Industrial and Corporate Change, 2020, Vol. 29, No. 3, 645–660
doi: 10.1093/icc/dtaa002
Advance Access Publication Date: 13 March 2020
Original article

Markets for data

Pantelis Koutroumpis,[1] Aija Leiponen,[2] and Llewellyn D. W. Thomas [3,*]

[1]Oxford Martin School, Oxford University, Oxford, OX1 3BD, UK. e-mail: pantelis.koutroumpis@oxfordmartin.ox.ac.uk, [2]Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY 14853, USA. e-mail: aija.leiponen@cornell.edu and [3]Department of Business and Technology, LaSalle Universitat Ramon Llull, Barcelona 08022, Spain. e-mail: llewellyn.thomas@salle.url.edu

*Main author for correspondence.

Abstract

Although datasets are abundant and assumed to be immensely valuable, 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 appropriability and quality concerns. We consider the requirements for efficient data exchange, comparing existing trading arrangements against efficient market models and show that it is possible to achieve either large markets with little control or small markets with greater control. We describe some future research directions.

JEL classification: O33, O34, D47

1. Introduction

Data are expected to become the fuel of the digital economy as they can be used to reduce information asymmetries, improve resource management, and identify causal relationships using artificial intelligence and statistical analyses.[1]Such data can encompass location, behavior, retail, health, administrative, and sensor-based industrial data (Manyika *et al.*, 2011). Because of its increasing volume and perceived importance, the data economy is a central element of the Digital Single Market policy framework as envisioned by the European Commission.[2] However, considering such high expectations of welfare impact, little is known about how data are shared and traded. This article explores the underpinnings of a data economy by focusing on different types of market designs for data trading, and the key issues that may cause market inefficiency or even failure. We thus attempt to lay the groundwork for a new stream of research on markets for data.

We focus on private observational data that have not yet been significantly processed or manipulated. Such data correspond to social, laboratory, and measurement data compiled by humans or machines (Uhlir and Cohen, 2011), and business data used for analytical purposes—collections of data items that have grouping, relatedness, and purpose (Borgman, 2012). As such, we do not consider markets for information goods such as software or content. We apply a general definition of markets as domains in which commercial exchange takes place as a result of buyers and sellers being in contact with one another (cf. Encyclopaedia Britannica, 2019). In our framework, markets for data thus encompass both spot and relational transactions for data that take place between organizations and that are

[1] https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy; accessed May 24, 2018.

[2] https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy; accessed November 28, 2017.

informed by some type of a price mechanism (cf. Baker *et al.*, 2002), though not necessarily a monetary price. This approach is consistent with the earlier literature on the markets for technology (Arora *et al.*, 2001), where spot transactions are rare and licensing relationships and joint ventures dominate.

Data are rarely valuable alone and are usually inputs into analytics (embedded in a software program) to generate insights that can become expressed as content-based information goods (such as a scientific report or advertisement). Data are thus primarily intermediate goods, produced with the intent of being combined and transformed to create other information goods (Koutroumpis *et al.*, 2019). Furthermore, data are experience goods, or even credence goods, which gives rise to challenges in verifying the quality and value of data. The value of an experience good is not observable before consumption. Most information goods, such as books or movies, are experience goods: a consumer will not know whether they like the good until they have consumed it. The quality of a credence good is difficult to evaluate even after consumption. For example, a consumer will not be able to assess the efficacy of dietary supplements, except possibly after a longer period of usage. Similarly, the quality and verity of data can only be evaluated by comparing its statistical properties against similar datasets, not directly by viewing or using the data. These characteristics of data goods need to be addressed through careful attention to market design.

We contribute to the emerging literature on data markets in three ways. First, we briefly review the institutional history of data trading and show that it is challenging to set up large-scale systems to trade data through open, multilateral markets in the same way we trade many other goods, including intangible goods such as content and even patented inventions. We suggest that markets for data operate differently from markets for other intangible assets, although none of these markets appear to work particularly efficiently.

Second, we review the markets for ideas literature (Gans and Stern, 2010) and show that, while there are some similarities, markets for data differ in that they require the establishment of rigorous provenance that tracks data from its origins to the destination (Simmhan *et al.*, 2005). Expressed through verifiable metadata for the data being traded, the importance of provenance arises from difficulties in assessing quality and maintaining appropriability. Traditional mechanisms ensuring quality in multilateral markets for experience goods, such as reputation systems (Pavlou and Gefen, 2004; Dellarocas, 2005; Moreno and Terwiesch, 2014), may be insufficient when even the sellers themselves may not be aware of the quality of their goods (or lack thereof). As a result, the full data provenance as evidenced through comprehensive metadata, including sources of data and the methods of collection and structuring, becomes the de facto proxy for data quality and legitimacy.

Furthermore, appropriation regimes for data are weak because it is difficult to define and enforce control rights to data. In particular, intellectual property rights do not appear to facilitate control of the use and dissemination of data (Wald, 2002; Mattioli, 2014; Duch-Brown *et al.*, 2017), and, therefore, data available through open markets are highly likely to be associated with significant knowledge spillovers. However, in markets seeking legitimate data trading (as opposed to unauthorized trading), comprehensive provenance can help clarify and verify the legal rights of the trading parties, thus partially alleviating appropriation problems.

Third, we describe the main data market matching mechanisms and present illustrative examples of actual data marketplaces that utilize these designs (Roth, 2002, 2008). Roth's (ibid.) market design framework allows us to qualitatively describe the benefits and shortcomings of each type of matching and thereby draw conclusions about the types of data and trades that can be completed via each. We show that with the currently available market mechanisms, it is only possible to achieve large markets with little control or small markets with somewhat greater, but not full, control.

This article is organized as follows. The following section briefly reviews the institutions and the history of data trading, and then compares markets for data against markets for ideas and patents. The penultimate section considers markets for data through the market design perspective of Roth (2002, 2008). We conclude by considering some future research directions.

2. The institutional context of data trading

Data have long been shared and traded: for example, academics share research data and businesses share household credit data. In recent years, the lower cost of data collection and the adoption of digital communication networks have dramatically increased volumes of collected data (Reinsel *et al.*, 2017). Much of the collected data are "exhaust data," created as a by-product of other activities such as online shopping or socializing, rather than specifically for an analytical purpose (Manyika *et al.*, 2011; Mayer-Schonberger and Cukier, 2013). Indeed, the purchasing patterns of consumers have become the first data market segment that has experienced significant commercial activity and raised privacy concerns regarding trading practices: the US Federal Trade Commission noted the near-complete lack of transparency in these markets for personal data, potentially harming consumers by breaching their privacy or enabling unfair marketing practices.[3] Digital platforms such as Apple, Amazon, Facebook, and Google enable trackers that collect and aggregate data from online sources, including mobile phones, and provide access or sell the data to third parties.[4] Furthermore, in the shadows of the digital economy, there have always been thriving marketplaces for stolen data (Holt and Lampke, 2010), such as credit card numbers or user profile data (Shulman, 2010). The growing amount of data has thus enabled highly controversial commercial practices.

The organizations and institutions in data markets are rapidly evolving. New regulations such as the General Data Protection Regulation (GDPR) of the European Union and the California Consumer Privacy Act have been implemented, as privacy has become a heightened concern. New types of data intermediaries have been envisioned that would either carry out data trading as their core activity, or trade data that arise from their core operations (Parmar *et al.*, 2014; Thomas and Leiponen, 2016). Such entities would allow third parties to upload and maintain datasets, with access, manipulation and use of the data by others, and regulated through varying licensing models (Schomm *et al.*, 2013). In principle, data marketplaces could resemble multisided platforms, where a digital intermediary connects data providers, data purchasers, and other complementary technology providers (Parker and Van Alstyne, 2005; Eisenmann *et al.*, 2006). Such platforms could generate value for both data buyers and sellers through lower transactional frictions, resource allocation efficiency, and improved matching between supply and demand (Bakos, 1991; Soh *et al.*, 2006).

However, in practice, data are rarely traded on a large scale through multilateral platforms (Borgman, 2012). There are large-scale open data repositories such as the London Datastore set up by the Greater London Authority that do not actually sell data. Commercial data "platforms" such as Acxiom (consumer data), Bloomberg (financial data), or LexisNexis (insurance data) operate as intermediaries that buy and sell data via bilateral and negotiated contractual relationships. Moreover, there are abundant examples of failed data platforms (Markl, 2014; Carnelley *et al.*, 2016): for instance, the Microsoft Azure DataMarket closed down in March 2017 after 7 years of poor performance.[5] It thus appears challenging to set up large-scale systems to trade data through open markets in the same way we trade many other goods, including intangible goods such as content and inventions. We next investigate the characteristics of data markets in detail to understand why this may be the case.

2.1 Markets for data versus ideas

Markets for data exhibit similar characteristics to those for ideas and patents. Ideas, patents, and data are intangible goods and therefore largely nonrival in use. An idea or a data point, if digitized, may be usable by many individuals and replicated at low marginal cost (Romer, 1990; Koutroumpis *et al.*, 2019). Even though the (strategic) value of an idea or data may diminish from wide dissemination, this will not prevent its application and use by many parties. Furthermore, ideas and patents need to be combined with complementary inputs for their commercialization (Teece, 1986; Bresnahan and Trajtenberg, 1995; Gans and Stern, 2010). Like inventions, data are intermediate goods and need to be further processed and combined with complementary inputs such as analytic technologies to become final goods and contribute to utility or productivity (Chebli *et al.*, 2015; Koutroumpis *et al.*, 2019).

Gans and Stern (2010) suggest that markets for ideas may exist in settings where intellectual property protection is sufficiently strong, which increases the likelihood that sellers appropriate enough of the value of an idea to justify the investment by excluding illegitimate trades and uses (Arrow, 1962; Teece, 1986). However, Hagiu and Yoffie (2013) have argued that multilateral digital marketplaces for patents are not viable due to the burdensome arrangements that would be required to ensure that high-quality patents are offered for sale. When the quality of the good is imperfectly observable, markets tend to be flooded with low-quality goods (Akerlof, 1970), and electronic markets

[3] See the Data Brokers report by the US Federal Trade Commission (Ramirez *et al.*, 2014).

[4] See https://www.washingtonpost.com/technology/2019/05/28/its-middle-night-do-you-know-who-your-iphone-is-talking/; retrieved May 29, 2019.

[5] https://social.msdn.microsoft.com/Forums/en-US/1005630f-a6da-4b00-ad4e-adfc968d9416/azure-datamarket-to-retire-on-march-31-2017; accessed November 6, 2019.

for such goods may function particularly poorly (Overby and Jap, 2009). Nevertheless, companies such as Ocean Tomo orchestrate both public and private auctions for intellectual property portfolios.[6]

Thus, scholarship into markets for ideas and patents implies that specific governance mechanisms may be needed for a data market to take off. This literature highlights that for market participants to safely transact, adequate protection, and quality assurance of the traded goods are essential. Next, we examine data governance from the perspectives of appropriability and quality assurance and demonstrate why the notion of provenance is central to data markets.

2.2 Appropriation regime

Appropriation of the returns to intangible goods, such as ideas and data, can be pursued through legal instruments that facilitate and protect control rights (Levin *et al.*, 1984; Teece, 1986). Intellectual property rights, such as patents, copyrights, and trademarks, are available to protect an idea, a technology, or expression (Gans and Stern, 2010). In contrast, the legal instruments that are available to protect data are less well defined. 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 observations it contains (unless the data themselves are characterized as creative content), provided there is an original contribution in putting the dataset together.

This weak appropriation regime is compounded by jurisdictional differences, with the United States 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 European Union, the database directive of 1995 sought to extend protection to the noncopyrightable aspects of databases, for example, when the data are provided in a different order or in a manipulated format, and even to parts of the database, so long as there has been a substantial investment to compile it. In the United States, despite some extensions of copyright to situations where the selection or arrangement of data required judgment,[7] 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 its reproduction or distribution (Mattioli, 2014). The goal of such datarights is a balance between protection and encouragement of innovation in data usage and practices.

However, designing data protections has proved difficult. The European database right appears to have had no measurable impact on the database industry,[8] and a limited number of legal cases have examined its boundaries. When data are observational records they 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," that is, the original structure of the database itself. The data may also be transformed by statistical analyses, and the results of the analyses are not subject to the original 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 may not be able to prove that a specific data source was utilized for an analytical output.

Therefore, data have a weak appropriation regime, and they are usually protected through trade secrecy and contractual means.[9] Data license agreements can be used to define rights for derivation, collection, reproduction, attribution, confidentiality, audit, and commercial use. These licenses tend to be lengthy and complicated, and the contract terms depend on laws, regulations, measurement units, and values of a particular jurisdiction (Truong *et al.*, 2012), seeking to define the admissible commercial utilization of data in explicit terms that depend on the market. Although

[6] https://www.oceantomo.com/auctions/; accessed June 18, 2019.

[7] 945 F. 2d 509—Key Publications, Inc. v. Chinatown Today Publishing Enterprises, Inc. 1991.

[8] 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.

[9] See for instance the Collateral Analytics v. Nationstar case, a trade secret lawsuit filed in US District Court, Northern District of California in Jan 2018: https://patentlyo.com/media/2018/01/CollateralAnalyticsComplaint.pdf; accessed November 12, 2018.

such terms are regularly stipulated in bilateral data license agreements, such as those of Bloomberg or Thomson Reuters, they are hard to define and enforce in a large-scale multilateral context.

2.3 Quality control

Most intangible goods are either experience goods or credence goods. The value of experience goods can only be verified during consumption, while that of credence goods can only be verified after longer-term usage or third-party certification. Quality assurance within markets for such goods is often addressed through verification services offered by a market intermediary for a fee (Dushnitsky and Klueter, 2011; Gefen and Pavlou, 2012; Catalini and Gans, 2016). Studies have shown that the reputation of the online marketplace itself can reduce the perceived risk of trading (Pavlou and Gefen, 2004; Gefen and Pavlou, 2012).

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 or establishing reputation systems that rate the quality of the participants. The reputation of the market participants themselves can influence the efficiency of the market (Dellarocas, 2005), for example, 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 being 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 multilateral markets require thorough screening and disclosure of the patents themselves to overcome the adverse selection problem by which only weak patents are offered. For data markets, participant-level quality verification by intermediaries may be necessary, as it is an effective means of ensuring market safety when there are high levels of moral hazard (Pavlou and Gefen, 2004; Dellarocas, 2005). However, product-level verification by intermediaries such as screening and disclosure is more difficult, given the vast heterogeneity in both the format and content of data.

A key data quality challenge is the legal status of data. Even the sellers themselves may not be aware of the legal status of their data. This is particularly true when the data includes personal (or customer) information. Personal data such as health records or mobile phone records permanently point to a specific individual (an characteristic termed "inalienability" by Koutroumpis *et al.*, 2019), and once several such data streams are integrated, the person in question can usually be identified despite anonymization. Computer scientists have convincingly demonstrated that they can rather easily "reidentify" or "deanonymize" individuals from anonymized data (Sweeney 2000; Ohm, 2010), highlighting that regulation of privacy is a crucial concern (although there is increasing effort to ensure such anonymization processes are effective, see Menon and Sarkar, 2016). Consequently, privacy protection for personal data is enacted primarily through regulations. For instance, credit rating data have been regulated in the United States since the 1970s through the Fair Credit Reporting Act of 1970, protecting consumers from unreasonable use of their financial information for credit, employment, insurance, housing, and other eligibility decisions (Federal Trade Commission, 2013).

However, the regulatory environment is complex. National regulations represent myriad solutions for collecting and using data in support of different institutional and corporate aims (Schwab *et al.*, 2011). In 2015, the European Union enacted the GDPR (fully in force since 2018) that mandates strict personal data protection practices and allows national jurisdictions to set up additional rights to other types of data. Meanwhile, various states of the United States have adopted or are in the process of developing data regulations, creating a veritable patchwork of state-level rights.[10] 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 across jurisdictions (Schwab *et al.*, 2011), with discrepant legislative structures, regulatory enforcement agencies, and jurisprudence (Perrin *et al.*, 2013). Efforts to enable interoperability across jurisdictions, for example the "safe harbor" principles developed between 1998 and 2000 to prevent private organizations within the European Union or United States which store customer data from accidentally disclosing or losing personal information, have only been partially

[10] https://www.dataprotectionreport.com/2018/07/u-s-states-pass-data-protection-laws-on-the-heels-of-the-gdpr/; accessed November 18, 2018.

successful. The original "safe harbor" agreement was overturned by the European Court of Justice in 2015 and its replacement, the "privacy shield," in force from 2016, has been contested.[11]

Taken together, the regulatory landscape suggests that the legality of data sales from certain sources, or for particular purposes, or across international borders may be unclear. Furthermore, when data have been combined into hybrid datasets, consisting of a variety of industries, jurisdictions, and contractual conditions, and used in a variety of corporate functions, the legal status of the hybrid product may be impossible to define. By not having certainty on the legal status of a dataset, the sellers themselves may be (perhaps unwittingly) offering a lower quality product. In such cases, verification processes that focus on the participant's credentials may only be partially effective. Furthermore, data-level verification processes such as disclosure and screening may be unfeasible due to the opacity of the original process that combined the data or the sources of the constituent data, resulting in prohibitively high verification costs.

2.4 Data provenance

Because of the weak appropriation regime and the substantial quality challenges, data quality and legality are judged to a large degree by where it originated. Rather than attempting to verify the status of the data goods directly, trading partners usually rely on the reputation and legal liability of the original source, potentially with their contractual commitment to correct any mistakes found in the data. Data thus need to have rigorous and comprehensive records of origin, characteristics, and history. Therefore, the value of data significantly depends on this complementary "metadata" about its provenance, making data and metadata strongly complementary in creating value (Mattioli, 2014). However, there may be significant barriers to disclosure of 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 deanonymize an otherwise anonymous pool of individuals; and methods of data preparation themselves can be valuable trade secrets (Mattioli, 2014).

There have been few institutional responses to the necessity for proving provenance, that is, disclosure of the sources and processes that created the data, although there have been 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: 5). Encouragingly, there are technical efforts to design provenance mechanisms, such as trust management tools for monitoring data consumers' contractual compliance (Moiso and Minerva, 2012; Noorian *et al.*, 2014; Schlegel *et al.*, 2014). At present, data provenance is typically shallow in the sense that data sellers claim provenance, but once the data leaves their control, provenance is lost. However, provenance is a key complement that contributes to the value of data.

3. Data market design and matching models

Considering the challenges to data trading posed by a weak appropriation regime, inscrutable quality, and the resulting need for provenance, we next investigate to what degree various market mechanisms can address the issues. We review the market design principles of Roth (2002, 2008) and examine how typical matching mechanisms available in data markets accommodate these. Markets for data are often based on exchanging access and services rather than explicit sales of specific data goods. For instance, Bloomberg sells access to financial market data on subscription basis, and Facebook provides access to user data for application creators in exchange for platform fees and a share of revenues.[12] Nevertheless, we consider these data sharing arrangements to be "markets" because data are used as a valuable exchangeable asset in commercial transactions. Similarly, the literature on the markets for technology considers cross-licensing and even contractual codevelopment arrangements to be a part of the "market," even though they might not involve outright sales of specific technologies (Arora and Gambardella, 2010). Such transactions are not "market-like" in the sense of being arm's length, anonymous, and involving exchange of a good for money. Instead they tend to occur under a variety of relational contracts (Gibbons and Henderson, 2012). Nevertheless,

[11] https://www.americanbar.org/groups/business_law/publications/blt/2016/05/09_alvarez/; https://en.wikipedia.org/wiki/EU%E2%80%93US_Privacy_Shield; both accessed April 1, 2019; we thank an anonymous reviewer for this suggestion.

[12] Due to a lack of competition and the high bargaining power of incumbents (e.g. Google and Facebook) these markets are often imperfect, as evidenced by ongoing EU regulatory action of Google and Facebook. This is further evidence of our assertion that these markets operate differently from markets for other intangible assets.

within industrial organization economics such transactions do constitute “markets” because they involve prices (monetary or otherwise) for (incompletely) substitutable goods or services that are affected by one another (Tirole, 1988: 12 and 13).

3.1 Market design principles

Markets 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 from continued trading. To do so it needs to offer low transaction costs and effective trading arrangements (pricing, contracting, and fulfillment) that support the 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.

Roth’s theory of market design (Roth, 2002, 2008, 2009) identifies several requirements that are associated with efficient market operation, in other words, markets where prices consistently reflect all the available information (Fama, 1970). Economic efficiency thus implies that valuable resources are in their best uses. First, 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. In markets for unique data that are valuable in highly specific contexts, a lack of thickness can be a major factor leading to inefficiency.

Second, while thickness is a necessary precondition for an efficient market, popularity can also create “congestion” by slowing down transaction times 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. In digital markets, congestion usually is a nonissue.

Third, the market needs to be perceived as “safe.” Safe markets are those where participants do not have opportunities to misrepresent information or undertake other strategic action that might reduce efficiency. The marketplace must be able to preclude behavior that influences the actions or preferences of other participants. For example, it would be important to prevent buyers from colluding and prevent sellers from making side contracts with buyers or other sellers or trade outside the market altogether. In the case of data, a safe marketplace will provide credible provenance information: if a buyer is unable to assess the origins (and thus the quality and legality) of the data, information asymmetries between the seller and the buyer are aggravated and the market becomes inefficient. Safety also requires that outsiders are excluded: the data are protected, and traders cannot share the data with outsiders.

Finally, the marketplace needs to respect the social and ethical norms associated with the underlying commodity and avoid engaging in transactions that Roth (2002) termed as “repugnant.” In this context, the limits of a marketplace mechanism may clash with social norms or legal restrictions that often nullify the effectiveness of pricing as an allocation mechanism; for instance, German citizens were repulsed by Google’s harvesting of Street View images for its Maps product.[13] Put differently, 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, the privacy and confidentiality implications of data can potentially limit the growth of marketplaces. Individuals or social groups may view trade in personal data as repugnant and seek to limit its legality and legitimacy. Not only is there increasing public interest in the societal impacts of data, privacy, and data trading,[14] there is also increasing regulatory interest in the transparency and quantity of the personal data that has been amassed and is being traded (Ramirez et al., 2014).

[13] See https://www.economist.com/europe/2010/09/23/no-pixels-please-were-german retrieved April, 2019.

[14] See for instance: Amnesty Global Insights, February 27, 2017, *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 March 4, 2017. 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 March 4, 2017.

Table 1. Types of data marketplaces by matching mechanism

Matching	Marketplace design	Terms of exchange	Examples
One-to-one	Bilateral	Negotiated	Personal data brokers; Acxiom
One-to-many	Dispersal	Standardized	Twitter API; Facebook API
Many-to-one	Harvest	Implicit barter	Google Waze; Google Search
Many-to-many	Multilateral	Standardized or negotiated	None

3.2 Matching models for data markets

We now characterize the four commonly observed distinct types of data markets with respect to the market design principles reviewed above. Table 1 classifies data marketplaces by the number of bargaining parties on each side and presents some examples of actual data marketplaces utilizing these designs. Conceptually, the matching of buyers to sellers for data is no different from any other type of market. Gans and Stern (2010), applying the market design approach of Roth (ibid.) to markets for ideas or technology, suggested that effective market design might be possible for some innovation markets. However, they warn that the nonrivalry of the goods, the need for complementary assets to create value, and weak intellectual property rights undermine the spontaneous and uncoordinated evolution of a market for ideas or technology. They note that when intellectual property rights are weak, the conditions for market thickness and market safety may not be met. Markets for data suffer from many of the same issues as the markets for ideas, but the governance remedies are different, as we explain below.

3.2.1 One-to-one

To begin, one seller can trade simultaneously with one or more buyers. One-to-one matching is a bilateral relationship that involves one buyer and one seller and is typically characterized by negotiated terms of exchange, usually setting up a relational contract (Macneil, 1978, 1985; Baker *et al.*, 2002; Gibbons and Henderson, 2012). Examples of bilateral data traders include personal or industrial data vendors and brokers, such as Acxiom, one of the consumer data brokers described in the report by the US Federal Trade Commission (2013). These firms buy, aggregate, and sell consumer data from hundreds of online and offline sellers of consumer goods and services. For example, Acxiom sells intricate profiles of US households including demographics, financial status, major purchases, political behavior, interests, and life events such as marriage, divorce, and birth of children.[15]

In a bilateral marketplace design, relational contracts often govern the repeated interaction within commercial relationships. The ongoing nature of the relationship enables some aspects of the exchange to be informally defined and not enforceable in court. For example, in an employment relationship, the employer could provide bonus payments to an employee based on a subjective assessment of "leadership" or "initiative." These are not terms that can be contractually defined and thus legally enforced. Nevertheless, the expectations of a long-term relationship provide incentives for both parties to act in good faith and thus, for the employer, to pay the bonus, and for the employee, to take initiative (cf. Levin, 2003). In a bilateral data market, licensing agreements often govern ongoing service relationships with subscription fees and even auditing clauses, the arrangement building a bilateral history and expectations about continuation. Furthermore, while bilateral data contracts may stipulate many aspects of the transaction, often these are very difficult to monitor and verify, and some critical issues may be impossible to anticipate or formalize in the agreement. As such, the success of the relational contract depends on the interests of both parties to maintain the relationship and their external reputations in the marketplace. Hence, bilateral data transactions are often governed by relational contracts that support trading but are associated with high transaction costs.

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 appropriation are easier to achieve due to more comprehensive contracts and their enforcement through monitoring. With a relatively small number of trades, congestion is unlikely to be a concern, but transaction costs will be high due to costs of search, negotiation, and relationship management, including contract enforcement. Furthermore, as bilateral markets often deal with confidential data, there are potentially greater issues with repugnance. For instance, the sale of medical data through opaque

[15] https://www.acxiom.com/what-we-do/infobase/; retrieved June 2, 2019.

intermediaries is increasingly considered as a socially unacceptable market.[16] Thus, even though the relational aspect of bilateral data markets can ameliorate some of the issues, these markets are still likely to feature significant failures whereby sellers with valuable assets are not able to trade with buyers willing to pay a positive price.

3.2.2 One-to-many

When a single seller transacts with many buyers for the same data, using 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 open data distributed through Application Programming Interfaces, or APIs,[17] such as the Twitter "firehose" data.[18] Much financial market data (e.g. securities or commodities data as provided by the New York Stock Exchange, NASDAQ, or the Chicago Mercantile Exchange) is accessed this way. Achieving 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 likely to lead to strategic behavior by some buyers. Buyers may thus use the data in ways that reduce the value of data for the seller and 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 and visible nature of these types of marketplaces, it is unlikely it will involve data that generate repugnance concerns.

3.2.3 Many-to-one

Many-to-one marketplaces involve many sellers but only one major buyer. These marketplaces are characterized by the harvesting of data, where users make their data available to a single service provider, under terms of exchange that often resemble barter: The user receives access to a "free" service in exchange for their data. An example of this is Waze, the map and traffic app now owned by Google, which users allow to harvest their location data in return for real-time mapping, routing, and transportation services. Waze aggregates the traffic data and uses it to provide travel predictions. It also monetizes the service by bundling other services such as advertising and music to the app.[19] Online social networks and many other mobile phone apps have similar harvesting arrangements. The harvested data is typically used internally for product development and commercialized externally via data brokers and other marketing companies. Data harvesting companies thus typically operate in two data markets, a many-to-one market to obtain data and either one-to-one or one-to-many market to monetize it.[20]

The thickness of such harvesting markets depends on the popularity of the adjacent market for bartered "free" services. If the services are highly desirable, such as search, 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 negotiation or relationship management. Transaction costs may quickly balloon, however, if data harvesting runs afoul of repugnance concerns such as norms related to privacy.[21] For example, the GDPR of the European Union gives users a "right to be forgotten." Should this become a popular right to exercise, it might become very costly to online service providers aiming to monetize user data. More generally, users of the adjacent service may find it repugnant that their behavioral data is exploited by the service provider for other purposes than those in which the users participate. As the imposition of the "right to be forgotten" directive and the GDPR suggest, there is a growing sense of repugnance related to harvesting markets. GDPR also stipulates a right for consumers to port their data from one

[16] https://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns; accessed April 1, 2019.

[17] Application Programming Interfaces are computer functions that permit the creation of software-based services that automatically access the underlying data of the service.

[18] The Twitter firehose is the complete stream of public messages on the Twitter service provided through an API: https://developer.twitter.com/en/pricing.html; accessed June 2, 2019.

[19] https://mashable.com/article/waze-spotify-pandora-music-audio-player/; accessed June 4, 2019.

[20] We thank an anonymous reviewer for this observation.

[21] Although Facebook and Google have been growing in recent years, this is due to a rather lax approach to privacy by regulators and users. However, if privacy concerns were to become much more salient then their transaction costs would rapidly increase, and this would be reflected in their service offerings and possibly pricing.

service to another, which can alter the competition landscape for service providers. 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.

In fact, the appropriation regime is likely to be weak for both data harvesters and data providers (service users) in harvesting markets, because, with the typically all-encompassing terms and conditions in large-scale settings,[22] the user retains little control over subsequent utilization and commercialization of their data. Data provenance will vary depending on its collection—data collected from browsing habits or mobility will have a clear provenance, while any data uploaded by users will have a shallow provenance, as standard terms and conditions may not be able to verify the origins beyond requesting a confirmation that users can legally share the data. Thus, appropriation is likely to be compromised in many-to-one (harvesting) markets, and provenance will vary but is likely to often be shallow.

3.2.4 Many-to-many

Finally, multilateral or many-to-many marketplaces are trading platforms upon which 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 these markets a platform potentially mediates transactions among participants from across the data ecosystem, including data creators, managers, analysts, service providers, and aggregators (Thomas and Leiponen, 2016). In its generic form, a many-to-many marketplace is a two-sided market (Parker and Van Alstyne, 2005; Hagiu, 2006).[23] 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). Multisided market theories (Rochet and Tirole, 2006; Bolt and Tieman, 2008; Weyl, 2009) appear to have straightforward implications for the potential structure and pricing of multilateral data marketplaces: data platform owners can in principle utilize pricing strategies to optimize participation and achieve profitability by internalizing the bulk of the network externalities.

Multilateral markets may provide several desirable features over other market designs, as they potentially enable economies of scale, scope, innovation, complementarity, transaction, and search (Tiwana *et al.*, 2010; Thomas *et al.*, 2014). In principle such digital platforms could generate value for data sellers and buyers through enhanced market efficiency due to high transaction volume, resource allocation efficiency, and stable matching (Eisenmann *et al.*, 2006; Thomas *et al.*, 2014). Although these platforms are costly to maintain, they will gain from scale effects where high volumes of data offset a fixed cost of meta-information. Due to such scale economies and network effects, it is conceivable that there are winner-take-all dynamics, meaning that only one or few data platforms would emerge for specific classes of data.

However, whereas digital technologies can mitigate direct transaction costs and facilitate stable matching, strategic behavior may present insurmountable governance problems for multilateral data platforms (Tiwana *et al.*, 2010; Wareham *et al.*, 2014). In particular, suppliers of data may not truthfully reveal the origins and quality of the data, and adverse selection may ensue with poor quality data flooding the market (Holmstrom and Weiss, 1985). This concern echoes Hagiu and Yoffie (2013) who argue that requirements to ensure that not only poor quality patents are offered, such as screening, listing fees, and disclosure, reduce the efficiency of the multilateral patent trading market. The main concern of data platforms, however, is that buyers of data may not respect the usage and access restrictions and consequently degrade the value, confidentiality, and security of the data. Designing technical or contractual systems that incentivize and enforce appropriate behavior of the participants on a multilateral platform, in the absence of relational contracting, may be difficult if not impossible. As a result, achieving market thickness can be very challenging (see Markl, 2014; Carnelley *et al.*, 2016 for some examples of failed data platforms).

We believe it is for these reasons that no "eBay for data" has emerged—the insurmountable concerns regarding strategic behavior, quality of data, and inadequate control over buyer usage of the data have hampered their development (Carnelley *et al.*, 2016). The Microsoft Azure Data Catalog, for example, allows data providers to list their available data and close the transaction via Azure platform. However, as of 2019, these services appear to be

[22] For example see Facebook terms of service: https://www.facebook.com/terms.

[23] There could be alternative types of transactions in cases where the traded dataset is auction-based pricing, or the trades are performed through high frequency (or machine learning based) 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).

Table 2. Characteristics of data marketplaces

Matching	Marketplace design	Liquidity	Transaction costs	Safety
One-to-one	Bilateral	Low	High	High
One-to-many	Dispersal	High	Low	Low
Many-to-one	Harvest	High	Low	Variable
Many-to-many	Multilateral	High	Low	Low

intended for internal use by large organizations.[24] As such, thus far we have no functioning examples of sustainable multilateral data platforms. The only instance where we can observe thriving multilateral data markets is the dark web. However, there are ongoing attempts to create legal alternatives.[25]

Table 2 summarizes the foregoing discussion of Roth's design principles for the four types of data markets. The bilateral market is likely to suffer from low liquidity, but the other three designs are expected, in principle, to be able to achieve market thickness. Bilateral markets also stand out in terms of their high transaction costs, but in return, they are expected to provide greater safety, in terms of provenance and protection from undesirable trades, thus reducing the strategic behavior of participants. The other marketplace designs are expected to suffer from limited safety, in terms of deficient provenance and appropriability, and are thus hampered by strategic behavior of participants. However, in some circumstances harvesting markets can be relatively safe if clear provenance is combined with a regulatory framework such as GDPR that restricts the data buyer's strategic behavior. However, such regulatory mechanisms are difficult to implement and enforce in dispersal or multilateral markets where there are large numbers of buyers.

Thus, with currently available market mechanisms, it seems possible to achieve either large markets with rather little control or small markets with greater control. However, when appropriability is not a critical issue such as in the context of highly time-sensitive financial data that loses much of its value within minutes, or personal data that only gains significant value when aggregated with millions of other data points, many-to-one or one-to-many markets may function reasonably well without the expectation that the data are tightly protected after the transaction. However, we are currently unaware what governance arrangements would enable multilateral platforms to accommodate commercial large-scale data trading.

4. Future research directions

We have investigated the nature of markets for data with a conceptual market design approach. Applying insights from the markets for ideas literature, we first demonstrated that markets for data require the establishment of rigorous provenance for the data being sold, expressed through verifiable metadata such as the origins, content, and the methods of collection of as well as the rights for the data, because of the difficulty of assessing quality or appropriating returns on data investment. Then, building upon the market design framework of Roth (2002, 2008) we characterized the main data market mechanisms and presented illustrative examples of actual data marketplaces that utilize these designs. We argued that with the currently available market mechanisms, it is possible to achieve either large but unsafe markets or small and somewhat safer markets.

Given the difficulties of ensuring the quality and appropriability of data, we suggest that large-scale multilateral data platforms are unlikely to succeed without additional governance innovations that strengthen the provenance of data for all parties. A generic digital marketplace with many suppliers and buyers of data would not be able to monitor and enforce usage restrictions, implying that participants would be able to strategically influence the behavior or valuation of their peers through such actions as trading bilaterally outside of the platform or sharing the data with unauthorized third parties. When contracts are highly incomplete, market failure may be prevented by relational governance through repeated interaction and trust building. Alternatively, market makers may attempt to write more complete contracts using sophisticated technologies such as "smart contracts" to mitigate strategic behavior in data trading.[26]

[24] See: https://azure.microsoft.com/en-us/services/data-catalog/; accessed June 9, 2019.

[25] https://asia.nikkei.com/Business/Business-trends/Big-data-trading-platform-to-launch-in-Japan-next-month; accessed June 10, 2019.

[26] See: https://en.wikipedia.org/wiki/Smart_contract; accessed June 11, 2019.

We next suggest directions for future research. One promising research opportunity is to assess technological solutions to improve provenance, such as distributed ledger technologies (DLTs) (Evans, 2014; Catalini and Gans, 2016). DLTs are distributed databases such as blockchain that underpins the cryptocurrency Bitcoin. They automatically track transactions in a trading network and offer an immutable provenance record, thus facilitating data quality assessment (Koutroumpis *et al.*, 2019). Furthermore, as a decentralized system, a DLT would enable trades to be directly executed and verified by market participants collectively rather than through a centralized intermediary. This means that the proof of provenance would be decentralized and automated, potentially allowing a larger market with greater control. However, a DLT would not completely remove the risk of strategic behavior. There remains the possibility of taking the data off-ledger and trading it bilaterally with third parties or even trying to take over the network consensus that is used to verify the legitimacy of each transaction. While this would sacrifice the benefits of tracking, provenance, and legitimacy, as well as break the rules of the marketplace and potentially subject transgressors to legal risks, there may be types of data that are sufficiently valuable even without the benefits of transaction verification. For instance, DLT-enabled data trading might not be able to prevent data breaches such as Cambridge Analytica, if parties are lured with substantial short-term profits and the likelihood of audit and enforcement is low. Technological solutions such as DLTs thus may not completely remove strategic behavior, but they point in the direction of potentially viable multilateral data market designs.

Another fruitful direction for future research is marketplace designs that involve a group of participants that collectively manage and share their data. Echoing notions of data as a common pool resource (Ostrom, 1990; Hess and Ostrom, 2003, 2011), these data collectives might adopt strong boundaries via elaborate vetting, establish clear rules through contracts and bylaws, have procedures to collectively change them, and use effective monitoring and enforcement through substantial investments in auditing, potentially by neutral third parties, to enable trading of data. While there are aspects of explicit contracting in such resource collectives, a significant degree of enforcement relies on the "shadow of the future" embedded in the implicit relational contract. However, freeriding problems are likely to be aggravated in a multilateral setting. Early examples of such data collectives or data consortia can be identified.[27] For instance, in the UK, the Insurance Fraud Register is a database where members of the Association of British Insurers share data of individuals who have been involved in fraudulent, bogus and exaggerated claims.[28] Insurers send their data to a central database managed by a not-for-profit entity controlled by the insurers which makes available the aggregate data to those market participants who opt to share their data. Thus, the private data from one member is available to all members. As only members of the not-for-profit Association of British Insurers are eligible to participate, there are effective boundaries and an organizational structure for collective rule-making and monitoring.

There are also emerging hybrid designs that incorporate elements of collective and bilateral designs. An example here is ID Analytics.[29] ID Analytics is a for-profit firm that provides fraud and credit risk assessment and other risk management solutions in exchange for clients' data. When clients subscribe to ID Analytics industry solutions, they commit to making available to others their transaction data. The data are packaged and made available to clients depending on their subscription. The difference between a genuinely multilateral data collective and a data intermediary is that the former allows users to retrieve each other's data directly whereas the intermediary collects and manipulates all client data before repackaging and serving it to other clients. However, it appears that these multilateral data collectives, regardless of form, are not necessarily easy to establish,[30] and further research into their dynamics and evolution would be useful.

Future research could also explore specific governance arrangements for the data collective model. For instance, some emerging financial service applications have proposed adopting DLT solutions in collective data sharing arrangements.[31] As we have suggested that relational contracting has an important role in some data marketplaces,

[27] Other examples include: 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/; and the Smart Steel initiative: https://www.ssab.us/ssab/newsroom/2018/05/23/07/00/smartsteel-10–the-first-step-toward-an-internet-of-materials; all accessed December 1, 2018.

[28] https://www.out-law.com/en/articles/2012/september/this-weeks-headline/; accessed January 2, 2019.

[29] https://www.idanalytics.com/our-business-model/; accessed December 15, 2018.

[30] https://www.insurancetimes.co.uk/insurance-fraud-register-now-live/1406541.article; accessed October 2, 2019.

[31] Major banks are working together to develop a decentralized platform for clearing transactions (Financial Times, August 2016): http://www.ft.com/cms/s/0/1a962c16-6952-11e6-ae5b-a7cc5dd5a28c.html); accessed September 3, 2016.

more research into how relational contracting operates in a multilateral context would be valuable, as at present, theory so far has assumed a bilateral relationship. Empirically, detailed case studies of specific multilateral data market models would illuminate how the governance issues have been operationalized and which features can be combined to enhance long-term viability.

We believe there are fruitful research opportunities also in considering how data governance might evolve in the industrial Internet of things. Here, emerging examples (such as industrial data enabled by the Predix platform of GE and platforms from other industrial firms) suggest the emergence of isolated pools of industrial data sharing rather than a global network of data connected across industries. Indeed, an important policy issue in the digital economy is the challenge of unlocking the value of private industrial data when its benefits depend on complementary private data held by many distinct parties potentially using different technologies. This is a classical anticommons dilemma, in other words, socially suboptimal information availability because of excessive privatization (Heller, 1998). It is possible that data collectives may sufficiently address such governance issues when the value of the data is substantial, there are significant complementarities among the participants' data, and the members of the collective have highly aligned and reasonably stable interests. Nevertheless, such arrangements for data sharing are unlikely to be global in reach or universal in nature. They are more likely to involve a limited number of partners in specific, narrowly defined contexts.

Finally, an important research question concerns how to structure the marketplace. The notion of a data collective implies that the marketplace is jointly created and governed. In contrast, the generic model of multilateral market design involves a platform leader who builds and operates the marketplace. One proposed approach is 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). Practitioners are considering alternatives to the harvest market model for personal data. For instance, the Solid initiative of Tim Berners-Lee seeks to give every user a choice about where their data is stored, which specific people and groups can access selected elements, and which applications can use them.[32] A data marketplace may thus be offered by a specialized platform that enables but does not engage in the data trading itself. The benefits and disadvantages of each model are yet unknown. More broadly, the challenges of trading data may underpin some of the shift toward platforms and ecosystems in the broader economy.[33] As platforms and ecosystems are new ways of coordinating enabled by modularization (Jacobides *et al.*, 2018), they can facilitate information and data exchange through means other than price (Baldwin and Clark, 2000). The interplay between data, platforms and privacy, a dynamic influenced by regulations such as the GDPR, can have also significant effects on market structure and competition.[34] The enforcement of these regulations is key for competition authorities around the world, predominantly due the constant need for adaptation in new technological and market conditions. There is much research to be done that considers the role of data exchange in platform ecosystem design, operation, and competition.

5. Conclusion

In closing, we note that the legal and regulatory environment for data markets is rapidly evolving. The very idea of data ownership is still debated. In response to various data scandals, the Financial Times has suggested that "a key part of the answer lies in giving consumers ownership of their own personal data."[35] In contrast, legal scholars (Evans, 2011) and data governance experts (Tisne, 2018) have argued that data ownership is either not feasible or is conceptually flawed as a mechanism to address the societal and economic challenges that the data economy presents. In the United States, a bill has been proposed to give individuals rights to how their data are used without requiring the individual to take ownership of them.[36] With such fundamental issues underlying the markets for data still being debated, the viability of any market design could dramatically change in the near future. Nevertheless, our analyses suggest that the questions of provenance and appropriability will be critical for the functioning of any market for

[32] https://medium.com/@timberners_lee/one-small-step-for-the-web-87f92217d085; accessed December 1, 2018.

[33] We thank Michael Jacobides for suggesting this.

[34] https://wayback.archive-it.org/12090/20191129193858/https:/ec.europa.eu/commission/commissioners/2014-2019/vestager/announcements/digital-power-service-humanity_en; accessed December 28, 2019.

[35] https://www.ft.com/content/a00ecf9e-2d03-11e8-a34a-7e7563b0b0f4; accessed December 21, 2018.

[36] Data Care Act 2018.

data, and therefore, markets for data will have different governance features than markets for ideas where valuation does not depend on provenance.

Acknowledgments

The authors acknowledge financial support from EPSRC project EP/K039504/1.

References

Akerlof, G. A. (1970), 'The market for "lemons": quality uncertainty and the market mechanism,' *Quarterly Journal of Economics*, **84**(3), 488–500.

Arora, A., A. Fosfuri and A. Gambardella (2001), 'Markets for technology and their implications for corporate strategy,' *Industrial and Corporate Change*, **10**(2), 419–451.

Arora, A. and A. Gambardella (2010), 'Chapter 15—The market for technology,' in B. H. Hall and N. Rosenberg (eds), *Handbook of the Economics of Innovation*, Vol. **1**. North-Holland: Amsterdam, NL, pp. 641–678.

Arrow, K. J. (1962), 'Economic welfare and the allocation of resources for invention,' in Universities-National Bureau Committee for Economic Research and Committee on Economic Growth of the Social Science Research Council (eds), *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton University Press: Princeton, NJ, pp. 609–626.

Baker, G., R. Gibbons and K. J. Murphy (2002), 'Relational contracts and the theory of the firm,' *The Quarterly Journal of Economics*, **117**(1), 39–84.

Bakos, J. T. (1991), 'A strategic analysis of electronic marketplaces,' *MIS Quarterly*, **15**(3), 295–310.

Baldwin, C. Y. and K. B. Clark (2000), *Design Rules: The Power of Modularity*, Vol. **1**. MIT Press: Cambridge, MA.

Bolt, W. and A. Tieman (2008), 'Heavily skewed pricing in two-sided markets,' *International Journal of Industrial Organization*, **26**(5), 1250–1255.

Borgman, C. L. (2012), 'The conundrum of sharing research data,' *Journal of the American Society for Information Science and Technology*, **63**(6), 1059–1078.

Bresnahan, T. F. and M. Trajtenberg (1995), 'General purpose technologies "Engines of growth"?,' *Journal of Econometrics*, **65**(1), 83–108.

Carnelley, P., H. Schwenk, G. Cattaneo, G. Micheletti and D. Osimo (2016), 'Europe's data marketplaces—current status and future perspectives,' in *European Data Market SMART 2013/0063 D.39*. IDC, http://datalandscape.eu/data-driven-stories/europe's-data-marketplaces-–-current-status-and-future-perspectives.

Catalini, C. and J. S. Gans (2016), 'Some simple economics of the blockchain,' *NBER Working Paper Series*, National Bureau of Economic Research: Cambridge, MA, p. 30.

Chebli, O., P. Goodridge and J. Haskel (2015), 'Measuring activity in big data: new estimates of big data employment in the UK market sector,' *Imperial College Business School Discussion Paper*. Imperial College Business School: London, UK, p. 28.

Dellarocas, C. (2005), 'Reputation mechanism design in online trading environments with pure moral hazard,' *Information Systems Research*, **16**(2), 209–230.

Duch-Brown, N., B. Martens and F. Mueller-Langer (2017), 'The economics of ownership, access and trade in digital data,' *JRC Digital Economy Working Paper; JRC Technical Reports*, European Commission: Seville, Spain.

Dushnitsky, G. and T. Klueter (2011), 'Is there an eBay for ideas? Insights from online knowledge marketplaces,' *European Management Review*, **8**(1), 17–32.

Eisenmann, T. R., G. Parker and M. W. Van Alstyne (2006), 'Strategies for two-sided markets,' *Harvard Business Review*, **84**(10), 92–101.

Encyclopaedia Britannica (2019), 'Market (economics)', https://www.britannica.com/topic/market.

Evans, B. J. (2011), 'Much ado about data ownership,' *Harvard Journal of Law Technology & Innovation*, **25**(1), 69–130.

Evans, D. S. (2014), 'Economic aspects of Bitcoin and other decentralized public-ledger currency platforms,' *Coase-Sandor Institute for Law & Economics Working Paper*, Coase-Sandor Institute for Law and Economics: Chicago, IL.

Fama, E. F. (1970), 'Efficient capital markets: a review of theory and empirical work,' *The Journal of Finance*, **25**(2), 383–417.

Federal Trade Commission (2013), 'What information do data brokers have on consumers, and how do they use it.' Presentation to the Committe on Commerce, Science, & Transportation, United States Senate, December 18, 2013, https://www.ftc.gov/sites/default/files/documents/public_statements/prepared-statement-federal-trade-commission-entitled-what-information-do-data-brokers-have-consumers/131218databrokerstestimony.pdf.

Gale, D. and L. S. Shapley (1962), 'College admissions and the stability of marriage,' *The American Mathematical Monthly*, **69**(1), 9–15.

Gans, J. S. and S. Stern (2010), 'Is there a market for ideas?,' *Industrial & Corporate Change*, **19**(3), 805–837.

Gefen, D. and P. A. Pavlou (2012), 'The boundaries of trust and risk: the quadratic moderating role of institutional structures,' *Information Systems Research*, **23**(**3-part-2**), 940–959.

Gibbons, R. and R. Henderson (2012), 'Relational contracts and organizational capabilities,' *Organization Science*, **23**(5), 1350–1364.

Hagiu, A. (2006), 'Pricing and commitment by two-sided platforms,' *RAND Journal of Economics*, **37**(3), 720–737.

Hagiu, A. and D. B. Yoffie (2009), 'What's your Google strategy?,' *Harvard Business Review*, **87**(4), 74–81.

Hagiu, A. and D. B. Yoffie (2013), 'The new patent intermediaries: platforms, defensive aggregators, and super-aggregators,' *Journal of Economic Perspectives*, **27**(1), 45–65.

Heller, M. A. (1998), 'The tragedy of the anticommons: property in the transition from marx to markets,' *Harvard Law Review*, **111**(3), 621–688.

Hess, C. and E. Ostrom (2003), 'Ideas, artifacts, and facilities: information as a common-pool resource,' *Law and Contemporary Problems*, **66**(1/2), 111–145.

Hess, C. and E. Ostrom (2011), 'Introduction: an Overview of the knowledge commons,' in C. Hess and E. Ostrom (eds), *Understanding Knowledge as a Commons*. The MIT Press: Cambridge, MA.

Holmstrom, B. and L. Weiss (1985), 'Managerial incentives, investment and aggregate implications: scale effects,' *Review of Economic Studies*, **52**(3), 403–425.

Holt, T. J. and E. Lampke (2010), 'Exploring stolen data markets online: products and market forces,' *Criminal Justice Studies*, **23**(1), 33–50.

Jacobides, M. G., C. Cennamo and A. Gawer (2018), 'Towards a theory of ecosystems,' *Strategic Management Journal*, **39**(8), 2255–2276.

Koutroumpis, P., A. Leiponen and L. D. W. Thomas (2019), 'The nature of data,' *Innovation and Entrepreneurship Working Papers*, Imperial College Business School: London, UK.

Levin, J. (2003), 'Relational incentive contracts,' *American Economic Review*, **93**(3), 835–857.

Levin, R. C., W. M. Cohen and D. C. Mowery (1984), 'R & D appropriability, opportunity, and market structure: new evidence on some Schumpeterian hypotheses,' *The American Economic Review*, **75**(2), 20–24.

Macneil, I. R. (1978), 'Contracts: adjustment of long-term economic relations under classical, neoclassical, and relational contract law,' *Northwestern University Law Review*, **72**(6), 854–906.

Macneil, I. R. (1985), 'Relational contract: what we do and do not know,' *Wisconsin Law Review*, **1985**(3), 483–526.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers (2011), 'Big data: the next frontier for innovation, competition, and productivity,' in M. G. Institute (ed.), *McKinsey Global Institute Report*, pp. 1–156, https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

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**, 534–584.

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

Menon, S. and S. Sarkar (2016), 'Privacy and big data: scalable approaches to sanitize large transactional databases for sharing,' *MIS Quarterly*, **40**(4), 963–981.

Moiso, C. and R. Minerva (2012), 'Towards a user-centric personal data ecosystem the role of the bank of individuals' data,' Paper presented at the 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications, Beijing, China.

Moreno, A. and C. Terwiesch (2014), 'Doing business with strangers: reputation in online service marketplaces,' *Information Systems Research*, **25**(4), 865–886.

Noorian, Z., J. Iyilade, M. Mohkami and J. Vassileva (2014), 'Trust mechanism for enforcing compliance to secondary data use contracts,' *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 University Press: Cambridge, UK.

Ohm, P. (2010), 'Broken promises of privacy: responding to the surprising failure of anonymization,' *UCLA Law Review*, **57**, 1701.

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

Overby, E. and S. Jap (2009), 'Electronic and physical market channels: a multiyear investigation in a market for products of uncertain quality,' *Management Science*, **55**(6), 940–957.

Parker, G. and M. W. Van Alstyne (2005), 'Two-sided network effects: a theory of information product design,' *Management Science*, **51**(10), 1494–1504.

Parmar, R., I. Mackenzie, D. Cohn and D. M. Gann (2014), 'The new patterns of innovation,' *Harvard Business Review*, **92**(1/2), 86–95.

Pavlou, P. A. and A. Dimoka (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), 392–414.

Pavlou, P. A. and D. Gefen (2004), 'Building effective online marketplaces with institution-based trust,' *Information Systems Research*, **15**(1), 37–59.

Perrin, B., F. Naftalski and R. Houriez (2013), 'Cultural behaviour and personal data at the heart of the big data industry,' in Ernst & Young Report. E&Y and Forum D'Avignon, pp. 1–52, http://ey.com/mediaentertainment.

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

Reinsel, D., J. Gantz and J. Rydning (2017), *Data Age 2025: The Evolution of Data to Life-Critical*. IDC: Framingham, MA. p. 25.

Rochet, J. C. and J. Tirole (2006), 'Two-sided markets: a progress report,' *RAND Journal of Economics*, **37**(3), 645–667.

Romer, P. M. (1990), 'Endogenous technological change,' *Journal of Political Economy*, **98**(**5, Part 2**), S71–102.

Roth, A. E. (2002), 'The economist as engineer: game theory, experimentation, and computation as tools for design economics,' *Econometrica*, **70**(4), 1341–1378.

Roth, A. E. (2008), 'What have we learned from market design?,' *The Economic Journal*, **118**(527), 285–310.

Roth, A. E. (2009), 'What have we learned from market design?,' in J. Lerner and S. Stern (eds), *Innovation Policy and the Economy*, Vol. **9**. The University of Chicago Press: Chicago, IL, pp. 79–112.

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

Schlegel, K., S. Bayerl, S. Zwicklbauer, F. Stegmaier, C. Seifort, M. Granitzer and H. Kosch (2014), 'Trusted facts: Triplifying primary research data enriched with provenance information'. Paper presented at the The Semantic Web: ESWC 2013 Satellite Events, Berlin, Heidelberg.

Schomm, F., F. Stahl and G. Vossen (2013), 'Marketplaces for data: an initial survey,' *SIGMOD Record*, **42**(1), 15–26.

Schwab, K., A. Marcus, J. R. Oyola, W. Hoffman and M. Luzi (2011), *Personal Data: The Emergence of a New Asset Class*. World Economic Forum: Geneva, Switzerland, p. 40.

Shulman, A. (2010), 'The underground credentials market,' *Computer Fraud & Security*, **2010**(3), 5–8.

Simmhan, Y. L., B. Plale and D. Gannon (2005), 'A survey of data provenance in e-science,' *ACM SIGMOD Record*, **34**(3), 31–36.

Soh, C., M. L. Markus and K. H. Goh (2006), 'Electronic marketplaces and price transparency: strategy, information technology, and success,' *MIS Quarterly*, **30**(3), 705–723.

Sweeney, L. (2000), *Uniqueness of Simple Demographics in the US Population. Laboratory for International Data Privacy*. Harvard University: Cambridge, MA.

Teece, D. J. (1986), 'Profiting from technological innovation: implications for integration, collaboration, licensing,' *Research Policy*, **15**(6), 285–305.

Thomas, L. D. W., E. Autio and D. M. Gann (2014), 'Architectural leverage: putting platforms in context,' *Academy of Management Perspectives*, **28**(2), 198–219.

Thomas, L. D. W. and A. Leiponen (2016), 'Big data commercialization,' *IEEE Engineering Management Review*, **44**(2), 74–90.

Tirole, J. (1988), *The Theory of Industrial Organization*. MIT Press: Cambridge, MA.

Tisne, M. (2018), 'It's time for a Bill of Data Rights,' *MIT Technology Review, https://www.technologyreview.com/s/612588/its-time-for-a-bill-of-data-rights/*.

Tiwana, A., B. Konsynski and A. A. Bush (2010), 'Research commentary—platform evolution: coevolution of platform architecture, governance, and environmental dynamics,' *Information Systems Research*, **21**(4), 675–687.

Truong, H.-L., M. Comerio, F. De Paoli, G. R. Gangadharan and S. Dustdar (2012), 'Data contracts for cloud-based data marketplaces,' *International Journal of Computational Science and Engineering*, **7**(4), 280–295.

Uhlir, P. F. and D. Cohen (2011), *Internal Document*. Board on Research Data and Information, Policy and Global Affairs Division, National Academy of Sciences, https://sites.nationalacademies.org/pga/brdi/index.htm.

Wald, J. (2002), 'Legislating the golden rule: achieving comparable protection under the European Union Database Directive,' *Fordham International Law Journal*, **25**(4), 987–1038.

Wareham, J., P. B. Fox and J. L. Cano Giner (2014), 'Technology ecosystem governance,' *Organization Science*, **25**(4), 1195–1215.

Weyl, E. G. (2009), 'Monopoly, Ramsey and Lindahl in Rochet and Tirole (2003),' *Economics Letters*, **103**(2), 99–100.

Zhu, H. and S. E. Madnick (2009), 'One size does not fit all: legal protection for non-copyrightable data,' *Communications of the ACM*, **52**(9), 123–128.

Zuiderwijk, A. and M. Janssen (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. Springer Verlag: Koblenz, Germany: pp. 38–49.