CESifo WORKING PAPERS

7523
2019

February 2019

Data brokers co-opetition

Yiquan Gu, Leonardo Madio, Carlo Reggiani

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7523*
Category 14: Economics of Digitization

Data brokers co-opetition

Abstract

Data brokers collect, manage, and sell customer data. We propose a simple model, in which data brokers sell data to downstream firms. We characterise the optimal strategy of data brokers and highlight the role played by the data structure for co-opetition. If data are "sub-additive", with the combined value lower than the sum of the values of the two datasets, data brokers share data and sell them jointly. When data are "additive" or "supra- additive", with the combined value equal to or greater than the sum of the two datasets, data brokers compete. Results are robust to several extensions.

JEL-Codes: D430, L130, L860, M310.

Keywords: data brokers, personal information, privacy, co-opetition.

Yiquan Gu
University of Liverpool
Management School
Chatham Street
United Kingdom – Liverpool, L69 7ZH
yiquan.gu@liv.ac.uk

Leonardo Madio
Université Catholique de Louvain
Center for Operations Research and Econometrics (CORE)
Voie du Roman Pays, 34 – L1.03.01
Belgium – 1348 Louvain-la-Neuve
leonardo.madio@uclouvain.be

Carlo Reggiani
University of Manchester
School of Social Sciences-Economics
United Kingdom – Manchester, M13 9PL
carlo.reggiani@manchester.ac.uk

February 2019
We are grateful to Paul Belleflamme, Federico Boffa, Emilio Calvano, Alexandre de Cornière, Flavio Delbono, Luca Ferrari, Juan-José Ganuza, Axel Gautier, Andreas Hervas-Drane, Johannes Johnen, Jan Krämer, Fabio Manenti, Francois Maniquet, Andrea Mantovani, Hassan Nosratabadi, David Ronayne, Daniel Schnurr, Juuso Välimäki, Tommaso Valletti, John Van Reenen, Patrick Waelbroeck, alongside seminar participants at CORE UCLouvain, HEC Liége, Liverpool, Manchester, the IX IBEO Workshop on Media Economics (Alghero), the 45th EARIE Conference (Athens), and the 11th TSE Digital Economics Conference, XVII SIEPI Conference (Rome) for useful comments and suggestions. The usual disclaimer applies. Leonardo acknowledges financial support from "MOVE-IN Louvain" Incoming post-doc Fellowship Programme.

"You may not know them, but data brokers know you." – Edith Ramirez, 2014

1 Introduction

If data are considered the fuel of the digital economy, "data brokers" are its catalyst.[1] A report of the US Federal Trade Commission (FTC) defines data brokers as "companies whose primary business is collecting personal information about consumers from a variety of sources and aggregating, analysing, and sharing that information" (Federal Trade Commission, 2014).[2]

This relatively under-explored and fast-growing sector presents several relevant features.[3] First, data brokers operate in the upstream market and do not usually have any contact with final customers, who often ignore their existence. Second, this sector is fairly concentrated and brokers specialise in different segments and services. For instance, companies like Acxiom and Datalogix possess information about millions of consumers worldwide and their data include almost every US consumer. Differently, Corelogic and eBureau focus on property and financial information. Third, they serve downstream firms who use data in a number of ways, including targeted advertising, personalised pricing, risk mitigation, algorithmic learning, product customisation, and other marketing purposes. Tech giants like Facebook, Amazon, Netflix, Google (also known as "Attention brokers") exploit the commercial value of data. However, their primary service is not data collection or data selling (Bergemann and Bonatti, 2019). Data brokers, on the other hand, are less visible entities and are even harder to track on how they use customer data.

A defining characteristic of this sector is that data brokers transact and exchange data with each other and more information is actually obtained this way than from direct sources. Moreover, the category of data they gather and manage is rather heterogeneous, including sensitive identifying information, social media usage, travel, purchase behaviour, and health

[1] The Economist (2017), "Fuel of the future: data is giving rise to a new economy", May 6, 2017.

[2] Although the focus of this report is on the US market, similar considerations also apply to the EU; see e.g., EPDS (2014).

[3] For instance, according to Transparency Market Research, the sector is highly lucrative and is expected to grow at an annual rate of 11.5 until 2026. Available at: https://www.transparencymarketresearch.com/data-brokers-market.html.

conditions.[4] Combining these data can lead to very different market values. In some cases, the overall commercial value is boosted, in others it increases only marginally.

This article delves deeper into the market for data and tackles the following research question: why do data brokers share data in some markets and compete in others? Major challenges in these markets are typically associated with privacy violations and anti-competitive practices stemming from access to data. The former challenge has been widely investigated in the law and economics literature. The latter has received comparatively less attention. Yet, it is relevant as data brokers can exploit their position of data suppliers to extract surplus from other market actors. Data sharing appears prominently as one of these anti-competitive practices.

To explore the above question, we present a simple yet rather general model of the data brokers sector. The economy consists of two data brokers and one downstream firm that supplies a product to customers further down the production chain.[5] The customer level information held by data brokers potentially allows the downstream firm to increase its profits. Data brokers can either share their data and produce a consolidated report or compete to independently supply the downstream firm.[6]

We identify the underlying incentives to exchange data. Specifically, the results crucially depend on the nature of the data structure held by the data brokers. For instance, data may be *sub-additive*. That is, aggregating two datasets leads to a data structure with an information power (i.e., its value to downstream firms) which is lower than the sum of the two separate datasets. This case is likely to arise in presence of overlaps between datasets, diminishing marginal returns of data, or correlated data points. A second possibility

[4]See Federal Trade Commission (2014), Appendix B: Illustrative List of Data Elements and Segments.

[5]This setting does not exclude competition in the downstream market. The only requirement is that selling data to one firm can create value and enhance profitability. For instance, this can happen through several mechanisms. See e.g., Brynjolfsson et al. (2011) on data-driven management, Mikians et al. (2012, 2013), Hannak et al. (2014), Shiller (2014), Dubé and Misra (2017), Dubé et al. (2017) on personalised prices, and Mikians et al. (2012) on search discrimination.

[6]The model presents a "snapshot" of a specific sub-market in which data brokers can supply a specific downstream firm. At the same time, one data broker can be selling or sharing datasets with other partners in other data markets (e.g., different segments, contexts, etc.). In a sense, these repeated interactions make sharing agreements credible and there is no possibility for side-transactions. Moreover, credibility may be enhanced by non-disclosure agreements or by merging datasets through external encrypted clouding services.

can be represented by an *additive* data structure: merging data provides an information power exactly equal to the sum of the two datasets, without adding or losing any value. For instance, one can imagine merging data regarding two different sets of consumers (e.g., young and senior people). A third and final configuration arises when merging two datasets results in a more powerful data structure. For instance, merging the browsing history with email addresses would provide a more detailed picture of the preferences of a certain consumer and providing more targeted offers: data create synergies and become more valuable. We refer to this data structure as *supra-additive*.

Intuitively, one might expect data brokers to have incentives to share data when these are supra-additive because of potential complementaries between datasets. The results from our benchmark model suggest the opposite: the incentive to share data only exists for sub-additive data. The mechanism is as follows: as data are (partially) overlapped, these can be seen as imperfect substitutes from the firm's perspective. This reduces the incentive for the firm to buy both datasets when these are sold independently. As a result, competition between data brokers is fierce and each data broker discounts the overlapped component from its entire value (i.e., its *intrinsic value*) and the firm appropriates some of the extra surplus generated by data. Clearly, data sharing arises as a dominant strategy for the data brokers to soften competition. In all other cases, datasets are not (partial) substitutes and data sharing becomes less appealing.

Overall, depending on the nature of the data, the brokers may compete to supply a client firm in a sub-market and, at the same time, cooperate and share data in another part of the market. In this sense, our model identifies *"co-opetition"* between data brokers as a characterising feature of the sector.

Data sharing may also relax market competition in another scenario, regardless of the data structure. If the firm faces a cost when handling independent datasets on its own, competition between data brokers intensifies. This happens as now each data broker sets a lower price and discounts the cost incurred by the firm. In this way data brokers try to ensure that the firm buys its dataset, when buying also the other. As a result, the comparative disadvantage leads to lower prices and the firm appropriates some extra surplus generated by data. Clearly, data sharing avoids granting the discount. All in all, these results indicate that also in this setting data brokers have

incentives to *share data* to transition from a competitive scenario to one of *co-opetition*, but the rationale now differs from the benchmark case.

The rest of the article is structured as follows. In the next section, we provide an overview of the relevant literature. In Section 3, we present our model of the data brokers sector. In Section 4, we present the main results and intuitions. In Section 5, we explore the implications of different sharing rules and further extend the model in several ways. Section 6 discusses the main managerial and policy implications, and provides some concluding remarks.

2 Related literature

This article aims to shed light on the potential anti-competitive effects of common practices in the data brokers sector. As such, our work relates to several streams of literature.

The recent developments of the digital economy and the volume of data available have further widened the range of activities enabled by these assets. These include risk mitigation (e.g., identity verification and fraud detection), marketing (e.g., customer lists, reports), product customisation, algorithm learning and price discrimination (Belleflamme and Vergote, 2018). Indeed, the literature on price discrimination has traditionally emphasised the role of information and data in enabling the practice that, under certain circumstances, can enhance firm profitability (Varian, 1989; Stole, 2007). Behaviour-based price discrimination involves firms gathering customer information through repeated interactions with them. Such information allows distinguishing past from new consumers and condition their prices accordingly. When done unilaterally, this usually increases firm profitability. Otherwise, it enhances competition (Fudenberg and Tirole, 2000; Villas-Boas, 1999).[7] Differently from our setting, firms create their own information and do not rely on data brokers.

A related and rapidly flourishing literature has tackled the impact of "big data" and consumer privacy on firms' competition (Conitzer et al., 2012; Casadesus-Masanell and Hervas-Drane, 2015; Choi et al., 2018, *inter alia*). Some studies explicitly model data sales. For example, Clavorà Braulin and Valletti (2016) and Montes et al. (2019) show that a data broker always pro-

[7] Fudenberg and Villas-Boas (2006) and Esteves (2010) provide reviews of this literature.

vides data exclusively to one of the downstream competitors. Belleflamme et al. (2017) and Bounie et al. (2018), instead, find that a data broker serves both competing firms, either by selling data with different precision or by ensuring some market segmentation. Gu et al. (2019) consider how access to a list of customers leads to price manipulation and affects the incentive to act as a price leader. A unifying characteristic, however, is that data are held by a unique broker and the emphasis is on the impact of data on downstream competition; we instead delve into the nature of the data brokers' strategies and, in particular, on the incentives for data sharing.

The related issue of strategic information-sharing has been examined in contexts such as oligopolistic competition (Raith, 1996; Kim and Choi, 2010) and financial intermediation (Pagano and Jappelli, 1993; Jappelli and Pagano, 2002; Gehrig and Stenbacka, 2007). Liu and Serfes (2006) study data sharing for price discrimination. They consider a duopoly model in which firms gather their own data and they can opt for a two-way sharing or a one-way data selling. Data selling only arises from the small to the larger firm in presence of sufficiently asymmetric market shares. Shy and Stenbacka (2013, 2016) present related models of competition with switching costs (with costly information acquisition or given customer information, respectively) on the incentives to share customers' data. In all the three latter articles, sharing a full dataset between firms never happens in equilibrium.

Krämer et al. (2019) set up a model in which a general-interest content provider (e.g., a platform like Facebook) offers a social login to special-interest content providers (e.g., a news site). If the social login is offered by the general interest provider and adopted by the special interest website, there are two effects: first, consumers enjoy and value the ease of logging in; second, both providers advertise more effectively due to the shared information. Data sharing is more likely in the presence of high competitive pressure and can lead to a prisoner's dilemma situation for the special interest websites. However, fierce competition can be avoided if the general interest provider does not offer a social login. In their two-sided market setting, all providers gather data and compete, whereas our data brokers serve a firm but are not in direct competition with it. Moreover, our results show instead that it is sharing, rather than avoiding it, that softens competition.

Although our model mainly focuses on mergers of datasets, data sharing can also be achieved through mergers of firms. In a dynamic setting, Esteves

and Vasconcelos (2015) show that a merger between price discriminating firms can be profitable. The reason is that, by merging the databases, the value of information increases, enabling better price discrimination. Their setting resembles our supra-additive case. However, in our benchmark, there is no incentive to share when data are supra-additive. Instead, it is only when data are sub-additive that data brokers merge datasets and relax competition. Kim et al. (2019), on the other hand, consider a spatial competition model in which downstream firms can acquire customer data to perfectly price discriminate. The results of a merger depend on both the downstream (merger to monopoly or duopoly) and upstream market structures. Their main focus is on the effects of a merger between the downstream firms, whereas we focus on data sharing between data brokers.

Prat and Valletti (2018) present a model in which two-sided social media platforms offer ads to consumers sponsored by an incumbent producer or an entrant. They conclude that a significant market overlap between platforms can generate more competition in the advertising market. As a result, social media mergers would lower social welfare, making entry less likely. This mechanism is similar to ours: sub-additive data encompasses the user overlap in their model. However, our results differ in several aspects. First, our study also considers other data structures and indicates that also supra-additive data may deserve attention. Second, anti-competitive effects arise upstream, as data brokers, unlike social media websites, are less visible and do not have direct contact with final users.

To a lesser extent, the issue we tackle shares some similarities with patent pools, i.e., firms agree on sharing patent licenses with one another.[8] This literature has shown that potential anti-competitive or pro-competitive effects may arise depending on whether patents are perfect substitutes or complements (Lerner and Tirole, 2004, 2007), with ambiguous effects in presence of imperfect substitution/complementarity. Our conclusions differ from this literature, as our benchmark model shows that an anti-competitive effect always arises when data are sub-additive, that is, also when data are only partial substitutes. Moreover, data brokers are mostly indifferent between sharing and not sharing data when some form of complementarity exists (i.e., supra-additive data).[9]

[8] For a recent survey, see, e.g., Comino et al. (2019).

[9] More nuanced conclusions, clearly, arise when enriching the model by assuming costly

3 The model

We consider a market with two data brokers, $k = 1, 2$, who possess information on consumers and can sell a dataset to a downstream firm.

3.1 The data brokers

Data brokers (DB) have independently collected data. In the economy, there are $N > 0$ consumers characterised by a set $M > 0$ of attributes (e.g., physical or email addresses, browsing history, etc.). Each DB may have independent access to a subset of attributes of consumers. For instance, this may result from a comparative advantage in different areas or from the different volume of data they gathered. This reflects the large heterogeneity existing in this market. For instance, according to Lambrecht and Tucker (2017), Acxiom has around 1,600 separate data points for 700 million consumers worldwide, whereas the cloud company Bluekai, which specialises in targeted ads, has data on a similar number of customers but with fewer data points (i.e., 10-15 attributes per user).

Let Λ_k be the $M \times N$ logical matrix that represents DB k's information. The element $\lambda_{kji} = 1(0)$ of the logical matrix implies that DB k has (no) information about consumer i's attribute j. These data can be sold to firms operating in the downstream market and give rise to additional surplus for the firm. Denote $f(\cdot)$ a function that maps $M \times N$ logical matrices to real numbers. Then, $f(\Lambda) \geq 0$ measures the extra surplus the firm can generate by using the data contained in Λ compared to a situation in which no data are available (i.e., $f(\mathbf{0}) = 0$). To an extent, this approach is consistent with that of Dalessandro et al. (2014) and the value function $f(\cdot)$ can be interpreted as the monetary evaluation of the dataset from the perspective of the data buyer.

We note that not all datasets are the same and the value of incremental data for a downstream firm is contextual. Consider, for instance, the well known example of the London butcher, Pendleton & Son (Marr, 2016; Claici, 2018). In response to price competition from a large grocery chain, the butcher had to rely on data to improve the product and service. In our setting, the butcher may need the assistance of a DB. The data gathered or owned by the DB on that segment of the market can have a different value

data handling.

for the butcher, say $f(\cdot)$, or for the supermarket, say $g(\cdot)$, with $f(\cdot) \neq g(\cdot)$. Indeed, the same data (or partitions of them, or marketing reports based on them), may have effects that differ across industries and even within industry. By the same token, the same butcher could obtain $h(\cdot) > f(\cdot)$ when using data in presence of an inelastic demand and only $f(\cdot)$ if in direct competition with the grocery chain. In this sense, the model provides a general set-up for any data-driven surplus generation stemming from different types of activities such as advertising, micro-targeting, price discrimination, etc.. This rather general specification encompasses different data structures and market configurations that can be present in an industry.

Data from different sources can be combined in a unique dataset. This assembling process affects the value function $f(\cdot)$ associated with the final dataset depending on the underlying data structure. Formally,

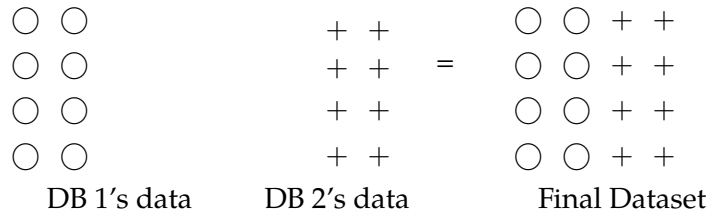
Definition 1. *The data structure is*

- *additive, if* $f(\Lambda_k|\Lambda_{-k}) = f(\Lambda_k) + f(\Lambda_{-k})$,
- *sub-additive, if* $f(\Lambda_k|\Lambda_{-k}) < f(\Lambda_k) + f(\Lambda_{-k})$, *and*
- *supra-additive, if* $f(\Lambda_k|\Lambda_{-k}) > f(\Lambda_k) + f(\Lambda_{-k})$,

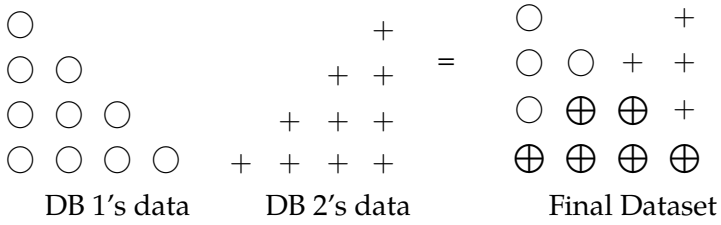
where $|$ *is the element-wise OR operator.*

The above taxonomy can be explained as follows. The data structure is additive when the value of the merged dataset is simply the sum of the individual values of Λ_k and Λ_{-k}. For instance, suppose $N = 4$ consumers populate an economy and each consumer is characterised by $M = 4$ attributes. This situation is illustrated by Figure 1, Panel A. DB 1 has data on all attributes of the first two consumers (e.g., students, represented by the '$\bigcirc$') and DB 2 possesses data regarding all attributes of the other two consumers (e.g., senior citizens, represented by '$+$'). When these datasets are combined, the resulting dataset contains all the relevant attributes of all the consumers. However, since these two segments of the market can be marketed effectively and independently, the value of the final dataset could simply be the sum of the values of the two individual ones.

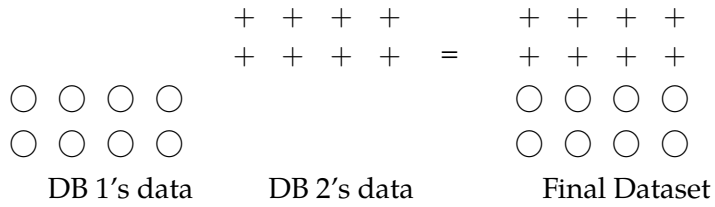
The data structure is sub-additive when the value of the merged dataset is lower than the sum of the values of each dataset Λ_k and Λ_{-k}. For instance, two DBs can have overlapping attributes for some consumers ('$\bigoplus$' in Figure



(A) Additive data structure



(B) Sub-additive data structure



(C) Supra-additive data structure

The figure presents some non-exhaustive examples of data structures with $N = 4$ consumers and $M = 4$ attributes. In example (A), DB 1 has all information on consumers $i = 1, 2$ and DB 2 has all information on consumers $i = 3, 4$. As data are additive, the value of the final dataset is simply the sum of the values of the two independent datasets. In example (B), data are sub-additive: DB 1 has partial information for all consumers and so does DB 2. As some data are owned by both DBs (e.g., both data brokers have information regarding attribute $j = 1$ for all consumers), the value of the final dataset is lower than the sum of the values of the two independent datasets. Overlapped data are reported in the final matrix with $\bigoplus$. In example (C), data are supra-additive. DB 1 possesses information for all consumers regarding attributes $j = 1, 2$ (e.g., browsing history) whereas DB 2 information regarding attributes $j = 3, 4$ (e.g., credit card purchases). Due to synergies across data, the resulting dataset has a greater value than the sum of the values of the two separate datasets.

Figure 1: Examples of Data Structure

1, Panel B). In this case, the value of the final dataset is lower than the sum of the individual values. For instance, many customer data are freely available online as Internet users leave some footprints when using social networks (e.g., Facebook, LinkedIn). As these data are not under exclusive control, DBs can effortlessly collect them, giving rise to potential overlaps when

two different datasets are combined. In an alternative example, this data structure also arises when the marginal contribution to the value of existing data is sufficiently low (Varian, 2018). For instance, Dalessandro et al. (2014) show that the combination of an existing dataset with data from a third party may lead to near-zero marginal contribution, thereby rendering the merger between datasets almost useless. This is consistent with some observations made in Lambrecht and Tucker (2017). The authors suggest that, in some circumstances, adding additional data may be detrimental, and predictions can be made with fewer data points. For instance, some customer attributes can be collinear or positively correlated (see e.g., Bergemann and Bonatti, 2019) and then lead to overlapping insights, whereas in some other cases data can be difficult to integrate (see e.g., health data discussed in Miller and Tucker, 2014).

A final case arises when the data structure is supra-additive. In this case, datasets are complements and their combination returns a final output whose value is higher than the sum of the individual values. There are indeed synergies in the data which lead to the creation of a more informationally powerful dataset. This may happen when the interaction between different types of data plays a crucial role. Offline data and online data seem to have a more significant synergy effect when used together. For example, online purchasing history combined with credit card data collected offline can lead to data complementarity as shown by the recent deal between Mastercard and Google, thereby enabling better and personalised offers.[10] Moreover, a supra-additive data structure may also arise in presence of missing values replaced by predicted ones, that is when a (missing) attribute j for customer i can be inferred from other customers and attributes. An example is the enhanced profiling enabled by matching a physical address with online browsing history, as the former is often correlated with income and other valuable socio-economic characteristics and the latter can predict shopping interests.[11]

DBs make revenues by selling their dataset. In our setting, we consider two ways through which this can happen. First, DBs sell their dataset independently and simultaneously to the downstream firm. In this case, profits

[10] Bloomberg (2018), "Google and Mastercard Cut a Secret Ad Deal to Track Retail Sales", August 30, 2018.

[11] In a recent study, Tucker (2018) reported how, thank to artificial intelligence, algorithms can provide very detailed pictures of individuals when different data are aggregated.

of DB k are given as:

$$\Pi_k = \begin{cases} p_k, & \text{if the firm buys } k\text{'s data,} \\ 0, & \text{if the firm does not buy } k\text{'s data,} \end{cases} \tag{1}$$

where p_k is DB k's price for its own data.

Second, DBs can share their data and construct a unique dataset. In this case, they jointly act as a monopolist data seller for that specific firm and the unique dataset. [12]

The firm buys the merged dataset and each DB obtains a share s_k of the joint profit that reflects individual bargaining power relative to the rival DB. Namely,

$$\Pi_k = s_k(\Lambda_k, \Lambda_{-k}) \cdot P_{\Lambda_k|\Lambda_{-k}}, \tag{2}$$

where s_k is the sharing rule adopted by the two DBs and $P_{\Lambda_k|\Lambda_{-k}}$ is the price jointly set by the two DBs for the merged dataset.

3.2 The firm

In the downstream market, the firm sells a product or service. When the firm does not buy data, it obtains a profit of $\pi^0 > 0$ through usual practice. When it buys data, it extracts an extra surplus $f(\cdot)$ from consumers. For instance, data may engender a market expansion or allow for more sophisticated pricing strategies. The data it can purchase depend on whether DBs have decided to share their datasets or not.

Specifically, when DBs do not share data, the firm's profits are as follows:

$$\Pi^r = \pi^0 + \begin{cases} f(\Lambda_k) - p_k, & \text{if the firm buys } k\text{'s data } \textit{only}, \\ f(\Lambda_k|\Lambda_{-k}) - p_1 - p_2, & \text{if the firm buys data from both.} \end{cases} \tag{3}$$

In the first case, the firm buys data only from one DB $k \in \{1, 2\}$, obtaining $f(\Lambda_k)$ and paying a price p_k. In the second case, the firm buys independently data from both DBs and merges them on its own. We assume that the firm merges the two datasets at no extra cost.[13] In turn, the firm obtains an extra surplus of $f(\Lambda_k|\Lambda_{-k})$ and pays both DBs. We assume that the firm buys data

[12] In another interpretation, also consistent with other model, DBs can be seen as merging into a unique entity (see discussion in Section 6).

[13] We relax this assumption in Section 5.

when it is indifferent between buying or not.

Alternatively, when DBs share their data and sell the merged dataset, the firm obtains the following profits:

$$\Pi^r = \pi^0 + f(\Lambda_k|\Lambda_{-k}) - P_{\Lambda_k|\Lambda_{-k}}, \tag{4}$$

where $f(\Lambda_k|\Lambda_{-k})$ is the value accruing to the firm because of the information contained in $\Lambda_k|\Lambda_{-k}$ and $P_{\Lambda_k|\Lambda_{-k}}$ is the price of the dataset set jointly by the DBs.

3.3 Timing

The timing is as follows. In the first stage, DBs simultaneously and independently decide whether or not to share their data on the basis of the sharing rule. Data sharing arises if and only if this strictly increases both DBs' profits. In the second stage, DBs jointly or independently set the price(s) for the dataset(s) conditional on the first stage outcome. Then, the firm decides whether or not to buy the offered dataset(s).

4 Analysis

We first consider the subgame when two DBs sell their data independently. We then investigate the case when the two DBs share and merge their datasets.

4.1 Independent data selling

Consider the second stage of the game when DBs simultaneously and independently set a price for their datasets. After observing the prices p_k, the firm decides whether to buy, and from whom, the dataset(s). That is, the firm maximises (3). We state the following results:

Proposition 1. *(i) If the data structure is additive, there exists a unique Nash equilibrium in this subgame in which $p_k^* = f(\Lambda_k)$, for $k = 1, 2$.*

(ii) If the data structure is sub-additive, there exists a unique Nash equilibrium in this subgame in which $p_k^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k})$, for $k = 1, 2$.

(iii) If the data structure is supra-additive, any pair of (p_1^*, p_2^*)*, such that* $p_1^* + p_2^* = f(\Lambda_1|\Lambda_2)$ *and* $p_k^* \geq f(\Lambda_k)$*, for* $k = 1, 2$*, constitutes a Nash equilibrium in this subgame.*

Proof. see Appendix A.1. □

The rationale of the above results is as follows. First, consider an additive data structure where the marginal value of DB k's data is independent of whether or not the firm has the other DB's data. Hence, the firm buys from DB k if, and only if, $f(\Lambda_k) \geq p_k$. In the unique equilibrium, $p_k = f(\Lambda_k)$, for $k = 1, 2$, and the firm buys both datasets and merge them on its own. The final price paid by the firm is equal to $p_k + p_{-k} = f(\Lambda_k) + f(\Lambda_{-k})$. Both DBs are indeed able to extract all the extra surplus they generate. The firm earns a profit of $\Pi^r = \pi^0$. Profits of DBs are exactly equal to the values of their respective dataset.

Second, consider a sub-additive data structure. There exists a unique equilibrium where the DBs set the following prices $p_k^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k})$. These prices reflect the marginal contribution of each DB to the value of the merged dataset, which is then used for surplus extraction by the firm. In this case, the firm is indifferent between buying from both DBs or from a single DB, so it buys from both. The firm's profits are $\Pi^r = \pi^0 + f(\Lambda_k) + f(\Lambda_{-k}) - f(\Lambda_k|\Lambda_{-k})$.

Finally, consider a supra-additive data structure. In this case, any pair of (p_k, p_{-k}) such that $p_k + p_{-k} = f(\Lambda_k|\Lambda_{-k})$ and $p_k \geq f(\Lambda_k)$ for $k = 1, 2$, constitutes an equilibrium, and hence, we have multiplicity of equilibria. In any of them, the firm buys data from both DBs and merge them on its own. It follows that all surplus generation enabled by data is fully extracted by DBs and the firm obtains a profit of $\Pi^r = \pi^0$.

The above discussion leads to the following corollary.

Corollary 1. *The firm's profits and the* combined profits *of the two DBs are as follows.*

(i) If the data structure is additive, $\Pi^r = \pi^0$ *and* $\Pi_1 + \Pi_2 = f(\Lambda_1) + f(\Lambda_2) = f(\Lambda_1|\Lambda_2)$*.*

(ii) If the data structure is sub-additive, $\Pi^r = \pi^0 + f(\Lambda_1) + f(\Lambda_2) - f(\Lambda_1|\Lambda_2) > \pi^0$ *and* $\Pi_1 + \Pi_2 = 2f(\Lambda_1|\Lambda_2) - f(\Lambda_1) - f(\Lambda_2) < f(\Lambda_1|\Lambda_2)$*.*

(iii) If the data structure is supra-additive, $\Pi^r = \pi^0$ *and* $\Pi_1 + \Pi_2 = f(\Lambda_1|\Lambda_2)$.

Taken together, Proposition 1 and Corollary 1 offer interesting insights. When the data structure is either additive or supra-additive, DBs can *independently* set their own prices to extract the entire surplus from the firm. When data are sub-additive, the downstream firm still buys data from both DBs but the data market is more *competitive*. Prices are set at a lower level relative to the intrinsic value of each dataset (i.e., $p_k^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) < f(\Lambda_k)$). Each DB sets a price equal to its marginal contribution to the final dataset. Hence, for the offer to be attractive for the firm, each DB has to discount the "overlapping" component. As a result, the firm partially appropriates the surplus generated by data. Paradoxically, when buying from both DBs, the firm would prefer to have redundant, less powerful data than more powerful ones. In the latter situation, the overall surplus is higher but the firm can appropriate some only in the former.

4.2 Data sharing

Consider the subgame when DBs share their data. In this case, they act as an exclusive supplier to the firm. This implies that the same dataset cannot be sold individually by any of the two parties.[14] They jointly make a take-it-or-leave-it offer to the firm. It follows then, the total profits the DBs can extract is $f(\Lambda_k|\Lambda_{-k})$, and individual profits are

$$\Pi_k = s_k(\Lambda_k, \Lambda_{-k}) \cdot f(\Lambda_k|\Lambda_{-k}),$$

where s_k is the sharing rule.

4.3 Data brokers' decision

In the first stage, data brokers make their decision on whether or not to share their data. We assume that they will share the data if and only if both agree. It implies that individual profits should be larger than those obtained when selling data independently. We let the sharing rule be

$$s_k(\Lambda_k, \Lambda_{-k}) = \frac{f(\Lambda_k)}{f(\Lambda_k) + f(\Lambda_{-k})}.$$

[14] For instance, data can be protected by non-disclosure agreements or DBs share data through an encrypted cloud.

This rule associates a bargaining power to each DB that reflects the relative value of its dataset. Note, however, that this does not represent their relative contribution to the merged dataset. An alternative sharing rule (the Shapley value implementation) is presented in Section 5. We state the following results.

Proposition 2. *A strictly positive incentive to share data by both DBs only exists when the data structure is sub-additive. In all other cases, there is no Pareto improvement for the DBs to share data.*

When the data structure is additive or supra-additive and data are sold independently, DBs already grab entirely the surplus generated by data for the firm (Corollary 1). It follows that sharing data does not bring about a Pareto improvement for the DBs. Hence, merging the two datasets would never result in a *strictly* higher joint payoff. Note that this result is compatible with any degree of asymmetry in the upstream market for data, i.e., $f(\Lambda_1) \neq f(\Lambda_2)$.

On the other hand, when the data structure is sub-additive, DB k obtains $p_k^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k})$ when sells its data independently. The difference in DB k's profits between data sharing and independent selling is

$$\begin{aligned}\Pi_k - p_k^* &= s_k(\Lambda_k, \Lambda_{-k}) \cdot f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k|\Lambda_{-k}) + f(\Lambda_{-k}) \\ &= f(\Lambda_{-k}) - s_{-k}(\Lambda_k, \Lambda_{-k}) \cdot f(\Lambda_k|\Lambda_{-k}) > 0,\end{aligned}$$

for $k = 1, 2$. Hence, DBs always find it optimal to share data. This is because sharing allows for a surplus extraction that they would otherwise fail to fully implement with independent selling. Specifically, the price jointly set by the DBs selling the shared data is $p^* = f(\Lambda_k|\Lambda_{-k})$ and this allows for a situation of Pareto improvement relative to when they compete.

When data are sub-additive, the actual competition in the market forces DBs to reduce their price relative to their intrinsic value. In other words, this type of data structure spurs market competition and DBs can only partially appropriate some of the surplus accruing to the firm. Taken together Proposition 1 and Proposition 2, indicate that forward looking DBs can anticipate such an outcome by sharing their data and hence soften the competition. This is as if they collude at the expenses of downstream firms and consumers.

A rather counter-intuitive result emerges from the above discussion. At first, one may expect that an incentive to share data would emerge when the

data structure is supra-additive. For instance, combining email addresses (or postal codes) with the browsing history would provide the two data brokers with a powerful information set to sell in the market for data. On the other hand, intuition may suggest that when data partially overlap, the incentive to share could decrease as the incremental benefit of the rival's DB is lower.

Our model leads to different conclusions. When the merged dataset can provide a comprehensive picture of consumers (i.e., supra-additive data), each DB is aware of the relative importance of its piece of information. Accordingly, each DB sets a price which has two properties. First, this price induces the firm to buy the data. Second, the price extracts the highest value from the firm. As DBs together extract the entire surplus from the firm, they are indifferent between sharing and competing to serve the firm.

When the data structure is sub-additive, the marginal contribution of each dataset to the merged dataset is lower than its standalone value and the firm can be in the condition to buy only from one. This forces DBs to engage in a fiercer competition and discount the evaluation of the overlapped component from its own price. Hence, sharing data would avert the *price war* and fully extract the surplus from the firm.

5 Extensions

5.1 Shapley Value implementation

To address the robustness of our results and intuitions, we consider a different specification for the sharing rule function s_k. We assume that the sharing rule follows the Shapley value implementation. The implementation of this rule captures the average marginal contribution of an agent to a given coalition. Assuming the merge dataset generates a total profit of $f(\Lambda_k|\Lambda_{-k})$ for the two DBs, the Shapley value of DB k, Sh_k, is

$$Sh_k = \frac{1}{2}\Big\{f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) + f(\Lambda_k)\Big\}. \tag{5}$$

The following result can be stated:

Proposition 3. *The implementation of the Shapley value as a sharing rule strictly supports data sharing when the data structure is sub-additive. In all other cases, data brokers are at best indifferent between sharing and competing.*

Proof. see Appendix A.2. □

The above proposition confirms in full the results presented in Section 4. Data sharing is an optimal strategy for DBs for sub-additive data structure as it softens competition in the downstream market.

5.2 Sequential pricing decision

We now investigate whether there is an incentive to share data when DBs set their prices sequentially. The timing is changed as follows. In the first stage of the game, DB k sets p_k. In the second stage, DB $-k$ sets p_{-k}. Given the resulting prices, the firm decides whether to buy the dataset(s) and from which seller.

Regardless of the order of moves, our main findings and intuitions remain unaltered. Data sharing emerges as a dominant strategy only with a sub-additive data structure to soften price competition. In all other cases, DBs have no strict incentives to share data. Interestingly, a first-mover advantage is identified with a supra-additive data structure, which leads to the possibility of naturally selecting one equilibrium from the multiplicity identified in the benchmark. See Appendix A.3 for a detailed proof of these results.

5.3 Sequential selling decision

Next, we discuss the sequential sale of datasets. The timing is as follows. In the first stage, DB k sets p_k. The firm decides to buy DB k's data or not. Then, DB $-k$ sets p_{-k} and the firm makes its purchasing decision.

No matter what happens in the first stage, DB $-k$ always has an incentive to sell its dataset because otherwise it gets zero profit. That is, DB $-k$'s best response is $p_{-k} = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k)$ if the firm has bought data from DB k in the first stage and $p_{-k} = f(\Lambda_{-k})$ if the firm has not bought data in the first stage. Anticipating this, DB k sets $p_k = f(\Lambda_k)$ to make the firm indifferent between buying and not buying in the first stage. To summarise, in the unique subgame perfect Nash equilibrium, $p_k^* = f(\Lambda_k)$, $p_{-k}^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k)$, and the firm's profits are $\Pi^r = \pi^0$.

This translates into a first-mover advantage when data are sub-additive, as the second-mover only gets its marginal contribution to the final dataset. In contrast, there is a second-mover advantage when data are supra-additive,

because in this case $p_{-k} = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k)$ which DB $-k$ can obtain in the second stage (when the firm has already bought k's data) is now larger than its intrinsic value $f(\Lambda_{-k})$. With additive data, instead, each DB sets a price equal to its intrinsic evaluation. These results indicate that, in all cases, the firm is left with no extra-surplus, $\Pi^r = \pi^0$, and data sharing never emerges as one DB would always lose out.

At first, this case seems to differ significantly from the benchmark case as together DBs are always able to extract all surplus from the firm. However, with sub-additive data, the incentive to data sharing may be restored if the timing of the move is endogenised. In fact, if DBs are allowed to choose whether to sell their data early or late, the first-mover advantage leads data brokers to move simultaneously. As a result, an incentive to share data emerges again as in the benchmark.

5.4 Costly data processing

In this subsection we relax the assumption that integrating, merging and managing the two datasets is costless for both DBs and the firm. In the real world, DBs may have a comparative advantage in this practice vis-à-vis most downstream firms.[15] Hence, we assume that the firm incurs an additional cost c when merging datasets on its own.[16] We present the following result:

Proposition 4. *When the firm incurs a cost for merging datasets on its own, the price competition between data brokers intensifies and there is an incentive to share data regardless of the data structure.*

Proof. see Appendix A.4. □

When the firm buys the datasets independently and then merges them on its own at a cost c, DBs take this into account when setting prices. This intensifies the competition as each DB offers a discount c to ensure that the firm buys also its own data, given that it is buying them from the rival. As a result, in equilibrium both DBs discount the merging cost c. This implies

[15] There is a large heterogeneity in the capabilities of firms to handle datasets. For many firms, this could involve costly activities as hiring external staff and/or run into the risk of obtaining sub-optimal output relative to a professional DB. In a similar manner, a firm may not have the same data analytics skills that DBs usually have.

[16] Clearly, the cost c cannot be prohibitive. See also Appendix A.4.

that the firm obtains:

$$\Pi^r = \pi^0 + c, \tag{6}$$

when the data structure is additive or supra-additive. In contrast, the firm obtains the following profits when data are sub-additive:

$$\Pi^r = \pi^0 + f(\Lambda_k) + f(\Lambda_{-k}) - f(\Lambda_k|\Lambda_{-k}) + c, \tag{7}$$

where the term $f(\Lambda_k) + f(\Lambda_{-k}) - f(\Lambda_k|\Lambda_{-k})$ indicates the gain that the firm obtains as a result of the competition generated by overlaps (see Corollary 1).

Interestingly, a more intense competition is not only typical of a sub-additive data structure. The above expressions show that regardless of the data structure the competitive disadvantage of the firm in handling data becomes a source of extra surplus (as shown in equations 6 and 7) as a result of a fiercer competition between the DBs.

As DBs are only partially able to extract the surplus that the dataset creates for the firm, data sharing represents a dominant strategy. This happens for two reasons. First, because of the apparent comparative advantage in merging data. Second, because of softened competition, which in this case is present under all data structures as a result of not granting a discount to the firm. In this sense, an additional mechanism through which data sharing arises is highlighted, which differs from the one identified in the benchmark model.

This result has interesting implications for firms. As the merging cost stimulates more competition in the upstream market, a firm may have an incentive to increase its costs. For instance, this may imply a firm holding on investments in data analytics or outsourcing this expertise.

6 Conclusion and discussion

This article sheds light on the quite obscure and relatively unexplored market for data. We present a model of data brokers and study their role as suppliers of valuable information to firms. A distinctive aspect of the sector, clearly transpiring from the Federal Trade Commission (2014)'s report, is the exchange and trade of data *between* brokers. We delve into the nature of "co-opetition" between data brokers and link their strategic incentives in

data sharing to different types of data structure.

We set-up a relatively simple but fairly general model of the sector in which two data brokers can supply a downstream firm with valuable information. This scenario is compatible with an economy where there are several markets for data, and different data brokers compete or have different partners in each sub-market. The firm can increase its profits by acquiring valuable customer level information. In this setting, we compare a scenario in which the data brokers sell their information independently, to one in which they can share their data and sell a unique consolidated dataset to the downstream firm.

We highlight how the incentives for data sharing are crucially related to the nature of the customer level information held by the brokers. Specifically, we find that data sharing can arise for two reasons.

First, when data are sub-additive, DBs compete fiercely due to the rather homogeneous nature of their datasets (e.g., overlapping information). Data sharing is valuable to both DBs as it softens such competition. Perhaps counter-intuitively, no incentive to share data arises when there are synergies between datasets (i.e., supra-additive data structure). We note that these findings are also consistent with the management literature on co-opetition, which has long held that companies may be collaborators with respect to value creation but become competitors when it comes to value capture (e.g., Nalebuff and Brandenburger, 1997).

Second, when the firm faces a cost in merging datasets, DBs would have to provide a discount to the firm if they sell their data independently. It follows that data sharing avoids duplicating the discount and enables a full extraction of the extra surplus generated by data. To summarise, whereas data sharing arises in both scenarios, the *drivers* are very different. Sharing incentives relate to either overlapping datasets, or asymmetric efficiency in handling data, or both.

Finally, we note that these conclusions are robust to the implementation of other sharing rules.

6.1 Managerial implications

This analysis has several important managerial implications. Our theoretical analysis rationalises the large heterogeneity in the contractual arrangements in this market, as also illustrated by the Federal Trade Commission (2014)'s

report.

We highlight that such heterogeneity is linked to the nature and value of the data. There is misalignment between the commercial value and the intrinsic value of data. This is because the intrinsic value of an output data (for data brokers) is not always identical to the value of data as an input (for the firm). Notably for managerial implications, this misalignment gives rise to different bargaining powers. For instance, the commercial value of data can be less than the intrinsic value of data when these are sub-additive. To partially mitigate the commercial devaluation of data, data brokers can follow two strategies. First, they can share data and be in a situation of Pareto improvement for both. This allows brokers to extract a surplus which they would fail to retain due to market competition. Second, sequential decisions may be helpful. For instance, one DB can gain by (partially or completely) internalising the externalities by acting as a leader when *setting prices* sequentially or as a follower when *selling data* sequentially.

For a data broker client, our results provide two rather counter-intuitive implications. First, a manager may prefer to buy "lower quality" (e.g. sub-additive, with overlapping information) data. This happens because competition between brokers intensifies and the firm can retain some of the surplus produced through the data. Second, costly data processing may prove to be an advantage as both data brokers grant the firm a discount. Of course, these conclusions hold provided that data brokers cannot share their data.

6.2 Policy implications

Given the importance of data brokers in today's digital economy and the high dynamism of the sector, our findings bear several relevant policy implications.[17]

First, so far the policy agenda has emphasised the customers protection and individual rights aspects of data brokerage, e.g., cyber-security and privacy regulation. Less attention has been devoted to the potential anti-competitive nature of data sharing. Our work suggests that data brokerage should be monitored by antitrust authorities. One implication of the model is that data sharing may be associated with the exertion of market power to downstream players. In some cases, data sharing does not bring about

[17] On the relevance of data brokerage in the current business environment, see e.g. Gartner (2016).

any additional surplus for the data brokers with respect to competition. For instance, when data are additive or supra-additive and the firm does not bear any merging cost. In some others, data sharing is a potentially anti-competitive practice and allows to soften the competition so as to extract all surplus from downstream firm(s). For instance, when data are sub-additive or the firm faces a cost to merge data bought independently in the market. Note that, in our simplified model, this anti-competitive behaviour does not generate inefficiency but losses may arise in richer environments. On top of that, regulators may be concerned about the reallocation of surplus across sides of the market.

Second, our model is also consistent with an interpretation of data sharing as a merger between two platforms gathering data about consumers and selling them to other firms (e.g., firms, advertisers). Our results can be therefore linked to the current discussion on attention brokers and the 2012 merger between Instagram and Facebook (see, e.g., Wu 2018, Prat and Valletti 2018). These suggest that a merger between two attention brokers could result in anti-competitive practices whenever there are significant overlaps in the customer data (e.g., sub-additive data resulting from customer multi-homing) or when data are supra-additive but downstream firms (or advertisers) may face a merging cost when buying these independently. Hence, antitrust enforcers should carefully take into account the nature of the data structure when scrutinising merger proposals.

Third and last, we shall note that the European Union and the United States have followed different regulatory approaches on how data should be managed by data brokers, third-parties and firms. Despite the initial attention (e.g., Federal Trade Commission, 2014), in the US there are only weak requirements when dealing with personal information (e.g., employment, credit, insurance), which can be shared and integrated without restrictions. The European Union has tackled the issue of privacy more strictly. The new EU General Data Protection Regulation (GDPR) has strengthened the conditions for consent by consumers, who need to be explicitly informed about the final use of the data collected. In other words, data sharing among different data brokers without the ex-ante authorisation of consumers is deemed illegal, to the point that such regulation is often emphatically evoked as the "death of third-party data".[18] In the light of our analysis, the EU GDPR may

[18]See, e.g., Wired (2018), "Forget Facebook, mysterious data brokers are facing GDPR

have some unintended effects for data brokers as, for example, generating a pro-competitive effect in the upstream market. Specifically, the need of the explicit consent of the consumers to data sharing should reduce the prevalence of this practice, with the further consequence of enabling firms to partially retain some of the data generated surplus.

A Appendix

A.1 Proof of Proposition 1

(i) Additive data structure. The firm buys from DB k, if, and only if, $f(\Lambda_k) \geq p_k$. Thus there exists a unique Nash equilibrium in which $p_k^* = f(\Lambda_k)$, for $k = 1, 2$. The firm buys both datasets and pays a combined price $p_1^* + p_2^* = f(\Lambda_1) + f(\Lambda_2) = f(\Lambda_k|\Lambda_{-k})$. Profits are $\Pi_k = f(\Lambda_k)$, and $\Pi^r = \pi^0$.

(ii) Sub-additive data structure. We first consider DB 1's best responses. Suppose $p_2 > f(\Lambda_2)$. In this case, the firm does not buy Λ_2 alone. DB 1 then has two ways of selling Λ_1. One is to set $p_1 = f(\Lambda_1)$ and the other is to set $p_1 = f(\Lambda_1|\Lambda_2) - p_2$. Since $f(\Lambda_1) > f(\Lambda_1|\Lambda_2) - f(\Lambda_2) > f(\Lambda_1|\Lambda_2) - p_2$, DB 1's best response is the former.

Now consider $f(\Lambda_1|\Lambda_2) - f(\Lambda_1) < p_2 \leq f(\Lambda_2)$. A DB again has two ways of selling Λ_1. The first is to set a price slightly lower than $f(\Lambda_1) - f(\Lambda_2) + p_2$ so that the firm finds it strictly better to buy Λ_1 alone than either buying Λ_2 alone or buying both. The other is to set it at $f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$, so that the firm finds buying both is at least as good as buying Λ_2 alone. Given the range of p_2 in this case, the former is better for DB 1. However, there exists no best response because no highest price that is strictly lower than $f(\Lambda_1) - f(\Lambda_2) + p_2$ can be found.

Finally, consider $p_2 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_1)$. DB 1 has the same two ways of selling Λ_1. However, now setting $p_1 = f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$ and let the firm buy both is better for DB 1 as p_2 is now lower. To summarise, DB

trouble", November 8, 2018.

1's best response function is

$$BR_1(p_2) = \begin{cases} f(\Lambda_1) & \text{if } p_2 > f(\Lambda_2) \\ \emptyset & \text{if } f(\Lambda_1|\Lambda_2) - f(\Lambda_1) < p_2 \leq f(\Lambda_2) \\ f(\Lambda_1|\Lambda_2) - f(\Lambda_2) & \text{if } p_2 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_1) \end{cases}. \tag{8}$$

Similarly, DB 2's best response function is

$$BR_2(p_1) = \begin{cases} f(\Lambda_2) & \text{if } p_1 > f(\Lambda_1) \\ \emptyset & \text{if } f(\Lambda_1|\Lambda_2) - f(\Lambda_2) < p_1 \leq f(\Lambda_1) \\ f(\Lambda_1|\Lambda_2) - f(\Lambda_1) & \text{if } p_1 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_2) \end{cases}. \tag{9}$$

By superimposing (8) and (9), one verifies that there exists a unique Nash equilibrium in which $p_k^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k})$ for $k = 1, 2$. Hence, the firm buys from both DBs and profits are $\Pi_k = p_k^*$ and $\Pi^r = \pi^0 + f(\Lambda_1) + f(\Lambda_2) - f(\Lambda_1|\Lambda_2)$.

(iii) Supra-additive data structure. As before, we first consider DB 1's best responses. Suppose $p_2 > f(\Lambda_1|\Lambda_2) - f(\Lambda_1)$, DB 2 cannot sell Λ_2 alone, whereas DB 1 has two ways of selling Λ_1. The first is to set $p_1 = f(\Lambda_1)$. The second is to set $p_1 = f(\Lambda_1|\Lambda_2) - p_2$ so that the firm buys both datasets. Given the range of p_2 in this case, the former is better for DB 1 and hence its best response is $p_1 = f(\Lambda_1)$.

Now consider $f(\Lambda_2) \leq p_2 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_1)$. Again DB 2 cannot sell Λ_2 alone, whereas DB 1 has two ways of selling Λ_1. In this case, however, $p_1 = f(\Lambda_1|\Lambda_2) - p_2$ is DB 1's best response as p_2 is now lower.

Finally, consider $p_2 < f(\Lambda_2)$. DB 1 can either set a price slightly lower than $f(\Lambda_1) - f(\Lambda_2) + p_2$ to beat what DB 2 alone can offer, or $f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$ so that the firm finds buying from both is better than buying from DB 2 alone. Given the range of p_2 in this case, $f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$

is strictly better. The below equation summarises the analysis:

$$BR_1(p_2) = \begin{cases} f(\Lambda_1) & \text{if } p_2 > f(\Lambda_1|\Lambda_2) - f(\Lambda_1) \\ f(\Lambda_1|\Lambda_2) - p_2 & \text{if } f(\Lambda_2) \leq p_2 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_1) \\ f(\Lambda_1|\Lambda_2) - f(\Lambda_2) & \text{if } p_2 < f(\Lambda_2) \end{cases}. \tag{10}$$

DB 2's best response function can be similarly constructed. With the best response functions, it is easy to verify that any pair of (p_1^*, p_2^*) such that $p_1^* + p_2^* = f(\Lambda_1|\Lambda_2)$ and $p_k^* \geq f(\Lambda_k)$ for $k = 1, 2$, constitutes a Nash equilibrium. The firm buys from both DBs, and the profits are $\Pi_k = p_k^*$, $\Pi_{-k} = f(\Lambda_1|\Lambda_2) - p_k^*$, where $p_k^* \in [f(\Lambda_k), f(\Lambda_1|\Lambda_2) - f(\Lambda_{-k})]$, and $\Pi^r = \pi^0$.

A.2 Proof of Proposition 3

We need to show that also under the Shapley value sharing rule there is an incentive to share data when the data structure is sub-additive, whereas DBs are indifferent in the other cases. To do so, we provide a comparison between DBs profits when sharing, Sh_k, and the profits obtained when not sharing data. Specifically,

(i) Sub-additive data structure. By Proposition 2, under price competition DB k obtains $p_k = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k})$. By using the Shapley value implementation, DB k finds it optimal data sharing if:

$$\begin{aligned} Sh_k = \frac{1}{2}\Big\{ f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) + f(\Lambda_k) \Big\} \geq & f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) \\ & f(\Lambda_{-k}) + f(\Lambda_k) \geq f(\Lambda_k|\Lambda_{-k}). \end{aligned} \tag{11}$$

The above condition is satisfied strictly for both DBs with sub-additive data. Hence, a sharing rule implementing the Shapley value always supports an equilibrium with data sharing.

(ii) Additive data structure. When independent selling, DB k sets a price equal to $p_k = f(\Lambda_k)$. Sharing data with Shapley value implementation

dominates price competition provided that:

$$\begin{aligned} Sh_k = \frac{1}{2}\Big\{ f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) + f(\Lambda_k) \Big\} \geq & f(\Lambda_k) \\ f(\Lambda_k|\Lambda_{-k}) \geq & f(\Lambda_k) + f(\Lambda_{-k}). \end{aligned} \tag{12}$$

There is a condition of indifference between sharing and not sharing data. Hence, data sharing never arises as strictly dominant.

(iii) Supra-additive data structure. With independent data selling, equilibrium profits are (p_1^*, p_2^*) such that $p_1^* + p_2^* = f(\Lambda_1|\Lambda_2)$ and $p_k^* \geq f(\Lambda_k)$ for $k = 1, 2$. On the other hand, $Sh_1 + Sh_2 = f(\Lambda_1|\Lambda_2)$. That is, if $Sh_k > p_k^*$, then $Sh_{-k} < p_{-k}^*$. Hence, with supra-additive data, DBs are, at best, indifferent between sharing data and selling them independently: data sharing never (strictly) emerges in equilibrium.

A.3 Sequential Pricing: Proof

Consider an additive data structure. Both DBs sell their data for sure whenever $f(\Lambda_k) \geq p_k$. Hence, no matter the timing, setting $p_k^* = f(\Lambda_k), k = 1, 2$ constitutes the unique Nash equilibrium of the sequential pricing game. As $f(\Lambda_k|\Lambda_{-k}) \geq p_k^* + p_{-k}^*$, both datasets are sold together and the firm obtains $\Pi^r = \pi^0$. DBs are indifferent between independent selling and data sharing. Hence, data sharing does not arise in equilibrium.

Consider a sub-additive data structure. For sake of simplicity and without loss of generality, let us identify the DB as follows: $k = 1$ and $-k = 2$. In the second stage, DB 2's best response function is given in (9). This means, DB 1 can only sell its data by setting $p_1 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$ and hence, in the unique SPNE, $p_1^* = f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$. By (9), $p_2^* = f(\Lambda_1|\Lambda_2) - f(\Lambda_1)$. As a result, the firm is left with a positive extra surplus $\Pi^r = \pi^0 + f(\Lambda_k) + f(\Lambda_{-k}) - f(\Lambda_k|\Lambda_{-k})$. Data sharing unambiguously dominates independent data selling.

Consider a supra-additive data structure. In the second stage, DB 2's best response function is, as in (10),

$$BR_2(p_1) = \begin{cases} f(\Lambda_2) & \text{if } p_1 > f(\Lambda_1|\Lambda_2) - f(\Lambda_2) \\ f(\Lambda_1|\Lambda_2) - p_1 & \text{if } f(\Lambda_1) \leq p_1 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_2) \\ f(\Lambda_1|\Lambda_2) - f(\Lambda_1) & \text{if } p_1 < f(\Lambda_1) \end{cases}. \tag{13}$$

Thus, DB 1 can sell its dataset in the first stage only by setting $p_1 \leq f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$ and hence, in the unique SPNE, $p_1^* = f(\Lambda_1|\Lambda_2) - f(\Lambda_2)$. By (13), $p_2^* = f(\Lambda_2)$. As a result, the firm pays a combined price $p_1^* + p_2^* = f(\Lambda_1|\Lambda_2)$ and obtains $\Pi^r = \pi^0$. The unique SPNE is characterised by a first-mover advantage for DB 1 and, at best, indifference between data sharing and independent selling can be achieved.

A.4 Proof of Proposition 4

Suppose that, other things equal, the firm faces a merging cost c. We let $c \leq \min\{f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}), f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k)\}$. This assumption ensures that, regardless of the data structure, there is always an incentive for each DB to sell its dataset.[19]

Consider an additive data structure. To ensure that the firm buys its dataset, DB k sets a price such that $f(\Lambda_k) - c \geq p_k$. The unique Nash equilibrium is $p_k^* = f(\Lambda_k) - c$, for any $k = 1, 2$. The firm buys data from both DBs and pays a combined price $p_k^* + p_{-k}^* = f(\Lambda_k) + f(\Lambda_{-k}) - 2c$. Profits of DB k are $\Pi_k^* = f(\Lambda_k) - c$, whereas the profits of the firm are $\Pi^r = \pi^0 + c$. Data sharing is a dominant choice as it avoids granting this discount.

Consider a sub-additive data structure. Suppose DB $-k$ charges a sufficiently high price $p_{-k} > f(\Lambda_{-k})$, the firm only buys from DB k at a price $p_k = f(\Lambda_k)$, and there is no market for DB $-k$. Suppose DB $-k$ charges now an intermediate price such that $f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k) - c < p_{-k} \leq f(\Lambda_{-k})$, then there is no best response for DB k (see also Appendix A.1). Finally, suppose the price set by DB $-k$ is sufficiently low, i.e., $p_{-k} \leq f(\Lambda_1|\Lambda_{-k}) - f(\Lambda_k) - c$, under the above assumptions on sufficiently low merging costs, DB k can sell its dataset by setting a price $p_k \in [0, f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) - c]$. As a result, $p_k^* = f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) - c$, so the combined price paid by the firm which is also the combined profits of the DBs, is $p_k^* + p_{-k}^* = 2f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k) - f(\Lambda_{-k}) - 2c$. The profits of the firm are $\Pi^r = \pi^0 + f(\Lambda_k) + f(\Lambda_{-k}) - f(\Lambda_k|\Lambda_{-k}) + c$. It follows that data sharing can generate higher profits for the DBs.

Finally, consider a supra-additive data structure. The best responses for

[19] Note that for an additive and supra-additive data structure, this assumption can be mildly relaxed, i.e., $c \leq \min\{f(\Lambda_k), f(\Lambda_{-k})\}$.

DB k are $BR_k(p_{-k}) =$

$$\begin{cases} f(\Lambda_k) & \text{if } p_{-k} \geq f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k) - c \\ f(\Lambda_k|\Lambda_{-k}) - p_{-k} - c & \text{if } f(\Lambda_{-k}) \leq p_{-k} < f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k) - c \\ f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) - c & \text{if } p_{-k} < f(\Lambda_{-k}) \end{cases}. \tag{14}$$

As in the benchmark case, the best responses are computed by considering the set of alternatives available to the firm: buying nothing, buying only from k, buying only from $-k$, or buying from both. In particular, when $p_{-k} \geq f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k) - c$, DB $-k$ practically prices itself out, and DB k can sell its data at $p_k = f(\Lambda_k)$. The firm only buys from k. Suppose now $f(\Lambda_{-k}) \leq p_{-k} < f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_k) - c$, DB k makes it indifferent for the firm to buy both or nothing at all. Hence, DB k's best response is $f(\Lambda_k|\Lambda_{-k}) - p_{-k} - c$. Finally, when $p_{-k} < f(\Lambda_{-k})$, DB k needs to ensure that the firm remains indifferent between buying only from $-k$ or both. It follows that DB k's best response is $f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) - c$.

Given these, the Nash equilibria involve setting prices such that $p_k^* + p_{-k}^* = f(\Lambda_k|\Lambda_{-k}) - c$ with $p_k^* \in [f(\Lambda_k), f(\Lambda_k|\Lambda_{-k}) - f(\Lambda_{-k}) - c]$. Hence, as compared to the benchmark case, the total profits of the DBs is lowered by c. In contrast, by merging their data the DBs can restore full surplus exaction, $f(\Lambda_k|\Lambda_{-k})$.

References

Belleflamme, P., Lam, W. M. W., and Vergote, W. (2017). Price discrimination and dispersion under asymmetric profiling of consumers. *AMSE - Document de travail 2017-13*.

Belleflamme, P. and Vergote, W. (2018). The intricate tale of demand and supply of personal data. *Concurrences*, 3:45–50.

Bergemann, D. and Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, 11:1–23.

Bloomberg (2018). Google and Mastercard cut a secret ad deal to track retail sales. https://www.bloomberg.com/news/articles/2018-08-30/google-and-mastercard-cut-a-secret-ad-deal-to-track-retail-sales. [Last accessed February 13, 2019].

Bounie, D., Dubus, A., and Waelbroek, P. (2018). Selling strategic information in competitive markets. *CESifo Working Paper No. 7078*.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: How does data-driven decision making affect firm performance? *SSRN Working Paper 1819486*.

Casadesus-Masanell, R. and Hervas-Drane, A. (2015). Competing with privacy. *Management Science*, 61(1):229–246.

Choi, J. P., Jeon, D.-S., and Kim, B.-C. (2018). Privacy and personal data collection with information externalities. *TSE Working Paper 887*.

Claici, A. (2018). Big data and competition policy. In *Economic Analysis of the Digital Revolution*. Funcas, Madrid.

Clavorà Braulin, F. and Valletti, T. (2016). Selling customer information to competing firms. *Economics Letters*, 149:10–14.

Comino, S., Manenti, F. M., Thumm, N., et al. (2019). The role of patents in information and communication technologies ICTs. a survey of the literature. *Journal of Economic Surveys*, forthcoming.

Conitzer, V., Taylor, C. R., and Wagman, L. (2012). Hide and seek: Costly consumer privacy in a market with repeat purchases. *Marketing Science*, 31(2):277–292.

Dalessandro, B., Perlich, C., and Raeder, T. (2014). Bigger is better, but at what cost? estimating the economic value of incremental data assets. *Big data*, 2(2):87–96.

Dubé, J.-P., Fang, Z., Fong, N., and Luo, X. (2017). Competitive price targeting with smartphone coupons. *Marketing Science*, 36(6):944–975.

Dubé, J.-P. and Misra, S. (2017). Scalable price targeting. *NBER Working Paper 23775*.

EPDS (2014). Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy. *Preliminary Opinion of the European Data Protection Supervisor*.

Esteves, R.-B. (2010). Pricing with customer recognition. *International Journal of Industrial Organization*, 28(6):669–681.

Esteves, R.-B. and Vasconcelos, H. (2015). Price discrimination under customer recognition and mergers. *Journal of Economics & Management Strategy*, 24(3):523–549.

Federal Trade Commission (2014). Data Brokers: A Call for Transparency and Accountability. May, 2014.

Fudenberg, D. and Tirole, J. (2000). Customer poaching and brand switching. *RAND Journal of Economics*, pages 634–657.

Fudenberg, D. and Villas-Boas, J. M. (2006). Behavior-based price discrimination and customer recognition. *Handbook on Economics and Information Systems*, 1:377–436.

Gartner (2016). How to choose a data broker. https://www.gartner.com/smarterwithgartner/how-to-choose-a-data-broker. [Last accessed February 13, 2019].

Gehrig, T. and Stenbacka, R. (2007). Information sharing and lending market competition with switching costs and poaching. *European Economic Review*, 51(1):77–99.

Gu, Y., Madio, L., and Reggiani, C. (2019). Consumer information, price competition and market leadership. *Mimeo*.

Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 305–318.

Jappelli, T. and Pagano, M. (2002). Information sharing, lending and defaults: Cross-country evidence. *Journal of Banking & Finance*, 26(10):2017–2045.

Kim, B.-C. and Choi, J. P. (2010). Customer information sharing: Strategic incentives and new implications. *Journal of Economics & Management Strategy*, 19(2):403–433.

Kim, J.-H., Wagman, L., and Wickelgren, A. (2019). The impact of access to consumer data on the competitive effects of horizontal mergers. *Journal of Economics and Management Strategy*, forthcoming.

Krämer, J., Schnurr, D., and Wohlfarth, M. (2019). Winners, losers, and facebook: The role of social logins in the online advertising ecosystem. *Management Science*, forthcoming.

Lambrecht, A. and Tucker, C. E. (2017). Can big data protect a firm from competition? *CPI Antitrust Chronicle*, 1(1).

Lerner, J. and Tirole, J. (2004). Efficient patent pools. *American Economic Review*, 94(3):691–711.

Lerner, J. and Tirole, J. (2007). Public policy toward patent pools. *Innovation Policy and the Economy*, 8:157–186.

Liu, Q. and Serfes, K. (2006). Customer information sharing among rival firms. *European Economic Review*, 50(6):1571–1600.

Marr, B. (2016). *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results*. John Wiley & Sons.

Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. (2012). Detecting price and search discrimination on the internet. *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, pages 79–84.

Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. (2013). Crowd-assisted search for price discrimination in e-commerce: First results. *Proceedings of the 9th ACM conference on Emerging networking experiments and technologies*, pages 1–6.

Miller, A. R. and Tucker, C. (2014). Health information exchange, system size and information silos. *Journal of Health Economics*, 33:28–42.

Montes, R., Sand-Zantman, W., and Valletti, T. (2019). The value of personal information in online markets with endogenous privacy. *Management Science*, forthcoming.

Nalebuff, B. J. and Brandenburger, A. M. (1997). Co-opetition: Competitive and cooperative business strategies for the digital economy. *Strategy & leadership*, 25(6):28–33.

Pagano, M. and Jappelli, T. (1993). Information sharing in credit markets. *Journal of Finance*, 48(5):1693–1718.

Prat, A. and Valletti, T. M. (2018). Attention oligopoly. *Mimeo*.

Raith, M. (1996). A general model of information sharing in oligopoly. *Journal of Economic Theory*, 71(1):260–288.

Shiller, B. R. (2014). First-degree price discrimination using big data. *Brandeis University Working Paper 58*.

Shy, O. and Stenbacka, R. (2013). Investment in customer recognition and information exchange. *Information Economics and Policy*, 25(2):92–106.

Shy, O. and Stenbacka, R. (2016). Customer privacy and competition. *Journal of Economics & Management Strategy*, 25(3):539–562.

Stole, L. A. (2007). Price discrimination and competition. *Handbook of Industrial Organization*, 3:2221–2299.

The Economist (2017). Fuel of the future: data is giving rise to a new economy. https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy. [Last accessed February 13, 2019].

Tucker, C. (2018). Privacy, algorithms, and artificial intelligence. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Varian, H. R. (1989). Price discrimination. *Handbook of Industrial Organization*, 1:597–654.

Varian, H. R. (2018). Artificial intelligence, economics, and industrial organization. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Villas-Boas, J. M. (1999). Dynamic competition with customer recognition. *RAND Journal of Economics*, pages 604–631.

Wired (2018). Forget Facebook, mysterious data brokers are facing GDPR trouble. https://www.wired.co.uk/article/gdpr-acxiom-experian-privacy-international-data-brokers. [Last accessed February 13, 2019].

Wu, T. (2018). Blind spot: The attention economy and the law. *Antitrust Law Journal*, 82.