Data sharing or analytics sharing?*

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

This version: December 2023
First version: September 2023

Abstract

Data combination and analytics can generate valuable insights for firms and society as a whole. Multiple firms can do so by means of new technologies that bring the analysis to the data ("analytics sharing") or, more conventionally, by sharing the data ("data sharing"). Analytics sharing technologies are gaining traction because of their advantages in terms of privacy, security, and environmental impact. We present a model that allows us to study the economic incentives generated by these technologies for both firms and a platform facilitating data combination. First, we find that the platform chooses data sharing unless the insights delivered by the analytics sharing technology are sufficiently superior to those associated to the data sharing technology for a given combination of datasets. Second, we show that analytics sharing results in a higher total data contribution than data sharing under general conditions. Third, we highlight scenarios in which, in presence of data externalities, there can be a misalignment between the choice of the platform and the preference of a social planner.

JEL Classification: D43; K21; L11; L13; L41; L86; M21; M31.
Keywords: data sharing, analytics sharing, data platforms, federated learning, data externalities.

*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 Vincenzo Denicolò and seminar participants at the 6th Unibg Industrial Organization Winter Symposium (Passo del Tonale) and the 12th EIEF-UNIBO-IGIER Bocconi (Milano) 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 and individuals in their every day activities have a great value, which goes beyond the economic impact. Combining and using data responsibly is essential for disparate matters such as predicting cancer or designing better products and services. To exploit data at their maximum potential, policy makers are experimenting the best ways to guarantee the necessary control without deterring their use for the benefit of society (AEPD, 2023).

Technological innovation can help achieve this objective. It is, in fact, increasingly common for multiple firms to combine their datasets in order to benefit from the economies of scale and scope in data aggregation that a common and richer dataset can bring about. One of the main drivers of inter-firm data combination is the training of artificial intelligence (AI) models, which can result in enhanced analytics. In this context, "federated learning" is a rising technology that allows to train algorithms by 'bringing the analysis to the data'. This innovative type of technological solution, which we label "analytics sharing", provides an alternative to more established practices such as "data sharing", which involves sharing the data with other parties.

Analytics sharing approaches have benefits that make them particularly appealing to policy makers. For example, there are well documented advantages in terms of privacy and security (Mothukuri et al., 2021; Blanco-Justicia et al., 2021; Ma et al., 2020), and these technological solutions can also reduce the environmental footprint of AI models training (Qiu et al., 2023; Yousefpour et al., 2023; Guler and Yener, 2021; Qiu et al., 2020).

While these novel data-combination technologies are being widely studied from a technical perspective, the economic drivers and implications of choosing a technology that brings the analysis to the data, like analytics sharing, over one that implies sharing the data remain understudied. What are the economic incentives driving firms to choose one technology over the other? Can one of these technologies prompt firms to combine more data than the other? In presence of externalities, positive or negative, that data may generate, under which conditions is the private choice of a data-combination technology socially optimal?

This article provides a first answer to these questions by analyzing how the characteristics of each type of data-combination technology act on firms' incentives to combine data, and the welfare implications of their choices. The analysis builds upon a model in which a platform can facilitate data combination by choosing between data sharing and analytics sharing. Firms, with possibly heterogeneous data endowments, decide whether or not to join the platform and, if so, how much data to contribute to it. The main benefit for firms is to access insights from enhanced analytics, whereas both firms and the platform face costs from combining data through the platform. Both benefits and costs can be technology-specific.

A fundamental difference between the two technologies is the following. Under data sharing, firms can exploit the joint dataset along with all the other data they have, regardless of how much data they decided to share. The dataset can be directly accessed so that analytics can be performed in-house. On the contrary, under analytics sharing, firms access analytics services from the platform, which extracts insights from the data made available by firms to train the platform's algorithm. In this case, there is no direct data transfer from firms to the platform; rather, it is the platform's centralized data analytics service that "travels" to the data firms have contributed. Given that a specialized platform's centralized analytics service are likely to be more efficient than firms' in-house analytics, we focus on the scenario in which analytics sharing provides better analytics than data sharing.

The analysis of the model delivers three major findings. First, a platform with complete information about firms' data endowments opts for analytics sharing only if it guarantees a sufficiently higher level of analytics than data sharing. This finding derives from the balance of two effects: the *endowment effect*, that favors data sharing, as all data in possession of the firm can be used to extract analytics, and the *data analytics effect* that favors analytics sharing. The baseline model considers a platform using personalized and public contracts. Similar findings hold if we consider alternative contractual agreements that preserve anonymity. Interestingly, we obtain 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, whereas analytics sharing requires that data be contributed to be combined with other firms' data. This results in a lower marginal benefit from contributing data under data sharing than under analytics sharing.

Third, we analyze the role of the externalities that combining data can generate, and show that the choice of a social planner may not be aligned with the platform's. For example, if there are positive data externalities not accounted for by the platform and the superiority of data analytics under analytics sharing is relatively weak, then data sharing may be privately chosen although the social planner has a preference for analytics sharing. The opposite may occur, *mutatis mutandis*, in case of negative externalities.

We also extend our analysis to account for the potential occurrence of data leakages. We consider two scenarios, depending on whether the leakage has a technology-specific impact on firms or on the externality. When data leakage negatively affects firms' profits, it is reasonable to assume that contributing data is more costly under data sharing, which requires data transfers from firms to the platform. As a consequence, analytics sharing

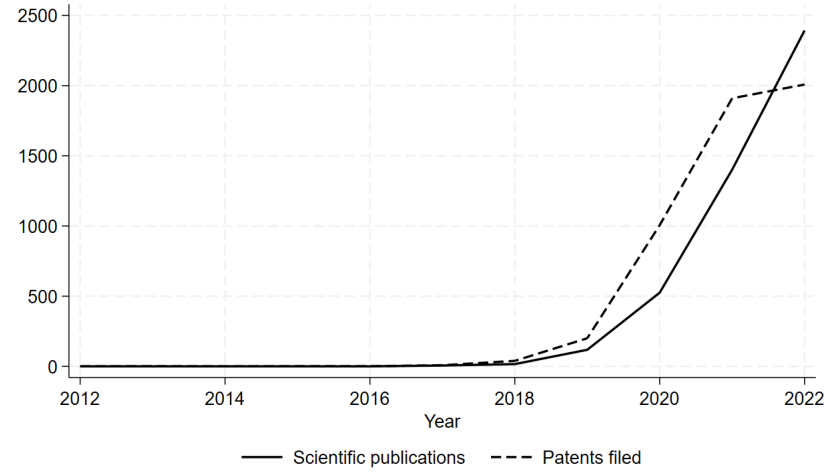
is more likely to be adopted and to generate more data contributions. Instead, when data leakage affects (negative) externalities, the private decision of the platform does not change, but analytics sharing is more likely to be favored by the social planner as it limits third-party access to data.

Our analysis has relevant policy implications. For given efficiency levels of data analytics, there is no one-size-fits-all optimal data-combination technology. Indeed, policymakers should encourage the adoption of analytics sharing in industries with high data endowments and strong positive data externalities. Conversely, they should incentivize data sharing technologies in industries with low data endowments and strong, negative data externalities. In the remaining cases, no intervention is needed, as the market will adopt the socially-optimal technology. Nevertheless, policymakers focused on strengthening privacy and security of data-combination technologies should design policies aimed at improving the efficiency of analytics sharing's generated insights with respect to data sharing. Doing so would increase the probability of analytics sharing being the socially- and privately-optimal technology whether data externalities are positive or negative.

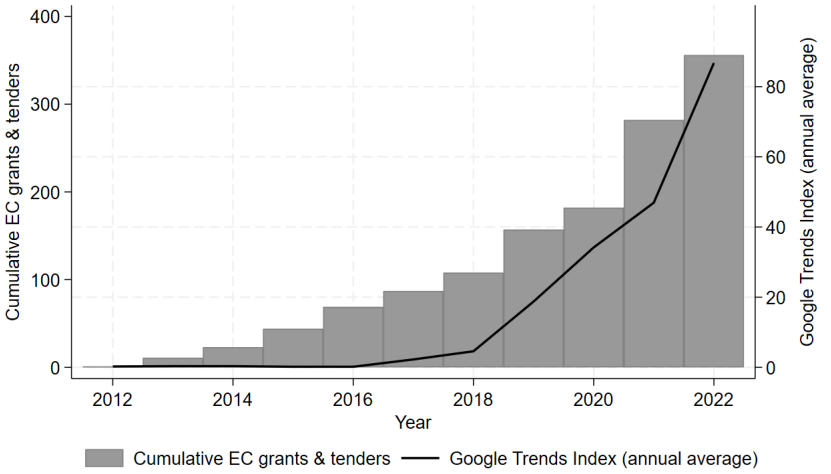
Analytics sharing and federated learning. Data combination solutions for data or analytics sharing have found promising applications in a number of sectors in recent years, ranging from mobility (e.g., to develop mobility-as-a-service solutions and to develop autonomous driving technologies), health (e.g., by advancing research on vaccines, treatments, and diagnosis) and to the aviation sector (e.g., to facilitate operational-, commercial-, and maintenance-related inter-firm coordination along the value chain).

The ascent of analytics sharing is tightly linked to technological advances such as Federated Learning, which is an innovative machine learning technique that trains an algorithm via multiple independent sessions. Unlike traditional machine learning, where all the data must be gathered together, Federated Learning makes it possible to extract knowledge from data distributed across different organizations. For this machine learning technique to be adopted, firms must make data available without sharing that data or moving it to a central location. This can be done by allowing data scientists and machine learning experts to run advanced analytics and train models in a federated (or distributed) manner.

Federated Learning can then be considered as the main driver of a potential shift in the data-combination paradigm from data sharing to analytics sharing. Figures 1a and 1b report the growth of Federated Learning in the last decade, in terms of scientific publications and patents, interest on Google search, and grants and tenders in the EU. The trends clearly illustrate how private firms and public actors are gaining interest in these technologies and are starting to implement it.



(a) Number of scientific publications and patents filed containing the term "federated learning" for years 2012-2022.
Source: own elaboration based on Web of Science and Google Patents.



(b) Cumulative European Commission grants and tenders containing the term "federated learning" (left axis) and annual average of the Google Trends Interest Index for the search term "federated learning" (right axis) for years 2012-2022.
Source: own elaboration based on European Commission and Google Trends. The Google Interest index equals 100 at the peak monthly popularity of the term.

For example, a consortium of firms including WeBank, Tencent, Huawei, and Intel collaborate through the Federated AI Technology Enabler (FATE) project to train AI models with their combined datasets while protecting data security and privacy. Another recent example is the ACCESS program funded by the European Medicines Agency. It uses federated learning on multiple countries' institutions to "monitor benefits, coverage and safety risks of new COVID-19 vaccines in the post-authorisation stage" at a continental scale. Still within the health sector, the MELLODDY consortium gathers ten pharmaceutical companies, universities start-ups and the tech company Nvidia to train a common drug-discovery model without sharing the confidential datasets of the individual partners.

One of the first experiments with Federated Learning may prove insightful to better understand the idea behind this new approach to data combination. The experiment involved sending algorithms to mobile phones in order to improve next-word prediction models of Google's Gboard keyboard (McMahan et al., 2017; Hard et al., 2018). The fundamental difference between this technique and more standard approaches used until then, is that messages no longer had to be sent to a central server. The experiment led to more accurate and faster predictions. This is a typical application of horizontal Federated Learning, where the algorithms improve when each device is used.

In other situations, as in health applications, different entities possess different records about the same individuals. Learning requires that these pieces of information are put together. Vertical Federated Learning can avoid that these data remain siloed. The institutions need to identify common patients without sharing names or other identifiers, using secure multi-part computing techniques (Bringer et al., 2013). Then, each entity has an algorithm that processes its data locally, extracting the necessary insights (Liu et al., 2022). Our approach is sufficiently general to encompass both scenarios.

Structure. The rest of the article is structured as follows. Section 2 discusses how the article relates to other streams of literature. Section 3 develops a baseline model of data combination with data sharing and analytics sharing technologies. Section 4 provides our main results. Section 5 extends the baseline model to analyze the impact of anonymity in contracting. Section 6 considers the possibility of data leakages. Section 7 discusses some policy implications of our findings. Section 8 concludes.

2 Related literature

Data gathering, use, sharing and re-use play a crucial part in today's highly digitized economy. For example, it is estimated that the contribution of the data market to the EU's economy in 2017 was 335.6 billion euros, corresponding to 2.4 per cent of total GDP (Frontier Technology Quarterly, 2019). As data are mostly non-rival, Jones and Tonetti (2020) study the positive externalities of sharing personal data, and how they can boost macroeconomic growth. Farboodi et al. (2019) and Farboodi and Veldkamp (2021) consider the implications of data generated by firms as information, and their implications for firms' growth and size distribution in both a partial and general equilibrium perspective.

Our article takes a different approach to focus on the microeconomic incentives that firms face when deciding whether to combine (some of) their data in order to exploit a joint dataset through a platform. In that respect, it mainly relates to the 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. Bergemann et al. (2022), Ichihashi (2021), Gu et al. (2022) and Abrardi et al. (2023), among others, study upstream

competition, or lack thereof, between data brokers. Data need not to be sold to one or to all firms: Abrardi et al. (2022) and Delbono et al. (2023) show that data may be sold only to a subset of competitors, whereas Bounie et al. (2021) identify the optimal partition of data to be sold to downstream competitors.

Horizontal sharing takes 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). Customer-specific data are gathered in a first market interaction and are used for price discrimination in future ones. Data sharing can be unilateral or bilateral. A related problem has been studied in credit markets (Padilla and Pagano, 1997; Pagano and Jappelli, 1993; Gehrig and Stenbacka, 2007, inter alia). When sharing data, lenders face a trade-off between reducing adverse selection and facing enhanced competition. Data are shared through a "credit bureau" that resembles our platform.

Platforms facilitate data sharing beyond the financial sector. Carballa Smichowski (2018) studies mobility-as-a-service (MaaS) platforms, that give consumers a one-stop shop to compare routes and purchase multimodal transportation services. Joining the platform, transportation service providers trade off a wider market reach with increased competition, particularly from providers with more overlapping routes. Martens et al. (2021) study data sharing between sellers and a hybrid platform, i.e., that acts both as an intermediary and as a seller. To tackle the platform's data advantage, they propose a new data sharing mode, "in-situ" data access, that involves the seller to use its own algorithms on its data generated on the platform, and compare it to data portability to a rival platform.

Despite its similarities with this body of work, our analysis differs in many respects. For example, our model is general enough to encompass the sharing of different types of data. On top of that, our focus is on the different properties of two data-combination technologies. Furthermore, we analyze multilateral data sharing through a platform rather than dyadic data sharing between firms. Finally, we abstract away from any form of downstream market competition between the firms contributing data to the platform.

There is a vast literature comparing alternative data-combination technologies in the computer science and engineering fields. Drainakis et al. (2023) and AbdulRahman et al. (2020), for example, provide reviews of the literature comparing federated learning and traditional centralized algorithm training techniques. 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, more limited is the number of studies on multi-firm data combination from an economics perspective. In this perspective, the closest work to ours is Calzolari et al. (2023). In both their model and ours, there is a monopolistic aggregator that has to create incentives for data-holding firms to decide whether to combine their data or not. If a firm decides to do so, it has to choose how much data to contribute, at a cost proportional

to this contribution. The aggregator, in turn, uses a per-unit-of-data transfer to incentivize contributions and a fixed fee to extract surplus. Notwithstanding these similarities, their article differs substantially from ours both in its focus and research questions. Indeed, they consider a situation in which producers own machine-generated non-personal data and there is value in combining datasets in order to obtain the same machine learning analytics. Our article, instead, focuses on the economic incentives resulting from different technological options for combining data and sheds light on the determinants of the choice between data sharing and analytics sharing and the potential inefficiencies associated to this choice.

Finally, our article assumes that firms' benefits from data analytics increase when they combine their datasets because doing so generates economies of scale and scope. In that respect, it relates to a recent strand of empirical literature that validates this assumption by studying the impact of the number of observations in a dataset on prediction accuracy in various setting such as search engines (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 (Claussen et al., 2023) and advertisement (Agrawal et al., 2018). Hocuk et al. (2022), in turn, provide an empirical measurement of "economies of scope in data aggregation". They study how the increase in the number of socio-economic variables in a dataset increases the accuracy of prediction of health-related outcomes while the number of observations (individuals) remains constant.

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

We consider an economy in which there are $N(\geq 2)$ firms and one data platform. The goal is to analyze the platform's and the social planner's choice between two alternative and technically feasible options for combining data and extracting valuable insights from it, and derive the conditions under which these choices may differ. 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 in detail in what follows.

Firms. The firms, $i = 1, \ldots, N$ are heterogeneous in their *data endowments*, y_i. The vector of data endowments is $\mathbf{y}$. Define as $\mathbf{x} = (x_1, \ldots, x_N)$ the vector of all data contributed by firms to the platform, and as $\mathbf{x_{-i}}$ the vector of all contributed data apart from the ones of 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 in detail below when introducing the two technology options for data combination.

All firms face a cost of contributing data to the platform that depends on their data contribution. 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. If a firm decides not to join the platform, its outside option is normalized to 0 and it is the same for all firms.

Platform. The platform aims to maximize its profits. To start with, we assume that the platform perfectly knows the firms' characteristics and can offer each firm a two-part-tariff contract. Under the chosen data-combination technology t, this amounts to offering to each firm: (i) f_i^t, a fixed fee to join the platform, and (ii) m_i^t, a transfer (that, a priori, can be positive or negative) per unit of data contributed to the platform. Denote as $\mathbf{f}$ and $\mathbf{m}$ the vectors of the fixed fees and per-unit transfers, respectively.

The platform faces an operational cost to manage all the data that has been contributed by the firms, which is captured by the function $G^t(\mathbf{x})$ increasing in all its arguments. The platform's profit function may be written as:

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

Technologies. We consider the two technologies of data combination introduced above. Both technologies allow to combine data and obtain insights from it. However, each generates different incentives to contribute data.

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, a firm can also use its *full data endowment* to extract value, even from the share that it has not contributed to the platform.

Therefore, the firms that join the platform obtain insights from the combination of two types of data. On the one hand, regardless of how much data they have contributed to the platform, all firms can exploit the data contributed by other firms to the platform, $\mathbf{x_{-i}}$, to get insights from it. On the other hand, the data contributed by other firms are combined with a firm's own data endowments, y_i. The combination of these two data components generates the benefits that can be harvested from data analytics. Bringing it all together, in presence of the data sharing technology, the firms' profit functions are:

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

Second, let us consider *analytics sharing*. In presence of this technology, the platform provides a *centralized data analytics* that is performed on the data contributed by the firms. Importantly, under analytics sharing, firms cannot access the data contributed by other firms. A joining firm can only use the platform's analytics, which is trained on the overall

data available to the algorithm, including the data endowment the firm has made available to the platform.

Therefore, on the one hand, as under the data sharing technology, all firms gain insights through the availability of the platform's data contributed by other firms, $\mathbf{x_{-i}}$. On the other hand, and contrary to the previously discussed technology, these data have to be combined with the share of its data endowment that it has contributed to the platform (x_i) in order to obtain data analytics' insights. Hence, in presence of an analytics-sharing technology, the firms' profit functions are:

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

Timing. The game unfolds as follows.

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

Welfare. Data can create (positive or negative) externalities that firms and the platform may fail to take into account. For example, shared health data that help developing a vaccine imply societal benefits that go beyond the ones accruing to the developing consortium.[1] Conversely, leakages of personal data due for example to cybersecurity attacks may lead to serious privacy damages. Sharing data and analytics has also energy intensive data storing and computational requirements, which can create severe environmental footprints.

In presence of externalities, the total surplus generated by data sharing and analytics sharing are respectively given by:

$$W^D(\mathbf{y}, \mathbf{x}) \equiv \Pi^D(\mathbf{y}, \mathbf{x}) + E^D(\mathbf{y}, \mathbf{x}),$$

and:

$$W^A(\mathbf{x}) \equiv \Pi^A(\mathbf{x}) + E^A(\mathbf{x}),$$

where $\Pi^t(\cdot)$ is the industry profit (i.e., the sum of the profit of the platform and of the firms) and $E^t(\cdot)$ is a function capturing the data externalities generated by data. Further note that:

[1] Survey evidence seems suggestive of a lack of internalization: in the case of targeted cancer treatment, Mattioli (2017) reports that between the most typical questions, from healthcare firms and professionals considering to contribute to data-combination projects, feature the following: "What is the value of sharing this data?" and "What is the value for me to share my data?". One interviewee, a prominent academic researcher, stated that the message "it's good for the World" is often not sufficient to convince potential contributors.

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

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

4 Analysis

Let us start by considering 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)

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

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:

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

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

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. 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 $t = D, A$. 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 maximizes 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

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

To focus on the role of asymmetric benefits, we assume that (data-related) costs endured by both the firms and the platform are the same under both technologies:

Assumption 1 (Cost Symmetry)**.** $g_i^D(\cdot) = g_i^A(\cdot) = g_i(\cdot)$ *and* $G^D(\cdot) = G^A(\cdot) = G(\cdot)$.

As per the asymmetric benefits related to the two different technologies, we assume that the performance of data analytics under analytics sharing is (weakly) superior than under data sharing. Specifically, we assume that:

Assumption 2 (Superiority of data analytics under analytics sharing)**.** $B_i^D(y_i, \mathbf{x_{-i}}) = B_i(y_i, \mathbf{x_{-i}})$ *and* $B_i^A(x_i, \mathbf{x_{-i}}) = (1 + \alpha^A) B_i(x_i, \mathbf{x_{-i}})$ *where* $\alpha^A \geq 0$.

Assumption 1 improves the tractability of our analysis by allowing us to focus on the role of the benefits of different technologies. In Assumption 2, instead, the parameter α^A can be interpreted as a measure of the performance of data analytics under analytics sharing relative to data sharing.

The assumption builds upon the fact that, under analytics sharing, the platform provides data combination and centralized data analytics services. Under data sharing, on the other hand, analytics is performed directly by the firms. Hence, under analytics sharing the platform gains more specialized knowledge regarding data analytics, that can grow with the number of data-combination consortia it manages.[2] Therefore, one can expect the platform's centralized analytics service to be more efficient than the analytics firms do in-house.

[2]For example, NVIDIA's Clara AI toolkit pilot program allows more than 38,000 doctors to use AI for diagnostic radiology building on federated learning. This AI-driven service's quality improves with the total data medical professionals contribute.

In some applications, Assumption 2 can also relate to the higher accuracy of data analytics that travel to the data vis-à-vis data analytics relied upon in conventional data sharing. The superior performance of analytics sharing has been documented, for example, in keyboard word prediction(Hard et al., 2018), image recognition (Chou et al., 2021), and various health applications (Lee and Shin, 2020; Xiong et al., 2021; Li et al., 2019). Finally, given the great policy and commercial interest in analytics sharing solutions, it is likely that their technical performance will improve further in the years to come.

We suppose hereafter that Assumptions 1 and 2 hold unless stated otherwise.

We can express (5) as:

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

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 (weakly) positive (i.e., it favors data sharing) and increasing in the data endowments y_i. The *data analytics effects* captures the superiority of data analytics under analytics sharing. This effect is (weakly) negative (i.e., it favors analytics sharing) and is increasing in α^A. To see why it is negative, note that:

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

The next proposition shows that analytics sharing is chosen if and only if data analytics under this technology is sufficiently more performant than under data sharing, and that this condition is more stringent the greater data endowments are.

Proposition 1 (**Technological Choice**). *There exists a threshold $\tilde{\alpha}^A(\mathbf{y}) \geq 0$ such that the platform chooses the analytics sharing technology if and only if $\alpha^A \geq \tilde{\alpha}^A(\mathbf{y})$. Moreover, $\tilde{\alpha}^A(\mathbf{y})$ is increasing in firms' data endowments y_i.*

Proof: See Appendix A.1

Data contributions. The marginal net benefits of increasing data contribution respectively under data sharing and under analytics sharing are given by:

$$\frac{\partial \Pi^D(\mathbf{y}, \mathbf{x})}{\partial x_i} = \sum_{j \neq i} \underbrace{\frac{\partial B_j^D}{\partial x_i}(y_j, \mathbf{x}_{-\mathbf{j}})}_{>0} - g_i'(x_i) - \frac{\partial G}{\partial x_i}(\mathbf{x});$$

$$\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} = \underbrace{\frac{\partial B_i^A}{\partial x_i}(x_i, \mathbf{x}_{-\mathbf{i}})}_{>0} + \sum_{j \neq i} \underbrace{\frac{\partial B_j^A}{\partial x_i}(x_j, \mathbf{x}_{-\mathbf{j}})}_{>0} - g_i'(x_i) - \frac{\partial G}{\partial x_i}(\mathbf{x}).$$

By substracting these two marginal benefits we obtain:

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

Under Assumption 2, it follows that:

$$\frac{\partial \Pi^A(\mathbf{x})}{\partial x_i} - \frac{\partial \Pi^D(\mathbf{y},\mathbf{x})}{\partial x_i} = (1+\alpha^A) \underbrace{\frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x_{-i}})}_{>0} + \sum_{j\neq i} \left[\underbrace{\frac{\partial B_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) - \frac{\partial B_j}{\partial x_i}(y_j, \mathbf{x_{-j}})}_{\gtrless 0} \right] + \alpha^A \sum_{j\neq i} \underbrace{\frac{\partial B_j}{\partial x_i}(x_j, \mathbf{x_{-j}})}_{>0}.$$

Notice that:

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

Moreover, $\frac{\partial^2 B_j}{\partial x_j \partial x_i}$ is negative when B_j is submodular in (x_i, x_j), and positive when it is supermodular in (x_i, x_j). As a consequence, if $\alpha^A > 0$ and B_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 be easily shown that this property also holds if the 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 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 assumptions: (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) B_j is not "too supermodular" in (x_i, x_j), i.e., $\frac{\partial^2 B_j}{\partial x_j \partial x_i} < \tilde{k} = \frac{\min_i \min_{\mathbf{x}} \frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x_{-i}})}{\sum_j y_j}$ for any x and any i and j such that $i \neq j$.*

Proof: See Appendix A.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 of transfer) remains true in equilibrium when the per-unit of 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[3] (i.e. there are no strong economies of scope in 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. Indeed, several studies have identified increasing returns to scale and scope in data combination only up to a certain amount of data contributed (Hocuk et al., 2022; Schäfer and Sapi, 2023; Lee and Wright, 2023).

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 this section that conditions (i) and (ii) of Proposition 2 hold, meaning that analytics sharing generates more data contribution than data sharing.

Welfare implications. As we have shown above, the privately optimal regime is determined by comparing $\Pi^D(\mathbf{y}, \mathbf{x}^{\mathbf{D}*})$ and $\Pi^A(\mathbf{x}^{\mathbf{A}*})$. The socially optimal (second-best) regime is instead determined by comparing $\Pi^D(\mathbf{y}, \mathbf{x}^{\mathbf{D}*}) + E^D(\mathbf{y}, \mathbf{x}^{\mathbf{D}*})$ and $\Pi^A(\mathbf{x}^{\mathbf{A}*}) + E^A(\mathbf{x}^{\mathbf{A}*})$.

We focus here on the special case in which $E^D(\mathbf{y}, \mathbf{x}) = E^A(\mathbf{x}) = E(\mathbf{x}) = e \cdot h(\mathbf{x})$ for any $\mathbf{x}$ where $e > 0$ is a scale parameter that captures the strength of data externalities and $h(\cdot)$ is monotonic in all x_i, with $h(0) = 0$.

Consider first the case of positive data externalities (i.e., $\frac{\partial h}{\partial x_i} > 0$ for all i). In this case, if the privately optimal technology is the one that generates the largest data contributions (that is, under our maintained assumption that conditions (i) and (ii) of Proposition 2 hold, analytics sharing), then it is also socially optimal. However, if the privately optimal technology is the one that generates the smallest data contributions (that is, under the same conditions, data sharing), then there is a potential for divergence between the privately optimal and socially optimal choice of technology. This is because the privately optimal technology generates less positive externalities than the other.

Consider now the case of negative data externalities (i.e., $\frac{\partial h}{\partial x_i} < 0$ for all i). In this case, if the privately optimal technology is the one that generates the smallest data contributions (that is, data sharing), then it is also socially optimal. However, if the privately optimal

[3] This encompasses both the case in which the functions are submodular and the one in which they are "moderately" supermodular.

technology is the one that generates the largest data contributions (that is, analytics sharing), then then there is again a potential for divergence between the privately optimal and socially optimal choices, this time because the privately optimal technology generates more negative externalities than the other one.

The following proposition provides necessary and sufficient conditions for the privately optimal technology to coincide with the socially optimal one.

Proposition 3. *If data externalities are positive, then there exists a threshold $e^+ > 0$ such that the privately and socially optimal technologies coincide if and only if $\alpha \geq \tilde{\alpha}^A$ or $e \leq e^+$. If data externalities are negative, then there exists a threshold $e^- > 0$ such that the privately and socially optimal technologies coincide if and only if $\alpha \leq \tilde{\alpha}^A$ or $e \leq e^-$.*

Proof: See Appendix A.3

Proposition 3 shows that there are two cases in which the socially optimal technology differs from the privately optimal one. First, if the superiority of data analytics under analytics sharing is relatively limited and data externalities are positive and strong enough, the platform chooses data sharing whereas the social planner prefers analytics sharing. Second, if data analytics are substantially superior under analytics sharing than under data sharing and data externalities are negative and strong enough, then the platform chooses analytics sharing whereas the social planner prefers data sharing.

5 Data sharing and analytics sharing under anonymity

In the baseline model, the platform uses personalized and public contracts. However, as discussed by Calzolari et al. (2023), firms producing data may value *anonymity* over whether they have joined the platform and the contractual details. If this information is public, data analytics might reveal information about firms' production strategies. Anonymity can be preserved either through secret personalized contracts or through a uniform public contract offered to all firms.

5.1 Secret personalized contracts

Assume that the platform offers secret personalized contracts and that 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 profit 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 by Calzolari et al. (2023). In our setting, if we denote $\mathbf{x}_{-i}^e$ 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_i^A is given by

$$(1+\alpha^A)\frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x}_{-i}^e) + m_i^A = g_i'(x_i).$$

Denoting $\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A)$ the solution to the equation above, the platform's maximization program can be written as

$$\max_{(\mathbf{f}^A, \mathbf{m}^A)} \sum_i [f_i^A - m_i^A \hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A)] - G((\hat{x}_i^A(\hat{\mathbf{x}}_{-i}, m_i^A))_{1\leq i\leq n})$$

subject to the participation constraints

$$(1+\alpha^A)B_i(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A), \mathbf{x}_{-i}^e) - g_i(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A)) + m_i^A \hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A) - f_i^A \geq 0,$$

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

$$\max_{\mathbf{m}^A} \sum_i [(1+\alpha^A)B_i(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A), \mathbf{x}_{-i}^e) - g_i(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A))] - G((\hat{x}_i^A(\hat{\mathbf{x}}_{-i}, m_i^A))_{1\leq i\leq n}).$$

The first-order condition with respect to m_i^A yields

$$\frac{\partial \hat{x}_i^A}{\partial m_i^A}\left\{(1+\alpha^A)\frac{\partial B_i}{\partial x_i}(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A), \mathbf{x}_{-i}^e) - \frac{\partial g_i}{\partial x_i}(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A)) - \frac{\partial G}{\partial x_i}((\hat{x}_i^A(\hat{\mathbf{x}}_{-i}, m_i^A))_{1\leq i\leq n})\right\} = 0$$

or, equivalently,

$$(1+\alpha^A)\frac{\partial B_i}{\partial x_i}(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A), \mathbf{x}_{-i}^e) - \frac{\partial g_i}{\partial x_i}(\hat{x}_i^A(\mathbf{x}_{-i}^e, m_i^A)) - \frac{\partial G}{\partial x_i}((\hat{x}_i^A(\hat{\mathbf{x}}_{-i}, m_i^A))_{1\leq i\leq n}) = 0.$$

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 profit under analytics sharing are (strictly) increasing in α^A and goes to infinity as α^A goes to infinity. 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 $\hat{\alpha}^A(\mathbf{y})$ increasing in data endowments y_i such that the platform chooses analytics sharing if and only if $\alpha \geq \hat{\alpha}^A(\mathbf{y})$. However, analytics sharing is less likely to be chosen than in the baseline model with public personalized contracts, i.e., $\hat{\alpha}^A(\mathbf{y}) > \tilde{\alpha}^A(\mathbf{y})$.*

5.2 Uniform public contract

Let us now assume that the platform preserves anonymity through a uniform public contract offered to all firms. 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., $B_i(\cdot) = B_j(\cdot)$ and $g_i(\cdot) = g_j(\cdot)$ for all $i \neq j$. This implies that the 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. This, combined with the fact that the platform's profit under data sharing increases with data endowments y_i and its profit under analytics sharing increases with α^A and goes to infinity as α^A goes to infinity, leads to the following result.

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

6 Data leakages

In this section, we consider how our results are affected by the potential occurrence of data leakages. A data leakage is the unauthorized transmission of data from within

a firm or organization to an external destination or recipient. Data can be transferred both electronically or physically, and such leakages constitute one of the most important cybersecurity threats for organizations(European Parliament, 2023). The damages caused, regardless of size of the firm or the industry of operation, can be extremely serious. They range from revenue loss to tarnished reputation, financial penalties and lawsuits. The damage can extend to external stakeholders, whose private and sensitive data can be exposed.

We distinguish between two scenarios. In the first one, we assume that the data leakage has a negative impact on firms' profits that is technology-specific, while its impact on data externalities is the same for both technologies. Conversely, in the second scenario, we assume that the data leakeage generates a negative externality that is technology-specific, while its impact on firms' profits is identical between the two technologies.

6.1 Technology-specific effects of data leakages on firms

A simple way of capturing this scenario in which, from the viewpoint of the firm, the consequences of data leakages are technology specific, is to drop Assumption 1 (i.e., cost symmetry). This amounts to assume that the expected cost of data leakages for firm i is part of the cost $g_i^t(x_i)$. Given that data sharing leads to giving the platform and other firms direct access to data while analytics sharing does not, it is reasonable to assume that $g_i^A(x_i) \leq g_i^D(x_i)$ for all firms i and all levels of data, x_i.

It is straightforward to show that Proposition 1 still holds in this case. However, the corresponding threshold is (weakly) lower than in the baseline model. This implies that it is more likely that analytics sharing is chosen over data sharing, relative to the baseline model, which is very intuitive given the lower overall costs of data contributions. Indeed, the lower expected risk of a data leakage, and its associated costs, imply that even a more moderate analytics advantage is sufficient to lead a platform towards the analytics sharing technology.

It is also clear that Proposition 2 still holds. The reason is that, under conditions (i) and (ii), analytics sharing generates more data than data sharing if $g_i^A(x_i) = g_i^D(x_i)$. Then, this is *a fortiori* true if $g_i^A(x_i) \leq g_i^D(x_i)$ because a decrease in data contribution costs, due to the lower risks of analytics sharing, leads to more data contributions under this technology. Finally, Proposition 3 is unaffected because we assume in this subsection that the effect of data leakages on externalities is not technology-specific.

6.2 Technology-specific effects of data leakages on externalities

Data leakages have effects that go beyond the firms and the platform that combine data. Private and sensitive information of partner firms, organizations and individuals may

be compromised, made public or used unlawfully. Breaches may even have negative market-wide effects on firms unrelated to the victim (Ashraf, 2021). As a result, we assume in this subsection that externalities are negative and that there exist e^D and e^A such that $E^D(\mathbf{y}, \mathbf{x}) = e^D h(\mathbf{x})$ and $E^A(\mathbf{x}) = e^A h(\mathbf{x})$. Given that analytics sharing limits third-party access to data relative to data sharing (Blanco-Justicia et al., 2021; Mothukuri et al., 2021), it is reasonable to assume that $0 < e^A \leq e^D$.

Note first that Propositions 1 and 2 remain unchanged because they do not depend on data externalities.

Let us now turn to the comparison between the privately optimal and socially optimal technologies. We need again to distinguish between the case in which the analytics under analytics sharing are sizably more performant than under data sharing, $\alpha^A \geq \tilde{\alpha}^A$, and the one in which they are not, $\alpha^A < \tilde{\alpha}^A$.

Consider the former case to begin with. If $\alpha^A \geq \tilde{\alpha}^A$, then analytics sharing is privately optimal. However, it is socially optimal if and only if

$$\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*}) + e^D h(\mathbf{x}^{D*}) - e^A h(\mathbf{x}^{A*}) \leq 0,$$

which can be rewritten as

$$e^A \leq \underbrace{\frac{\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*})}{h(\mathbf{x}^{A*})}}_{>0} + e^D \underbrace{\frac{h(\mathbf{x}^{D*})}{h(\mathbf{x}^{A*})}}_{>0} \equiv e^-(e^D).$$

In the first inequality above, the first term on the right hand side is positive, as both the numerator and the denominator are negative, and so is the second one, which is also increasing in e^D.

Turn then to the second case. If $\alpha^A < \tilde{\alpha}^A$ then data sharing is privately optimal. However, it is socially optimal if and only if

$$\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*}) + e^D h(\mathbf{x}^{D*}) - e^A h(\mathbf{x}^{A*}) \geq 0,$$

which can be rewritten as

$$e^A \geq \max\left\{0, \underbrace{\frac{\Pi^D(\mathbf{y}, \mathbf{x}^{D*}) - \Pi^A(\mathbf{x}^{A*})}{h(\mathbf{x}^{A*})}}_{<0} + e^D \underbrace{\frac{h(\mathbf{x}^{D*})}{h(\mathbf{x}^{A*})}}_{>0}\right\} \equiv e^-(e^D).$$

Consider the second expression in curly brackets on the right hand side of the previous inequality. The first term is now negative, as the numerator is positive and the denominator negative. The second term is instead positive, and increasing in e^D. If the first term dominates, then the whole expression is negative, and data sharing is also socially optimal. If, instead, the second term is relatively large, there is a positive threshold value of e^A

above which data sharing is both privately and socially preferred.

Figure 2: Analytics sharing or data sharing in presence of negative externalities. The effect of technology-specific data leakages.

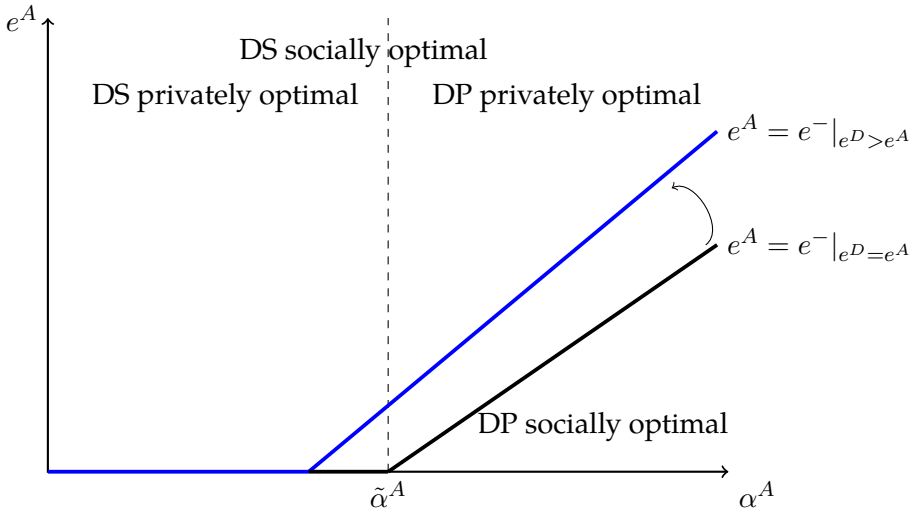


Figure 2 provides a visualization of the results in the previous discussion. In intuitive terms, if data sharing is sufficiently profitable compared to analytics sharing (i.e., α^A is low), then the profit effect may more than compensate for the stronger negative externalities compared to analytics sharing. This can be seen on the left hand side of the figure. As the profit advantage of data sharing becomes thinner (i.e., α^A increases towards $\tilde{\alpha}^A$), then this technology can still be socially optimal, but that is the case only if it does not generate too much (negative) externalities compared to analytics sharing. This happens if the externality rate under analytics sharing, e^A, is relatively large with respect to the respective rate under data sharing, e^D. In our illustration, this takes place above the increasing line $e^A = e^-(e^D)$.

A similar intuition explains a social preference for analytics sharing when this technology is also privately optimal (i.e., α^A is high). Indeed, the planner would prefer this technology in the lower-right part of the figure, below the line $e^A = e^-(e^D)$. In other words, analytics sharing is both privately and socially desirable when the externality rate , e^A, is not too high compared to data sharing, and the negative data externality does not offset the relative profit advantage of the technology.

The figure also shows that in our baseline model data sharing is necessarily socially optimal when leading to higher profits ($\alpha^A < \tilde{\alpha}^A$). Through the movement of the threshold line $e^A = e^-(e^D)$, it can be seen that this is not the case in this extension, which captures the reduction in the risks of externalities connected to data breaches under analytics sharing.

7 Policy implications

The analysis in sections 4 to 6 sheds some light on the economic drivers and implications of choosing a data-combination technology that brings the analysis to the data, analytics sharing, over conventional data sharing. In this section we discuss two policy implications of these findings.

Analytics sharing and the privacy-efficiency trade-off

The vast literature on the economics of privacy shows that, in many different contexts, there is a trade-off between privacy-reducing data combination and the efficiency gains that combining (consumer) data can bring about to firms, consumers and society. As several authors have pointed out, the advent of privacy-enhancing technologies such as analytics sharing have the potential of relaxing and even eliminating this trade-off (Acquisti et al., 2016; Johnson, 2022).

We have shown that data sharing has an intrinsic efficiency premium with respect to analytics sharing, as the former allows firms to interact the platform's data with all their data endowments. Therefore, in order for analytics sharing to be adopted, the performance of its data analytics has to be sufficiently superior. This result of our model, which puts forward the platform's choice of a data-combination technology, inverts the causality of the privacy-efficiency trade-offs discussed in the literature. In the literature, increased data combination reduces privacy and increases efficiency. In our model, by increasing analytics sharing's efficiency with respect to data sharing's, the likelihood of adopting a privacy-enhancing technology (analytics sharing) raises, and so does the amount of data being combined, as analytics sharing leads to higher contributions than data sharing.

This has a direct policy implication. Policymakers concerned about strengthening privacy protection should design technological policies aiming at encouraging the improvement of the efficiency of analytics sharing's data analytics capacity with respect to data sharing's. As shown in Section 4, doing so would increase the probability of analytics sharing being the socially- and privately-optimal technology whether data externalities are positive or negative. In this case, policies aiming at enhancing the capacity of analytics sharing technologies to perform analytics without actual data sharing are to be favored. As the literature in computer science and related disciplines shows, this is one of the major areas of interest in the current research on federated learning (Drainakis et al., 2023; AbdulRahman et al., 2020; Nilsson et al., 2018).

Industry-specific social desirability of a data-combination technology

We have shown that the threshold of the performance of data analytics under analytics sharing relative to data sharing above which analytics sharing is chosen by the platform is

increasing in firms' data endowments. Moreover, we have identified conditions suggesting that is likely that more data is combined under analytics sharing than under data sharing. Hence, *ceteris paribus*, under positive (negative) data externalities, analytics sharing is preferred (disfavored) by the social planner over data sharing. By combining these two findings, we can provide some insights on how, in the short run (i.e., when policymakers cannot influence the efficiency of data analytics), the social desirability of the private choice of a data-combination technology depends on sector characteristics.

In industries with low data endowments and strong positive data externalities, it is more likely that the analytics sharing technology will be chosen by both the platform and the social planner. Conversely, in sectors with high data endowments and strong negative data externalities both the platform and the social planner would more likely opt for data sharing. A possible example of the latter is the banking sector, in which the increasing digitization of payments and the advent of fintech have substantially increased the amount and the granularity of the data collected. At the same time, the nature of finance poses important cybersecurity and privacy risks. The implication is that in both type of industries policymakers should be a priori technology-agnostic, as it is more likely that firms will choose the socially-optimal technology.

On the contrary, if either (i) data endowments are low and data externalities strongly negative, or, (ii) data endowments are high and data externalities strongly positive, it is more likely that the platform's and the social planner's choice of technology differ. A possible example of the latter case is the mobility sector, for which vasts amounts of data (e.g., multimodal ticketing, routing queries, GPS tracking, etc.) exist and, as shown by many open data initiatives, the availability of large, comprehensive datasets has positive spillovers in terms of competition and innovation in the mobility sector. In such cases, policymakers should be more prone to do industry-specific interventions that create incentives for firms to adopt analytics sharing technologies (if data endowments are high and positive externalities strong and positive) or, on the contrary, data sharing technologies (if data endowments are low and data externalities strong and negative), depending on the industry's characteristics.

8 Conclusion

In this article we have studied how the choice of a data-combination technology affects firms' incentives to combine data, and the welfare consequences of such a choice. We have shown that even if firms internalize the security and privacy advantages of technologies that bring the analysis to the data (i.e., analytics sharing) compared to standard data sharing, the former is chosen only if its data analytics performs sufficiently better than the latter's. However, because analytics sharing leads to higher data contributions than data sharing, and given the existence of positive and negative externalities generated

by data, the privately-optimal technology can differ from the socially-optimal one. This is the case when either (i) the superiority of data analytics under analytics sharing is relatively limited and data externalities are positive and strong enough, or, (ii) if data analytics are substantially superior under analytics sharing than under data sharing and data externalities are negative and strong enough. Finally, we have shown that our findings hold qualitatively in presence of secret contracting or if the platform cannot condition contracts to the firms' data endowments, as well as in presence of technology-specific effects of data leakages that can be internalized by firms or not.

Our setting can be extended to investigate other interesting policy issues related to multi-firm data-combination technologies. First, by considering that data-holding firms might be competitors, we can study the competition-relaxing and competition-enhancing effects of data combination. This would allow us to extend our analysis to understand the antitrust implications of the choice of a data-combination technology, and the tensions that might arise between privacy and security protection, on the one hand, and competition, on the other hand. Second, we can consider 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 in the connected car industry, the NINDS Parkinson's Disease Biomarkers Program in the health sector), will likely face the threat of entry using analytics sharing technologies.

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. (2022). User data and endogenous entry in online markets. *SSRN Working Paper 4256544*.

Abrardi, L., Cambini, C., and Pino, F. (2023). Data brokers' competition and downstream market entry. *Mimeo*.

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

AEPD (2023). Federated learning: Artificial intelligence without compromising privacy. *Agencia Española de Protección de Datos*, `https://www.aepd.es/en/prensa-y-comunicacion/blog/federated-learning-artificial-intelligence-without-compromising-privacy`. [Last accessed December 27, 2023].

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

Ashraf, M. (2021). The market-wide externalities of cyber risk: Evidence from customers and data breaches. *SSRN Working Paper 3802846*.

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. *The RAND Journal of Economics*, 53(2):263–296.

Blanco-Justicia, A., Domingo-Ferrer, J., Martínez, S., Sánchez, D., Flanagan, A., and Tan, K. E. (2021). Achieving security and privacy in federated learning systems: Survey, research challenges and future directions. *Engineering Applications of Artificial Intelligence*, 106:104468.

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

Bringer, J., Chabanne, H., and Patey, A. (2013). Privacy-preserving biometric identification using secure multiparty computation: An overview and recent trends. *IEEE Signal Processing Magazine*, 30(2):42–52.

Calzolari, G., Cheysson, A., and Rovatti, R. (2023). Cooperative data-analytics: a market for machine-data. *CEPR Discussion Paper DP17842*.

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*.

Chou, L., Liu, Z., Wang, Z., and Shrivastava, A. (2021). Efficient and less centralized federated learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pages 772–787. Springer.

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

Delbono, F., Reggiani, C., and Sandrini, L. (2023). Strategic data sales with partial segment profiling. *SSRN Working Paper 4046961*.

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 Parliament (2023). Cybersecurity: main and emerging threats. `https://www.europarl.europa.eu/news/en/headlines/society/20220120STO21428/cybersecurity-main-and-emerging-threats`. [Last accessed December 27, 2023].

Farboodi, M., Mihet, R., Philippon, T., and Veldkamp, L. (2019). Big data and firm dynamics. *American Economic Association: Papers and Proceedings*, 109:38–42.

Farboodi, M. and Veldkamp, L. (2021). A model of the data economy. *NBER Working Paper 28427*.

Frontier Technology Quarterly (2019). Data economy: Radical transformation or dystopia? `https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/FTQ_1_Jan_2019.pdf`. [Last accessed December 27, 2023].

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.

Guler, B. and Yener, A. (2021). Sustainable federated learning. *arXiv preprint arXiv:2102.11274*.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

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*.

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*.

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

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.

Lee, G. H. and Shin, S.-Y. (2020). Federated learning on clinical benchmark data: performance assessment. *Journal of Medical Internet Research*, 22(10):e20891.

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., and Feng, A. (2019). Privacy-preserving federated brain tumour segmentation. In *Machine Learning in Medical Imaging*, volume 10, pages 133–141. Springer.

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

Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q., and Yang, Q. (2022). Vertical federated learning. *arXiv preprint arXiv:2211.12814*.

Ma, C., Li, J., Ding, M., Yang, H. H., Shu, F., Quek, T. Q., and Poor, H. V. (2020). On safeguarding privacy and security in the framework of federated learning. *IEEE Network*, 34(4):242–248.

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.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640.

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*.

Nilsson, A., Smith, S., Ulm, G., Gustavsson, E., and Jirstrand, M. (2018). A performance evaluation of federated learning algorithms. In *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning*, pages 1–8.

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.

Qiu, X., Parcollet, T., Beutel, D. J., Topal, T., Mathur, A., and Lane, N. D. (2020). Can federated learning save the planet? *arXiv preprint arXiv:2010.06537*.

Qiu, X., Parcollet, T., Fernandez-Marques, J., de Gusmao, P. P., Gao, Y., Beutel, D. J., Topal, T., Mathur, A., and Lane, N. D. (2023). A first look into the carbon footprint of federated learning. *Journal of Machine Learning Research*, 24(129):1–23.

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

Xiong, Z., Cheng, Z., Lin, X., Xu, C., Liu, X., Wang, D., Luo, X., Zhang, Y., Jiang, H., Qiao, N., and Zheng, M. (2021). Facing small and biased data dilemma in drug discovery with enhanced federated learning approaches. *Science China Life Sciences*, pages 1–11.

Yousefpour, A., Guo, S., Shenoy, A., Ghosh, S., Stock, P., Maeng, K., Krüger, S.-W., Rabbat, M., Wu, C.-J., and Mironov, I. (2023). Green federated learning. *arXiv preprint arXiv:2303.14604*.

A Proofs

A.1 Proof of Proposition 1

Given that: (i) $\Pi^A(\mathbf{x}^{A*})$ is continuous and increasing in α^A; (ii) $\Pi^A(\mathbf{x}^{A*}) \leq \Pi^D(\mathbf{y}, \mathbf{x}^{D*})$ for $\alpha^A = 0$ because in this case $\Pi^A(\mathbf{x}) \leq \Pi^D(\mathbf{y}, \mathbf{x})$ for any $\mathbf{x}$ (which implies that $\max_{\mathbf{x}} \Pi^A(\mathbf{x}) \leq \max_{\mathbf{x}} \Pi^D(\mathbf{y}, \mathbf{x})$); (iii) $\Pi^A(\mathbf{x}^{A*}) \to \infty$ if $\alpha^A \to \infty$, then $\exists\ \tilde{\alpha}^A(\mathbf{y}) > 0$ such that $\Pi^A(\mathbf{x}^{A*}) \geq \Pi^D(\mathbf{y}, \mathbf{x}^{D*}) \Leftrightarrow \alpha^A \geq \tilde{\alpha}^A$. 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 $\tilde{\alpha}^A(\mathbf{y})$ is increasing in y_i. *Q.E.D.*

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

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

Therefore,

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

Hence,

$$\begin{aligned}(1+\alpha^A)\frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x_{-i}}) + \sum_{j \neq i} \left[\frac{\partial B_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) - \frac{\partial B_j}{\partial x_i}(y_j, \mathbf{x_{-j}}) \right] + \alpha^A \sum_{j \neq i} \frac{\partial B_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) \\ \geq (1+\alpha^A)\frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x_{-i}}) - k_i \sum_j y_j + \alpha^A \sum_{j \neq i} \frac{\partial B_j}{\partial x_i}(x_j, \mathbf{x_{-j}}).\end{aligned}$$

If $\alpha^A > 0$, then

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

because $\frac{\partial B_j}{\partial x_i}(x_i, \mathbf{x_{-i}}) > 0$ and $\frac{\partial B_j}{\partial x_i}(x_j, \mathbf{x_{-j}}) > 0$. 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 $k_i \leq \frac{\frac{\partial B_i}{\partial x_i}(x_i, \mathbf{x_{-i}})}{\sum_j y_j}$, which holds if $k_i \leq \frac{\min_i \min_{\mathbf{x}} \frac{\partial B_j}{\partial x_j}(x_j, \mathbf{x_{-j}})}{\sum_j y_j} \equiv \tilde{k}$, 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

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

it follows that

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

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

$$H^A_{N+1}(x^{A*}_{N+1}) = x^{A*}_{N+1}(x_1^{A*}, x_2^{A*}, ..., x_N^{A*}) - x^{A*}_{N+1} = 0.$$

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 $i = 1, 2, ..., N$,

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

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 $x_i^{A*} > x_i^{D*}$ for all $i = 1, 2, ..., N, N + 1$ which completes the proof of Step 2. *Q.E.D.*

A.3 Proof of Proposition 3

Consider the scenario in which data externalities are positive. Assume first that $\alpha \geq \tilde{\alpha}^A$. In this case, analytics sharing is at the same time privately optimal and socially optimal. Assume now that $\alpha < \tilde{\alpha}^A$. Then, data sharing is privately optimal, but it is socially optimal if and only if

$$\Pi^D(\mathbf{y}, \mathbf{x^{D*}}) - \Pi^A(\mathbf{x^{A*}}) + e(h(\mathbf{x^{D*}}) - h(\mathbf{x^{A*}})) \geq 0,$$

From $x_i^{D*} < x_i^{A*}$ and $\frac{\partial h}{\partial x_i} > 0$ for all i it follows that

$$h(\mathbf{x^{D*}}) - h(\mathbf{x^{A*}}) < 0.$$

Therefore, data sharing is socially optimal if and only if

$$e \leq \frac{\Pi^D(\mathbf{y}, \mathbf{x^{D*}}) - \Pi^A(\mathbf{x^{A*}})}{h(\mathbf{x^{A*}}) - h(\mathbf{x^{D*}})} \equiv e^+.$$

Consider now the scenario in which data externalities are negative. Assume first that $\alpha \leq \tilde{\alpha}^A$. In this case, data sharing is at the same time privately optimal and socially optimal. Assume now that $\alpha > \tilde{\alpha}^A$. Then, analytics sharing is privately optimal, but it is socially optimal if and only if

$$\Pi^D(\mathbf{y}, \mathbf{x^{D*}}) - \Pi^A(\mathbf{x^{A*}}) + e(h(\mathbf{x^{D*}}) - h(\mathbf{x^{A*}})) \leq 0,$$

From $x_i^{D*} < x_i^{A*}$ and $\frac{\partial h}{\partial x_i} < 0$ for all i it follows that

$$h(\mathbf{x^{D*}}) - h(\mathbf{x^{A*}}) > 0.$$

Therefore, analytics sharing is socially optimal if and only if

$$e \leq \frac{\Pi^A(\mathbf{x^{A*}}) - \Pi^D(\mathbf{y}, \mathbf{x^{D*}})}{h(\mathbf{x^{D*}}) - h(\mathbf{x^{A*}})} \equiv e^-.$$

Q.E.D.