WORKING PAPERS

N° 1615

February 2025

"Data Sharing or Analytics Sharing?"

Bruno Carballa-Smichowski, Yassine Lefouili,
Andrea Mantovani and Carlo Reggiani

Toulouse
School of
Economics

Data sharing or analytics sharing?[*]

Bruno Carballa-Smichowski[†] Yassine Lefouili[‡]
Andrea Mantovani[§] Carlo Reggiani[¶]

February 2025

Abstract

Data combination and analytics can generate valuable insights for firms and society as a whole. Firms can seize these opportunities by joining platforms that either allow them to access the data contributed by other firms or provide the result of the analytics performed on such data, depending on whether the platform adopts ”data sharing” or “analytics sharing” technologies. The former technology enables firms to exploit their data endowment together with the data contributed by others, whereas the latter offers advantages in terms of privacy and security by reducing data transmission. We present a model that allows us to study the economic and managerial incentives generated by these technologies for both firms and a platform. First, we find that the platform chooses analytics sharing only when the security advantage of this technology is sufficiently large. Second, we show that analytics sharing results in a higher total data contribution than data sharing under general and reasonable conditions. Third, we determine the optimal data-combination technology from the perspective of consumers and discuss potential misalignments between the platform's and consumers' preferred technology. Our findings carry relevant policy and managerial implications, offering a pathway to enhance both data provision and security.

JEL Classification: D43; K21; L11; L13; L41; L86; M21; M31.

Keywords: data sharing, analytics sharing, data platforms.

[*] The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Joint Research Centre or the European Commission. We thank Hemant Bhargava, Giacomo Calzolari, Vincenzo Denicolò, Lucas Eustache, Alessandro Fedele, Aaron Kaye, Laura Lasio, Bertin Martens, Patrick Rey, Andrew Rhodes, Sandro Shelegia, Robert Somogyi, Volker Stocker, Greg Taylor, and participants at the 6th Unibg Industrial Organization Winter Symposium (Passo del Tonale, 2023), the 12th EIEF-UNIBO-IGIER Bocconi (Milano, 2023), the 16th TSE Digital Economics Conference (Toulouse, 2024), the PLAMADISO Talks (online, 2024), the TBS-TSE Digital Health Workshop (Toulouse, 2024), the Machine Learning for Data & Algorithm-sharing in Social Science Workshop (Lucca, 2024), the Spanish IO day (Barcelona, 2024), the 51st EARIE Conference (Amsterdam, 2024), the XXXVIII Jornadas de Economia Industrial (Seville, 2024), the 3rd Workshop on Digital Economics (CMA, London), and Université Paris Dauphine for helpful comments and suggestions. We are grateful to the NET Institute (www.netinst.org) for financial support. Funding from Agence Nationale de la Recherche (ANR) under the TEPREME project (ANR–21–CE26–0014–01) and under grant ANR-17-EURE-0010 (Investissements d'Avenir program) is gratefully acknowledged. The usual disclaimer applies.

[†] European Commission's Joint Research Centre, Seville, Spain. E-mail: bruno.carballa-smichowski@ec.europa.eu.

[‡] Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France. E-mail: yassine.lefouili@tse-fr.eu.

[§] TBS Business School, Toulouse, France. E-mail: a.mantovani@tbs-education.fr.

[¶] European Commission's Joint Research Centre, Seville, Spain and Department of Economics, University of Manchester, Manchester, UK. E-mail: carlo.reggiani@ec.europa.eu.

1 Introduction

The data generated by businesses, individuals and other actors holds significant economic and societal value. Given the proliferation of data sources, many applications, such as improving cancer predictions or designing better products, require data from multiple firms. Consequently, policymakers are setting up policies and experimenting with technologies to encourage and facilitate inter-firm data combination (Farrell et al., 2023). For instance, the European Commission is developing Common European Data Spaces to facilitate data combination in several strategic sectors such as healthcare, energy, mobility, finance, manufacturing, and agriculture (European Commission, 2024).

In practice, data can be combined in several ways, and specialized platforms have emerged to help firms in this endeavor. We focus on two classes of technologies that data-combination platforms may adopt, depending on whether they prioritize providing access to data or delivering analytics services. In our setting, when the platform chooses a *data sharing* technology, firms use the platform to access the data contributed by other firms while conducting data analytics in-house. Alternatively, when the platform opts for an *analytics sharing* technology, firms contribute their data to the platform and, in return, receive analytics or insights generated from the combined dataset. Importantly, under this model, firms do not have direct access to the data contributed by others.

As an illustrative example, Snowflake is a cloud-based platform primarily focused on data sharing and warehousing, whereas Databricks' core services include big data analytics, machine learning, and advanced analytics workflows.[1] Compatible with our distinction, Snowflake emphasizes sharing and querying data directly, enabling firms to access and utilize data contributed by others. In contrast, Databricks focuses on providing analytics outputs, where firms contribute data and receive insights or processed results without direct access to the underlying data. This distinction highlights the fundamental difference in how these platforms manage access to data and analytics services.

Despite the growing importance of these two types of data-combination technologies, they have received little attention in the management and economics literature. In particular, the economic drivers and managerial implications of choosing a technology where firms share data with each other (data sharing) over one where they share data-driven analytics but not the data (analytics sharing), remain understudied. In this article, we contribute to this research area by addressing three main questions: What are the incentives driving a platform to choose one technology over the other? Does one of these technologies prompt firms to contribute more data than the other? And finally, under what conditions does the data combination technology chosen by the platform differ from the one preferred by consumers?

[1] For more details, see https://www.snowflake.com and https://www.databricks.com/. We further discuss these examples in Section 2.

To address these questions, we build a model in which a platform can facilitate inter-firm data combination and analytics by relying on either a data sharing or an analytics sharing technology. Firms, with possibly heterogeneous data endowments, decide whether to join the platform and, if so, how much data to contribute to it. Data contribution entails a potential security cost due to the risk of data leakages. In the baseline model, we assume that the platform has complete information about the firms' data endowments and offers them personalized and public two-part tariff contracts. We also extend our analysis to alternative contractual arrangements that may arise due to incomplete information or firms' taste for anonymity.

Under data sharing, firms obtain direct access to the contributed data, which is made available to all participating firms. They conduct in-house analytics, deriving value from the combined dataset as well as from any data they choose not to contribute to the platform. In contrast, under analytics sharing, firms are not granted access to the contributed data and derive value exclusively from the platform's analytics, with the value of these insights increasing as the amount of data they contribute grows.

Additionally, if firms can store the contributed data in-house under data sharing, the number of potential access points for a cyberattack increases compared to analytics sharing. Consequently, analytics sharing enhances data security by restricting both the transfer and the access to data, thereby offering a "security advantage" that translates into a reduced likelihood of data breaches. This security benefit provides analytics sharing with a cost advantage over data sharing from the perspective of firms.

The analysis of the model delivers three main findings. First, the platform opts for analytics sharing only if it guarantees a sufficiently higher level of data security than data sharing. Moreover, the critical level of security advantage above which analytics sharing is preferred by the platform is increasing in the firms' data endowments. This finding derives from the balance of two effects: the *endowment effect*, which favors data sharing, as firms can use all the data they possess to extract analytics in-house, and the *data security effect*, which favors analytics sharing. Similar findings hold when considering alternative contractual agreements. We further show that analytics sharing is more likely to be adopted in the presence of uniform public contracts, while the opposite holds for secret personalized contracts.

Second, we compare data contributions under data sharing and analytics sharing. We find that analytics sharing leads to higher equilibrium data contributions under general conditions that appear to be consistent with recent studies investigating economies of scale and scope in data combination. A key mechanism behind this finding is that, under data sharing, firms can benefit from the data contributed by other firms even if they do not contribute data themselves. In contrast, analytics sharing requires firms to contribute data in order to benefit from the combined data. This results in a lower marginal benefit from contributing data under data sharing compared to analytics sharing.

Third, we determine the optimal data-combination technology from the perspective of consumers to evaluate whether it aligns with the technology chosen by the platform. To do this, we analyze how the choice of a data-combination technology affects the consumer surplus generated by the consumption of a product sold by a given firm when it can combine its own data with that of all other firms that also join the platform. We find situations where the platform's and consumers' choices are aligned, and others where they are not. The technology preferred by consumers depends on how their surplus from consuming a specific product is affected by the producing firm's own data and the data contributed by other firms.

To enhance the economic intuition behind our findings, we provide a microfoundation where non-competing firms can use data for both quality enhancement and price personalization, each having opposite effects on consumer surplus. As a result, multiple cases are possible. When data is mostly used to enhance product quality, both a firm's own data and that of other firms positively impact consumer surplus. Under these conditions, both the platform and consumers prefer analytics sharing when this technology is sufficiently more secure than data sharing. However, this no longer holds when data is primarily used to profile consumers and implement personalized pricing. In this case, consumers are harmed by the availability of more data from the firm selling the product, as well as from other firms contributing to the platform. As a result, consumers prefer data sharing, whereas the platform prefers analytics sharing when it offers significantly greater security than data sharing.

Structure. Section 2 offers examples of platforms whose business model is centered around either data sharing or analytics sharing, illustrating the practical relevance of these two data-combination technologies. Section 3 discusses how the article relates to the existing literature. Section 4 develops a baseline model of data combination with data sharing and analytics sharing technologies. Section 5 analyzes the platform's optimal data-combination technology and compares total data contribution under the two technologies. Section 6 studies the data-combination technology preferred by consumers and compares it with the one chosen by the platform. Section 7 extends the baseline model to analyze alternative contractual arrangements between the platform and firms. Section 8 discusses some policy and managerial implications of our findings and concludes.

2 Data-combination platforms and technologies

Data-combination platforms have become increasingly important in helping firms unlock the value of their data by enabling efficient collaboration and integration. As firms increasingly rely on data-driven strategies, the role of these platforms in supporting diverse business needs has grown significantly. Among the various technologies these platforms employ, two key approaches, exemplified in our distinction between data sharing and analytics sharing, have emerged, each offering distinct ways to harness the power of data. This distinction underscores the trade-off between direct data access and analytics-driven value, shaping the roles and incentives of firms and platforms in data-driven markets.

Our driving example relates to the distinction between Snowflake and Databricks introduced in the previous section. The Snowflake platform allows organizations to share and access data across multiple clouds and data providers, enabling firms to perform their own data analytics. Figure A.1 further illustrates these features of the Snowflake business model. The figure identifies the central role of the platform in a typical data pipeline, which is to provide solutions for sharing the data that clients can then access and exploit.

Conversely, Databricks' core business model focuses on providing multiple organizations with a unified analytics platform that provides data engineering, collaborative data science, and machine learning services. The platform offers scalable cloud-based solutions for processing and analyzing large datasets, enabling organizations to derive insights and build data-driven applications efficiently using internal and external data. Figure A.2, with reference to their flagship Lakehouse Platform, showcases the various analytics services it offers to clients.

However, it is important to note that both platforms offer features that may blur this distinction to some extent. For instance, Snowflake provides basic analytics services, such as SQL-based analytics, while Databricks also supports some limited direct data access. The key difference lies in their primary focus and approach to data sharing and analytics. Snowflake is primarily a data warehouse that excels in storing, querying, and sharing structured and semi-structured data, whereas Databricks offers a unified analytics platform designed for data engineering, big data, and machine learning tasks.

While acknowledging that platforms may offer both services, albeit to varying extents, this paper focuses on the platform's decision regarding which mechanism to adopt, as this choice can have significant implications for data provision and the welfare of market participants. To further illustrate this distinction, several additional examples demonstrate how data-combination platforms prioritize either data sharing or analytics sharing technologies.

For data sharing, SAP Data Intelligence allows organizations to share data across systems and integrate various data sources for advanced analytics.[2] Similarly, Palantir Foundry integrates, analyzes, and visualizes data from multiple sources, with data sharing capabilities that allow organizations to share data across departments or with external partners, while still maintaining control over access.[3] These platforms support sharing datasets between organizations or departments, providing direct access to data for in-house analytics.

[2] Additional information is available at: https://www.sap.com/products/technology-platform/data-intelligence.html.

[3] See also: https://www.palantir.com/platforms/foundry/.

In terms of analytics sharing, Salesforce Einstein Analytics collects data from different sources and provides firms with analytics and insights, typically without granting them access to the underlying data.[4] IBM Watson Studio, designed for data science tasks such as model training and AI application development, enables businesses to contribute data for processing, generating predictive analytics, machine learning insights, or other outputs.[5] However, firms do not have direct access to the data contributed by others; instead, they contribute data and receive insights or models based on it.

Data-combination solutions are often tailored to specific sectors. For example, in healthcare both the potential benefits and the challenges of data combination are extremely large. The Health-X platform, part of the Gaia-X consortium, is primarily a data-sharing initiative, although it also provides the infrastructure and standards necessary for stakeholders to derive insights from shared data, potentially facilitating both independent and collaborative analytics efforts.[6] Moreover, the Observational Health Data Sciences and Informatics (OHDSI) platform, while enabling analytics, has a primary emphasis on sharing and standardizing observational health data for collaborative research.[7]

On the other hand, a number of platforms offer data combination through analytics sharing. The US Patient-Centered Outcomes Research Institute (PCORI), through its PCORnet and PScanner platforms, allows health researchers to query health data from hundreds of hospital and clinics to perform national-level health studies. The PCORnet-enabled approach shares only the minimum necessary information needed to answer a question. Researchers' queries are sent to the data — and answers, not data, are sent back to researchers.[8] Similarly, Flatiron offers health facilities analytics services to conduct oncology studies, but does not give access to third-party data.[9]

In the agricultural sector, Farmers Edge offers tools for precision agriculture, including soil, climate, and field performance data. It emphasizes enabling farmers to share their data (such as yield, soil health, and weather data) with agricultural service providers or technology partners.[10] Conversely, on platforms like Bayer's Climate FieldView, farmers usually share data collected by sensors on their devices.[11] The data from multiple farms is then processed, and analytics are returned to help each farmer optimize irrigation schedules, adjust fertilizer application rates, and detect pests.

[4] See: https://www.salesforce.com/news/press-releases/salesforce-launches-einstein-analytics.

[5] See: https://www.ibm.com/products/watson-studio.

[6] See also: https://gaia-x.eu/news-press/health-x-a-common-data-space-for-the-health-sector.

[7] See: https://www.ohdsi.org/data-standardization/.

[8] See, for example: https://pcornet.org/data/.

[9] See also: https://flatiron.com/about-us.

[10] See: https://farmersedge.ca/about-us/.

[11] See: https://climate.com/.

3 Related literature

Our article focuses on the economic incentives that firms face when deciding whether to combine (some of) their data to exploit a joint dataset through a platform that can opt for different data-combination technologies. In that respect, it connects to several papers that have studied related problems, although without focusing on the choice of a data-combination technology.

In this vein, a seminal contribution and the work most closely related to ours is Calzolari et al. (2024). Both their baseline model and ours feature an aggregator that has to create incentives for firms to combine their data. Firms must then decide how much data to contribute. Despite these similarities, their article differs significantly from ours in both focus and research questions. More precisely, Calzolari et al. (2024) focus on a single data-combination technology —analytics sharing— while exploring various inefficiencies, related to ownership rights, information asymmetry, contracting, and competition at both downstream and upstream levels. By contrast, we investigate the economic incentives driving a data platform's choice between two data-combination technologies, and the implications of this choice. In doing so, we identify an additional source of inefficiency: the potential misalignment between the data-combination technology preferred by the platform and the one preferred by consumers.

There is a vast literature comparing the technical properties and performances of alternative data-combination technologies in the computer science and engineering fields (AbdulRahman et al., 2020; Drainakis et al., 2023; Ramírez et al., 2023). In technology law, Mattioli (2017) has posed the "data pooling problem", and has shown that potential pooling contributors may be impeded by reputational and professional concerns, even if the goal is as high and socially valuable as optimizing cancer treatment. At the same time, there are fewer studies on multi-firm data combination from an economics and management perspective.

At a more general level, our work is also connected to the broader literature on data sharing between firms. Data sharing can be "vertical" or "horizontal". It is vertical when it occurs through sales by data brokers to downstream firms. Ichihashi (2021), Bergemann et al. (2022), Gu et al. (2022) and Abrardi et al. (2024b), among others, study upstream competition, or lack thereof, between data brokers. Data can be optimally partitioned before being sold to downstream firms (Bounie et al., 2021), and need not to be sold to only one or to all the firms in the market (Abrardi et al., 2024a; Delbono et al., 2024).

Horizontal sharing takes instead place at the same market level. Information sharing between competing firms has been studied, for example, in duopoly retail markets (Liu and Serfes, 2006; Jentzsch et al., 2013, inter alios), in credit markets where data are exchanged through a "credit bureau" (Padilla and Pagano, 1997; Pagano and Jappelli, 1993; Gehrig and Stenbacka, 2007, inter alios) and in the transport sector through mobility-as-a-service

(MaaS) platforms (Carballa Smichowski, 2018). Large hybrid platforms can also enjoy a huge data advantage (Gu et al., 2024): as a possible solution, Martens et al. (2021) propose "in-situ" data sharing between sellers and a platform that also operates as a retailer.

Furthermore, our article assumes that firms' benefits from data analytics increase when they combine their datasets. Hagiu and Wright (2023) provide a microfoundation to anchor these benefits, as they study how firms can improve their products by learning from customer data. In doing so, they distinguish between within-user learning (learning from a user's repeated usage of a product) and across-user learning (learning from pooling multiple users' data).

Firms can benefit from data analytics through economies of scale and scope in data aggregation. Our article relates to a recent empirical literature that studies the impact of dataset size on prediction accuracy across various settings, such as search engines (Ullrich et al., 2024; Schäfer and Sapi, 2023; Klein et al., 2022; Chiou and Tucker, 2017; McAfee et al., 2015), sales forecast (Bajari et al., 2019), user jokes rating (Lee and Wright, 2023), consumer profiling (Neumann et al., 2019), news recommendations (Peukert et al., 2023) and advertisement (Agrawal et al., 2018). Additionally, Hocuk et al. (2022) empirically measure "economies of scope in data aggregation", showing that adding socio-economic variables to a dataset improves health outcome predictions, even with a constant number of observations.

To a lesser extent, our article is also related to an earlier literature that has studied information sharing in oligopoly. In these models, each firm receives a private signal containing information about the intercept of a stochastic demand function or a stochastic marginal cost. They can decide whether to reveal (part of) this information to other firms, and then they compete in the final market. This literature has studied the incentives to share information and its impact on prices, quantities and profits in various settings: Cournot competition and demand uncertainty (Novshek and Sonnenschein, 1982; Clarke, 1983; Vives, 1984; Li, 1985; Sakai, 1986; Kirby, 1988; Hviid, 1989), Bertrand competition and demand uncertainty (Vives, 1984; Sakai, 1986), Cournot competition and cost uncertainty (Fried, 1984; Li, 1985; Gal-Or, 1985; Shapiro, 1986) and Bertrand competition and cost uncertainty (Gal-Or, 1986), among others.[12] Our model focuses on data, from which the various types of information studied by that literature can potentially be extracted. However, we study the properties of two data-combination technologies and the economic incentives of a platform to adopt one or the other.

[12] In Vives (1990) monopolistic competition à la Chamberlin is introduced. Moreover, these models also differ in other important aspects such as whether there is product differentiation or whether private signals are noisy. See Raith (1996) for a general model.

4 A model of data combination with data sharing and analytics sharing

In this section we present a model in which a platform allows firms to combine data using either a data sharing or an analytics sharing technology. The model is general enough to encompass different business models (e.g., business-to-consumers or business-to-business), industries and types of data (e.g., personal and non-personal data).

We consider a setting with $N(\geq 2)$ firms and one data platform. A primary goal of our analysis is to determine the platform's choice between alternative and technically feasible options for combining data and extracting valuable insights from it. To this end, we focus on the previously introduced *two technologies* for data combination, *data sharing* and *analytics sharing*. We index them as $t = D, A$, and describe each of them in detail in what follows.

Firms. The firms, $i = 1, \dots, N$ are heterogeneous in their *data endowments*, y_i. The vector of data endowments is $\mathbf{y}$. Define as $\mathbf{x} = (x_1, \dots, x_N)$ the vector of data contributed by all firms to the platform, and as $\mathbf{x_{-i}}$ the vector of data contributed by all firms but firm i. Firms joining the platform benefit from insights from the application of analytics on the combined data. We capture these benefits through the function $B_i^t(\cdot)$, increasing in all its arguments, which we discuss below when introducing the two technology options for data combination.

All firms incur a cost associated with contributing data to the platform, which depends on the extent of their data contribution. We interpret this cost as an expected security cost, stemming from the potential risks of data leakages due to security breaches or other vulnerabilities. For example, under the European Union' General Data Protection Regulation (GDPR), firms are held accountable as data controllers and processors for any damage related to personal data. Hence, in our setting, the cost of a breach would be borne by the data contributor. Alternatively, for non-personal data, the expected damage may be linked to the risk of industrial secrets being revealed. In the model, this cost is captured by the function $g_i^t(x_i)$, which is increasing in x_i. In principle, the technology can affect both the effectiveness and the costs of each firm, which explains the superscript t.[13] The profit of a firm deciding not to join the platform is normalized to 0 (regardless of its data endowment).[14].

Platform. The platform aims to maximize its profits. In our baseline model we assume that the platform perfectly knows the firms' characteristics and can offer each firm a two-part

[13] We note that there are other costs for firms that are related to the amount of data shared, x_i, such as data handling, homogenization, and storage costs. However, since these costs are likely to be similar across technologies, we do not explicitly model them.

[14] The assumption that firms' outside options are independent of their data endowments is primarily made for the sake of exposition. It only plays a role—by simplifying the analysis—in the extension discussed in Section 7.2

tariff contract. Under the chosen data-combination technology t, this involves offering each firm: (i) a fixed fee f_i^t to join the platform, and (ii) a transfer m_i^t (which, a priori, can be positive or negative) per unit of data contributed to the platform. Denote as $\mathbf{f}^t$ and $\mathbf{m}^t$ the vectors of the fixed fees and per-unit transfers, respectively.

The platform faces an operational cost to manage the data contributed by the firms, captured by the function $G^t(\mathbf{x})$, which is increasing in all its arguments.

Technologies. We consider the two data-combination technologies introduced above. Both technologies allow to combine data and obtain insights from it, but they have distinct characteristics and give rise to different business models.

Let us first consider *data sharing*. Through this technology, the platform provides a data-combination service (i.e., access to the contributed data from other firms) to firms. The characterizing feature is that, since firms can access the contributed data from other firms, they can do data analytics in-house with it. Hence, firm i can gain insights from the combination of its *full data endowment* y_i and the data contributed by other firms $\mathbf{x_{-i}}$, which generates a benefit $B_i^D(y_i, \mathbf{x_{-i}})$. The cost of performing data analytics is incurred by the firms and is denoted by $C_i^D(y_i, \mathbf{x_{-i}})$ where $C_i^D(\cdot)$ is increasing in all its arguments.

Bringing it all together, in presence of the data sharing technology, the firms' profit functions are:

$$\pi_i^D(y_i, \mathbf{x}, f_i^D, m_i^D) \equiv B_i^D(y_i, \mathbf{x_{-i}}) - C_i^D(y_i, \mathbf{x_{-i}}) - g_i^D(x_i) + m_i^D x_i - f_i^D. \tag{1}$$

Second, let us consider *analytics sharing*. Under this technology, the platform provides a *centralized data analytics* service that is performed on the data contributed by the firms. A firm joining the platform thus benefits from the platform's analytics, which improves with the total amount of data contributed to the platform, including the data the firm itself has provided. The benefit function associated with analytics sharing can therefore be expressed as: $B_i^A(x_i, \mathbf{x_{-i}})$. Importantly, it is the platform that bears the cost of performing data analytics, which we denote by $C^A(\mathbf{x}) = \sum_i C_i^A(x_i, \mathbf{x_{-i}})$, where $C_i^A(x_i, \mathbf{x_{-i}})$ is the cost of analytics to generate insights for firm i. As in the case of data sharing, we assume that $C_i^A(\cdot)$ is increasing in all its arguments. It follows that, when the platform adopts an analytics-sharing technology, the firms' profit functions are:

$$\pi_i^A(\mathbf{x}, f_i^A, m_i^A) \equiv B_i^A(x_i, \mathbf{x_{-i}}) - g_i^A(x_i) + m_i^A x_i - f_i^A. \tag{2}$$

The profits of the platform under data sharing and analytics sharing are therefore respectively given by:

$$\pi_0^D(\mathbf{x}, \mathbf{f}, \mathbf{m}) \equiv \sum_i \left[f_i^D - m_i^D x_i\right] - G^D(\mathbf{x}),$$

and:

$$\pi_0^A(\mathbf{x}, \mathbf{f}, \mathbf{m}) \equiv \sum_i \left[f_i^A - m_i^A x_i - C_i^A(x_i, \mathbf{x_{-i}})\right] - G^A(\mathbf{x}).$$

The industry profits (i.e., the sum of the profit of the platform and of the firms) under data sharing and analytics sharing are respectively given by:

$$\Pi^D(\mathbf{y},\mathbf{x}) \equiv \sum_i \left[B_i^D(y_i,\mathbf{x_{-i}}) - C_i^D(y_i,\mathbf{x_{-i}}) - g_i^D(x_i)\right] - G^D(\mathbf{x}),$$

$$\Pi^A(\mathbf{x}) \equiv \sum_i \left[B_i^A(x_i,\mathbf{x_{-i}}) - C_i^A(x_i,\mathbf{x_{-i}}) - g_i^A(x_i)\right] - G^A(\mathbf{x}).$$

Timing. The game unfolds as follows.

1. The platform chooses the data-combination technology t, $t = D, A$.
2. The platform chooses the terms of its contracts, i.e., $(\mathbf{f}^t, \mathbf{m}^t)$.
3. The firms simultaneously decide whether to join the platform or not and, if they do, the amount of data to contribute x_i^t.

5 Analysis

We begin by analyzing stage 3. Under the data sharing technology, recall that the firms' profits are as in Equation (1). Conditional on joining the platform, firm i contributes an amount $x_i^D(m_i^D)$—independent of $\mathbf{x_{-i}}$—which solves the following first order condition (henceforth, FOC)

$$m_i^D = g_i^{D'}(x_i), \tag{3}$$

whenever it is interior. Firm i participates if and only if:

$$\pi_i^D(y_i, x_i^D(m_i^D), \mathbf{x_{-i}}, f_i^D, m_i^D) \geq 0.$$

Consider now the analytics sharing technology. The firms' profits are as in Equation (2). Conditional on joining the platform, firm i contributes an amount $x_i = BR_i^A(\mathbf{x_{-i}}, m_i^A)$, which solves the following FOC:

$$\frac{\partial B_i^A}{\partial x_i}(x_i,\mathbf{x_{-i}}) + m_i^A = g_i^{A'}(x_i), \tag{4}$$

whenever it is interior. Assume that the system of equations $x_i = BR_i^A(\mathbf{x_{-i}}, m_i^A)$, $i = 1,\dots,N$ has a unique solution $(x_i^A(m_i^A))_{1\leq i\leq N}$.

Comparing FOCs (3) and (4) reveals that, for a *given* per-unit transfer $m_i^D = m_i^A = m_i$, a firm's marginal benefit from contributing data (i.e., the left-hand side of the FOC) is higher under analytics sharing than under data sharing (whenever the data contributions are interior). The intuition behind this is as follows. Under data sharing, the only benefit that a firm derives from contributing data is the payment received from the platform. Under analytics sharing, there is a second benefit stemming from the fact that a necessary condition for a firm's own data to be combined with other firms' data is that this data is contributed to the platform. This makes the benefit from an additional unit of contributed data greater under analytics sharing.

Consider now stage 2. Let us assume that the industry profit function Π^t is concave in $\mathbf{x}$ and denote $\mathbf{x}^{t*}$ the unique vector of data contributions that maximizes Π^t under technology $t = D, A$. We assume that this vector is interior, that is, $0 < x_i^{t*} < y_i$ for all $i = 1, ..., N$. Note that the inequalities $x_i^{t*} < y_i$ are satisfied if and only if the data endowments y_i are sufficiently large.[15]

Under complete information, it is straightforward that the optimal two-part contract ($\mathbf{f}^{t*}$, $\mathbf{m}^{t*}$) is such that: $x_i^t(m_i^{t*}) = x_i^{t*}$, and the participation constraints of all firms are binding. The two-part contract chosen by the platform induces data contributions that maximize the industry profits, and such profits are fully captured by the platform.

At stage 1, the platform compares

$$\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) = \max_{\mathbf{x}} \Pi^D(\mathbf{y}, \mathbf{x})$$

with

$$\Pi^A(\mathbf{x}^{A*}) = \max_{\mathbf{x}} \Pi^A(\mathbf{x}).$$

This boils down to determining the sign of

$$\begin{aligned}\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*}) &= \sum_i \left[B_i^D(\mathbf{y}, \mathbf{x}^{D*}) - C_i^D(\mathbf{y}, \mathbf{x}^{D*}) - \left(B_i^A(\mathbf{x}^{A*}) - C_i^A(\mathbf{x}^{A*})\right)\right] \\ &- \sum_i \left[g_i^D(x_i^{D*}) - g_i^A(x_i^{A*})\right] - \left[G^D(\mathbf{x}^{D*}) - G^A(\mathbf{x}^{A*})\right]. \qquad (5)\end{aligned}$$

In the following analysis, we emphasize the role of asymmetric (data-related) costs and, to that end, we assume that analytics sharing is more cost-efficient from the perspective of the firms than data sharing:

Assumption 1 (Security advantage). *$g_i^A(\cdot) = g_i(\cdot)$ and $g_i^D(\cdot) = (1 + \gamma^D)g_i(\cdot)$ where $\gamma^D > 0$.*

As per the benefits related to the two different technologies, as well as platform and analytics costs, we posit symmetry between both technologies. Specifically, we assume that:

[15] This holds because x_i^{t*} does not depend on y_i whenever it is interior, which immediately follows from the first-order condition defining x_i^{t*}.

Assumption 2 (Symmetry)**.** (*i*) *Analytics benefits:* $B_i^D(\cdot) = B_i^A(\cdot) = B_i(\cdot)$*;*

(*ii*) *Analytics costs:* $C_i^D(\cdot) = C_i^A(\cdot) = C_i(\cdot)$*;*

(*iii*) *Platform costs:* $G^D(\cdot) = G^A(\cdot) = G(\cdot)$*.*

In Assumption 1, we assume that, from the perspective of the firms, the two technologies differ in their security-related expected costs. In particular, the parameter γ^D can be interpreted as a measure of the security advantage of analytics sharing relative to data sharing.

This assumption builds on the fact that, in data sharing, as firms access the contributed data directly, data transfers may take place between the platforms and all the contributing firms. In contrast, under analytics sharing, the platform gives firms access to analytics, not data. Additionally, under data sharing, if firms can store the contributed data in-house, the number of potential access points for a cyberattack increases compared to analytics sharing. Therefore, data sharing is more vulnerable to data breaches than analytics sharing. It follows that the (expected) security-related costs for firms are higher under data sharing than under analytics sharing.

Assumption 2 improves the tractability of our analysis by allowing us to focus on the role of the heterogeneity in the security-related costs of different technologies.

Under Assumptions 1 and 2, we can express (5) as:

$$\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*}) = \underbrace{\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^D(\mathbf{x}^{D*}, \mathbf{x}^{D*})}_{\text{data endowment effect>0}} + \underbrace{\Pi^D(\mathbf{x}^{D*}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*})}_{\text{data security effect<0}}.$$

The *data endowment effect* captures the fact that the data sharing technology allows firms to combine the data shared by other firms with all their data endowment, while the analytics sharing technology does not. This effect is positive (i.e., it favors data sharing) and increasing in the data endowments y_i.[16] The *data security effect* captures the superiority of analytics sharing, driven by its lower data-related security costs. This effect is negative (i.e., it favors analytics sharing) and, in absolute value, it is increasing in γ^D. To see why it is negative, note that:

$$\Pi^D(\mathbf{x}^{D*}, \mathbf{x}^{D*}) = \sum_i \left[B_i(x_i^{D*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}*}) - C_i(\mathbf{x}^{D*}, \mathbf{x}^{D*}) - (1+\gamma^D)g_i(x_i^{D*})\right] - G(\mathbf{x}^{\mathbf{D}*})$$

$$< \sum_i \left[B_i(x_i^{D*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}*}) - C_i(\mathbf{x}^{D*}, \mathbf{x}^{D*}) - g_i(x_i^{D*})\right] - G(\mathbf{x}^{\mathbf{D}*}) \leq \max_{\mathbf{x}} \Pi^A(\mathbf{x}) = \Pi^A(\mathbf{x}^{A*}).$$

[16] Note that this effect would not exist if all firms contributed all their data under the analytics sharing technology (i.e., $\mathbf{x}^{D*} = \mathbf{y}$). This would require the firms' data endowments to be sufficiently small, which we ruled out.

The next proposition shows that analytics sharing is chosen if and only if it is sufficiently more secure than data sharing, and that this condition is more stringent the greater firms' data endowments are.

Proposition 1 (**Technological Choice**). *There exists a threshold* $\tilde{\gamma}^D(\mathbf{y}) > 0$ *such that the platform prefers the data sharing technology* (*resp. analytics sharing*) *technology if* γ^D *is below* (*resp. above*) $\tilde{\gamma}^D(\mathbf{y})$. *Moreover,* $\tilde{\gamma}^D(\mathbf{y})$ *is increasing in firms' data endowments* y_i.

Proof: See Appendix B.1.

Proposition 1 implies that the mere existence of a security advantage for the analytics sharing technology is not sufficient for this technology to be chosen by the platform. It is only if that security advantage is strong enough that the platform prefers the analytics sharing technology.

Data contributions. Under Assumption 2, the marginal benefits (net of all costs) of increasing data contribution respectively under data sharing and under analytics sharing are given by:

$$\frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i} = \sum_{j\neq i} \underbrace{\frac{\partial \left[B_j(y_j,\mathbf{x_{-j}}) - C_j(y_j,\mathbf{x_{-j}})\right]}{\partial x_i}}_{>0} - g_i^{\prime D}(x_i) - \frac{\partial G}{\partial x_i}(\mathbf{x}); \tag{6}$$

$$\begin{aligned}\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} = {} & \underbrace{\frac{\partial[B_i(x_i,\mathbf{x_{-i}}) - C_i(x_i,\mathbf{x_{-i}})]}{\partial x_i}}_{>0} + \sum_{j\neq i} \underbrace{\frac{\partial \left[B_j(x_j,\mathbf{x_{-j}}) - C_j(x_j,\mathbf{x_{-j}})\right]}{\partial x_i}}_{>0} \\ & - g_i^{\prime A}(x_i) - \frac{\partial G}{\partial x_i}(\mathbf{x}).\end{aligned} \tag{7}$$

Denote $NB_i(\cdot) \equiv B_i(\cdot) - C_i(\cdot)$ the *net benefit of analytics* for firm i, and assume that it is increasing in all its arguments.[17]

Then, by subtracting the two marginal benefits, (6) and (7), we obtain:

$$\begin{aligned}\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} - \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i} = {} & \underbrace{\frac{\partial NB_i(x_i,\mathbf{x_{-i}})}{\partial x_i}}_{>0} + \sum_{j\neq i} \left[\underbrace{\frac{\partial NB_j(x_j,\mathbf{x_{-j}})}{\partial x_i} - \frac{\partial NB_j(y_j,\mathbf{x_{-j}})}{\partial x_i}}_{\gtrless 0} \right] \\ & + g_i^{\prime D}(x_i) - g_i^{\prime A}(x_i).\end{aligned}$$

[17] We note that under data sharing, the firm both reaps the benefits and bears the costs of analytics, whereas in the case of analytics sharing the costs of conducting analytics are borne by the platform. Still, for the industry profits (which is full captured by the platform in equilibrium), it is the net benefits of analytics, as defined above, that play an important role.

Under Assumption 1, it follows that:

$$\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} - \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i} = \underbrace{\frac{\partial NB_i(x_i, \mathbf{x_{-i}})}{\partial x_i}}_{>0} + \sum_{j \neq i} \left[\underbrace{\frac{\partial NB_j(x_j, \mathbf{x_{-j}})}{\partial x_i} - \frac{\partial NB_j(y_j, \mathbf{x_{-j}})}{\partial x_i}}_{\gtreqless 0} \right] + \underbrace{\gamma^D g_i^{'}(x_i)}_{>0}.$$

The above equation shows that the difference between the marginal benefits of increasing data contribution under analytics sharing and data sharing can be decomposed into three terms. The first term, which is positive, is the net benefit of a marginal increase in a firm's own data on the benefit it derives from analytics under the analytics sharing technology. The third term is also positive and captures the marginal gain induced by the security advantage of analytics sharing. The second term represents the difference in the marginal cross-data effects between analytics sharing and data sharing and has a generally ambiguous sign. However, notice that:

$$\sum_{j \neq i} \left[\frac{\partial NB_j(x_j, \mathbf{x_{-j}})}{\partial x_i} - \frac{\partial NB_j(y_j, \mathbf{x_{-j}})}{\partial x_i} \right] > 0 (< 0) \text{ if } \frac{\partial^2 NB_j}{\partial x_j \partial x_i} < 0 (> 0) \text{ for any } i \neq j.$$

This means that the difference in the marginal cross-data effects between analytics sharing and data sharing is negative when the net benefit of analytics NB_j is submodular in (x_i, x_j), and positive when it is supermodular in (x_i, x_j). As a consequence, if $\gamma^D > 0$ and NB_j is weakly submodular for all j, then $\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} \geq \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i}$. This, combined with the fact that Π^t is concave in $\mathbf{x}$ for $t = D, A$, implies that, for a given $\mathbf{x_{-i}}$, the optimal data contribution for firm i under analytics sharing $x_i^A(\mathbf{x_{-i}})$ is greater than its counterpart under data sharing $x_i^D(\mathbf{y}, \mathbf{x_{-i}})$. It can easily be shown that this property also holds if the net benefit functions are supermodular as long as they are not "too supermodular".

The following proposition goes one step beyond the above analysis by showing that analytics sharing leads to more *equilibrium* data contributions than data sharing under two relatively mild assumptions on the profit and net benefit functions.

Proposition 2 (**Data Contributions**). *Analytics sharing generates more equilibrium data contributions than data sharing (i.e., $x_i^{A*} > x_i^{D*}$ for any i) under the following conditions: (i) $\Pi^A(\mathbf{x})$ and $\Pi^D(\mathbf{y}, \mathbf{x})$ are supermodular in (x_i, x_j) for any $i \neq j$, i.e., $\frac{\partial^2 \Pi^A(\mathbf{x})}{\partial x_j \partial x_i} > 0$ and $\frac{\partial^2 \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_j \partial x_i} > 0$, ($ii$) NB_j is not "too supermodular" in (x_i, x_j), i.e., $\frac{\partial^2 NB_j}{\partial x_j \partial x_i} < \tilde{k} = \frac{\min_i \min_{\mathbf{x}} \frac{\partial NB_i}{\partial x_i}(x_i, \mathbf{x_{-i}})}{\sum_j y_j}$ for any x and any i and j such that $i \neq j$.*

Proof: See Appendix B.2.

Proposition 2 provides conditions under which the intuition that contributing a unit of data is more valuable under analytics sharing than under data sharing discussed in the analysis of stage 3 (for a given per-unit transfer) remains true in equilibrium when the per-unit transfer is endogenized under both regimes. The reasoning behind this proposition proceeds in two steps.

First, as shown in the analysis preceding the proposition, the condition that the benefit functions are not too supermodular (i.e., there are no strong returns to data combination) ensures that contributing an additional unit of data leads to a higher increase in profits under analytics sharing than under data sharing for *given* contributions by the other firms. The requirement that the benefit function is not too supermodular appears to be in line with recent empirical evidence on data combination and the returns to machine learning analytics. Several studies find returns to scale and scope in data combination to be either decreasing or increasing only up to a certain amount of data contributed (Schäfer and Sapi, 2023; Hocuk et al., 2022; Lee and Wright, 2023; Yoganarasimhan, 2020; Azevedo et al., 2020; Peukert et al., 2023; Junqué de Fortuny et al., 2013).

Once it is established that the marginal increase in profits from contributing data is higher under analytics sharing than data sharing, it remains to show that this leads to higher *equilibrium* data contributions. A sufficient condition for this to be true is that the profit functions satisfy the supermodularity property.

We assume in the remainder of the paper that conditions (i) and (ii) of Proposition 2 hold.

6 Implications for consumer surplus

In this section we characterize the data-combination technology preferred by consumers and compare it with the one chosen by the platform.

6.1 Reduced-form setting

For the sake of simplicity, we assume that the consumers of products sold by firms $i = 1, ..., N$ do not suffer any harm in case of data leakages. This is the case, for instance, if the liability regime is such that consumers are fully compensated by the firms sharing their data in case they are harmed due to a data leakage. Consistent with this, we suppose that the consumer surplus generated by the consumption of the product of firm i is given by $CS_i(z_i, \mathbf{x_{-i}})$ where z_i is the amount of firm i's *combined* data: $z_i = y_i$ under data sharing and $z_i = x_i$ under analytics sharing.

To compare the consumer surplus generated by the consumption of product i under the two data-combination technologies in equilibrium, denote $CS_i^{D^*} = CS_i(y_i, \mathbf{x}_{\mathbf{-i}}^{\mathbf{D}^*})$, and

$CS_i^{A^*} = CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{A}^*})$. We have

$$CS_i^{D^*} - CS_i^{A^*} = \underbrace{\Big[CS_i(y_i, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}^*}) - CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}^*})\Big]}_{\text{Firm's own data effect}} + \underbrace{\Big[CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}^*}) - CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{A}^*})\Big]}_{\text{Other firms' data effect}}.$$

The above equation shows that the difference between consumer surplus under data sharing and under analytics sharing can be decomposed into a term that captures the impact of the difference in firm i's combined data under the two technologies on consumer surplus (i.e., $CS_i(y_i, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}^*}) - CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}^*})$), and a term that captures the effect of the difference in other firms' combined data under the two technologies on consumer surplus (i.e., $CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{D}^*}) - CS_i(x_i^{A^*}, \mathbf{x}_{-\mathbf{i}}^{\mathbf{A}^*})$). We call the former the *firm's own data effect* and the latter the *other firms' data effect*.

Notice that the firm's own data effect is weakly positive (resp. negative) if $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ is increasing (resp. decreasing) in z_i because $y_i \geq x_i^{A^*}$, whereas the other firms' data effect is positive (resp. negative) if $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ is decreasing (resp. increasing) in all the components of $\mathbf{x}_{-\mathbf{i}}$ because $x_{-i}^{D^*} < x_{-i}^{A^*}$. Using the above decomposition, we establish the following result:

Proposition 3. *The preferred technology from the perspective of consumers of product i depends on the monotonicity properties of $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ in the following way:*

(*i*) *If $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ is increasing in z_i and decreasing in $\mathbf{x}_{-\mathbf{i}}$, then data sharing always dominates analytics sharing;*

(*ii*) *If $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ is decreasing in z_i and increasing in $\mathbf{x}_{-\mathbf{i}}$, then analytics sharing always dominates data sharing;*

(*iii*) *If $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ is decreasing in both z_i and $\mathbf{x}_{-\mathbf{i}}$, then there exists $\tilde{\gamma}^{CS}(y_i) \in [0, +\infty[\cup\{+\infty\}$ (weakly) increasing in y_i such that data sharing (weakly) dominates analytics sharing if and only if $\gamma \geq \tilde{\gamma}^{CS}(y_i)$;*

(*iv*) *If $CS_i(z_i, \mathbf{x}_{-\mathbf{i}})$ is increasing in both z_i and $\mathbf{x}_{-\mathbf{i}}$, then there exists $\hat{\gamma}^{CS}(y_i) \in [0, +\infty[\cup\{+\infty\}$ (weakly) decreasing in y_i such that analytics sharing (weakly) dominates data sharing if and only if $\gamma \geq \hat{\gamma}^{CS}(y_i)$.*

Proof: See Appendix B.3.

In cases (i) and (ii), the way consumer surplus generated by the consumption of product i is affected by the amount of firm i's data that gets combined (z_i) is opposite to the way it is affected by the amounts of other firms' data that get combined ($\mathbf{x}_{-\mathbf{i}}$). This implies that the firm's own data effect and the other firms' data effect have the same sign. This sign is positive in case (i) and negative in case (ii), which implies that data sharing is preferred by consumers in case (i), whereas analytics sharing is preferred by consumers in case (ii).

In cases (iii) and (iv), the monotonicity of consumer surplus generated by the consumption of product i with respect firm i's data is the same as its monotonicity with respect to other firms' data. This implies that the firm's own data effect and the other firms' data effect have opposite signs and, therefore, the comparison between consumer surplus under the two technologies is *a priori* ambiguous.

However, it is possible to show that there exists a (potentially infinite) threshold of the security advantage parameter γ above which data sharing is preferred in case (iii), and another threshold above which analytics sharing is preferred in case (iv). The statement regarding case (iii) follows from the fact that consumer surplus under data sharing increases with γ if consumer surplus is decreasing in other firms' amounts of contributed data because the latter are decreasing in γ. Similarly, the statement in case (iv) follows from the fact that consumer surplus under data sharing decreases with γ if consumer surplus is increasing in other firms' amounts of contributed data because the latter are decreasing in γ.

Propositions 1 and 3 show that there are several cases in which the technology chosen by the platform is not the one that is preferred by consumers. The potential misalignment stems from the fact that an increase in the amount of combined data may benefit the platform while harming consumers. We further clarify this in the next section.

6.2 Microfoundation

We now present a microfoundation for our reduced-form setting. This microfoundation is simple but rich enough to generate all four scenarios arising in Proposition 3. Suppose that firms operate in separate markets and that each firm i is a monopolist in its market (also denoted i). We assume that firm i produces at a constant marginal cost c_i. The use of insights from combined data allows a given firm to (i) offer a product of higher quality to its customers and/or (ii) profile some of its customers and engage in price discrimination by charging the profiled customers personalized prices (see, e.g., de Cornière and Taylor, 2025).

In each market i, there is a unit mass of consumers who have heterogeneous valuations for the good sold by firm i. Specifically, a consumer's gross utility from consuming product i is given by:

$$u_i(z_i, \mathbf{x_{-i}}) = v_i + q_i(z_i, \mathbf{x_{-i}}),$$

where v_i, the stand-alone valuation for the good, is distributed over an interval $[\underline{v}_i; \overline{v}_i]$ with a c.d.f. $F_i(\cdot)$ and a p.d.f. $f_i(\cdot)$, and $q_i(z_i, \mathbf{x_{-i}})$, the data-driven quality, is (weakly) increasing in all its arguments. This implies that the quality of firm i's product is (weakly) increasing in the amount of combined data, which can be justified by the ability of firms to improve their products by learning from their and other firms' customer data (see, e.g., Hagiu and Wright, 2023).

Moreover, firm i can find out the valuation of a consumer by using data. The probability that a consumer is profiled and her valuation becomes known to firm i is $\lambda_i(z_i, \mathbf{x_{-i}})$. Profiled consumers are charged personalized prices equal to their valuations, while non-profiled consumers are charged a non-personalized (uniform) price. We assume that the function $\lambda_i(\cdot)$ is (weakly) increasing in both its arguments, that is, greater amounts of combined data improve the firm's ability to profile consumers. Note that the probability of a consumer being profiled is independent of her valuation.

The demand from non-profiled consumers for a given non-personalized price p_i is

$$D_i(p_i, z_i, \mathbf{x_{-i}}) = [1 - \lambda_i(z_i, \mathbf{x_{-i}})][1 - F_i(p_i - q_i(z_i, \mathbf{x_{-i}}))].$$

Assuming that the profit $[1 - \lambda_i(z_i, \mathbf{x_{-i}})](p_i - c_i)[1 - F_i(p_i - q_i(z_i, \mathbf{x_{-i}}))]$ derived from the non-profiled segment of the market is quasi-concave in p_i and denoting $\tilde{p}_i(z_i, \mathbf{x_{-i}})$ the non-personalized price that maximizes it, i.e.,

$$\tilde{p}_i(z_i, \mathbf{x_{-i}}) = \arg\max_{p_i}(p_i - c_i)[1 - F_i(p_i - q_i(z_i, \mathbf{x_{-i}}))],$$

firm i's benefit function (i.e., gross profits) can be written as:

$$B_i(z_i, \mathbf{x_{-i}}) = \underbrace{\lambda_i(z_i, \mathbf{x_{-i}}) \int_{\underline{v}_i}^{\overline{v}_i} v_i dF_i(v_i)}_{\text{Profiled segment}} + \underbrace{[1 - \lambda_i(z_i, \mathbf{x_{-i}})] (\tilde{p}_i(z_i, \mathbf{x_{-i}}) - c_i)[1 - F_i(\tilde{p}_i(z_i, \mathbf{x_{-i}}) - q_i(z_i, \mathbf{x_{-i}}))]}_{\text{Non-profiled segment}}.$$

In this setting, profiled consumers do not get any surplus from consuming the product, while non-profiled infra-marginal consumers do. Specifically, consumer surplus in market i is given by:

$$CS_i(z_i, \mathbf{x_{-i}}) = [1 - \lambda_i(z_i, \mathbf{x_{-i}})] \int_{\tilde{p}_i(z_i, \mathbf{x_{-i}}) - q_i(z_i, \mathbf{x_{-i}})}^{\overline{v}_i} [v_i + q_i(z_i, \mathbf{x_{-i}}) - \tilde{p}_i(z_i, \mathbf{x_{-i}})] dF_i(v_i),$$

where the term $\tilde{p}_i(z_i, \mathbf{x_{-i}}) - q_i(z_i, \mathbf{x_{-i}})$ can be interpreted as the quality-adjusted price set by firm i for non-profiled consumers. It is easy to check that this term is decreasing in all the components of $(z_i, \mathbf{x_{-i}})$ under the assumption that the profit derived from the non-profiled segment is quasi-concave. This implies that, if firms are not able to use data to profile consumers (i.e., $\lambda_i(z_i, \mathbf{x_{-i}}) = 0$), consumer surplus is (weakly) increasing in the amounts of combined data, i.e.,

$$\int_{\tilde{p}_i(z_i, \mathbf{x_{-i}}) - q_i(z_i, \mathbf{x_{-i}})}^{\overline{v}_i} [v_i + q_i(z_i, \mathbf{x_{-i}}) - \tilde{p}_i(z_i, \mathbf{x_{-i}})] dF_i(v_i)$$

is (weakly) increasing in all its arguments. This, combined with the fact that profiling has a negative effect on consumer surplus (i.e., the term $1 - \lambda_i(z_i, \mathbf{x_{-i}})$ is weakly decreasing in its arguments), shows that $CS_i(z_i, \mathbf{x_{-i}})$ could be either increasing or decreasing in each of its arguments depending on how an increase in this argument affects the relative magnitudes of the two forces at play (higher quality vs more first-degree price discrimination). In other words, all scenarios in Proposition 3 can arise. For instance, if both own data and other firms' data are mostly used to offer higher-quality products (resp., to price discriminate) then scenario (iv) (resp., (iii)) is likely to arise . If own data (resp., other firms' data) is mostly used for the purpose of offering higher quality product, while other firms' data (resp,. own data) is mostly used for the purpose of price discrimination then scenario (i) (resp., (ii)) is likely to arise.

7 Extension: Technological choice under alternative contracts

In the baseline model, the platform uses personalized and public contracts. However, firms may require the platform to use secret contracts[18] or the platform may lack the necessary information to offer personalized contracts. This extension investigates the platform's choice of data-combination technology in these two scenarios and compares it to its choice in the baseline model.

7.1 Secret personalized contracts

Assume that the platform offers secret personalized contracts, and firms hold passive beliefs (i.e., if they receive an off-equilibrium contract offer, they believe that the platform did not change the offers made to the other firms).

Note first that assuming that contracts are secret does not affect the outcomes of stages 2 and 3 under data sharing. The reason is that, under this technology, firms' optimal data contributions in stage 3 do *not* depend on other firms' data contributions. This implies that, under data sharing, the platform is still able to induce data contributions that maximize industry profits (and capture these profits). Thus, the platform's profit under data sharing remains the same as in the baseline model.

In contrast, if the platform offers secret personalized contracts, its profits under analytics sharing are (strictly) lower than its counterpart in the baseline model. The reason is that firms' optimal data contributions now depend on other firms' data contributions, which prevents the platform from inducing industry-profit-maximizing data contributions. This is due to a classic opportunism problem in vertical contracting with multiple firms (see, e.g., McAfee and Schwartz, 1994) and has been illustrated in the case of data transactions

[18] As discussed by Calzolari et al. (2024), firms producing data may value *anonymity* over whether they have joined the platform and the contractual details.

by Calzolari et al. (2024). In our setting, if we denote $\mathbf{x}^e_{-i}$ firm i's beliefs about other firms' data contributions, the first-order condition determining firm i's optimal data contribution under analytics sharing for a given m^A_i is given by

$$\frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x}^e_{-i}) + m^A_i = g'_i(x_i).$$

Denoting $\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)$ the solution to the equation above, and assuming that the platform still finds it optimal to induce participation of all firms, the platform's maximization program can be written as

$$\max_{(\mathbf{f}^A,\mathbf{m}^A)} \sum_i [f^A_i - m^A_i \hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)] - \sum_i C_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \hat{\mathbf{x}}_{-i}) - G((\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i))_{1\leq i\leq n})$$

subject to the participation constraints

$$B_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \mathbf{x}^e_{-i}) - g_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)) + m^A_i \hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i) - f^A_i \geq 0,$$

$i = 1, ..., N$. Since the participation constraints must be binding at the optimum, i.e., the fixed fees must be given by $f^A_i = B_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \mathbf{x}^e_{-i}) - g_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)) + m^A_i \hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)$, the platform's maximization program with respect to $\mathbf{m}^A$ can be rewritten as

$$\begin{aligned}\max_{\mathbf{m}^A} & \sum_i [B_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \mathbf{x}^e_{-i}) - g_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i))] \\ & - \sum_i C_i(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \hat{\mathbf{x}}_{-i}) - G((\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i))_{1\leq i\leq n}).\end{aligned}$$

The first-order condition with respect to m^A_i yields

$$\begin{aligned}\frac{\partial \hat{x}^A_i}{\partial m^A_i} \bigg\{ & \frac{\partial B_i}{\partial x_i}(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \mathbf{x}^e_{-i}) - \frac{\partial g_i}{\partial x_i}(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)) \\ & - \frac{\partial C_i}{\partial x_i}(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \hat{\mathbf{x}}_{-i}) - \frac{\partial G}{\partial x_i}((\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), m^A_i))_{1\leq i\leq n}) \bigg\} = 0\end{aligned}$$

or, equivalently,

$$\begin{aligned}& \frac{\partial B_i}{\partial x_i}(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \mathbf{x}^e_{-i}) - \frac{\partial g_i}{\partial x_i}(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i)) \\ & - \frac{\partial C_i}{\partial x_i}(\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), \hat{\mathbf{x}}_{-i}) - \frac{\partial G}{\partial x_i}((\hat{x}^A_i(\mathbf{x}^e_{-i}, m^A_i), m^A_i))_{1\leq i\leq n}) = 0.\end{aligned}$$

This first-order condition is different from the one that maximizes industry profits with respect to data contributions. More specifically, the terms capturing the effect of a firm's data endowments on other firms in the first-order condition associated to industry-profit maximization are missing. This shows that the platform's profit under analytics sharing are (strictly) lower than in the baseline model. Note, however, that it is still the case that the platform's profits under data sharing are (strictly) decreasing in γ^D and goes to zero as γ^D grows large. It is also still the case that the platform's profit under data sharing is increasing in data endowments y_i. Therefore, we have the following result.

Proposition 4. *Assume that the platform offers secret personalized contracts. There exists a threshold $\hat{\gamma}^D(\mathbf{y})$ increasing in data endowments y_i such that the platform chooses analytics sharing if and only if $\gamma \geq \hat{\gamma}^D(\mathbf{y})$. However, analytics sharing is less likely to be chosen than in the baseline model with public personalized contracts, i.e., $\hat{\gamma}^D(\mathbf{y}) > \tilde{\gamma}^D(\mathbf{y})$.*

7.2 Uniform public contract

Let us now assume that the platform offers a uniform public contract to all firms (i.e., $f_i^t = f_j^t$ and $m_i^t = m_j^t$ for all $i \neq j$). In this case, the result obtained in the previous section can be reversed, i.e., analytics sharing can become more likely to be chosen than in the baseline model.

We show this in a stark way, by assuming that all firms have the same benefit and cost functions under a given technology, i.e., $B_i(\cdot) = B_j(\cdot)$, $C_i(\cdot) = C_j(\cdot)$, and $g_i(\cdot) = g_j(\cdot)$ for all $i \neq j$. It follows that firms are heterogeneous only with respect to their data endowments. This heterogeneity is payoff-relevant in the case of data sharing but is *not* in the case of analytics sharing. This in turn implies that the platform's profit under analytics sharing remains the same whether contracts are personalized and public (as in the baseline model) or uniform and public.

However, if contracts are uniform and public, the platform's profit under data sharing is lower than with personalized and public contracts because the platform is no longer able to capture the whole industry profit as soon as there is some heterogeneity in firms' data endowments. This, combined with the fact that the platform's profit under data sharing increases with data endowments y_i, decreases with γ^D, and it goes to zero as γ^D grows large, leads to the following result.

Proposition 5. *Assume that the platform offers a uniform public contract to all firms and that all firms have the same benefit and cost functions under a given technology. There exists a threshold $\check{\gamma}^D(\mathbf{y})$ increasing in data endowments y_i such that the platform chooses analytics sharing if and only if $\gamma \geq \check{\gamma}^D(\mathbf{y})$. Moreover, analytics sharing is more likely to be chosen than in the baseline model with public personalized contracts, i.e., $\check{\gamma}^D(\mathbf{y}) < \tilde{\gamma}^D(\mathbf{y})$.*

8 Concluding remarks

In this article, we have studied a platform's choice of a data-combination technology and its implications for firms' incentives to combine data as well as for consumers. To investigate these issues, we have considered a setting in which a platform facilitating inter-firm data combination chooses between data sharing and analytics sharing technologies. Data sharing enables firms to exploit the joint dataset along with all their own data, whereas analytics sharing offers a security advantage by reducing the amount of data transmission and points of data access.

We have demonstrated that the choice between the two technologies depends on how the platform navigates the above trade-off. In particular, since data sharing inherently benefits from a data endowment advantage over analytics sharing, the platform prefers analytics sharing only if its security advantage is sufficiently large. Moreover, our analysis has shown that analytics sharing leads to higher equilibrium data contributions than data sharing. This holds under relatively general conditions that appear to be consistent with recent empirical studies investigating economies of scale and scope in data combination.

We have also highlighted a potential misalignment between the platform's and consumers' preferred technology, which stems from the fact that an increase in the amount of combined data may harm consumers while benefiting the platform. This occurs because consumers' preferred technology results from how the consumer surplus generated by the consumption of the product of a firm depends on data contributions. This effect can be either positive (e.g., if data is used to provide higher-quality products) or negative (e.g., if data enables price discrimination). The platform, in turn, prefers analytics sharing if and only if its security advantage over data sharing is sufficiently strong.

Our findings have a number of relevant implications. To begin with, extensive literature shows that, in many different contexts,there is a trade-off between privacy and the efficiency gains stemming from (consumer) data combination. In our model, there is a potential to invert the standard causality of the privacy-efficiency trade-offs. In particular, an increase in the security advantage of analytics sharing over data sharing raises the likelihood of adopting a more secure technology, and leads to an increase in the amount of data being combined. This is in line with the observation that the advent of privacy-enhancing technologies has the potential of relaxing and even eliminating this trade-off (Acquisti et al., 2016; Johnson, 2022).

From a policy perspective, these findings suggest that policymakers aiming to strengthen privacy protection while fostering data contribution for societal benefits should design policies aimed at promoting improvements in the data security of analytics sharing technologies. While this would increase the probability of analytics sharing being the privately chosen technology, its impact on consumers depends on how data combination influences their surplus. When security-enhancing technologies fail to benefit consumers, policies supporting analytics-sharing should be complemented by measures to prevent data leaks and price discrimination.

From a managerial perspective, the choice of data combination technology is delicate and potentially hard to reverse. This holds true for both the platforms and for firms seeking to extract insights from their data. While data sharing offers the possibility of exploiting in-house, non-shared data, it also requires firms to invest in adequate capabilities to perform their own data analytics. This can be particularly costly for firms with limited ICT skills and experience, reinforcing arguments in favor of adopting analytics sharing, beyond its increased data security.

A further managerial insight of our analysis is that the nature of the contractual arrangements between the platform and the firms can affect the choice between data sharing and analytics sharing. If the platform uses secret personalized contracts, analytics sharing becomes less attractive than when it uses public personalized contracts. By contrast, if the platform uses a uniform public contract, analytics sharing becomes more attractive to the platform compared to the scenario where it uses public personalized contracts.

Finally, since analytics sharing can lead to higher equilibrium data contributions, platform managers might prefer this option to maximize the amount of data available for analysis. This is particularly relevant if the platform's business model benefits significantly from the volume of data that it holds and processes.

Our setting can be extended to investigate other interesting issues related to multi-firm data-combination technologies. First, we could provide a microfoundation for our reduced-form model in which data-holding firms compete in the same final market. This would enable us to explore the competitive implications of selecting a data-combination technology and the potential tensions between privacy and security protection on one hand and competition on the other.

Finally, the model could also be extended to encompass competition between platforms offering data-combination services. As analytics sharing technologies mature and become widespread, incumbent data combination platforms based on data sharing (e.g., Caruso[19] in the connected car industry, the NINDS Parkinson's Disease Biomarkers Program[20] in the health sector), will likely face the threat of entry using analytics sharing technologies.

[19] See: https://www.caruso-dataplace.com/.

[20] See: NINDS boosts research for biomarkers in Parkinson's disease.

References

AbdulRahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C., and Guizani, M. (2020). A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. *IEEE Internet of Things Journal*, 8(7):5476–5497.

Abrardi, L., Cambini, C., Congiu, R., and Pino, F. (2024a). User data and endogenous entry in online markets. *Journal of Industrial Economics*, forthcoming.

Abrardi, L., Cambini, C., and Pino, F. (2024b). Data brokers competition, synergic datasets, and endogenous information value. *SSRN Working Paper 4901441*.

Acquisti, A., Taylor, C., and Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, 54(2):442–492.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: the simple economics of Artificial Intelligence*. Harvard Business Press.

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J., and Weyl, E. G. (2020). A/B testing with fat tails. *Journal of Political Economy*, 128(12):4614–000.

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.

Bergemann, D., Bonatti, A., and Gan, T. (2022). The economics of social data. *RAND Journal of Economics*, 53(2):263–296.

Bounie, D., Dubus, A., and Waelbroek, P. (2021). Selling strategic information in competitive markets. *RAND Journal of Economics*, 52(2):283–313.

Calzolari, G., Cheysson, A., and Rovatti, R. (2024). Machine data: market and analytics. *Management Science*, forthcoming.

Carballa Smichowski, B. (2018). Determinants of coopetition through data sharing in MaaS. *Management & Data Science*, 2(3).

Chiou, L. and Tucker, C. (2017). Search engines and data retention: Implications for privacy and antitrust. *NBER Working Paper 23815*.

Clarke, R. N. (1983). Collusion and the incentives for information sharing. *Bell Journal of Economics*, pages 383–394.

de Cornière, A. and Taylor, G. (2025). Data and competition: A simple framework. *RAND Journal of Economics*, forthcoming.

Delbono, F., Reggiani, C., and Sandrini, L. (2024). Strategic data sales with partial segment profiling. *Information Economics & Policy*, 68(101102).

Drainakis, G., Pantazopoulos, P., Katsaros, K. V., Sourlas, V., Amditis, A., and Kaklamani, D. I. (2023). From centralized to federated learning: Exploring performance and end-to-end resource consumption. *Computer Networks*, 225:109657.

European Commission (2024). Common european data spaces. https://digital-strategy.ec.europa.eu/en/policies/data-spaces. [Last accessed February 4, 2025].

Farrell, E., Minghini, M., Kotsev, A., Soler-Garrido, J., Tapsall, B., Micheli, M., Posada, M., Signorelli, S., Tartaro, A., Bernal, J., Vespe, M., Di Leo, M., Carballa-Smichowski, B., Smith, R., Schade, S., Pogorzelska, K., Gabrielli, L., and De Marchi, D. (2023). European data spaces – scientific insights into data sharing and utilisation at scale. *Joint Research Centre - Science for Policy Report*, JRC129900.

Fried, D. (1984). Incentives for information production and disclosure in a duopolistic environment. *Quarterly Journal of Economics*, 99(2):367–381.

Gal-Or, E. (1985). Information sharing in oligopoly. *Econometrica*, pages 329–343.

Gal-Or, E. (1986). Information transmission—Cournot and Bertrand equilibria. *Review of Economic Studies*, 53(1):85–92.

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. (2022). Data brokers co-opetition. *Oxford Economic Papers*, 74(3):820–839.

Gu, Y., Madio, L., and Reggiani, C. (2024). Data-driven market leadership and price coordination. *SSRN Working Paper 4002261*.

Hagiu, A. and Wright, J. (2023). Data-enabled learning, network effects, and competitive advantage. *RAND Journal of Economics*, 54(4):638–667.

Hocuk, S., Martens, B., Prufer, P., Carballa Smichowski, B., and Duch-Brown, N. (2022). Economies of scope in data aggregation: Evidence from health data. *TILEC Discussion Paper DP2022-020*.

Hviid, M. (1989). Risk-averse duopolists and voluntary information transmission. *Journal of Industrial Economics*, pages 49–64.

Ichihashi, S. (2021). Competing data intermediaries. *RAND Journal of Economics*, 52(3):515–537.

Jentzsch, N., Sapi, G., and Suleymanova, I. (2013). Targeted pricing and customer data sharing among rivals. *International Journal of Industrial Organization*, 31(2):131–144.

Johnson, G. (2022). Economic research on privacy regulation: Lessons from the GDPR and beyond. *NBER Working Paper 30705*.

Junqué de Fortuny, E., Martens, D., and Provost, F. (2013). Predictive modeling with big data: is bigger really better? *Big Data*, 1(4):215–226.

Kirby, A. J. (1988). Trade associations as information exchange mechanisms. *RAND Journal of Economics*, pages 138–146.

Klein, T. J., Kurmangaliyeva, M., Prüfer, J., Prüfer, P., and Park, N. N. (2022). How important are user-generated data for search result quality? Experimental evidence. *TILEC Discussion Paper DP2022-016*.

Lee, G. and Wright, J. (2023). Recommender systems and the value of user data. *National University of Singapore*, Mimeo.

Li, L. (1985). Cournot oligopoly with information sharing. *RAND Journal of Economics*, pages 521–536.

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

Martens, B., Parker, G., Petropoulos, G., and Van Alstyne, M. W. (2021). Towards efficient information sharing in network markets. *TILEC Discussion Paper DP2021-014*.

Mattioli, M. (2017). The data-pooling problem. *Berkeley Technology Law Journal*, 32(1):179–236.

McAfee, R. P., Rao, J., Kannan, A., He, D., Qin, T., and Liu, T. (2015). Measuring scale economies in search. *LEAR Conference*.

McAfee, R. P. and Schwartz, M. (1994). Opportunism in multilateral vertical contracting: Nondiscrimination, exclusivity, and uniformity. *American Economic Review*, pages 210–230.

Neumann, N., Tucker, C. E., and Whitfield, T. (2019). How effective is black-box digital consumer profiling and audience delivery? Evidence from field studies. *SSRN Working Paper 3203131*.

Novshek, W. and Sonnenschein, H. (1982). Fulfilled expectations Cournot duopoly with information acquisition and release. *Bell Journal of Economics*, pages 214–218.

Padilla, A. J. and Pagano, M. (1997). Endogenous communication among lenders and entrepreneurial incentives. *Review of Financial Studies*, 10(1):205–236.

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

Peukert, C., Sen, A., and Claussen, J. (2023). The editor and the algorithm: Recommendation technology in online news. *Management Science*, forthcoming.

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

Ramírez, D. H., Díaz, L. P., Rahimian, S., García, J. M. A., Peña, B. I., Al-Khazraji, Y., Alarcón, Á. J. G., Fuente, P. G., Garrido, J. S., and Kotsev, A. (2023). Technological enablers for privacy preserving data sharing and analysis. *Joint Research Centre - Technical Report*, JRC134350.

Sakai, Y. (1986). Cournot and Bertrand equilibria under imperfect information. *Journal of Economics*, 46(3):213–232.

Schäfer, M. and Sapi, G. (2023). Complementarities in learning from data: Insights from general search. *Information Economics and Policy*, page 101063.

Shapiro, C. (1986). Exchange of cost information in oligopoly. *Review of Economic Studies*, 53(3):433–446.

Ullrich, H., Hannane, J., Peukert, C., Aguiar, L., and Duso, T. (2024). Returns to data: Evidence from web tracking. *Discussion Papers of DIW Berlin 2091*.

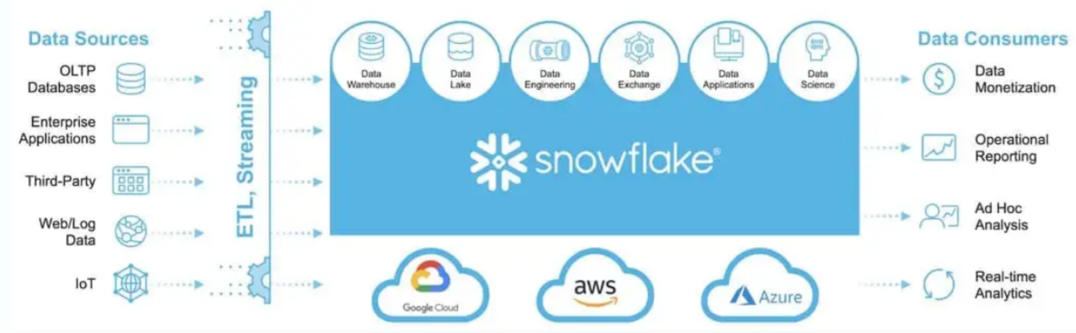
Vives, X. (1984). Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory*, 34(1):71–94.

Vives, X. (1990). Trade association disclosure rules, incentives to share information, and welfare. *RAND Journal of Economics*, 21(3):409–430.

Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science*, 66(3):1045–1070.

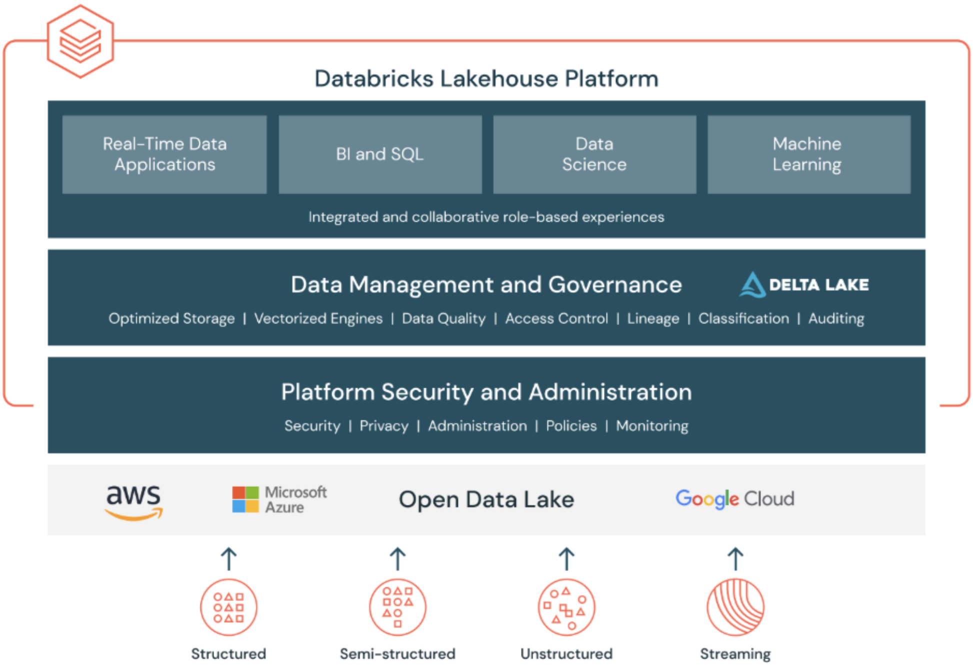
A Data sharing vs analytics sharing: motivating examples

Figure A.1: Data pipeline with Snowflake technology as part of it.



Source: snowflake.com

Figure A.2: Data pipeline and Databricks services.


Databricks Lakehouse Platform
Real-Time Data Applications
BI and SQL
Data Science
Machine Learning
Integrated and collaborative role-based experiences
Data Management and Governance
DELTA LAKE
Optimized Storage | Vectorized Engines | Data Quality | Access Control | Lineage | Classification | Auditing
Platform Security and Administration
Security | Privacy | Administration | Policies | Monitoring
aws
Microsoft Azure
Open Data Lake
Google Cloud
Structured
Semi-structured
Unstructured
Streaming

Source: databricks.com

B Proofs

B.1 Proof of Proposition 1

Given that: (i) $\Pi^D(\mathbf{y}, \mathbf{x}^{D*})$ is continuous and decreasing in γ^D; (ii) $\Pi^A(\mathbf{x}^{A*}) < \Pi^D(\mathbf{y}, \mathbf{x}^{D*})$ when γ^D goes to 0 because in this case $\Pi^A(\mathbf{x}) < \Pi^D(\mathbf{y}, \mathbf{x})$ for any $\mathbf{x}$ (which implies that $\max_{\mathbf{x}} \Pi^A(\mathbf{x}) < \max_{\mathbf{x}} \Pi^D(\mathbf{y}, \mathbf{x})$); (iii) $\Pi^D(\mathbf{y}, \mathbf{x}^{D*})$ goes to 0 as γ^D grows large (because $\mathbf{x}^{D*}$ goes to $\mathbf{0}$), while $\Pi^A(\mathbf{x}^{A*})$ is positive and does not depend on γ. Then $\exists\, \tilde{\gamma}^D(\mathbf{y}) > 0$ such that $\Pi^A(\mathbf{x}^{A*}) \geq \Pi^D(\mathbf{y}, \mathbf{x}^{D*}) \Leftrightarrow \gamma^D \geq \tilde{\gamma}^D$. Since $\Pi^D(\mathbf{y}, \mathbf{x})$ increases with y_i, it follows that $\Pi^D(\mathbf{y}, \mathbf{x}^{D*})$ also increases with y_i, which implies that $\tilde{\gamma}^D(\mathbf{y})$ is increasing in y_i. *Q.E.D.*

B.2 Proof of Proposition 2

We proceed in two steps.

Step 1: Let us show that $\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} > \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i}$ if condition (ii) holds.

Denote $k_i = \max_{j \neq i} \max_{\mathbf{x}} \frac{\partial^2 NB_j}{\partial x_i \partial x_j}(x_j, \mathbf{x_{-i}})$. Notice that

$$\frac{\partial NB_j}{\partial x_i}(y_j, \mathbf{x_{-j}}) - \frac{\partial NB_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) \leq (y_j - x_j)k_i \leq y_j k_i.$$

Therefore,

$$\sum_{j \neq i} \left[\frac{\partial NB_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) - \frac{\partial NB_j}{\partial x_i}(y_j, \mathbf{x_{-j}}) \right] > -k_i \sum_{j \neq i} y_j > -k_i \sum_j y_j.$$

Hence,

$$\begin{aligned} \frac{\partial NB_i}{\partial x_i}(x_i, \mathbf{x_{-i}}) + \sum_{j \neq i} \left[\frac{\partial NB_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) - \frac{\partial NB_j}{\partial x_i}(y_j, \mathbf{x_{-j}}) \right] + \gamma^D g_i'(x_i) \\ \geq \frac{\partial NB_i}{\partial x_i}(x_i, \mathbf{x_{-i}}) - k_i \sum_j y_j + \gamma^D g_i'(x_i). \end{aligned}$$

If $\gamma^D > 0$, then

$$\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} - \frac{\partial \Pi^D(\mathbf{y}, \mathbf{x})}{\partial x_i} \geq \frac{\partial NB_i}{\partial x_i}(x_i, \mathbf{x_{-i}}) - k_i \sum_j y_j.$$

because $\frac{\partial NB_j}{\partial x_i}(x_i, \mathbf{x_{-i}}) > 0$ and $\frac{\partial NB_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) > 0$. Hence, a sufficient condition for $\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} > \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i}$ is $k_i \leq \frac{\frac{\partial NB_i}{\partial x_i}(x_i, \mathbf{x_{-i}})}{\sum_j y_j}$, which holds if $k_i \leq \frac{\min_i \min_{\mathbf{x}} \frac{\partial NB_j}{\partial x_j}(x_j, \mathbf{x_{-j}})}{\sum_j y_j} \equiv \tilde{k}$, which completes the first step.

Step 2: Let us now show that, under condition (i), the set of inequalities $\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} > \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i}$ for any $\mathbf{x}$ and any $\mathbf{y}$ implies that $x_i^{A*} > x_i^{D*}$ for any i. Notice that the supermodularity of $\Pi^A(\mathbf{x})$ and $\Pi^D(\mathbf{y},\mathbf{x})$ with respect to (x_i, x_j) for any $i \neq j$ implies that $x_i^A(\mathbf{x_{-i}})$ and $x_i^D(\mathbf{y}, \mathbf{x_{-i}})$ are increasing in all x_j's. To see why, recall that the first order condition defining $x_i^A(\mathbf{x_{-i}})$ is:

$$\frac{\partial \Pi^A}{\partial x_i}(x_i^A(\mathbf{x_{-i}}), \mathbf{x_{-i}}) = 0.$$

Differentiating the above equation with respect to x_j and rearranging terms, we obtain:

$$\frac{\partial x_i^A}{\partial x_j} = \frac{\frac{\partial^2 \Pi^A}{\partial x_i \partial x_j}}{-\frac{\partial^2 \Pi^A}{\partial x_i^2}} > 0$$

because the numerator is positive due to supermodularity of Π^A and the denominator is positive due to the concavity of Π^A. A similar reasoning applies to $x_i^D(\mathbf{y}, \mathbf{x_{-i}})$.

We establish Step 2 recursively, that is, we show that the result holds for $N = 2$ and, whenever it holds for a given $N \geq 2$, it holds for $N + 1$ too.

Let us first show that the result holds for $N = 2$. Define $H_2^A(\cdot)$ and $H_2^D(\mathbf{y}, \cdot)$ as follows:

$$H_2^A(x_2) = x_2^A(x_1^A(x_2)) - x_2 \text{ and } H_2^D(\mathbf{y}, x_2) = x_2^D(y_2, x_1^D(y_1, x_2)) - x_2.$$

Since $x_2^{A*} = x_2^A(x_1^{A*}) = x_2^A(x_1^A(x_2^{A*}))$ and $x_2^{D*} = x_2^D(y_2, x_1^{D*}) = x_2^D(y_2, x_1^D(y_1, x_2^{D*}))$ we have that:

$$H_2^A(x_2^{A*}) = 0 \text{ and } H_2^D(\mathbf{y}, x_2^{D*}) = 0.$$

Since $H_2^A(0) > 0$, the uniqueness of (x_1^{A*}, x_2^{A*}) as a maximizer of $\Pi^A(x_1, x_2)$ ensures that $H_2^A(x_2) > 0$ for any $x_2 < x_2^{A*}$ and $H_2^A(x_2) < 0$ for any $x_2 > x_2^{A*}$. Similarly, $H_2^D(\mathbf{y}, x_2) > 0$ for any $x_2 < x_2^{D*}$ and $H_2^D(\mathbf{y}, x_2) < 0$ for any $x_2 > x_2^{D*}$.

Let us now show that $x_2^{A*} > x_2^{D*}$. In order to do so, assume that the reverse holds, i.e. $x_2^{A*} \leq x_2^{D*}$. Then, this implies that $H_2^A(x_2^{s*}) \leq 0$. Moreover,

$$\begin{aligned} H_2^D(\mathbf{y}, x_2^{D*}) - H_2^A(x_2^{D*}) &= x_2^D(y_2, x_1^D(y_1, x_2^{D*})) - x_2^A(x_1^A(x_2^{D*})) \\ &< x_2^D(y_2, x_1^D(y_1, x_2^{D*})) - x_2^D(y_2, x_1^A(x_2^{D*})) < 0. \end{aligned}$$

The first inequality results from $x_2^A(x_1^A(x_2^{D*})) > x_2^D(y_2, x_1^A(x_2^{D*}))$ (which itself results from $\frac{\partial \Pi^A}{\partial x_2} > \frac{\partial \Pi^D}{\partial x_2}$ and the concavity of Π^A and Π^D). The second inequality results from $x_1^A(x_2^{D*}) > x_1^D(y_1, x_2^{D*})$ (which results itself from $\frac{\partial \Pi^A}{\partial x_1} > \frac{\partial \Pi^D}{\partial x_1}$ and the concavity of Π^A and Π^D) and the fact that $x_2^D(y_2, x_1)$ is increasing in x_1 (which results itself from the supermodularity of Π^D, as shown before). Therefore, we have $H_2^D(\mathbf{y}, x_2^{D*}) < H_2^A(x_2^{D*}) \leq 0$ which leads to a contradiction because $H_2^D(\mathbf{y}, x_2^{D*}) = 0$. This proves that $x_2^{A*} > x_2^{D*}$. We can show that $x_1^{A*} > x_1^{D*}$ in a similar way, which completes the proof for $N = 2$.

Let us now assume that the result stated in Step 2 holds for a given N and show that it holds for $N+1$ too. Denote

$$(\tilde{x_1}^A(x_{N+1}), \tilde{x_2}^A(x_{N+1}), ..., \tilde{x_N}^A(x_{N+1})) = \arg\max_{(x_1,...,x_N)} \Pi^A(x_1, ..., x_N, x_{N+1})$$

and

$$(\tilde{x_1}^D(y_1, x_{N+1}), \tilde{x_2}^D(y_2, x_{N+1}), ..., \tilde{x_N}^D(y_N, x_{N+1})) = \arg\max_{(x_1,...,x_N)} \Pi^D(\mathbf{y}, x_1, ..., x_N, x_{N+1}).$$

The fact that the result holds for N implies that $\tilde{x_i}^A(x_{N+1}) > \tilde{x_i}^D(y_i, x_{N+1})$ for any $i = 1, ..., N$ and any x_{N+1}.

Now define $H^A_{N+1}(\cdot)$ and $H^D_{N+1}(\cdot)$ as follows:

$$H^A_{N+1}(x_{N+1}) = x^A_{N+1}(\tilde{x_1}^A(x_{N+1}), \tilde{x_2}^A(x_{N+1}), ..., \tilde{x_N}^A(x_{N+1})) - x_{N+1}$$

and

$$H^D_{N+1}(\mathbf{y}, x_{N+1}) = x^D_{N+1}(y_{N+1}, \tilde{x_1}^D(y_1 x_{N+1}), \tilde{x_2}^D(y_2, x_{N+1}), ..., \tilde{x_N}^D(y_N, x_{N+1})) - x_{N+1}.$$

From

$$(x^{A*}_1, x^{A*}_2, ..., x^{A*}_N, x^{A*}_{N+1}) = \arg\max_{(x_1,...,x_N,x_{N+1})} \Pi^A(x_1, ..., x_N, x_{N+1})$$

it follows that

$$(x^{A*}_1, x^{A*}_2, ..., x^{A*}_N) = \arg\max_{(x_1,...,x_N)} \Pi^A(x_1, ..., x_N, x_{N+1}),$$

which implies (by uniqueness of the maximizer) that $x^{A*}_i = \tilde{x_i}^A(x^{A*}_{N+1})$ for all $i = 1, 2, ..., N$. Using this we obtain the following:

$$H^A_{N+1}(x^{A*}_{N+1}) = x^{A*}_{N+1}(x^{A*}_1, x^{A*}_2, ..., x^{A*}_N) - x^{A*}_{N+1} = 0.$$

Similarly, we get that $H^D_{N+1}(\mathbf{y}, x^{D*}_{N+1}) = 0$. As in the proof for $N = 2$, the uniqueness of the maximizer ensures that $H^A_{N+1}(x_{N+1}) > 0$ if $x_{N+1} < x^{A*}_{N+1}$ and $H^A_{N+1}(x_{N+1}) < 0$ if $x_{N+1} > x^{A*}_{N+1}$, and, similarly, $H^D_{N+1}(\mathbf{y}, x_{N+1}) > 0$ if $x_{N+1} < x^{D*}_{N+1}$ and $H^D_{N+1}(\mathbf{y}, x_{N+1}) > 0$ if $x_{N+1} > x^{D*}_{N+1}$.

Using a reasoning by contradiction, similar to the one used in the proof of the case $N = 2$, we can show that $x^{A*}_{N+1} > x^{D*}_{N+1}$. Moreover, for any $i = 1, 2, ..., N$,

$$x^{A*}_i = \tilde{x_i}^A(x^{A*}_{N+1}) > x^D_i(y_i, x^{A*}_{N+1}) > \tilde{x_i}^D(y_i, x^{D*}_{N+1}) = x^{D*}_i,$$

where the first inequality follows from the fact that the result we want to show for $N+1$ is assumed to hold for N, and the second inequality follows from $x^{A*}_{N+1} > x^{D*}_{N+1}$ and the fact that $\tilde{x_i}^D(y_i, x_{N+1})$ is increasing in x_{N+1} (due to supermodularity of Π^D in (x_i, x_j)).

Thus, we have shown that $x^{A*}_i > x^{D*}_i$ for all $i = 1, 2, ..., N, N+1$ which completes the proof of Step 2. *Q.E.D.*

B.3 Proof of Proposition 3

Recall that we have the following decomposition:

$$CS_i^{D^*} - CS_i^{A^*} = \left[CS_i(y_i, \mathbf{x_{-i}^{D^*}}) - CS_i(x_i^{A^*}, \mathbf{x_{-i}^{D^*}})\right] + \left[CS_i(x_i^{A^*}, \mathbf{x_{-i}^{D^*}}) - CS_i(x_i^{A^*}, \mathbf{x_{-i}^{A^*}})\right].$$

Cases (i) and (ii) are straightforward because both terms in the above decomposition have the same sign.

Consider now case (iii). In this case, $CS_i(y_i, \mathbf{x_{-i}^{D}})$ is increasing in γ^D because $CS_i(y_i, \mathbf{x_{-i}})$ is decreasing in x_j for all $j \neq i$ and x_j^D is decreasing in γ^D for all $j \neq i$. This implies that $CS_i^{D^*} - CS_i^{A^*} = CS_i(y_i, \mathbf{x_{-i}^{D^*}}) - CS_i(x_i^{A^*}, \mathbf{x_{-i}^{A^*}})$ is increasing in γ^D. Therefore, there exists $\hat{\gamma}^{CS}(y_i) \in [0, +\infty[\cup \{+\infty\}$ such that $CS_i^{D^*} \leq CS_i^{A^*}$ if and only if $\gamma \leq \hat{\gamma}^{CS}$. Moreover, $\hat{\gamma}^{CS}(y_i)$ is (weakly) decreasing in y_i because $CS_i^{D^*} - CS_i^{A^*}$ is (weakly) increasing in y_i.

Finally consider case (iv). In this case, $CS_i(y_i, \mathbf{x_{-i}^{D^*}})$ is decreasing in γ^D because $x_j^{D^*}$ is decreasing in γ^D for all $j \neq i$, and $CS_i(y_i, \mathbf{x_{-i}})$ is increasing in x_j for all $j \neq i$. This implies that $CS_i^{D^*} - CS_i^{A^*} = CS_i(y_i, \mathbf{x_{-i}^{D^*}}) - CS_i(x_i^{A^*}$ is decreasing in γ^D. Therefore, there exists $\tilde{\gamma}^{CS}(y_i) \in [0, +\infty[\cup \{+\infty\}$ such that $CS_i^{D^*} \leq CS_i^{A^*}$ if and only if $\gamma \geq \tilde{\gamma}^{CS}$. Moreover, $\tilde{\gamma}^{CS}(y_i)$ is (weakly) increasing in y_i because $CS_i^{D^*} - CS_i^{A^*}$ is (weakly) increasing in y_i.

Q.E.D.