DLTs is currently hotly debated. Although there is much IS research into increasing the scalability of DLTs (for instance, see Croman et al. 2016; Eyal et al. 2016), congestion may be the greatest obstacle to an efficient decentralized data market. For IS scholars there are research opportunities in considering the architecture of the decentralized registry itself. At one extreme, the mechanism could be driven by a fully decentralized trust network, where the DLT is not controlled by any one centralized entity. IS scholars have already begun to consider such approaches, with Zyskind et al. (2015) investigating the use of blockchain to secure personal data. How could such approaches be extended to other types of data, such as those gathered by sensors? Such research could leverage emerging thoughts into how blockchain can be adopted for the internet of things (Bhaga and Madisetti 2016; Huh et al. 2017).

Alternatively, what different types of market architectures are possible with the decentralized layer? Could the development of multilateral data markets be undertaken as a closed community of trusted participants (as some emerging financial service applications of blockchain have proposed), adopting DLT solutions in collective governance arrangements?[12] Whereas this approach may not achieve a large scale, it could provide effective management of the common-pool resource with strict boundaries, rules, and monitoring. For industrial internet development where a group of known and stable industrial

[12] Major banks are working together to develop a decentralized platform for clearing transactions (Financial Times, Aug 2016 http://www.ft.com/cms/s/0/1a962c16-6952-11e6-ae5b-a7cc5dd5a28c.html); retrieved 03/09/2016.

partners' shares data over a closed network, the collective approach may be more viable. This may be one of the most fruitful approaches for IS scholars in the short term.

Another solution is to establish the data ledger (data control and traceability functions) with a centralized "trust" platform operator, providing a centralized service for the other data marketplaces. In many ways, this would be akin to the role of the central bank in issuing money within an economy. Indeed, this notion has previously been proposed as a "Bank of Individuals' Data" where a centrally organized "personal data management service" enables consumers to exploit their personal data through the provision of secure and trusted space (Moiso and Minerva 2012). What mechanisms would be required so that the other participants trust this centralized registry?

A second issue for IS scholars is the technical requirements and practicalities of linking individual data with the decentralized layer. Given the vast quantity, heterogeneity and speed with which data is being generated, a scalable mechanism that "tags" data within the decentralized governance function is required to ensure efficient and effective operation. What would such a protocol look like? How would an authentication mechanism that controls whether content owners accept the terms of each exchange function? Scholars could leverage existing applications of blockchain to data goods to begin to address these issues. Existing practical applications include proof of ownership for digital assets and content including images, arts and pictures (for example, such as those provided by MyPowers, Blockai, Bitproof, ascribe and Artplus). Other practical applications have included the authenticity of reviews or endorsements for employee peer review (The World Table, Asimov, TRST.im). Management of sensitive data such as electronic health records or business contracts in a safe and decentralized manner is also being addressed (BitHealth, Colored Coins), while other applications covering the proof of ownership in app development, content storage, and the Internet of Things are also emerging. More broadly, there is much for IS scholars to rigorously consider the more generalized application of blockchain governance to digital artifacts (Kallinikos et al. 2013).

Data contract management

An important technical component of the services provided by data marketplaces will be a data contract management service. The distributed layer provides a mechanism to record and enforce the contractual terms and conditions for an individual transaction involving data, which will include the intellectual property rights to data. However, it does not address the complex issue of combining different data contracts to create hybrid data products that are sold through the marketplace. Although a significant share of the data transacted through a marketplace will be original sourced data with (relatively) simple intellectual property protection and provenance data, in reality we expect much of the data will have been processed in some way to add value (Thomas and Leiponen 2016). At its most basic this will require update to provenance data, when the data is “cleaned” or individual proprietary data sources (non 3rd party) are combined to create a more comprehensive data product for sale. More likely however, much of the value adding may involve the combination of other 3rd party datasets, with their own terms and conditions of use as well as provenance data, which will need to be reflected into the subsequent terms and conditions. This will become particularly true as data products are registered within the decentralized marketplace, requiring their usage and provenance to be maintained.

To manage the integration of data contracts, as well as to manage the subsequent contractual requirement of such data products will require a data contract clearance service. The following example illustrates the requirements for the integration of data contracts. Supplier A provides data within an open (free) data contractual framework that requires all derivative data products to acknowledge the original data source; supplier B sells data under a contractual agreement that requires a once off fee for use and a share of any derivative revenues. Supplier C takes their own proprietary data and combines it with the data from supplier A and supplier B to offer data product C. In order for C to legally use the data of A and B, the data contractual terms of C need to ensure compliance with those contractual terms of A and B, clearly detail the provenance of each of the data sources, as well as specify their own contractual demands for sale. We can extend this example by adding supplier D who then purchases the data product from C, adds some of their own proprietary data, and then offers a subsequent data product for sale. Firstly, the sale from C to D will require a transfer of money to supplier B; furthermore, any sale of the data product D must now also result in a transfer of money to B.

As the above example demonstrates, a data contract engine will need to interrogate the contractual status (rights and provenance) of data product, provide information as to access and usage rights a data product, and create derivative data contracts of hybrid data products so that the rights and provenance of the hybrid data product accurately reflect the data contracts of the constituent data products. On the one hand, there will be requirements to homogenize the contract management, including the IP aspects and create a common framework, in order to enable efficiency. On the other hand, given the vast heterogeneity of data and data owner preferences, others will demand more customized rights and terms and conditions. Data rights and access are thus contentious issues that need to be carefully defined and adjudicated. It may be difficult to do this using a blanket agreement, and there may arise needs for more tailored contracts and protections. The ability to technically meet this data contracting management requirement will enable the data marketplace to scale, and will also have an important enabling role for economies of scope, for instance allowing analytic and other derivative products to be made available through the platform.

To date, however, there has been little IS research that has directly considered the need for such data contract management. An exception is Truong et al. (2011), who have proposed a service for composing, managing, analyzing data agreements in cloud environments and data marketplaces. Truong et al. (2012) have also considered an abstract model for data contracts that can be used to build different types of data contracts for specific types of data, and propose several techniques for evaluating data contracts. Furthermore, there are some tangential early patents that are beginning to consider technical implementations.[13] As such, there are several key questions that IS scholars need to begin considering, given the decentralized model suggested above. How should data contracts be abstracted, and more particularly, how should provenance be represented within a data contract? What algorithmic principles should underpin the combination of hybrid data contracts? What mechanisms can be implemented to ensure continuity of provenance and terms and conditions in successive data contracts? What type of architecture is best suited to this technical challenge?

REFERENCES

Akerlof, G. A. 1970. "The Market for "Lemons": Quality Encertainty and the Market Mechanism," *Quarterly Journal of Economics* (84:3), pp. 488-500.

Arrow, K. J. 1962. "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors,* Universities-National Bureau Committee for Economic Research and Committee on Economic Growth of the Social Science Research Council (eds.). Princeton, NJ: Princeton University Press, pp. 609-626.

Bakos, J. T. 1991. "A Strategic Analysis of Electronic Marketplaces," *MIS Quarterly* (15:3), pp. 295-310.

Bhaga, A., and Madisetti, V. K. 2016. "Blockchain Platform for Industrial Internet of Things," *Journal of Software Engineering and Applications* (9:10), pp. 533-546.

Bolt, W., and Tieman, A. 2008. "Heavily Skewed Pricing in Two-Sided Markets," *International Journal of Industrial Organization* (26), pp. 1250-1255.

Borgman, C. L. 2012. "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* (63:6), pp. 1059-1078.

Bresnahan, T. F., and Trajtenberg, M. 1995. "General Purpose Technologies: Engines of Growth?," *Journal of Econometrics* (65:1), pp. 83-108.

[13] US Patent Number US20160085544 A1, “Data management system”, which considers the improvement the management of data contracts during a software development lifecycle. US Patent Number US 20170053295 A1, “Platform data marketplace”, which considers the provision, aggregation and analysis of data through a marketplace with no reference to quality or IP control.

Carnelley, P., Schwenk, H., Cattaneo, G., Micheletti, G., and Osimo, D. 2016. "Europe's Data Marketplaces - Current Status and Future Perspectives," IDC.

Catalini, C., and Gans, J. S. 2016. "Some Simple Economics of the Blockchain," in: *NBER Working Paper*. Cambridge, MA: National Bureau of Economic Research, p. 30.

Chebli, O., Goodridge, P., and Haskel, J. 2015. "Measuring Activity in Big Data: New Estimates of Big Data Employment in the Uk Market Sector," in: *Imperial College Business School Discussion Paper*. London, UK: Imperial College Business School, p. 28.

Cohn, E. G. 1993. "The Prediction of Police Calls for Service: The Influence of Weather and Temporal Variables on Rape and Domestic Violence," *Journal of Environmental Psycholog* (13:1), pp. 71-83.

Croman, K., Decker, C., Eyal, I., Gencer, A., Juels, A., Kosba, A., Miller, A., Saxema, P., Shi, E., Sirer, E. G., Song, D., and Wattenhofer, R. 2016. "On Scaling Decentralized Blockchains," *International Conference on Financial Cryptography and Data Security,* J. Clark, S. Mieklejohn, P. Ryan, D. Wallach, M. Brenner and K. Rohloff (eds.), Barbados: Springer, pp. 106-125.

Dellarocas, C. 2005. "Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard," *Information Systems Research* (16:2), pp. 209-230.

Duch-Brown, N., Martens, B., and Mueller-Langer, F. 2017. "The Economics of Ownership, Access and Trade in Digital Data," European Commission, Sevilla, Spain.

Dushnitsky, G., and Klueter, T. 2011. "Is There an Ebay for Ideas? Insights from Online Knowledge Marketplaces," *European Management Review* (8:1), pp. 17-32.

Eisenmann, T. R., Parker, G., and Van Alstyne, M. W. 2006. "Strategies for Two-Sided Markets," *Harvard Business Review* (84:10), pp. 92-101.

Eisenmann, T. R., Parker, G., and Van Alstyne, M. W. 2009. "Opening Platforms: How, When and Why?," in *Platforms, Markets and Innovation,* A. Gawer (ed.). Cheltenham, UK: Edward Elgar, pp. 131-162.

Evans, D. S. 2014. "Economic Aspects of Bitcoin and Other Decentralized Public-Ledger Currency Platforms," in: *Coase-Sandor Institute for Law & Economics Working Paper*. Chicago, IL: Coase-Sandor Institute for Law and Economics.

Eyal, I., Gencer, A., Sirer, E. G., and van Renesse, R. 2016. "Bitcoin-Ng: A Scalable Blockchain Protocol," *13th USENIX Symposium on Networked Systems Design and Implementation*, Santa Clara, CA.

Faber, H. S. 2014. "Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers," w20604, National Bureau of Economic Research, Washington, DC.

Gale, D., and Shapley, L. S. 1962. "College Admissions and the Stability of Marriage," *The American Mathematical Monthly* (69:1), pp. 9-15.

Gans, J. S., and Stern, S. 2010. "Is There a Market for Ideas?," *Industrial & Corporate Change* (19:3), pp. 805-837.

Gefen, D., and Pavlou, P. A. 2012. "The Boundaries of Trust and Risk: The Quadratic Moderating Role of Institutional Structures," *Information Systems Research* (23:3-part-2), pp. 940-959.

Gil, N., and Baldwin, C. Y. 2014. "Sharing Design Rights: A Commons Approach for Developing Infrastructure." Cambridge, MA: Harvard Business School.

Gopalkrishnan, V., Steier, D., Lewis, H., Guszcza, J., and Lucker, J. 2013. "Big Data 2.0: New Business Strategies from Big Data," in: *Deloitte Review*. pp. 54-69.

Hagiu, A. 2006. "Pricing and Commitment by Two-Sided Platforms," *RAND Journal of Economics* (37:3), pp. 720-737.

Hagiu, A., and Yoffie, D. B. 2009. "What's Your Google Strategy?," *Harvard Business Review* (87:4), pp. 74-81.

Hagiu, A., and Yoffie, D. B. 2013. "The New Patent Intermediaries: Platforms, Defensive Aggregators, and Super-Aggregators," *Journal of Economic Perspectives* (27:1), pp. 45-65.

Hess, C., and Ostrom, E. 2003. "Ideas, Artifacts, and Facilities: Information as a Common-Pool Resource," *Law and Contemporary Problems* (6:1&2), pp. 111-145.

Holmstrom, B., and Weiss, L. 1985. "Managerial Incentives, Investment and Aggregate Implications: Scale Effects," *Review of Economic Studies* (52:3), pp. 403-425.

Holt, T. J., and Lampke, E. 2010. "Exploring Stolen Data Markets Online: Products and Market Forces," *Criminal Justice Studies* (23:1), pp. 33-50.

Huh, S., Cho, S., and Kim, S. 2017. "Managing Iot Devices Using Blockchain Platform," in: *19th International Conference on Advanced Communication Technology*. Pyeongchange, South Korea.

Kallinikos, J., Aaltonen, A., and Marton, A. 2013. "The Ambivalent Ontology of Digital Artifacts," *MIS Quarterly* (37:2), pp. 357-370.

Koutroumpis, P., Leiponen, A., and Thomas, L. D. W. 2016. "The Simple Economics of Data Goods," in: *Innovation & Entrepreneurship Department Working Papers*. London, UK: Imperial College Business School.

Koutroumpis, P., and Leiponen, A. E. 2013. "Understanding the Value of (Big) Data," *2013 IEEE International Conference on Big Data*, pp. 38-42.

Levinson, A. 2012. "Valuing Public Goods Using Happiness Data: The Case of Air Quality," *Journal of Public Economics* (96:9), pp. 869-880.

Loughran, T., and Shultz, P. 2004. "Weather, Stock Returns, and the Impact of Localized Trading Behavior," *Journal of Financial and Quantitative Analysis* (39:2), pp. 343-364.

Maccini, S. L., and Yang, D. 2008. "Under the Weather: Health, Schooling, and Economic Consequences of Early-Life Rainfall," w14031, National Bureau of Economic Research, Washington, DC.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. 2011. "Big Data: The Next Frontier for Innovation, Competition, and Productivity," pp. 1-156.

Markl, V. 2014. "Project Final Report: Data Supply Chains for Pools, Services and Analytics in Economics and Finance," TU Berlin, Berlin, Germany.

Mattioli, M. 2014. "Disclosing Big Data," *Minnesota Law Review* (99), pp. 534-584.

Mayer-Schonberger, V., and Cukier, K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London, UK: John Murrary (Publishers).

Menon, S., and Sarkar, S. 2016. "Privacy and Big Data: Scalable Approaches to Sanitize Large Transactional Databases for Sharing," *MIS Quarterly* (40:4), pp. 963-981.

Moiso, C., and Minerva, R. 2012. "Towards a User-Centric Personal Data Ecosystem the Role of the Bank of Individuals' Data," *2012 16th International Conference on Intelligence in Next Generation Networks, ICIN 2012, October 8, 2012 - October 11, 2012*, Berlin, Germany: IEEE Computer Society, pp. 202-209.

Moreno, A., and Terwiesch, C. 2014. "Doing Business with Strangers: Reputation in Online Service Marketplaces," *Information Systems Research* (25:4), pp. 865-886.

Nissenbaum, H. 2004. "Privacy as Contextual Integrity," *Washington Law Review* (79), pp. 101-139.

Noorian, Z., Iyilade, J., Mohkami, M., and Vassileva, J. 2014. "Trust Mechanism for Enforcing Compliance to Secondary Data Use Contracts," in: *2014 IEEE 13th*

International Conference on Trust, Security and Privacy in Computing and Communications. pp. 519-526.

North, D. C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge, UK: Cambridge University Press.

Ohm, P. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* (57), p. 1701.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Overby, E., and Jap, S. 2009. "Electronic and Physical Market Channels: A Mulityear Investigation in a Market for Products of Uncertain Quality," *Management Science* (55:6), pp. 940-957.

Parker, G., and Van Alstyne, M. W. 2005. "Two-Sided Network Effects: A Theory of Information Product Design," *Management Science* (51:10), pp. 1494-1504.

Parmar, R., Mackenzie, I., Cohn, D., and Gann, D. M. 2014. "The New Patterns of Innovation," *Harvard Business Review* (92:1/2), pp. 86-95.

Pavlou, P. A., and Dimoka, A. 2006. "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research* (17:4), pp. 392-414.

Pavlou, P. A., and Gefen, D. 2004. "Bulding Effective Online Marketplaces with Institution-Based Trust," *Information Systems Research* (15:1), pp. 37-59.

Perrin, B., Naftalski, F., and Houriez, R. 2013. "Cultural Behaviour and Personal Data at the Heart of the Big Data Industry," E&Y and Forum D'Avignon, pp. 1-52.

President's Council of Advisors on Science and Technology. 2014. "Big Data and Privacy: A Technological Perspective," Executive Office of the President, Washington, DC.

Ramirez, R., Brill, J., Ohlhausen, M. K., Wright, J. D., and McSweeney, T. 2014. "Data Brokers: A Call for Transparency and Accountability," Federal Trade Commission, Washington, DC.

Rochet, J. C., and Tirole, J. 2006. "Two-Sided Markets: A Progress Report," *RAND Journal of Economics* (37:3), pp. 645-667.

Romer, P. M. 1990. "Endogenous Technological Change," *Journal of Political Economy* (98:5), pp. S71-S102.

Roth, A. E. 2002. "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics," *Econometrica* (70:4), pp. 1341-1378.

Roth, A. E. 2007. "The Art of Designing Markets," *Harvard Business Review* (85:10), pp. 118-125.

Roth, A. E. 2008. "What Have We Learned from Market Design?," *The Economic Journal* (118:527), pp. 285-310.

Royal Academy of Engineering. 2015. "Connecting Data: Driving Productivity and Innovation," Royal Academy of Engineering, London, UK.

Schlegel, K., Bayerl, S., Zwicklbauer, S., Stegmaier, F., Seifort, C., Granitzer, M., and Kosch, H. 2014. "Trusted Facts: Triplifying Primary Research Data Enriched with Provenance Information." pp. 1-2.

Schomm, F., Stahl, F., and Vossen, G. 2013. "Marketplaces for Data: An Initial Survey," *SIGMOD Record* (42:1), pp. 15-26.

Schwab, K., Marcus, A., Oyola, J. R., Hoffman, W., and Luzi, M. 2011. "Personal Data: The Emergence of a New Asset Class," p. 40.

Shulman, A. 2010. "The Underground Credentials Market," *Computer Fraud & Security*), pp. 5-8.

Soh, C., Markus, M. L., and Goh, K. H. 2006. "Electronic Marketplaces and Price Transparency: Strategy, Information Technology, and Success," *MIS Quarterly* (30:3), pp. 705-723.

Swan, M. 2015. *Blockchain: Blueprint for a New Economy*. Sebastopol, CA: O'Reilly Meida Inc.

Sweeney, L. 2000. "Uniqueness of Simple Demongraphics in the Us Population." Cambridge, MA: Laboratory for International Data Privacy, Harvard University.

Teece, D. J. 1986. "Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing," *Research Policy* (15:6), pp. 285-305.

Thomas, L. D. W., Autio, E., and Gann, D. M. 2014. "Architectural Leverage: Putting Platforms in Context," *Academy of Management Perspectives* (28:2), pp. 198-219.

Thomas, L. D. W., and Leiponen, A. E. 2016. "Big Data Commercialization," *IEEE Engineering Management Review* (42:2), pp. 74-90.

Tiwana, A., Konysnski, B., and Bush, A. A. 2010. "Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics," *Information Systems Research* (21:4), pp. 675-687.

Truong, H.-L., Comerio, M., De Paoli, F., Gangadharan, G. R., and Dustdar, S. 2012. "Data Contracts for Cloud-Based Data Marketplaces," *International Journal of Computational Science and Engineering* (7:4), pp. 280-295.

Truong, H.-L., Dustdar, S., Gotze, J., Fleuren, T., Muller, P., Tbahriti, S.-E., Mrissa, M., and Ghedira, C. 2011. "Exchanging Data Agreements in the Daas Model," *2011 IEEE Asia-Pacific Services Computing Conference, APSCC 2011, December 12, 2011 - December 15, 2011*, Jeju Island, Korea, Republic of: IEEE Computer Society, pp. 153-160.

Uhlir, P. F., and Cohen, D. 2011. "Internal Document. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences."

Wald, J. 2002. "Legislating the Golden Rule: Achieving Comparable Protection under the European Union Database Directive," *Fordham International Law Journal* (25:4), pp. 987-1038.

Walport, M. 2016. "Distributed Ledger Technology: Beyond Block Chain," UK Government Office for Science, London, UK.

Wareham, J. D., Fox, P. B., and Cano Giner, J. L. 2014. "Technology Ecosystem Governance," *Organization Science* (25:4), pp. 1195-1215.

Weyl, E. G. 2009. "Monopoly, Ramsey and Lindahl in Rochet and Tirole (2003)," *Economics Letters* (103), pp. 99-100.

Zhu, H., and Madnick, S. E. 2009. "One Size Does Not Fit All: Legal Protection for Non-Copyrightable Data," *Communications of the ACM* (52:9), pp. 123-128+110.

Zuiderwijk, A., and Janssen, M. 2013. "A Coordination Theory Perspective to Improve the Use of Open Data in Policy-Making," *12th IFIP WG 8.5 International Conference on Electronic Government, EGOV 2013, September 16, 2013 - September 19, 2013*, Koblenz, Germany: Springer Verlag, pp. 38-49.

Zyskind, G., Nathan, O., and Pentland, A. 2015. "Decentralizing Privacy: Using Blockchain to Protect Personal Data," in: *2015 IEEE CS Security and Privacy Workshops*. San Jose, CA.