Working Paper in Economics

# 202101

January 2021

Data Brokers Co-Opetition

Yiquan Gu
Leonardo Madio
Carlo Reggiani

https://www.liverpool.ac.uk/management/people/economics/
© authors

Data brokers co-opetition[*]

Yiquan Gu[†] Leonardo Madio[‡] Carlo Reggiani[§]

January 2021

Abstract

Data brokers share consumer data with rivals and, at the same time, compete with them for selling. We propose a "co-opetition" game of data brokers and characterise their optimal strategies. When data are "sub-additive" with the merged value net of the merging cost being lower than the sum of the values of individual datasets, data brokers are more likely to share their data and sell them jointly. When data are "super-additive", with the merged value being greater than the sum of the individual datasets, competition emerges more often. Finally, data sharing is more likely when data brokers are more efficient at merging datasets than data buyers.

Keywords: data brokers, consumer information, co-opetition, data sharing.
JEL codes: D43, L13, L86, M31.

1 Introduction

In today's highly digitised economy, data have become particularly valuable and have attracted the attention of policymakers and institutions. To mention some examples, in

[*] We are grateful to Paul Belleflamme, Nina Bobkova, Marc Bourreau, Emilio Calvano, Elias Carroni, Alexandre de Cornière, Flavio Delbono, Antoine Dubus, Juan-José Ganuza, Axel Gautier, Andreas Hervas-Drane, Johannes Johnen, Yassine Lefouili, Jan Krämer, Fabio Manenti, Andrea Mantovani, David Ronayne, Lorien Sabatino, Daniel Schnurr, Juuso Välimäki, Tommaso Valletti, John Van Reenen, Wouter Vergote, Patrick Waelbroeck, alongside participants at the IX IBEO Workshop on Media Economics (Alghero), the 45th EARIE Conference (Athens), the 11th TSE Digital Economics Conference, the X NERI Workshop (Rome), the 9th Florence School of Regulation Seminar on Media and the Digital Economy, the 2019 EEA-ESEM Conference (Manchester), the 2019 IIOC (Boston), the 2020 MACCI/EPoS Virtual IO Seminars, and seminar participants at different universities, for useful comments and suggestions. The usual disclaimer applies. Leonardo acknowledges financial support from the "MOVE-IN Louvain" Incoming Fellowship Programme during his stay at the Université Catholique de Louvain.

[†] Email: yiquan.gu@liv.ac.uk. Address: University of Liverpool, Management School, Chatham Street, Liverpool, L69 7ZH, UK.

[‡] Email: leonardo.madio@unipd.it. Address: University of Padova, Department of Economics and Management, Via del Santo, 33, 35123 Padova, Italy. Also affiliated with CESifo.

[§] Email: carlo.reggiani@manchester.ac.uk. Address: University of Manchester, School of Social Sciences-Economics, Manchester, M13 9PL, UK.

2018 the EU General Data Protection Regulation (GDPR) to protect personal data was promulgated, and the State of California followed suit with the California Consumer Privacy Act. In 2020, the European Commission announced the EU Data Strategy (EU, 2020) to boost data sharing among firms and the recently proposed Digital Market Act includes mandatory data sharing as a crucial competition tool. The conventional view is that being non-rival, data can generate positive externalities, and the EU data strategy's vision is that data sharing has to be incentivised or even mandated.

If data are considered the fuel of the digital economy, "data brokers" are its catalyst.[1] These often unknown actors are "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) and engage mostly in business-to-business relations. As they do not usually have any contact with final consumers, the latter are often unaware of their existence. A defining characteristic of this sector is that data brokers transact and exchange data with each other and more information is obtained this way than from direct sources. The Federal Trade Commission (2014) reports that seven out of nine data brokers were buying and selling consumer data to each other. For instance, Acxiom has partnerships with other data brokers, including Corecom (specialised in entertainment data) and Nielsen (a global data company).

Yet, these sharing practices might not necessarily be consistent with the positive social role envisioned in the current regulatory debate and, more worryingly, may hide anti-competitive behaviours. As little is known about the behaviours of these data brokers, investigations worldwide are taking place. For instance, the French authority CNIL carried out an in-depth investigation in the period 2017-2019 auditing fifty data brokers and ad-tech companies (Financial Times, 2019).

In this context, our main research question is to identify the incentives of data brokers to share data in some markets and compete in others and how these relate to the nature of the data a data broker has. This is relevant as, on the one hand, these companies compete to provide customers with specialised data, analytics, and market research; on the other hand, they also cooperate through partnerships and data sharing agreements. Moreover, data brokers may be particularly strong in different areas and specialise in some services, rendering the nature and type of data crucial for their strategies. For example, Acxiom and Datalogix profile consumers for targeting purposes, collecting information such as demographics, sociographics, and purchasing behaviours. Data brokers like Corelogic and eBureau mostly sell in-depth financial and property data analytics.

To this end, we present a simple yet rather general model to analyse how the nature of data and merging costs shape data brokers' decisions. Our economy consists of two data brokers, and one data buyer who supplies a product or a service to consumers

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

downstream. Throughout the paper, we use “the (data) buyer” and “the downstream firm” interchangeably. The consumer level information held by data brokers potentially allows the downstream firm to increase its profits in its own market. For instance, a firm can use data to facilitate targeted advertising, to engage in price discrimination, or to adopt data-driven management practices.[2] Data brokers, on the other hand, can either share data and produce a consolidated report or compete to independently supply the downstream firm. If the data brokers share data, they incur an upstream merging cost. If the data brokers compete and the buyer acquires both datasets, then the buyer needs to merge them incurring a downstream merging cost.

We find that the underlying incentives to engage in either data sharing or competition crucially depend on whether the value of the merged dataset, net of the merging costs, shows forms of complementarities or substitutabilities. Indeed, data may be *super-additive* when combining two data sources, net of the merging costs, results in a more valuable dataset than the sum of the individual components. Combining the browsing history with email addresses, for example, would provide a detailed picture of the preferences of a certain consumer and enable targeted offers. In this example, data create synergies and become more valuable when merged.

Data are *sub-additive* when aggregating two datasets leads to a new value, net of the merging costs, that is lower than the sum of the two separate datasets. For examples, datasets might present overlapping information, diminishing marginal returns of data, correlated data points, or high merging costs. Finally, when combining two different data sources is extremely costly, a sharp reduction in the merged dataset’s net value may occur. This represents a case of *extreme sub-additivity* and the value of the merged dataset is lower than the stand-alone value of its components.

Data sharing arises for two main reasons. First and foremost, to soften competition between data brokers; second, to enable data brokers to internalise potential merging cost inefficiencies on the buyer’s side. The balance of these two effects drives our results. The former contrasts with the pro-competitive vision of data sharing, whereas the latter is consistent with the socially valuable perspective permeating the regulatory debate.

Suppose data brokers are more efficient than the buyer in handling data. Then, when the data structure is sub-additive or extreme sub-additive both effects favour sharing. By merging sub-additive datasets, data brokers can avoid granting the buyer the discount that results from competition and reflects the overlapping information and the buyer’s merging cost. In the presence of an extreme sub-additive data structure resulting from a high merging cost, the mechanism is similar: as the buyer is only interested in one dataset,

[2] Note that our stylised setting does not exclude competition in the product market. Essentially, we assume that consumer level data creates extra value for the downstream firm and enhances its profitability in a given market environment.

sharing avoids an intense, Bertrand-like, competition. When data complementaries are present, there exists a multiplicity of equilibria under competition and these render sharing less likely to occur: one data broker may prefer to veto a sharing agreement when it expects to grab a larger share of the surplus than the sharing rule prescribes.

However, not always are data brokers more efficient than buyers in merging datasets. For example, as a former partnership between Facebook and Acxiom suggests, a tech company may acquire information from data brokers, and the former can be more efficient in handling data, given its expertise and computational capabilities.[3] In this case, the cost internalisation incentive is clearly not present. However, an incentive to share data does exist when the value of the combined dataset is limited. Specifically, sharing avoids fierce competition when the datasets are extreme sub-additive. When instead the datasets are sub-additive, the two forces driving the incentives to share are now in contrast. On the one hand, data brokers may be willing to share to soften competition and avoid discounting the overlapping component of the datasets. On the other hand, independent selling avoids the high merging cost facing the data brokers.

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

Our modelling of data intermediaries is consistent with some distinguishing characteristics of the data market. First, our model captures that the value of data is contextual. For example, the same two datasets can be substitutes or complements depending on their final use and downstream market circumstances (Sarvary and Parker, 1997). While our model abstracts away from the specifics of the downstream market and sheds light on both substitute and complementary data, it is compatible with a market where data brokers repeatedly interact to supply downstream buyers in different sub-markets and with buyer-specific projects. Second, combining and sharing data sources can be substantially more costly than bundling other products. This highlights a crucial difference between data, that can be merged and disposed, and product bundling.[4] For instance, merging datasets requires resource intensive preparation of the data, and this may result in a very low net value of the final dataset. We highlight the importance of merging costs in shaping the data market outcome and characterise conditions for sharing to emerge in the unique subgame perfect Nash equilibrium. Finally, we discuss the possibility of data partitioning as, unlike many other products, a data broker may be able to partly control the potential complementarity and substitutability when selling data.

[3] This partnership was in place between 2015 and 2018 (Acxiom, 2015).

[4] For the potential anti-and pro-competitive effects of bundling see, e.g., Choi (2008).

Contribution to the literature. This article focuses on the market for data and the role of data intermediaries. The main contribution of our paper is to capture the co-existence of competition and co-opetition between data brokers, and identify the determinants of the transition between these. The closest papers to ours are Sarvary and Parker (1997), Bergemann et al. (2020) and Ichihashi (2020a). Sarvary and Parker (1997) focus on the incentives of information sellers (e.g., consultancy, experts) to sell reports about uncertain market conditions to downstream firms, interested in finding the real state of the world. A crucial role is played by the reliability of information, data complementarity or substitutability. In our framework, complementarity and substitutability are mediated by the presence of downstream and upstream merging costs, and data refer to individual characteristics rather than their reliability about the correct state of the world.

Instead, Bergemann et al. (2020) and Ichihashi (2020a) analyse competition between data brokers in obtaining data from consumers which can then be sold downstream. Similarly to ours, Ichihashi (2020a) considers a setting in which data intermediaries compete to serve a downstream with consumer data. However, he focuses on the welfare implications of data collection, whereas we explicitly study the incentives of data sharing and its implications for market actors.

Other studies have concentrated on related issues as privacy violations and anti-competitive practices stemming from access to data (Conitzer et al., 2012; Casadesus-Masanell and Hervas-Drane, 2015; Clavorà Braulin and Valletti, 2016; Choi et al., 2019; Montes et al., 2019; Belleflamme et al., 2020; Gu et al., 2019; Bounie et al., 2020; Ichihashi, 2020b, *inter alios*), strategic information sharing and signal jamming in oligopoly (Vives, 1984; Raith, 1996; Kim and Choi, 2010) and more recently, the impact of data-driven mergers (Kim et al., 2019; Prat and Valletti, 2019; Chen et al., 2020; De Cornière and Taylor, 2020).

Our study also contributes to the recent law and economics literature on data sharing. In line with recent regulatory developments, this literature takes a mostly favourable view of the practice, based on the premise that, from a social perspective, there is not enough data sharing. For example, in Prüfer and Schottmüller (2017), data sharing might prevent tipping outcomes in data-driven markets. Graef et al. (2018) argue that the right to data portability, which enhances personal data sharing, should be seen as a new regulatory tool to stimulate competition and innovation in data-driven markets. Borgogno and Colangelo (2019) underline that data sharing via APIs requires a costly implementation process and to leverage their pro-competitive potential a regulatory intervention is necessary. Our results, instead, point to the possibility of excessive data sharing, through a harmful use of data to soften competition between data holding firms. This adds to other negative aspects of data sharing, as the overutilisation of data pools or the reduced incentives for data gathering (Graef et al., 2019; Martens et al., 2020).

To a lesser extent, the issue we tackle shares similarities with patent pools (Lerner and Tirole, 2004, 2007) and how substitutability/complementarity might engender anti- or pro-competitive effects. In our framework, merging costs play an important role and interact with other forces in inducing data sharing. Moreover, a relevant difference between data and patent pools is that the latter can be considered as a structured combination of ideas whereas the former is a factor of production (Jones and Tonetti, 2020).

Outline. The rest of the paper is organised as follows. Section 2 outlines the model. Our main results are presented in Section 3. Section 4 explores several extensions to our main model and Section 5 concludes with final remarks. A microfoundation of the data structure and all proofs can be found in the Appendix.

2 The model

The data brokers. Consider an economy with two data brokers, $k = 1, 2$, who are endowed with data on different individuals and attributes. Each data broker (DB) may have independent access to a subset of the attributes.[5]

To fix ideas, let Λ_k be the $M \times N$ logical matrix that represents DB k's information, where N is the number of consumer profiled and M their attributes. Denote a function $f(\Lambda) \geq 0$ that measures the extra surplus the buyer in question can generate by using the data contained in Λ, compared to a situation in which no data are available (i.e., $f(\mathbf{0}) = 0$). The value function $f(\cdot)$ can be interpreted as the monetary evaluation of the dataset from the perspective of the data buyer.

Data from different sources can be combined in a single dataset. This assembling process affects the value of the final dataset depending on the underlying data structure, as defined below. In the *absence of merging costs*, a data structure is super-additive if $f(\Lambda_k|\Lambda_{-k}) \geq f(\Lambda_k) + f(\Lambda_{-k})$ and sub-additive if $f(\Lambda_k|\Lambda_{-k}) < f(\Lambda_k) + f(\Lambda_{-k})$, where $|$ is the element-wise OR operator.[6] In the following, for ease of notation, we use f_k to refer to $f(\Lambda_k)$ and f_{12} for $f(\Lambda_1|\Lambda_2)$.

The data structure identifies a continuum of cases depending on the value of the merged dataset. It is super-additive when datasets are complements and their combination returns a final output whose value is at least as large as the sum of the individual components. 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

[5] For instance, this may result from a comparative advantage in different areas or from the different volumes of data they gathered. For more details, see, e.g., Lambrecht and Tucker (2017).

[6] More details about the microfoundation of the data structure can be found in Appendix A.1.

different types of data plays a crucial role. 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.[7]

The data structure is sub-additive when the value of the merged dataset is lower than the sum of the values of individual datasets but is at least as large as either of the individual datasets. This happens when the two merging datasets have overlapping information.

The data structure is extreme sub-additive when the value of the merged dataset is lower than the value of an individual dataset. For instance, Dalessandro et al. (2014) suggest that, in some circumstances, adding additional data may be detrimental, and predictions can be made with fewer data points. This is consistent with the seminal findings of Radner and Stiglitz (1984) who show theoretically that information can have a negative marginal net value. Moreover, some customer attributes can be collinear or positively correlated (see, e.g., Bergemann and Bonatti, 2019) and then lead to overlapping insights, whereas in other cases data can be difficult to integrate (see, e.g., health data in Miller and Tucker, 2014). Similar decreasing returns to scale are present in the recent literature on algorithms (Bajari et al., 2019; Claussen et al., 2019; Schaefer and Sapi, 2020).

Data brokers obtain revenues by selling their dataset. This can happen in two ways. First, data brokers can sell their own dataset independently and simultaneously to the buyer. DB k's profit is then

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

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

Alternatively, data brokers can share their data and sell a single dataset. In this case, they jointly act as the unique data seller and make a take-it-or-leave-it offer to that specific buyer. In case of a sale, their joint profit is $P_{12} - c_{db}$, where P_{12} identifies the price jointly set by the two data brokers, and $c_{db} > 0$ is the data brokers' merging cost in the upstream. Let $s_k \in [0,1]$ be k's share of the joint profit given by an exogenously fixed sharing rule. For our main analysis, we use a proportional sharing rule, to be specified in Section 3.3.1, that reflects the data brokers' respective bargaining power. However, other desirable sharing rules, such as the Shapely value sharing rule can also be accommodated. We discuss this possibility in Section 4.1. DB k's individual profit when sharing is then

$$\Pi_k = \begin{cases} 0 & \text{if the downstream firm does not buy the merged data} \\ s_k \cdot (P_{12} - c_{db}) & \text{if the downstream firm buys the merged data} \end{cases}. \tag{2}$$

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

The data buyer. When data brokers do not share data, the buyer's profits are as follows:

$$\Pi^b = \pi^0 + \begin{cases} 0 & \text{if the downstream firm does not buy data} \\ f_k - p_k & \text{if the downstream firm buys } k\text{'s data } \textit{only} \\ f_{12} - p_k - p_{-k} - c_b & \text{if the downstream firm buys data from both} \end{cases}, \quad (3)$$

where π^0 is the profit the buyer can make without data and c_b is the buyer's downstream merging cost.

Alternatively, when data brokers share their data and sell the merged dataset, the buyer obtains the following profit:

$$\Pi^b = \pi^0 + \begin{cases} 0 & \text{if the downstream firm does not buy the merged data} \\ f_{12} - P_{12} & \text{if the downstream firm buys the merged data} \end{cases}. \quad (4)$$

Timing. The timing of the game is as follows. In the first stage, the two data brokers simultaneously and independently decide whether or not to share their data. Data sharing arises if, and only if, both data brokers choose to share data. In the second stage, data brokers jointly or independently set the price(s) for the dataset(s). Then, in the third stage, the buyer decides whether or not to buy the offered dataset(s). The equilibrium concept is Subgame Perfect Nash Equilibrium (SPNE).

3 Analysis

Before the analysis is presented, we first need to define the data structure taking into account the merging cost, occurring either at the upstream (data brokers) or the downstream (the buyer) level. That is, our definition focuses on the net value of the final dataset when two different data sources are combined.

Assume, without loss of generality, that $f_2 \geq f_1$. We categorise the data structure as follows:

Definition 1. *Under a given downstream merging cost c_b facing the buyer, the data structure is*

- *downstream super-additive, if $f_{12} - c_b \geq f_1 + f_2$,*
- *downstream sub-additive, if $f_2 \leq f_{12} - c_b < f_1 + f_2$, and finally*
- *downstream extreme sub-additive, if $f_{12} - c_b < f_2$.*

The corresponding upstream *data structure can be analogously defined by replacing c_b by c_{db}.*

We note that the net benefit entailed by the combination of two datasets does not necessarily mirror the data structure in the absence of merging costs. For instance, a super-additive data structure without a merging cost may result in an extreme sub-additive data structure if the sharing activity takes place and its related cost is extremely high.

3.1 Independent data selling

We solve the game by backward induction. First, consider a second stage subgame where at least one data broker has decided not to share data in the first stage and hence they simultaneously and independently set a price for their own data.

After observing the prices (p_1, p_2), the downstream firm decides whether to buy, and from whom, the dataset(s) so to maximise its profit (3). This gives rise to the demand and revenue facing each data broker for any given strategy profile (p_1, p_2).

Proposition 1. (i) *If the data structure is downstream super-additive, any pair of (p_1^*, p_2^*), such that $p_1^* + p_2^* = f_{12} - c_b$ and $p_k^* \geq f_k$, for $k = 1, 2$, constitutes a Nash equilibrium in this subgame. The downstream firm buys both datasets and merge them.*

(ii) *If the data structure is downstream sub-additive, there exists a unique Nash equilibrium in this subgame in which $p_k^* = f_{12} - c_b - f_{-k}$, for $k = 1, 2$. The downstream firm buys both datasets and merge them.*

(iii) *If the data structure is downstream extreme sub-additive, there exists a unique Nash equilibrium in this subgame in which $p_1^* = 0$ and $p_2^* = f_2 - f_1$. The downstream firm does not merge the two datasets even when it buys both.*

Proof. see Appendix A.2. □

The rationale of the above results is as follows. First, consider the data structure is downstream super-additive. In this case, the two datasets are characterised by strong synergies and complementarities persist even when considering merging costs c_b. This implies that rather than trying to pricing the rival out, each data broker prefers the rival to sell its dataset too. This way, each data broker hopes to appropriate some of the (positive) externalities the datasets produce downstream. As a result, in equilibrium the buyer acquires data from both data brokers and merge them on its own.

We note that in this case of downstream super-additivity, there is a continuum of competitive equilibria in which the data brokers always extract the entire surplus from the buyer, i.e., $\Pi_k^* + \Pi_{-k}^* = f_{12} - c_b$. This leaves the buyer 0 net benefit. Note also that the merging cost that the downstream firm faces is passed upstream because, in any equilibrium, the downstream firm will pay no more than $f_{12} - c_b$ in total.

Consider now the case where merging two datasets leads to downstream sub-additivity. In contrast to the super-additivity case, the data brokers prefer undercutting the rival than accepting its own marginal value to the rival's dataset, an observation common in Bertrand type price competition models. As a result, the unique equilibrium in (ii) emerges. Note that even if the downstream merging cost was negligible, the prices set by the data brokers are limited by the substitutability of the datasets when the structure is sub-additive (e.g., overlapping information or high correlation between datasets).

In equilibrium, the buyer purchases from both data brokers and pays a composite price of $p_1^* + p_2^* = 2f_{12} - f_1 - f_2 - 2c_b$, with a net benefit of $f_1 + f_2 - f_{12} + c_b > 0$. As a result, the buyer is better off: in competition, data brokers have to discount the merging costs, which are incurred by the buyer only once, and also the overlapping component.

Finally, merging costs can be large for the buyer such that the data structure gets extreme sub-additive. This implies that combining different data sources becomes less appealing and the buyer would only need the most valuable dataset. Under the assumption of $f_2 \geq f_1$, only DB 2 sells its data in equilibrium for sure. Its equilibrium price in this case equals the difference in the datasets' intrinsic values, whereas the rival is forced to set a zero price, as a result of competition. The buyer obtains a net benefit of f_1.

The following corollary summarises the downstream firm's surplus and, for comparison, the industry profit of the data brokers.

Corollary 1. (*i*) *If the data structure is downstream super-additive,* $\Pi^b = \pi^0$ *and* $\Pi_1^c + \Pi_2^c = f_{12} - c_b$.

(*ii*) *If the data structure is downstream sub-additive,* $\Pi^b = \pi^0 + f_1 + f_2 - f_{12} + c_b$ *and* $\Pi_1^c + \Pi_2^c = 2f_{12} - f_1 - f_2 - 2c_b$.

(*iii*) *If the data structure is downstream extreme sub-additive,* $\Pi^b = \pi^0 + f_1$ *and* $\Pi_1^c + \Pi_2^c = f_2 - f_1$,

where Π_k^c *denotes DB k's profit under competition.*

Figure 1 illustrates the buyer's surplus in relation to the gross value of the merged dataset, f_{12}. It is clear from the figure, and surprisingly, that the buyer is weakly worse off as the value of the merged dataset increases. It starts off with a positive net benefit of f_1 when the datasets are downstream extreme sub-additive and ends up with zero net surplus in the case of downstream super-additivity. Remarkably, the more synergy between the individual datasets, the worse it is for the downstream firm.

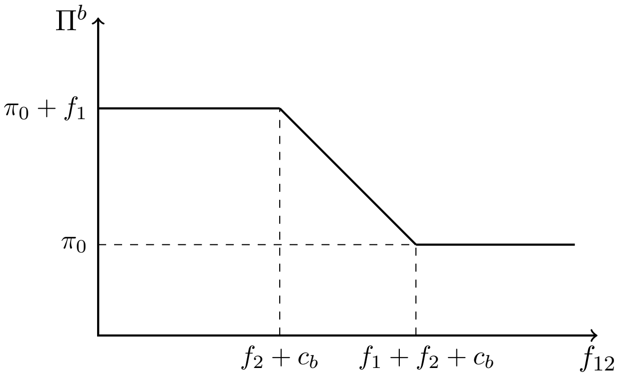


Figure 1: The data buyer's surplus and the value of the merged dataset in the absence of a merging cost, f_{12}.

3.2 Data sharing

Consider the subgame when both data brokers agreed to share their data. In this case, they act as an exclusive supplier to the downstream firm for its specific project.[8] As they jointly make a take-it-or-leave-it offer to the buyer, if the data structure is upstream super- or sub- additive, the total profit the data brokers can obtain is $f_{12} - c_{db}$. If the data structure is upstream extreme sub-additive, data brokers would not proceed to merging the datasets and simply sell the most valuable one to the buyer, jointly obtaining f_2. To sum up, under a given sharing rule s_k individual data broker's profits are, for $k = 1, 2$,

$$\Pi_k^s = s_k \cdot \max\{f_{12} - c_{db}, f_2\}, \tag{5}$$

where Π_k^s denotes DB k's profit under data sharing.

3.3 Data brokers' decision

We now analyse data brokers' decision on data sharing. Figure 2 presents the normal form representation at the first stage of the game. To simplify the presentation, we assume $|c_b - c_{db}| \leq f_1$. That is, we exclude the less relevant cases where the cost difference is larger than the value of DB1's dataset.[9]

For data sharing to occur as a SPNE, the joint profit of the data brokers when sharing their data has to be no less than those under competition, i.e., $\Pi_1^s + \Pi_2^s \geq \Pi_1^c + \Pi_2^c$. Otherwise, sharing cannot be a mutual best response at the first stage.

[8] Being an exclusive supplier of data for a specific project implies that the same dataset cannot be sold individually by any of the two parties. For instance, data can be protected by non-disclosure agreements or data brokers share data through an encrypted cloud or a sandbox (OECD, 2013, p.33).

[9] If $|c_b - c_{db}| > f_1$, DB1 is very much disadvantaged and cooperation becomes a moot point.

		DB 2	
		Share	Compete
DB 1	Share	Π_1^s, Π_2^s	Π_1^c, Π_2^c
	Compete	Π_1^c, Π_2^c	Π_1^c, Π_2^c

Figure 2: The normal form game at the first stage

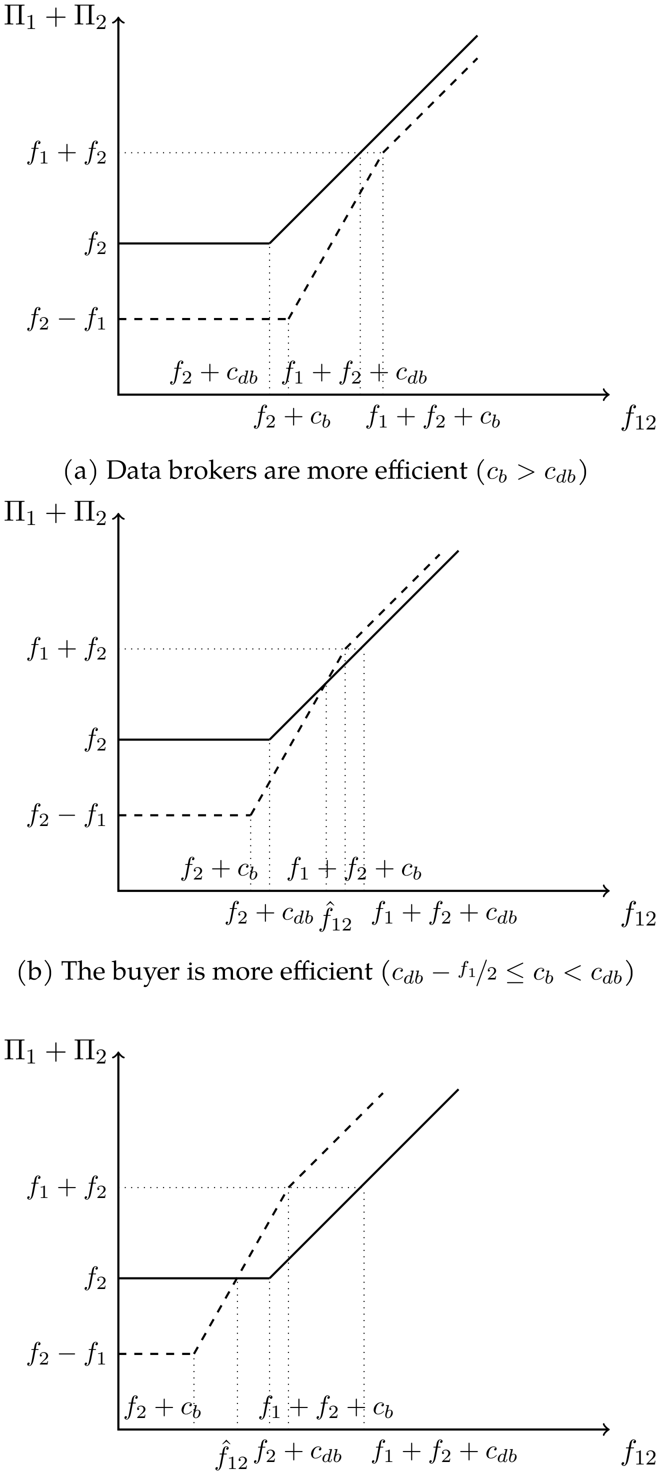
Proposition 2 (Joint Profits)**.** (*i*) *Suppose $c_b \geq c_{db}$. The joint profits of the data brokers under data sharing are no less than those under independent selling, irrespective of the nature of the data structure.*
(*ii*) *Suppose instead $c_b < c_{db}$. The joint profits of the data brokers under data sharing are no less than those under independent selling if $f_{12} \leq \hat{f}_{12}$, where*

$$\hat{f}_{12} = \begin{cases} f_1 + f_2 + 2c_b - c_{db} & \text{if } c_{db} - f_1/2 \leq c_b < c_{db} \\ f_1/2 + f_2 + c_b & \text{if } c_b < c_{db} - f_1/2 \end{cases} . \tag{6}$$

Proof. See Appendix A.3. □

Figure 3 provides a graphical representation of the findings presented in Proposition 2. Figure 3a focuses on the more natural case in which the buyer is less efficient than the data brokers in merging the datasets, $c_b > c_{db}$. For example, a supermarket acquires consumer level information and these are merged with internal data such as loyalty card and browsing shelves data. The solid line (joint profits under sharing) is always above the dashed line (joint profits under competition). As a result, data brokers are collectively better off when sharing data as it helps internalising downstream inefficiencies as well as avoiding competition when their datasets overlap.

Figures 3b and 3c consider the cases where the buyer is more efficient than the data brokers, $c_b < c_{db}$. For example, a dot com company, particularly effective in handling data, acquires new information from the data brokers. Sharing in such cases is only an option if $f_{12} < \hat{f}_{12}$, that is, when the value of the merged datasets is sufficiently small. Intuitively, without the benefit of internalising downstream merging inefficiencies, sharing only helps to increase joint profit when information overlapping is sufficiently severe. The graphs also illustrate how the cut-off value $\hat{f}_{12}$ is derived in these two scenarios, i.e., when the downstream merging cost is relatively high or low compared to $c_{db} - f_1/2$.



(a) Data brokers are more efficient ($c_b > c_{db}$)

(b) The buyer is more efficient ($c_{db} - {}^{f_1}/_2 \leq c_b < c_{db}$)

(c) The buyer is much more efficient ($c_b < c_{db} - {}^{f_1}/_2$)

Figure 3: Data brokers' joint profits from sharing (solid line) and from individual sales (dashed line), and the joint value of the datasets.

3.3.1 Proportional sharing rule

Data sharing may not necessarily emerge even if joint profits are larger when sharing than under competition. For sharing to be a mutual best response, individual sharing profits must be no less than individual competition profits for both data brokers. To compare these, we assume the following sharing rule that assigns a share of the joint profits to a data broker that is proportional to the stand-alone value of its dataset. Namely, for $k = 1, 2$,

$$s_k = \frac{f_k}{f_k + f_{-k}}.$$

On the other hand, when data are downstream super-additive, competition between data brokers leads to a multiplicity of equilibria and, similarly, only joint profits are identified. To enable the comparison, we introduce a parameter $\alpha \in [0, 1]$ to index the Nash equilibria in the competitive subgame when data are downstream super-additive. α captures the data brokers' (common) belief about the share of the extra surplus assigned to DB2. Formally, we select the equilibrium where

$$p_1^* = \alpha f_1 + (1-\alpha)\left(f_{12} - c_b - f_2\right) \text{ and } p_2^* = (1-\alpha) f_2 + \alpha\left(f_{12} - c_b - f_1\right).$$

In this way, we capture all possible equilibria, ranging from the one in which the extra-surplus is allocated equally across data brokers ($\alpha = 0.5$) to the ones characterised by a very asymmetric surplus reallocation ($\alpha = 1$ or $\alpha = 0$).

We are now ready to present the main result of our analysis.

Proposition 3 (Equilibrium Sharing). (*i*) *Suppose $c_b \geq c_{db}$. Data sharing emerges in the unique Subgame Perfect Nash Equilibrium of the game, if and only if, $f_{12} < \tilde{f}_{12}$ where*

$$\tilde{f}_{12} = f_1 + f_2 + c_b + \begin{cases} \frac{(c_b - c_{db}) f_1}{(1-\alpha) f_2 - \alpha f_1} & \text{if } \alpha < \frac{f_2}{f_1+f_2} \\ \infty & \text{if } \alpha = \frac{f_2}{f_1+f_2} \text{ and } c_b > c_{db} \\ 0 & \text{if } \alpha = \frac{f_2}{f_1+f_2} \text{ and } c_b = c_{db} \\ \frac{(c_b - c_{db}) f_2}{\alpha f_1 - (1-\alpha) f_2} & \text{if } \alpha > \frac{f_2}{f_1+f_2} \end{cases}. \tag{7}$$

(*ii*) *Suppose instead $c_b < c_{db}$. Data sharing emerges in the unique Subgame Perfect Nash Equilibrium of the game, if and only if, $f_{12} < \approx{f}_{12}$ where*

$$\approx{f}_{12} = f_1 + c_b + \begin{cases} \frac{f_2^2}{f_1+f_2} & \text{if } c_{db} - c_b < \frac{f_1^2}{f_1+f_2} \\ f_2 - \frac{f_2}{f_1}(c_{db} - c_b) & \text{if } c_{db} - c_b > \frac{f_1^2}{f_1+f_2} \end{cases}. \tag{8}$$

Proof. see Appendix A.4. □

Consider the case where data brokers are more efficient than the buyer in handling data, i.e., $c_b \geq c_{db}$. Suppose first that the data structure features some complementarities. The previous proposition established that sharing could be industry-efficient, but this does not necessarily arise. As under competition, data brokers may make very asymmetric profits (given the multiplicity of equilibria), and sharing would make one of them better off but penalise the other. In other words, for either a large or a small α, one data broker vetoes a sharing agreement provided that the joint profits are sufficiently large. Only in the special case where the expected competitive profit shares are exactly in line with the sharing rule, do both brokers agree to share their data for any value of the joint dataset. To obtain the uniqueness result, we differentiate whether $c_b > c_{db}$ or $c_b = c_{db}$ as in the latter case for any $f_{12} \geq f_1 + f_2 + c_b$, competition can also be an equilibrium outcome. The above discussion is reflected in the critical value of $\tilde{f}_{12}$ and in the conclusion that data sharing arises for $f_{12} < \tilde{f}_{12}$ as defined by (7).

Turning to a sub-additive data structure, data sharing allows for a surplus extraction that they would otherwise fail to implement fully with independent selling. Because competition leads data brokers to provide a discount to the buyer (equal to downstream merging cost and the overlapping component of the datasets), sharing data can restore full surplus extraction. This way, data brokers can soften competition and internalise downstream inefficiencies. A similar argument applies to an extreme sub-additive data structure. In this case, data sharing is optimal for data brokers as it always allows them to coordinate on "throwing away" DB1's dataset and extract all surplus generated by the most valuable dataset. Importantly, both data brokers are better off with sharing under the assumed sharing rule than under competition.

Suppose now that the buyer is more efficient than the data brokers. Note that in this case, the benefit of internalising inefficient merging costs through sharing is absent and hence, at least one data broker objects sharing when the data structure is super-additive.

When the data structure is sub-additive or extreme sub-additive, sharing can help data brokers to appropriate some surplus otherwise left because of the overlapping component between their datasets. However, this appealing strategy constitutes an equilibrium only when the loss from the higher merging cost outweighs each data broker's loss under competition. When the value of the merged dataset is sufficiently low, meaning substantial overlapping information, then sharing would be optimal for both data brokers. As a result, there exists a critical value such that only for lower values of the joint dataset both data brokers agree to share and to take on the higher upstream merging cost. This critical value is denoted by $\tilde{\tilde{f}}_{12}$.

A somewhat counter-intuitive result emerges from the above discussion. At first, one may expect that an incentive to share data would emerge when complementarities between data are strong. For instance, combining email addresses (or postal codes) with the browsing history would provide the two data brokers with powerful information to be sold in the market for data. This is, for example, the rationale of patent pooling agreements (Lerner and Tirole, 2004, 2007). On the other hand, intuition may suggest that when data partially overlap or lead to quality deterioration, the incentive to share would decrease as the incremental benefit of the rival's database decreases too.

Our model leads to different conclusions. Data sharing is most likely to arise when datasets present forms of substitutability and data brokers are more efficient than buyers in handling data. On the contrary, competition arises more often when datasets are complements and there are upstream inefficiencies in merging data.

4 Extensions

4.1 Alternative sharing rules

The sharing rule adopted in the previous section is just one among several possible alternatives. For example, s_k can follow the Shapley value implementation. Unlike the proportional rule, the Shapley value captures the average marginal contribution of a data broker to a given coalition, i.e., in our context, a data sharing proposition.

The results obtained prove very robust. Also in this context, data sharing arises for relatively low values of the combined dataset, whereas competition prevails if combining datasets generates high values. Moreover, sharing is more likely if data brokers are relatively more efficient in handling the data and if the competitive equilibrium share of profits is expected to be balanced, i.e., when α is close to the Shapley sharing rule.

4.2 Data can be partitioned

A key feature of data is its divisibility. That is, a dataset containing information regarding N consumers and M attributes can be "repackaged" to contain information on alternative sets $\hat{N}$ of consumers and $\hat{M}$ of attributes. One may wonder whether data brokers have an incentive to operate strategically such partitions when competition occurs. A rationale for partitioning might be that data brokers try to soften the very harsh competition that occurs when data are sub-additive. In other words, if the original datasets feature some overlaps or correlation, the data may be restructured prior to competition in a way that eliminates or minimises such issues.

We note, however, that this would not affect the conclusions of our previous analysis for two reasons. First, selectively repackaging some information can be particularly costly. Second, as part (ii) of Proposition 1 demonstrates, the data broker that considered removing some overlapping information from its own dataset still obtains a profit equal to its net marginal contribution, whereas the other data broker would now obtain a higher profit. This suggests that absent anti-competitive side-transfers, a data broker may not have incentives to unilaterally reduce overlaps.

4.3 Sequential pricing

We also investigate whether data brokers' incentive to share data changes when they set their prices sequentially. The timing is changed as follows. DB k first sets p_k and then DB $-k$ sets p_{-k} after observing p_k. Given the resulting prices, the downstream firm decides whether to buy the dataset(s) and from which data broker. Regardless of the order of moves, our main findings and intuitions remain qualitatively similar: data sharing emerges as a tool to soften the competition between data brokers. However, as compared to the case in which prices are set simultaneously, sharing arises less often.

The intuition is as follows. A first-mover advantage is identified with a downstream super-additive data structure, which leads to the possibility of naturally selecting one equilibrium from the multiplicity identified in the benchmark. Formally, this implies selecting the equilibrium with $\alpha = \{0, 1\}$ from the benchmark model with $c_{db} \leq c_b$, and, hence, the most asymmetric surplus divisions. As a result, the first-mover has an incentive to veto any sharing agreement, rendering competition the most likely scenario.

5 Conclusion and discussion

This article sheds light on the quite obscure and relatively unexplored market for data. We present a model of data intermediaries and study their role as suppliers of valuable information to downstream firms. A distinctive aspect of the sector, prominently transpiring from the Federal Trade Commission (2014)'s report, is the exchange and trade of data *between* brokers and how this relates to the particular properties of data, as compared to other products (contextual value, merging costs, complementarities).

Our framework is compatible with a market for data in which data brokers repeatedly interact to supply buyers in different sub-markets, and in which projects are buyer-specific. We highlight how the incentives for data sharing are crucially related to the nature of the data held by the brokers. Specifically, we find that data sharing can arise for two reasons. First, data brokers can soften competition when data present some form of substitutability. Second, it allows data brokers to internalise downstream inefficiencies, as buyers may be less efficient than data brokers in merging multiple datasets. In turn, we identify a possible trade-off between the positive effects of cost internalisation, consistent with the spirit of the EU Data Strategy (EU, 2020), and the negative effects of data sharing linked to reduced competition in this opaque market.

In particular, our analysis highlights the importance of the sub- or super-additive data structures, the data merging costs, and the selection of the competitive equilibrium for their decisions to co-operate on a shared project. These insights are also partly consistent with the 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). In the context of our model, collaboration may go beyond situations of value creation (efficiency savings) and can soften competition between data brokers at the expense of their clients.

Our theoretical analysis rationalises the large heterogeneity in the contractual arrangements and collaborations in this market, as also illustrated by the Federal Trade Commission (2014). For a client, our results provide two rather counter-intuitive implications. First, a firm 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, downstream cost inefficiencies may prove to be an advantage as competition leads data brokers to grant a discount to a downstream firm. This suggests that downstream firms may not have incentives to develop their digital skills when there is a functioning data market.

The sector is not particularly transparent and reliable information to conduct a proper empirical analysis of data brokers' strategies is not easy to access. If data were available, however, our model delivers testable predictions. For example, the probability that data brokers may exchange a dataset required by a buyer should positively relate to their relative efficiency in handling data compared to the buyers. The probability should also increase in the data homogeneity, and decrease when composite information from a variety of sources are usually in demand. At the same time, it might be inferred from highly asymmetric revenues in competitive segments of the market that data sharing has failed due to the profitable firm anticipating its dominant role.

Moreover, we shall note that the European Union and the United States have followed different regulatory approaches on how data should be managed by intermediaries, third-parties, and retailers. The European Union has tackled the issue of privacy more strictly. More specifically, the EU 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 prior authorisation of consumers is deemed illegal, to the point that such regulation is often emphatically evoked as the "death of third-party data".[10] In the light of our analysis, the EU GDPR may

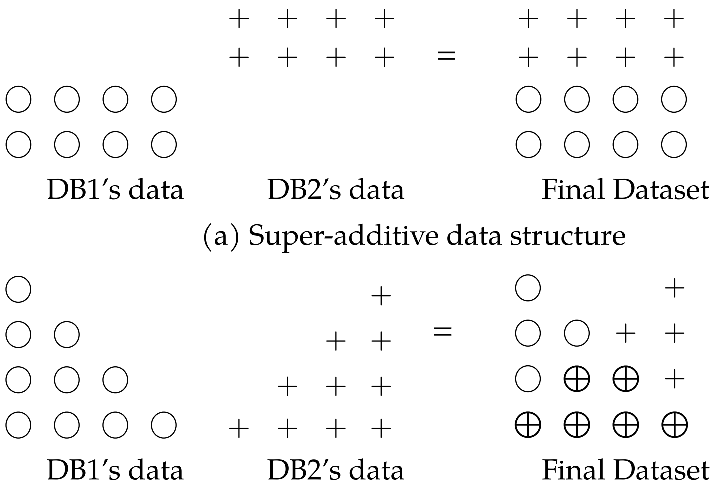
[10] See, e.g., Wired (2018), "Forget Facebook, mysterious data brokers are facing GDPR trouble", November 8, 2018.

have some unintended pro-competitive effect in the upstream data 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 downstream firms to partially retain some of the data generated surplus.

Finally, most of the attention of the policymakers has been devoted to the final use of data and on how data sharing might create positive externalities and pro-competitive effects. Nevertheless, little attention has been given to data as an input, produced, managed and traded by data brokers. Our analysis highlights that the co-opetitive practices of data brokers might require additional scrutiny from a regulator.

A Appendix

A.1 Microfoundation of the data structure



(a) Super-additive data structure

(b) Sub-additive data structure

The figure presents some non-exhaustive examples of data structures with $N = 4$ consumers and $M = 4$ attributes. In example (a), data are super-additive. DB1 possesses information for all consumers regarding attributes $j = 3, 4$ (e.g., browsing history) whereas DB2 information regarding attributes $j = 1, 2$ (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. In example (b), data are sub-additive: DB1 has partial information for all consumers and so does DB2. As some data are owned by both data brokers (e.g., both data brokers have information regarding attribute 4 for all consumers), the value of the final dataset is lower than the sum of the values of the two independent datasets. Overlapped entries are indicated in the final matrix with $\bigoplus$.

Figure A.1: Examples of Data Structure

Consider an economy with $N > 0$ individuals, each characterised by a set $M > 0$ of attributes (e.g., physical or email addresses, browsing history, etc.). Let Λ_k be the $M \times N$ logical matrix that represents DB k's information. The element $\lambda_{kji} = 1(0)$ of the matrix implies that DB k has (no) information about consumer i's attribute j. These data can be sold to buyers and give rise to additional surplus in the market where the buyers operate. Denote a function $f(\Lambda_k) \geq 0$ that 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., $f(\mathbf{0}) = 0$). We can define the data structure, in the absence of merging costs, as follows

- super-additive, if $f(\Lambda_k|\Lambda_{-k}) \geq f(\Lambda_k) + f(\Lambda_{-k})$,
- sub-additive, if $f(\Lambda_k|\Lambda_{-k}) < f(\Lambda_k) + f(\Lambda_{-k})$,

where | is the element-wise OR operator. Figure A.1 visually outlines an example in each case.

A.2 Proof of Proposition 1

Proof. Depending on the buyer's merging cost, c_b, we can have three *downstream* data structures.

(i) *Downstream super-additivity.* Consider first DB1's best response. Suppose DB2's price is high, i.e., $p_2 > f_{12} - c_b - f_1$. DB2 cannot sell its dataset alone, whereas DB1 has two ways of selling its dataset. The first is to set $p_1 = f_1$. The second is to set $p_1 = f_{12} - c_b - p_2$ so that the buyer buys both datasets. Given the range of p_2 in this case, the former is better for DB1 and, hence, its best response is $p_1 = f_1$.

Next, consider $f_2 \leq p_2 \leq f_{12} - c_b - f_1$. Again DB2 cannot sell its dataset alone, whereas DB1 has two ways of selling its dataset. In this case, however, $p_1 = f_{12} - c_b - p_2$ is DB1's best response as p_2 is now lower.

Finally, consider $p_2 < f_2$. DB1 can either set a price slightly lower than $f_1 - f_2 + p_2$ to undercut what DB2 alone can offer, or $f_{12} - c_b - f_2$ so that the buyer finds buying both datasets is better than buying from DB2 alone. Given the range of p_2 in this case, $f_{12} - c_b - f_2$ is strictly better. The below equation summarises the analysis:

$$BR_1(p_2) = \begin{cases} f_1 & \text{if } p_2 > f_{12} - c_b - f_1 \\ f_{12} - c_b - p_2 & \text{if } f_2 \leq p_2 \leq f_{12} - c_b - f_1 \\ f_{12} - c_b - f_2 & \text{if } p_2 < f_2 \end{cases} . \tag{A.1}$$

DB2'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 $p_1^* + p_2^* = f_{12} - c_b$ and $p_k^* \geq f_k$ for $k = 1, 2$, constitutes a Nash equilibrium in this subgame. The buyer buys from both data brokers, and the profits are $\Pi_k = p_k^*$, $\Pi_{-k} = f_{12} - c_b - p_k^*$, where $p_k^* \in [f_k, f_{12} - c_b - f_{-k}]$, and $\Pi^b = \pi^0$.

(ii) *Downstream sub-additivity.* Consider again DB1's best response. Suppose $p_2 > f_2$. In this case, the buyer does not buy dataset 2 alone. DB1 then has two ways of selling its dataset. One is to set $p_1 = f_1$ and the other is to set $p_1 = f_{12} - c_b - p_2$. Since $f_1 > f_{12} - c_b - f_2 > f_{12} - c_b - p_2$, DB1's best response is the former.

Now consider $f_{12} - c_b - f_1 < p_2 \leq f_2$. DB1 again has two ways of selling its dataset. The first is to set a price slightly lower than $f_1 - f_2 + p_2$ so that the buyer finds it strictly better to buy dataset 1 alone than either buying dataset 2 alone or buying both. The other is to set it at $f_{12} - c_b - f_2$, so that the buyer finds buying both is at least as good as buying dataset 2 alone. Given the range of p_2 in this case, the former is better for DB1. However, technically there exists no best response because no highest price that is strictly lower than $f_1 - f_2 + p_2$ can be found.

Finally, consider $p_2 \leq f_{12} - c_b - f_1$. DB1 has the same two ways of selling its dataset. However, now setting $p_1 = f_{12} - c_b - f_2$ and let the buyer buy both is better for DB1 as p_2 is now lower. To summarise, DB1's best response function is

$$BR_1(p_2) = \begin{cases} f_1 & \text{if } p_2 > f_2 \\ \emptyset & \text{if } f_{12} - c_b - f_1 < p_2 \leq f_2 \\ f_{12} - c_b - f_2 & \text{if } p_2 \leq f_{12} - c_b - f_1 \end{cases} . \tag{A.2}$$

Similarly, DB2's best response function is

$$BR_2(p_1) = \begin{cases} f_2 & \text{if } p_1 > f_1 \\ \emptyset & \text{if } f_{12} - c_b - f_2 < p_1 \leq f_1 \\ f_{12} - c_b - f_1 & \text{if } p_1 \leq f_{12} - c_b - f_2 \end{cases} . \tag{A.3}$$

By superimposing (A.2) and (A.3), one verifies that there exists a unique Nash equilibrium in which $p_k^* = f_{12} - c_b - f_{-k}$ for $k = 1, 2$. Hence, the buyer buys from both data brokers and profits are $\Pi_k = p_k^*$ and $\Pi^b = \pi^0 + f_1 + f_2 - f_{12} + c_b$.

(iii) *Downstream extreme sub-additivity.* Once more, consider again DB1's best response. Suppose $p_2 > f_2$. In this case, the buyer does not buy dataset 2 alone. Since $f_1 > 0 > f_{12} - c_b - f_2$, DB1 then sets $p_1 = f_1$ as inducing the buyer to buy both datasets requires making a loss.

Next, consider $f_2 - f_1 < p_2 \leq f_2$. DB1 can set a price slightly lower than $f_1 - f_2 + p_2$ so that the buyer finds it strictly better to buy dataset 1 alone than buying dataset 2 alone. Note that also in this case, as $f_{12} - c_b \leq p_2$, there is no non-negative price

p_1 that allows the buyer to buy both datasets. Indeed, there exists no best response because no highest price, that is strictly lower than $f_1 - f_2 + p_2$, can be found.

Finally, focus on $p_2 \leq f_2 - f_1$. Also in this case, it is never an option to induce the buyer to buy both datasets. Alternatively, DB1 can try to sell dataset 1 with $p_1 \leq f_1 + p_2 - f_2$. However, as the latter price is negative, any price in the interval $[0, +\infty)$ a best response. To summarise, DB1's best response function is:

$$BR_1(p_2) = \begin{cases} f_1 & \text{if } p_2 > f_2 \\ \emptyset & \text{if } f_2 - f_1 < p_2 \leq f_2 \\ [0, +\infty) & \text{if } p_2 \leq f_2 - f_1 \end{cases} . \tag{A.4}$$

DB2's best response function is only slightly different from that of DB1. Specifically, when DB1 reaches the lowest price, i.e., $p_1 = 0$, DB2's best response is $f_2 - f_1$. To sum up:

$$BR_2(p_1) = \begin{cases} f_2 & \text{if } p_1 > f_1 \\ \emptyset & \text{if } 0 < p_1 \leq f_1 \\ f_2 - f_1 & \text{if } p_1 = 0 \end{cases} . \tag{A.5}$$

With the above best response functions, one verifies that the pair $(p_1^*, p_2^*) = (0, f_2 - f_1)$ represents a Nash equilibrium in this subgame. The buyer buys from DB2 and obtains profits $\Pi^b = \pi^0 + f_2 - (f_2 - f_1) = \pi^0 + f_1$. Data brokers' profits are $\Pi_1 = 0$ and $\Pi_2 = f_2 - f_1$, respectively for DB1 and DB2. As $p_1^* = 0$ the buyer is indifferent between buying or not buying dataset 1. However, even when the buyer buys both datasets, it does not merge them.

□

A.3 Proof of Proposition 2

Proof. In light of Corollary 1 and the focus on joint profits, we need to differentiate which downstream and upstream data structure the information value of the merged dataset f_{12} gives rise to. To this end, Table A.1 summarises all relevant scenarios, distinguishing $c_b \geq c_{db}$ and$c_b < c_{db}$.

Consider first $c_b \geq c_{db}$. In region (i), data are both upstream and downstream super-additive and $f_{12} - c_{db} \geq f_{12} - c_b$. In regions (ii) and (iii) data are upstream sub-additive. By the definition of upstream sub-additivity ($f_{12} - c_b < f_1 + f_2$), for given f_{12} in these regions, $2f_{12} - f_1 - f_2 - 2c_b < f_{12} - c_b \leq f_{12} - c_{db}$. Hence, data sharing also warrants higher joint profits in regions (ii) and (iii). Finally, consider regions (iv) and (v), in which data are upstream extreme sub-additive. For any given f_{12} in these regions, it can

Region ($c_b \geq c_{db}$)	Values of f_{12}	$\Pi_1 + \Pi_2$ sharing	$\Pi_1 + \Pi_2$ competition
(i)	$[f_1 + f_2 + c_b, +\infty)$	$f_{12} - c_{db}$	$f_{12} - c_b$
(ii)	$[f_1 + f_2 + c_{db}, f_1 + f_2 + c_b)$	$f_{12} - c_{db}$	$2f_{12} - f_1 - f_2 - 2c_b$
(iii)	$[f_2 + c_b, f_1 + f_2 + c_{db})$	$f_{12} - c_{db}$	$2f_{12} - f_1 - f_2 - 2c_b$
(iv)	$[f_2 + c_{db}, f_2 + c_b)$	$f_{12} - c_{db}$	$f_2 - f_1$
(v)	$[0, f_2 + c_{db})$	f_2	$f_2 - f_1$
Region ($c_b < c_{db}$)	Values of f_{12}	$\Pi_1 + \Pi_2$ sharing	$\Pi_1 + \Pi_2$ competition
(i)	$[f_1 + f_2 + c_{db}, +\infty)$	$f_{12} - c_{db}$	$f_{12} - c_b$
(ii)	$[f_1 + f_2 + c_b, f_1 + f_2 + c_{db})$	$f_{12} - c_{db}$	$f_{12} - c_b$
(iii)	$[f_2 + c_{db}, f_1 + f_2 + c_b)$	$f_{12} - c_{db}$	$2f_{12} - f_1 - f_2 - 2c_b$
(iv)	$[f_2 + c_b, f_2 + c_{db})$	f_2	$2f_{12} - f_1 - f_2 - 2c_b$
(v)	$[0, f_2 + c_b)$	f_2	$f_2 - f_1$

Table A.1: Data brokers joint profits under sharing and competition.

be verified that the joint profits of sharing $f_{12} - c_{db}$ and f_2, respectively, exceed the joint profits when competing, i.e., $f_2 - f_1$. Overall, the joint profits of sharing always exceed those of competition.

Turn then to $c_b < c_{db}$. In regions (i) and (ii), where the data are downstream super-additive, the relation between the merging costs implies that competition generates higher joint profits. In region (v), however, where data are both upstream and downstream extreme sub-additive, sharing leads to higher joint profits than competition. Both joint profit functions are continuous and non-decreasing in f_{12}, and they only cross once in either region (iii) or region (iv). The intersection point in region (iii) is found by solving: $f_{12} - c_{db} = 2f_{12} - f_1 - f_2 - 2c_b$, or $f_{12} = f_1 + f_2 + 2c_b - c_{db}$. The intersection takes place within the boundaries of region (iii) provided that the efficiency gap between the upstream and downstream is not too large, i.e., $c_{db} - c_b \leq f_1/2$. The intersection point is in region (iv) if $c_{db} - c_b > f_1/2$, and it is found when $f_2 = 2f_{12} - f_1 - f_2 - 2c_b$, or $f_{12} = f_2 + c_b + \frac{f_1}{2}$. Hence, defining $\hat{f}_{12}$ as in equation (6) data sharing increases joint profits if $f_{12} \leq \hat{f}_{12}$. □

A.4 Proof of Proposition 3

Proof. For sharing to be a mutual best response at the first stage,

$$\Pi_k^s \geq \Pi_k^c, \quad k = 1, 2 \tag{A.6}$$

under a given data structure. The profit functions are:

$$\Pi_1^s = \begin{cases} \frac{f_1}{f_1+f_2}(f_{12}-c_{db}) & \text{if } f_{12} \geq f_2 + c_{db} \\ \frac{f_1 f_2}{f_1+f_2} & \text{if } f_{12} < f_2 + c_{db} \end{cases},$$

$$\Pi_2^s = \begin{cases} \frac{f_2}{f_1+f_2}(f_{12}-c_{db}) & \text{if } f_{12} \geq f_2 + c_{db} \\ \frac{f_2^2}{f_1+f_2} & \text{if } f_{12} < f_2 + c_{db} \end{cases},$$

$$\Pi_1^c = \begin{cases} \alpha f_1 + (1-\alpha)(f_{12} - c_b - f_2) & \text{if } f_{12} \geq f_1 + f_2 + c_b \\ f_{12} - c_b - f_2 & \text{if } f_2 + c_b \leq f_{12} < f_1 + f_2 + c_b \\ 0 & \text{if } f_{12} < f_2 + c_b \end{cases}, \text{ and}$$

$$\Pi_2^c = \begin{cases} (1-\alpha) f_2 + \alpha(f_{12} - c_b - f_1) & \text{if } f_{12} \geq f_1 + f_2 + c_b \\ f_{12} - c_b - f_1 & \text{if } f_2 + c_b \leq f_{12} < f_1 + f_2 + c_b \\ f_2 - f_1 & \text{if } f_{12} < f_2 + c_b \end{cases}.$$

All four functions are continuous, piecewise linear and non-decreasing in f_{12}. Note first that independent of the merging cost, c_{db} and c_b, in region (v) defined by Table A.1, as f_{12} becomes smaller, condition (A.6) is satisfied and both data brokers prefer sharing. Evaluating the profits at the lower bound of f_{12},

$$\Pi_1^s = \frac{f_1 f_2}{f_1 + f_2} > 0 = \Pi_1^c, \text{ and } \Pi_2^s = \frac{f_2^2}{f_1 + f_2} > f_2 - f_1 = \Pi_2^c.$$

On the other extreme, in region (i), as f_{12} becomes larger, condition (A.6) fails for at least one data broker.

Consider first $c_b \geq c_{db}$. At the lower bound of region (i) of Table A.1, $f_{12} = f_1 + f_2 + c_b$, sharing still leads to more profits for both data brokers than competition: $\text{sign}\{\Pi_i^s - \Pi_i^c\}|_{f_{12}=f_1+f_2+c_b} = \text{sign}\{f_i(c_b - c_{db})\} > 0$. At that point, the slope of the profit functions of DB1 are ${f_1}/{f_1+f_2}$ if sharing and $(1-\alpha)$ if competing; the respective slopes for DB2's profit functions are ${f_2}/{f_1+f_2}$ and α. This implies that if ${f_2}/{f_1+f_2} \neq \alpha$, the profit functions of one of the data brokers will eventually cross, as f_{12} increases. It follows that there is a value $\tilde{f}_{12}$ above which one of the data brokers objects sharing and (strictly) below which data sharing arises as a unique Subgame Perfect Nash Equilibrium of the game. Note that when $\alpha < (>) {f_2}/{f_1+f_2}$, DB1 (DB2) is the one to object sharing. By comparing respective profits, this critical value $\tilde{f}_{12}$ is characterised in (7).

However, if $\alpha = {f_2}/{f_1+f_2}$, sharing is weakly better than competition for both data brokers for all values of f_{12}: sharing is strict better if $c_b > c_{db}$ and equally good when $c_b = c_{db}$. To obtain the uniqueness result, we differentiate these two cases in (7).

We now note that by imposing $f_{12} < \tilde{f}_{12}$, we ruled out all cases where both data brokers are indifferent between sharing and competition, and hence all cases where both sharing and competition can feature in a SPNE.

Turning to $c_b < c_{db}$, as in regions (i) and (ii) the joint profits are lower, data sharing never arises. Given the initial results in region (v), the profit functions of at least one of the data brokers must intersect in either region (iii) or (iv). It can be shown that, DB2 objects sharing before DB1 does as f_{12} increases from $f_2 + c + b$ to $f_1 + f_2 + c + b$. It follows that data sharing emerges in the unique Subgame Perfect Nash Equilibrium of the game, if and only if, $f_{12} < \tilde{\tilde{f}}_{12}$ where $\tilde{\tilde{f}}_{12}$ is obtained by comparing DB2's profits and is defined in (8). Otherwise, DB2 or both data brokers object sharing. $\square$

References

Acxiom (2015). Acxiom becomes audience data provider Facebook marketing partner program. https://www.acxiom.co.uk/news/acxiom-becomes-audience-data-provider-facebook-marketing-partner-program. [Last accessed January 16, 2021].

Bajari, P., Chernozhukov, V., Hortaçsu, A., and Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. *AEA Papers and Proceedings*, 109:33–37.

Belleflamme, P., Lam, W. M. W., and Vergote, W. (2020). Competitive imperfect price discrimination and market power. *Marketing Science*, 39(5).

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

Bergemann, D., Bonatti, A., and Gan, T. (2020). The economics of social data. *Cowles Foundation Discussion Paper*.

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 January 16, 2021].

Borgogno, O. and Colangelo, G. (2019). Data sharing and interoperability: Fostering innovation and competition through APIs. *Computer Law & Security Review*, 35(5):105314.

Bounie, D., Dubus, A., and Waelbroeck, P. (2020). Selling strategic information in digital competitive markets. *RAND Journal of Economics*. Forthcoming.

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

Chen, Z., Choe, C., Cong, J., Matsushima, N., et al. (2020). Data-driven mergers and personalization. *Institute of Social and Economic Research Discussion Papers*, 1108:1–14.

Choi, J. P. (2008). Mergers with bundling in complementary markets. *Journal of Industrial Economics*, 56(3):553–577.

Choi, J. P., Jeon, D.-S., and Kim, B.-C. (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics*, 173:113–124.

Claussen, J., Peukert, C., and Sen, A. (2019). The editor vs. the algorithm: Targeting, data and externalities in online news. *SSRN Working Paper*.

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

Conitzer, V., Taylor, C., 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.

De Cornière, A. and Taylor, G. (2020). Data and competition: a general framework with applications to mergers, market structure, and privacy policy. *CEPR Discussion Paper No. DP14446*.

EU (2020). A European strategy for data.

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

Financial Times (2019). Data brokers: regulators try to rein in the 'privacy deathstars'. https://www.ft.com/content/f1590694-fe68-11e8-aebf-99e208d3e521. [Last accessed January 16, 2021].

Graef, I., Husovec, M., and Purtova, N. (2018). Data portability and data control: lessons for an emerging concept in eu law. *German Law Journal*, 19(6):1359–1398.

Graef, I., Tombal, T., and De Streel, A. (2019). Limits and enablers of data sharing. an analytical framework for eu competition, data protection and consumer law. *TILEC Discussion Paper No. DP 2019-024*.

Gu, Y., Madio, L., and Reggiani, C. (2019). Exclusive data, price manipulation and market leadership. *CESifo Working Paper*.

Ichihashi, S. (2020a). Competing data intermediaries. *Mimeo*.

Ichihashi, S. (2020b). Online Privacy and Information Disclosure by Consumers. *American Economic Review*, 110(2):569–595.

Jones, C. I. and Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58.

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. L. (2019). The impact of access to consumer data on the competitive effects of horizontal mergers and exclusive dealing. *Journal of Economics & Management Strategy*, 28(3):373–391.

Lambrecht, A. and Tucker, C. (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.

Martens, B., De Streel, A., Graef, I., Tombal, T., and Duch-Brown, N. (2020). Business-to-business data sharing: An economic and legal analysis. *EU Science Hub*.

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*, 65(3):1342–1362.

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.

OECD (2013). Enhancing access to and sharing of data. *Report*.

Prat, A. and Valletti, T. (2019). Attention oligopoly. *Mimeo*.

Prüfer, J. and Schottmüller, C. (2017). Competing with big data. *TILEC Discussion Paper*.

Radner, R. and Stiglitz, J. (1984). A nonconcavity in the value of information. In Boyer, M. and Kihlstrom, R., editors, *Bayesian Models of Economic Theory*, pages 33–52. Elsevier, Amsterdam.

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

Sarvary, M. and Parker, P. M. (1997). Marketing information: A competitive analysis. *Marketing Science*, 16(1):24–38.

Schaefer, M. and Sapi, G. (2020). Data network effects: The example of Internet search. *DIW Berlin Discussion Paper No. 1894*.

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 January 16, 2021].

Vives, X. (1984). Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory*, 34(1):71–94.

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 January 16, 2021].