Data brokers competition, synergic datasets, and endogenous information value

International Journal of Industrial Organization xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Industrial Organization

journal homepage: www.elsevier.com/locate/ijio

Data brokers competition, synergic datasets, and endogenous information value ☆

Laura Abrardi, Carlo Cambini *, Flavio Pino

Politecnico di Torino, Department of Management, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy

ARTICLE INFO

JEL classification:
L12
L41
L86

Keywords:
Data broker
Competition
Data sales
Price discrimination

ABSTRACT

Data Brokers (DBs) aggregate vast amounts of data and sell them to downstream firms for customer profiling. Firms can decide to purchase data from multiple DBs, to leverage synergies that enhance profiling accuracy. We study how competition between DBs and the synergies between their datasets influence the price and quantity of data sold, and the effects in the downstream market in terms of prices and incentives to purchase from multiple DBs. We find that DBs can coordinate over the quantity and price of data sold to endogenously increase the value of data synergies and induce firms to purchase multiple datasets, even when synergies are relatively weak. As synergies increase, DBs reduce the quantity of data to temper downstream competition and charge higher prices for the data. If the number of firms is endogenous, higher data prices lead to reduced market entry and consumer harm.

1. Introduction

Data Brokers (DBs), also known as information brokers or information intermediaries, are companies whose core business is to harvest and resell massive amounts of data about individuals. The data collected include sensitive data on demographic, financial, health, interests, and purchases information, which they gather by relying on a wide range of sources, such as social media, search services, apps, customer loyalty programs, card payment providers, and public records (FTC, 2014). The coverage and depth of DBs' datasets is unmatched. In 2021, Acxiom, one of the largest consumer DBs, advertised to collect up to 11,000-plus data attributes per person for 2.5 billion consumers, representing 68% of the world's digital population.[1] DBs apply proprietary heuristics or ma-

☆ This paper supersedes a previous version that circulated under the title "Data Broker Competition and Downstream Market Entry". We thank Paul Belleflamme, Alberto Bennardo, Paolo Bertoletti, Alessandro Bonatti, Luigi Buzzacchi, Joan Calzada, Jiajia Cong, Antoine Dubus, Tomaso Duso, Clara Graziano, Matthias Hunold, Jan Krämer, Gaston Llanes, Leonardo Madio, Andrea Mantovani, Sarit Markovich, José L. Moraga, Federico Navarra, Martin Peitz, Salvatore Piccolo, Carlo Reggiani, Patrick Rey, Lorien Sabatino, Luca Sandrini, Shiva Shekhar, Frank Verboven, Patrick Waelbroeck as well as participants to the 2024 SIEPI workshop (Bergamo), the 2023 Industrial Organization Winter Symposium (Bergamo), the 2023 European Association for Research in Industrial Economics (EARIE, Rome), the 2023 Conference of the Competition Law and Economics European Network (CLEEN), the Workshop on digital markets of the Universidad Complutense de Madrid, the 2023 CRESSE Conference (Rhodes), the 2023 Jornadas de Economia Industrial (JEI, Bilbao), the sixth UniBG Industrial Organization Winter Symposium for useful comments on previous versions of this manuscript. This study was carried out within the "Cyber resilience: markets, investments and regulation" project - funded by European Union - Next Generation EU within the PRIN 2022 PNRR program (D.D.1409 del 14/09/2022 Ministero dell'Università e della Ricerca). This manuscript reflects only the authors' views and opinions and the Ministry cannot be considered responsible for them.

* Corresponding author.
E-mail address: carlo.cambini@polito.it (C. Cambini).

[1] https://marketing.acxiom.com/rs/982-LRE-196/images/Fact_Sheet_Global_Data_Navigator.pdf.

https://doi.org/10.1016/j.ijindorg.2025.103146

Available online 13 February 2025

Please cite this article as: Laura Abrardi et al., *International Journal of Industrial Organization*, https://doi.org/10.1016/j.ijindorg.2025.103146

chine learning to make inferences, bundle the information into consumer categories as fine as "Affluent Baby Boomers", "Downtown Dwellers",[2] or "Metro Parents" [3] (FTC, 2014), and then resell the consumer profiles to firms for tracking and targeting.

Differently from attention platforms, DBs have no direct interaction with consumers, who are often unaware of their existence (Christl, 2017). Yet, many aspects of daily life are now mediated by DBs, ranging from online advertising, pricing, credit scoring, and risk management. It is estimated that DBs are used by 90% of landlords to perform background checks on would-be tenants (Kirchner, 2020) and 47% of employers to check the credit score of applicants (Traub and McElwee, 2016). First-degree price discrimination, once a theoretical possibility, has now become a reality (CEA, 2015; FTC, 2025). For example, Rhodes and Zhou (2024) report the case of the dating app Tinder showing different prices for its premium version to different users, depending on their gender, age, and sexual orientation.[4] Moreover, Aparicio et al. (2021) finds that the same retailer in the online grocery market sets different prices for identical products across different delivery zip-codes. In the airline industry, data-powered recommendation engines deliver personalized offers based on customer's history and preferences (Tuttle, 2013).

Yet, the accuracy of these data is far from perfect. Neumann et al. (2019) document that several DBs correctly predict a user's gender less than 50% of the time, although they also observe a large heterogeneity in audience accuracy across data brokers, thus highlighting the importance to select the right data supplier. To improve accuracy through the significant economies of scope in data aggregation (Hocuk et al., 2022), firms often trade with multiple DBs and combine their different datasets. For example, in 2013 Facebook began a partnership with four DBs – Acxiom, Epsilon, Datalogix and BlueKai – to improve its user tracking and profiling even further by using data collected outside the platform.[5]

Mainly because of the opacity of DBs' operations and the black box of their algorithms, their behavior is still largely unexplored. In this paper, we study how the accuracy of DBs' information and the synergies between their datasets influence DBs' competition and selling strategies. We distinguish between two dimensions of data. A quality dimension –the data accuracy– reflects the effectiveness of the attributes collected by DBs on each person to make inferences about her preferences/characteristics. In addition, a quantity dimension –the size of data partitions– reflects the coverage of the datasets sold by DBs in terms of consumers. For example, a data partition may be composed by data about consumers in a specific geographic region.

In our model, two competing DBs decide the price and size of data partitions, which they sell to firms in a downstream Salop (1979) market, where data allow firms to operate first-degree price discrimination on identified consumers. Firms offer a personalized price to identified consumers, and a basic price to unidentified consumers, so that each consumer observes only one price. Price discrimination is imperfect, i.e., a particular consumer can be identified with a probability lower than one, representing the information accuracy (Belleflamme et al., 2020); To explore the role of market power in the DB market, we assume that the two DBs exogenously have different levels of data accuracy (i.e., different data quality). Firms may decide to purchase both the data partitions offered by the two DBs, only one, or neither. In case they accept both offers, the accuracy of the combined dataset can be lower than the sum of the individual accuracies (*sub-modular* datasets), or higher (*super-modular* datasets). Sub-modularity emerges when, for example, the two datasets have large overlaps of information, while super-modularity might depend on strong complementarities between the different data attributes present in the two datasets.[6] For example, Equifax and Acxiom, two major players in the DB industry, provide datasets that exhibit some common data points, such as the income distributions of families within an area and the average age of tenants.[7] However, the two datasets also include different data points, such as information on houses' valuation and transaction history (in the Equifax dataset) and demographic data (in the Acxiom dataset). The combination of these data paints a much richer picture of a given geographical zone, generating synergies between the two datasets that sharply increase the targeting accuracy. Overall, the extent of the overlapping and the strength of complementarities determine the modularity of the two datasets. As our main focus is to understand the data effects on the downstream market, we abstract from the DBs' data collection phase, assuming that both the accuracies and the synergy level are exogenous.

Our first result is that DBs are always able to induce firms to purchase both datasets, even if synergies are very weak. They do so by adjusting the size of data partitions, decreasing their size as synergies become stronger. This occurs because stronger synergies increase the accuracy of the combined dataset and, in turn, downstream competition. The more intense competition curtails firms' profits and thus their willingness to pay for data. To temper this competition effect, DBs must reduce the quantity of data sold. By always inducing the sale of the combined datasets regardless of the level of synergies, DBs raise the price for data, ultimately reducing downstream entry and causing consumer harm. Although the modularity of datasets reflects the technical gains of combining different datasets, it says little about the resulting *economic* gains, i.e., the profit obtained by firms through the combined data. Our second result is that DBs, by properly choosing the price and size of data partitions, can make the value of the combined dataset for firms always higher than the sum of the individual values (*super-additive* datasets), even if there are overlaps between datasets (sub-modular datasets).

[2] "Lower-income, single, downtown-metro dwellers", that are "upper-middle-aged" and with a "high-school" or "vocational/technical" degree working to "make[] ends meet with low-wage clerical or service jobs".

[3] Consumers, "primarily in high school or vocationally educated," "handling single parenthood and the stresses of urban life on a small budget".

[4] Prices may vary significantly, ranging from 6.99$ to 34.37$ per month. See https://www.abc.net.au/news/2020-08-12/tinder-price-setting-more-expensive-for-older-people-looking-to/12549186.

[5] https://shorturl.at/gkwLR.

[6] The recent empirical literature has also highlighted how the performance of data-driven services usually exhibits diminishing returns to data. Examples include the utility created by a recommendation system (Lee and Wright, 2023), the quality of search results (Klein et al., 2022), the accuracy of sales forecasts (Bajari et al., 2019) and, most importantly, the accuracy of customer profiling (Neumann et al., 2019). Hocuk et al. (2022) finds that a 1% increase in the number of predictor variables improves prediction accuracy by 0.087-0.132%.

[7] Samples of these datasets are freely available on Amazon Web Services Marketplace and are from the Equifax datasets "Commercial Real Estate - AMP Insights", "Consumer Credit Trends", and "Analytic Dataset" and the Acxiom dataset "Germany - Statistical Data for Marketing Areas (refined Postcode)".

This implies that firms always purchase both datasets, allowing DBs to fully extract the value created by synergies. A problem with selling both datasets is that synergies could make datasets too costly for firms to purchase. To overcome this, DBs sell smaller data partitions, the stronger the synergies. By doing so, they soften firms' competition and thus raise their willingness to pay for the combined data.

Although synergies increase the value of data for firms, this does not mean that in equilibrium firms obtain higher profits, despite competition in the upstream market of DBs. Our third result is that DBs appropriate the value of synergies by setting higher partition prices, the stronger the synergies.

We extend the model in two directions. First, we consider endogenous entry, showing that the reduction in firms' profits when synergies are high lowers the downstream number of firms, possibly causing consumer harm. Then, a positive outcome for consumers can emerge only if synergies are sufficiently weak and competition between DBs is strong enough. Second, we assume that datasets can certify whether a certain consumer belongs to the dataset or not. Firms can thus distinguish between consumers outside the data partitions, consumers inside the partition but with unknown locations, and consumers inside with precisely identified locations. Thus, this setup allows us to consider first- and third-degree price discrimination simultaneously. Our results show that DBs charge higher prices for these more informative datasets, even though coordination strategies and super-additivity emerge only for higher levels of synergies.

The previous literature has highlighted the downstream effects of DBs in a context of monopoly (Montes et al., 2019; Bounie et al., 2021b; Abrardi et al., 2024; Delbono et al., 2024) or competition (Ichihashi, 2021; Bounie et al., 2021a), but it typically factors out the multiple dimensions of data (quantity vs. accuracy), or the existence of potential data synergies (see Section 2 for details). The first contribution of this paper concerns the role of data synergies. In a context in which the size of data partitions and downstream profits are exogenous, Gu et al. (2022) highlight the DBs' incentive to coordinate their sales only when the accuracy of the combined dataset is higher than the sum of individual accuracies.[8] We contribute to this literature by showing that additivity is an endogenous feature of data, which DBs can influence through the size of data partitions. This creates an incentive for DBs to coordinate their sale even if synergies are relatively weak.

A second contribution concerns the welfare effects of the data sale. A general conclusion of the aforementioned literature is that data have two opposite effects on consumer surplus. On the one hand, data intensify downstream competition, which is good for consumers (Bounie et al., 2021a). On the other hand, with a monopolistic DB data originate an entry barrier effect that harms consumers and dominates the competition effect (Abrardi et al., 2024). We show that competition by DBs, while not eliminating the entry barrier, can weaken it enough to reestablish a positive outcome for consumers, but only if DB competition is sufficiently intense. If not, consumers are better off without data. Moreover, synergies between DBs' datasets reduce competition between DBs, negatively affecting consumer surplus.

A third contribution of this paper is to investigate the possibility for firms to adopt both first- and third-degree price discrimination. We show that in this case coordination between DBs is less likely and stronger synergies are necessary to achieve super-additivity, as the extra value generated by combining different datasets declines.

Our work sheds light on an industry that is still as obscure as it is pervasive. Investment in third-party targeting services and solutions is estimated at USD 19.2 billion for the United States alone (Biegel et al., 2018). The data brokerage ecosystem generates revenues of over USD 200 billion a year (Crain, 2018), and it is dominated by a few very large companies which, as of now, are virtually unregulated from the competitive point of view (FTC, 2014; ACCC, 2023). The European Union has recently introduced new regulations for digital markets, such as the Digital Service Act, the Digital Markets Act, and the Data Act, but their effect on the DB's market is rather limited as DBs fall outside of their scope (Ruschemeier, 2022).[9] Nevertheless, competition authorities have recently started to investigate on DBs and their data sale practices.[10]

The remainder of the paper is organized as follows: Section 2 describes the relevant literature and our contribution. Section 3 describes the institutional framework of the DB industry. Section 4 presents our model. Section 5 analyzes the effects of DBs' competition in the downstream market. Section 6 studies the DB's equilibrium strategies and how they are affected by the data synergies. Section 7 extends the baseline model by endogenizing the number of downstream firms and by allowing firms to operate both first- and third-degree price discrimination. Finally, Section 8 concludes.

2. Related literature

A scant literature has investigated the effects of DB competition on downstream markets, and for the most part it factors out data synergies. Bounie et al. (2021a) analyze competition between DBs who sell data to a number of duopolistic downstream markets, where data are used for third-degree price discrimination. They show that competition between DBs reduces the price of data and provides the incentive to increase the accuracy of data. Ichihashi (2021) focuses instead on consumers' incentive to share their data with multiple DBs. Because of the non-rivalrous nature of data, the decision to share them with several DBs decreases their value,

[8] Using our terminology, they find that super-additivity emerges if and only if data are super-modular.

[9] The Digital Service Act and the Digital Markets Act focus on large online intermediaries and gatekeeper platforms, which directly interact with consumers. The Data Governance Act provides a framework for the use of public sector data. Finally, the Data Act, while giving the right to a consumer to require the sharing of her data with third parties, forbids the use of these data to compete with the product from which data originated (Article 6.2(e)) and in practice limits the effects of data sharing on competition (Krämer et al., 2023).

[10] See the press releases of the 2022 FTC case against the DB Kochava, or its 2024 case against the DB X-mode at https://shorturl.at/pwJL3 and https://shorturl.at/kCRY4.

allowing DBs to sustain a monopoly outcome in the downstream market when consumer data are used to extract consumer surplus. Gu et al. (2022) analyze the role played by the complementarity of different datasets and show under which conditions competing DBs would be better off by merging their datasets and selling them in a bundle. However, they assume that the value of the DBs' datasets is exogenous by factoring out downstream competition, and only focus on DBs' strategic price setting. We depart from Gu et al. (2022) in two ways. First, we model competition in the downstream market, so that the price of data in our setup is endogenously affected by the profits achieved by downstream firms (and thus, their willingness to pay for data). Second, we allow DBs to choose the partition size, which influences the value of data for downstream firms. Then, differently from Gu et al. (2022), in our model both the price and the quantity of data are endogenously determined by DBs.

Other works have analyzed the sale of data by a monopolistic DB. Abrardi et al. (2024) focus on its effect on downstream entry, highlighting how the presence of the DB originates a downstream entry barrier, which reduces consumer surplus. Net of entry dynamics, Delbono et al. (2024) highlight how a DB may have the incentive to sell data only to a subset of downstream firms, particularly to those that do not benefit the most from it, that can preempt another firm from gaining a strong competitive advantage. Other studies analyze the effects of a monopolistic DB on a downstream duopolistic market, finding that consumer surplus is proportional to the consumers' cost to hide from being targeted (Montes et al., 2019), and that the DB may prefer to only sell a subset of consumer data to stifle downstream competition and increase firms' willingness to pay for data (Bounie et al., 2021b). The DB may also choose different data allocations (exclusive or non-exclusive) based on whether its data are fully or partially informative (Braulin, 2023), or depending on the selling mechanism it uses (Bounie et al., 2022). Belleflamme et al. (2020) specifically focus on the effect of data accuracy when the data sold by a monopolistic DB can be used by two firms to imperfectly price discriminate in a homogeneous goods market. They show that data allow positive profits to downstream firms only if the latter obtain data with different accuracy, as this allows firms to escape the Bertrand paradox. Haberer et al. (2022) focus on the role of Personal Data Brokers (PDBs), who directly offer financial compensation to consumers in exchange for data they generate by interacting with content providers. They find that PDBs can generate a positive outcome for consumers only if the PDB is more efficient than content providers in generating value from consumer data.

Another stream of literature focuses on the effects of data in the downstream market in the absence of DBs. Data may be exogeneously available to firms (Thisse and Vives, 1988; Shaffer and Zhang, 1995; Liu and Serfes, 2004; Taylor and Wagman, 2014; Chen et al., 2020), or firms may directly obtain data from consumers (Villas-Boas, 2004; Bergemann and Bonatti, 2011; Hagiu and Wright, 2023). In such cases, the use of data for price discrimination intensifies competition downstream, typically making consumers better off.[11] Similar conclusions hold even when the personalized offers are discounts over a basic price (Baik and Larson, 2022), which is always observable by consumers. Yu and Zhang (2024) extend this framework by showing that, if an incumbent can offer targeted discounts, it can use the basic price to signal its targeting cost to the rival, tempering competition. In both cases, tailored discounts, similar to tailored prices, induce a prisoner's dilemma: firms' profits decrease once they can target consumers, but they do so because not targeting would put them at a disadvantage against their rivals. The prisoner's dilemma implies lower firm profits, which would translate into a decrease in entry. However, if access to data is exclusive, the informed firm can strategically share some of the available data with the uninformed rival to influence the rival's pricing strategy and, in turn, increase his own profits (Choe et al., 2023). A similar result is obtained by Bhargava et al. (2024) in a different setting, where data improve product quality: an incumbent platform prefers to partially share data with an entrant platform, even for free, in order to temper competition among them. All in all, the previous literature highlights that the use of data for price discrimination may be beneficial for consumers, as long as the downstream market concentration is not affected. Moreover, when synergies between datasets are particularly strong, competition between DBs can be softened by the non-rivalrous nature of data or by strategic price setting by DBs that allows them to extract more surplus from downstream firms. Conversely, recent empirical contributions (Bajari et al., 2019; Neumann et al., 2019; Lee and Wright, 2023; Klein et al., 2022) have often highlighted diminishing returns to data. Our study thus contributes to the literature by highlighting a novel channel through which DBs can temper competition between themselves, even when synergies between datasets are weak: by strategically choosing how much data to sell to downstream firms, DBs can endogenously increase the value of information, leading to a coordinated sale even under diminishing returns to data.

3. Institutional framework

The regulatory framework governing data collection and sharing is primarily anchored in the privacy legislation. In the European Union, the General Data Protection Regulation (GDPR) sets stringent guidelines on how personal data can be collected, processed, and shared. Under GDPR, DBs are required to obtain explicit consent from individuals and maintain transparency regarding data usage. However, in practice DBs often operate in a legal grey area where the collection, aggregation, and sale of personal data frequently occur without individuals' explicit consent, exploiting exceptions such as relying on "legitimate interests" (Ruschemeier, 2022). Additionally, many brokers sidestep GDPR's stringent requirements by dealing with anonymized data, which technically falls outside the regulation's scope, despite the persistent risks of re-identification (Ohm, 2010).

Also at the enforcement level, the complexity of modern data ecosystems, where data is continuously shared and re-shared across industries and borders, exacerbates the challenge of maintaining clear origins and consent statuses for personal data. This creates enforcement difficulties, particularly when individuals seek to exercise their rights under GDPR. For instance, Ruschemeier (2022) documents a case where an Austrian data subject filed an access request under Article 15 of the GDPR, asking a DB to disclose

[11] For recent surveys regarding data markets, refer to Bergemann and Bonatti (2019) and Goldfarb and Tucker (2019).

the origin and recipients of their data. The DB, an address publisher, claimed ignorance about where the data had been obtained, highlighting a fundamental issue in data traceability and accountability. A further enforcement challenge stems from the fact that enforcement of the GDPR varies significantly across EU member states, resulting in inconsistent application of its provisions (European Commission, 2024). This fragmented regulatory landscape is even more pronounced in the United States, where there is no federal equivalent to the GDPR. Instead, privacy regulations are governed by a patchwork of state laws, which often lack the rigor needed to effectively regulate the activities of DBs. This regulatory fragmentation grants DBs considerable freedom, raising persistent concerns about privacy, data security, and the ethical use of personal information.

In addition to privacy regulations, the European Union has introduced a suite of new laws to address the complexities of the digital economy. The Digital Markets Act (DMA) and the Digital Services Act (DSA) are designed to regulate large online platforms, ensuring fair competition, transparency, and accountability in digital markets. The Data Act seeks to foster a more competitive data economy by establishing rules for data access and sharing, particularly between businesses and governments, while the Data Governance Act provides a framework for the use of public sector data. Together, these laws form a comprehensive regulatory architecture aimed at balancing innovation with the protection of individual rights and promoting fair competition.

Despite these advances, DBs often operate beyond the direct reach of these frameworks. While the DMA and DSA primarily focus on large digital platforms that act as gatekeepers in the online ecosystem, they do not directly address the activities of DBs, leaving this sector in a regulatory blind spot (Ruschemeier, 2022). This gap has been noted by competition authorities, concerned that the market power of DBs is underestimated and that they might benefit from fragmented oversight (ACCC, 2023; FTC, 2014). Furthermore, while the Data Act promotes transparency and access to data, its focus is primarily on business-to-business and business-to-government data sharing, not the consumer data routinely handled by brokers. As a result, DBs continue to operate with minimal regulation, raising persistent concerns about privacy and data security (Mishra, 2021).

4. The model

We study two interconnected markets. In the upstream market, two DBs (DB_1 and DB_2) have data regarding individual consumers' preferences in a downstream market. In the downstream market, horizontally differentiated firms can purchase the data from either, or both, DBs and offer tailored prices to consumers.

4.1. Consumers, firms and data brokers

In the downstream market, we consider a circular city (Vickrey, 1964; Salop, 1979), with a number $n > 2$ of firms. Firms (he), indexed by $i \in \{0, 1, 2, ..., n-1\}$, sell competing products to consumers. Firms are equally spaced in the circle, such that a generic firm i is located in $\frac{i}{n}$. Firms' marginal costs are normalized to zero.

Consumers (she) are uniformly distributed over the circle, and their mass is normalized to 1. Their location is indexed by $x \in [0, 1)$ in counter-clockwise order, and each of them buys at most one unit of the product. Gross utility derived from consumption is v, and consumers face a linear transportation cost t.

In the upstream market, two DBs (it) have datasets containing customers' information that allow firms to identify consumers with a given probability. In line with the previous literature (Montes et al., 2019; Bounie et al., 2021b) we assume that DBs do not face any cost of data collection or processing, and their reservation profit is equal to zero. Following Belleflamme et al. (2020), DB_1's dataset contains information that grants firms a probability $\alpha \in [0, 1]$ of identifying consumers and operating first-degree price discrimination. Instead, DB_2's dataset contains information granting firms a probability $\beta\alpha, \beta \in [0, 1)$ of identifying consumers. We can interpret α as the data accuracy and β as the degree of vertical differentiation between DBs. The combination of the two datasets provides an accuracy γ, with $\alpha < \gamma \leq 1$ over those consumers. The accuracy of the combined datasets γ can be interpreted as the level of synergies between the two datasets. The values of α, β and γ are exogenously given and common knowledge for both DBs and downstream firms. For example, DBs offer free samples of their datasets, allowing firms to learn the data attributes contained there.[12] Depending on the level of synergies $\gamma \in (\alpha, 1]$ of the combined dataset, we have two cases. If $\gamma < \alpha + \beta\alpha$, the two datasets are *sub-modular*, e.g., they contain some overlapping information. If instead $\gamma \geq \alpha + \beta\alpha$, the two datasets are *super-modular*, i.e., there are complementarities between the information in the two datasets.

DBs can sell partitions of their datasets to downstream firms. A partition $d_{i,k} \in \left[0, \frac{1}{n}\right]$, offered by DB_k to firm i, with $k \in \{1, 2\}$ provides information about the location of consumers in a segment of the market. Note that DBs can choose not to offer data to any firm i by setting $d_{i,k} = 0$.[13] In line with the previous literature (Bounie et al., 2021b; Abrardi et al., 2024), we assume that

[12] Given that information can be replicated at no cost, the sale of data could give rise to the famous Arrow paradox (Arrow, 1972), whereby the actual value of information cannot be determined until after it has been received by a prospective buyer, but by then the information becomes effectively worthless to sell because the buyer already knows it. While Arrow's information paradox diagnoses a genuine problem in trading information and constitutes a theoretical pillar of intellectual property management, a competitive market for information still exists in many situations. For instance, the appropriation problem on the supply side can be solved if the seller can communicate beforehand the value of knowledge for the buyer without data disclosure (Leppälä, 2013, 2015). In our setup, firms learn the value of the datasets by observing α, β, γ and the partition sizes, but the datasets are proprietary and their content can be accessed only after paying the price of data.

[13] Moreover, by imposing $d_{i,k} \leq \frac{1}{n}$, we assume that the partitions do not overlap, as in Bounie et al. (2021b). This implies that each consumer is identified by at most one firm. A direct implication of this assumption is that downstream firms directly compete only on unidentified consumers. If instead more than one firm has information on a consumer they will engage in price wars, reducing their profits and their willingness to pay for data (Thisse and Vives, 1988; Montes et al., 2019; Delbono et al., 2024).

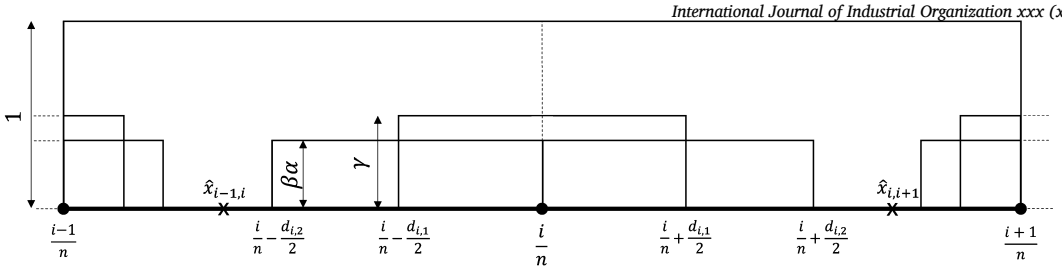

Fig. 1. Firm i's market share when buying from both DB_1 and DB_2, assuming $d_{i,2} > d_{i,1}$.

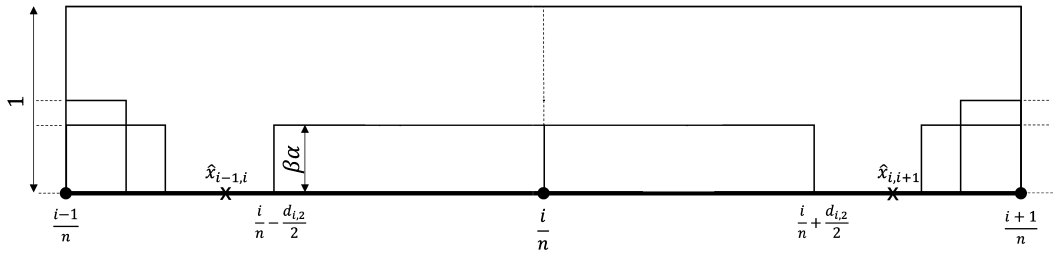

Fig. 2. Firm i's market share when only buying from DB_2.

the partitions offered to each firm are centered on his location, i.e., a partition sold by DB_k to firm i contains information about consumers located on the segment $\left[\frac{i}{n}-\frac{d_{i,k}}{2}, \frac{i}{n}+\frac{d_{i,k}}{2}\right]$.[14] Although the assumption of centered partitions allows us to significantly streamline the analysis, we relax it in Appendix A, and show that it entails no loss of generality.

The data sold by the two DBs have different accuracy. In particular, the partition sold by DB_1 allows the generic firm i to identify a segment of size $d_{i,1}$ of consumers with probability α, whereas the partition sold by DB_2 allows to identify $d_{i,2}$ consumers with probability $\beta\alpha$.[15] Consumers located inside both $d_{i,1}$ and $d_{i,2}$ are identified with probability γ.

A given firm i offers a basic price $p^B_{i,k'} \geq 0$, to all unidentifed consumers, and location-specific tailored prices $p^T_{i,k'}(x) \geq 0$ to identified ones, where $k' \in \{0, 1, 2, 12\}$ indicates whether the firm has not purchased data ($k' = 0$), or has purchased data from DB_1 ($k' = 1$), DB_2 ($k' = 2$), or both ($k' = 12$).

Following the relevant literature (Montes et al., 2019; Bounie et al., 2021a,b; Abrardi et al., 2024), we assume that each consumer only observes one price from a given firm, and, as a tie-breaking rule, that consumers prefer tailored prices over basic prices when indifferent. This is the case, for example, of firms operating exclusively online, so that a consumer landing on the website is shown only one price, either tailored or basic.[16]

A consumer utility is defined as

$$U(x,i) = v - p_i(x) - tD(x,i),$$

where $p_i(x) \in \{p^B_{i,k'}, p^T_{i,k'}(x)\}$ and $D(x,i)$ is the shortest arch between the consumer's location x and firm's location $\frac{i}{n}$. We denote the location of the indifferent consumer between firms i and $i+1$ as $\hat{x}_{i,i+1}$.

Fig. 1 illustrates the key features of our setup when firm i buys partitions from both DBs and $d_{i,1} < d_{i,2}$ (i.e., DB_2's partition contains information about a larger set of consumers than DB_1's). By purchasing both datasets, firm i has a probability γ of identifying consumers on the arch $[\frac{i}{n}-\frac{d_{i,1}}{2}, \frac{i}{n}+\frac{d_{i,1}}{2}]$, which are contained in both data partitions. Moreover, it identifies with probability $\beta\alpha$ consumers on the arches $[\frac{i}{n}-\frac{d_{i,2}}{2}, \frac{i}{n}-\frac{d_{i,1}}{2}]$ and $[\frac{i}{n}+\frac{d_{i,1}}{2}, \frac{i}{n}+\frac{d_{i,2}}{2}]$, whose information are only contained in DB_2's partition. Finally, it does not identify consumers located on arches $[\hat{x}_{i-1,i}, \frac{i}{n}-\frac{d_{i,2}}{2}]$ and $[\frac{i}{n}+\frac{d_{i,2}}{2}, \hat{x}_{i,i+1}]$, whose information are outside both data partitions. Firm i offers a tailored price to identified consumers, and the basic price to unidentified ones. Fig. 2 illustrates the key features of our setup when firm i only buys a partition from DB_2.

Firm i's profits when buying from both DBs ($k' = 12$) and $d_{i,2} \geq d_{i,1}$, before paying for data, are:

[14] Previous literature in marketing has stressed how targeting consumers with strong preferences is more beneficial to firms (Iyer et al., 2005). From a theoretical viewpoint, Bounie et al. (2021a) show in a Hotelling market that in equilibrium partitions contain the firms' locations and cannot be disjoint, i.e., partitions must be a continuous segment. Although their result is derived in a linear setting, it extends to our Salop circular city model because the competitive structure of equidistant firms remains analogous in both frameworks.

[15] The level of information accuracy α could also be interpreted as the share of consumers, uniformly distributed, that a firm can identify over the arch $[\frac{i}{n}-\frac{d_{i,k}}{2}, \frac{i}{n}+\frac{d_{i,k}}{2}]$ once he buys a partition from DB_k.

[16] Given that consumers can only observe one price, tailored prices could be higher than basic prices.

$$\pi_{i,12} = \gamma \left(\int_{\frac{i}{n}-\frac{d_{i,1}}{2}}^{\frac{i}{n}+\frac{d_{i,1}}{2}} p^{\mathrm{T}}_{i,12}(x)\,dx \right) + (1-\gamma) d_{i,1} p^{\mathrm{B}}_{i,12} +$$

$$+ \beta\alpha \left(\int_{\frac{i}{n}-\frac{d_{i,2}}{2}}^{\frac{i}{n}-\frac{d_{i,1}}{2}} p^{\mathrm{T}}_{i,12}(x)\,dx + \int_{\frac{i}{n}+\frac{d_{i,1}}{2}}^{\frac{i}{n}+\frac{d_{i,2}}{2}} p^{\mathrm{T}}_{i,12}(x)\,dx \right) + (1-\beta\alpha)(d_{i,2}-d_{i,1}) p^{\mathrm{B}}_{i,12} +$$

$$+ p^{\mathrm{B}}_{i,12} \left(\hat{x}_{i,i+1} - \hat{x}_{i-1,i} - d_{i,2} \right), \quad (1)$$

where the first and second term on the right-hand side are firm i's profits from consumers it identifies with accuracy γ, which are identified through both DB_1 and DB_2's partitions, the third and fourth term are the profits from consumers it identifies only through DB_2's partitions with accuracy $\beta\alpha$, and the fifth term are profits from unidentified consumers. Note that the second and third terms in equation (1) are independent of each other. In other words, if a consumer is not identifiable by firm i through $d_{i,1}$, he will not infer whether the consumer belongs to $d_{i,2}$ with some probability. This follows from the assumption that the partitions are bought simultaneously by firms. Analogously, firm i's profits when buying from both DBs and $d_{i,2} < d_{i,1}$, before paying for data, are[17]:

$$\pi_{i,12} = \gamma \left(\int_{\frac{i}{n}-\frac{d_{i,2}}{2}}^{\frac{i}{n}+\frac{d_{i,2}}{2}} p^{\mathrm{T}}_{i,12}(x)\,dx \right) + (1-\gamma) d_{i,2} p^{\mathrm{B}}_{i,12} +$$

$$+ \alpha \left(\int_{\frac{i}{n}-\frac{d_{i,1}}{2}}^{\frac{i}{n}-\frac{d_{i,2}}{2}} p^{\mathrm{T}}_{i,12}(x)\,dx + \int_{\frac{i}{n}+\frac{d_{i,2}}{2}}^{\frac{i}{n}+\frac{d_{i,1}}{2}} p^{\mathrm{T}}_{i,12}(x)\,dx \right) + (1-\alpha)(d_{i,1}-d_{i,1}) p^{\mathrm{B}}_{i,12} +$$

$$+ p^{\mathrm{B}}_{i,12} \left(\hat{x}_{i,i+1} - \hat{x}_{i-1,i} - d_{i,1} \right). \quad (2)$$

4.2. Data sale and timing

DBs simultaneously sell data partitions through non-renegotiable Take It Or Leave It (TIOLI) offers (Bergemann et al., 2018). We denote by $w_{i,k}$ the price offered by DB k for partition $d_{i,k}$.

The timing of the model is as follows[18]:

Stage 1. Each DB $k \in \{1,2\}$ chooses a partition $d_{i,k}$ for each firm and offers it to that firm at a price $w_{i,k}$.
Stage 2. Each firm chooses whether to accept or decline the DBs' offers.
Stage 3. Firms set basic prices $p^{B}_{i,k'}$ for unidentified consumers, i.e. consumers located outside the data partition, and consumers located within the data partition in the case data do not allow their identification.
Stage 4. Firms that purchased a partition set tailored prices $p^{T}_{i,k}(x)$ for the identified consumers, i.e. consumers located within the data partition in the case data allow their identification.

We solve the model through backward induction by first analyzing downstream equilibrium prices (in Section 5), and then considering DBs' strategies in the upstream market (in Section 6). As a useful benchmark, we refer to the standard Salop (1979) model with an exogenous number of firms. In equilibrium firms set prices equal to $\widetilde{p}_i = \frac{t}{n}$, and obtain profits equal to $\widetilde{\pi}_i = \frac{t}{n^2}$.

5. Equilibrium downstream prices

We first analyze the firms' equilibrium pricing strategies in the subgame in which all firms acquire data from both DB_1 and DB_2, and then consider the subgames in which a generic firm i instead buys data from only DB_1 or only DB_2.[19]

When all firms purchase data from both DB_1 and DB_2, indifferent consumers' locations only depend on basic prices and thus correspond to those in the standard Salop model. The indifferent consumers between firm i and firms $i-1$ and $i+1$, respectively, are:

[17] We will show in Section 6 that the case $d_{i,2} < d_{i,1}$ never arises in equilibrium.

[18] Prices in stages 3 and 4 are set sequentially to ensure the existence of Pure Strategy Nash Equilibria (see also Montes et al. (2019); Bounie et al. (2021b,a) among others).

[19] In Section 6 we will show that firms' profits are always lower than their rivals if they are less informed than them. Therefore, we can rule out equilibria in which *all* firms only purchase one dataset, due to the firm's incentive to deviate and buy from both DBs.

$$\hat{x}_{i-1,i} = \frac{2i-1}{2n} + \frac{p^B_{i,12} - p^B_{i-1,12}}{2t} \quad \text{and} \quad \hat{x}_{i,i+1} = \frac{2i+1}{2n} + \frac{p^B_{i+1,12} - p^B_{i,12}}{2t}. \tag{3}$$

Firm i, which buys a partition from both DBs, can offer tailored prices $p^T_{i,12}(x)$ to the identified consumers. The tailored prices match the direct rivals' basic prices in utility levels, resulting in

$$p^{\mathrm{T}}_{i,12}(x) = \begin{cases} p^{\mathrm{B}}_{i-1,12} + 2tx - \frac{t}{n}(2i-1) & \text{for } x \in [\frac{i}{n} - \frac{\max\{d_{i,1},d_{i,2}\}}{2}, \frac{i}{n}] \\ p^{\mathrm{B}}_{i+1,12} - 2tx + \frac{t}{n}(2i+1) & \text{for } x \in [\frac{i}{n}, \frac{i}{n} + \frac{\max\{d_{i,1},d_{i,2}\}}{2}]. \end{cases} \tag{4}$$

Using (3) and (4), we obtain firm i's FOC of Equation (1) with respect to $p^B_{i,1}$, which provides firm i's reaction function in basic price:

$$p^{\mathrm{B}}_{i,12} = \begin{cases} \frac{t}{2n} - \frac{t}{2}\left(\gamma d_{i,1} + \beta\alpha\left(d_{i,2} - d_{i,1}\right)\right) + \frac{p^{\mathrm{B}}_{i+1,12} + p^{\mathrm{B}}_{i-1,12}}{4} & \text{for } d_{i,2} \geq d_{i,1} \\ \frac{t}{2n} - \frac{t}{2}\left(\gamma d_{i,2} + \alpha\left(d_{i,1} - d_{i,2}\right)\right) + \frac{p^{\mathrm{B}}_{i+1,12} + p^{\mathrm{B}}_{i-1,12}}{4} & \text{for } d_{i,2} < d_{i,1}. \end{cases} \tag{5}$$

From the system of reaction functions of all firms, we obtain firms' equilibrium prices, whose properties are described in the following lemma.

Lemma 1. *When all firms purchase data from both DB_1 and DB_2, firms' basic and tailored prices in equilibrium are decreasing in $\alpha, \gamma, d_{i,1}, d_{i,2}$ and weakly decreasing in β, $\forall i \in \{0, ..., n-1\}$.*

Proof. See Appendix A. ■

As the quantity of data or their accuracy increases, or synergies become stronger, a firm identifies a larger share of consumers. Then, the equilibrium basic price is offered to consumers who are, on average, farther from the firm's location, and have a lower willingness to pay for its product. This reduces the basic price, as highlighted also by (5). In turn, the reduction of basic prices causes a downward pressure on tailored prices as well as firms' profits (the so called *competition effect* highlighted by the literature). At the same time, however, data allow firms to offer tailored prices that more effectively extract consumer surplus and increase firm profits (referred to in the literature as *surplus extraction effect*). Overall, the opposite signs of the competition and surplus extraction effect have an ambiguous effect on firms' profits (Thisse and Vives, 1988).

The effect of β on equilibrium prices is strictly negative, but only when $d_{i,2} > d_{i,1}$, because in this case some consumers are only identified through DB_2's partitions with accuracy $\beta\alpha$. If instead $d_{i,2} \leq d_{i,1}$, no consumer is identified by DB_2's partition alone, so that β does not affect equilibrium prices.

We now focus on the subgame where firm i buys data from only DB_2, while all other firms buy data from both DBs. In this case, firm i's profit function is given by

$$\pi_{i,2} = \beta\alpha \int_{\frac{i}{n} - \frac{d_{i,2}}{2}}^{\frac{i}{n} + \frac{d_{i,2}}{2}} p^{\mathrm{T}}_{i,2}(x)\,dx + (1-\beta\alpha) d_{i,2} p^{\mathrm{B}}_{i,2} + p^{\mathrm{B}}_{i,2}\left(\hat{x}_{i,i+1} - \hat{x}_{i-1,i} - d_{i,2}\right), \tag{6}$$

while all other firms' profits are given by Equations (1) and (2). Analogously, firm i's profit when he buys data from only DB_1 is

$$\pi_{i,1} = \alpha \int_{\frac{i}{n} - \frac{d_{i,1}}{2}}^{\frac{i}{n} + \frac{d_{i,1}}{2}} p^{\mathrm{T}}_{i,1}(x)\,dx + (1-\alpha) d_{i,1} p^{\mathrm{B}}_{i,1} + p^{\mathrm{B}}_{i,1}\left(\hat{x}_{i,i+1} - \hat{x}_{i-1,i} - d_{i,1}\right). \tag{7}$$

The properties of firms' equilibrium prices in this subgame are described in the following lemma.

Lemma 2. *(i) If firm i buys data either from DB_1 or DB_2, whereas all other firms buy data from both DBs, all firms' basic and tailored prices in equilibrium are higher than the prices in the subgame where all firms buy from both DBs. (ii) Equilibrium basic prices are higher in the subgame where firm i buys from DB_2 than when he buys from DB_1 iff $\beta d_{i,2} < d_{i,1}$.*

Proof. See Appendix A. ■

As firm i buys less data than other firms, his equilibrium prices are higher and, by the complementarity of pricing strategies, also other firms' prices are higher. In particular, firm i's basic price is decreasing in the total mass of identified consumers. If firm i buys from DB_1, he identifies $\alpha d_{i,1}$ consumers, whereas if he buys from DB_2 he identifies $\alpha\beta d_{i,2}$ consumers. Thus, if $\beta d_{i,2} < d_{i,1}$, firm i identifies fewer consumers when buying from DB_2, and thus sets a higher basic price in that subgame.

6. Equilibrium data partitions and entry

We now focus on the upstream market for data. DBs compete in prices and in the sizes of the partitions they sell to downstream firms. DBs have different data points, so that combining the two datasets grants synergies to the purchasing firms. For example, DB_1 has information about consumers' gender, income, and occupation, whereas DB_2 has information on consumers' occupation and address.

Recall that the two datasets are *sub-modular* if $\gamma < \alpha + \beta\alpha$, i.e., they contain some overlapping information. If instead $\gamma \geq \alpha + \beta\alpha$, the two datasets are *super-modular*, i.e., there are synergies between the two datasets. The "technological" concept of modularity affects the economic value of data. Let us denote with $\Delta_{i,k'}(d_{i,1}, d_{i,2}) = \pi_{i,k'} - \pi_{i,0}$ the extra profits firm i obtains when purchasing a partition from either or both DBs, where $\pi_{i,k'}$ are firm i's profits (gross of the price of data) when purchasing partitions from $k' \in \{0, 1, 2, 12\}$. We define datasets as *sub-additive* if $\Delta_{i,12}(d_{i,1}, d_{i,2}) < \Delta_{i,1}(d_{i,1}, d_{i,2}) + \Delta_{i,2}(d_{i,1}, d_{i,2})$, and *super-additive* if $\Delta_{i,12}(d_{i,1}, d_{i,2}) \geq \Delta_{i,1}(d_{i,1}, d_{i,2}) + \Delta_{i,2}(d_{i,1}, d_{i,2})$.

The implications of data additivity have been explored by Gu et al. (2022) in a setting where the value of the datasets are exogenous and DBs always sell the whole dataset. In their context, the concept of modularity corresponds to that of additivity. Conversely, in our framework, DBs endogenously choose the partition sizes, leading to endogenous valuations of datasets. We will show that the concepts of modularity and additivity do not coincide when the value of data is endogenous.

6.1. Equilibrium price and size of data partitions

In the presence of synergies, firms may purchase data from both DBs. Thus, DBs have only two possible pricing strategies in equilibrium.[20] First, they could set the price of their data equal to the additional value that their dataset provides relative to the dataset of the rival, i.e., $w_{i,k} = \Delta_{i,12} - \Delta_{i,-k}$. DB_1 and DB_2 solve the following problems, respectively[21]:

$$\max_{d_{0,1}, d_{1,1}, \ldots, d_{n-1,1}} \pi_{DB_1} = \sum_{i=0}^{n-1} \left(\Delta_{,0i^*,1.2} - \Delta_{i,2} \right), \tag{8}$$

$$\max_{d_{0,2}, d_{1,2}, \ldots, d_{n-1,2}} \pi_{DB_2} = \sum_{i=0}^{n-1} \left(\Delta_{i,12} - \Delta_{i,1} \right). \tag{9}$$

Second, they could set their data price to extract the value of the *combined* dataset, i.e., $w_{i,1} + w_{i,2} = \Delta_{i,12}$. Then, the DBs maximize their joint profits:

$$\max_{d_{0,1}, \ldots, d_{n-1,1}, d_{0,2}, \ldots, d_{n-1,2}} \pi_{DB_1} + \pi_{DB_2} = \sum_{i=0}^{n-1} \Delta_{i,12}. \tag{10}$$

Denoting with $d^*_{i,k}, k \in \{1, 2\}$, the equilibrium partition of DB_k, with $\Delta^*_{i,k'} = \Delta_{i,k'}(d^*_{i,1}, d^*_{i,2})$, $k' \in \{1, 2, 12\}$, the equilibrium extra profits, and with $w^*_{i,k}$ the equilibrium partition prices, the following Proposition highlights the conditions under which datasets are sub-additive and summarizes the DBs' strategies.

Proposition 1. *There exists a threshold $\gamma^* \in (\alpha, \alpha + \beta\alpha)$ such that datasets are sub-additive iff $\gamma < \gamma^*$. If datasets are sub-additive, in equilibrium, $\Delta^*_{i,k'} = \Delta^*_{k'}$, $w^*_{i,k} = w^*_k$ and $d^*_{i,k} = d^*_k$, $\forall i \in \{0, \ldots, n-1\}$. Moreover, there exists a unique Nash equilibrium such that $d^*_2 > d^*_1$ and $w^*_k = \Delta^*_{12} - \Delta^*_{-k}, k \in \{1, 2\}$. Both d^*_1 and d^*_2 are decreasing in γ.*

Proof. See Appendix A. ■

Proposition 1 highlights that, if data synergies are weak, the joint value of the datasets is lower than the sum of the individual values, i.e., datasets are sub-additive. Then, DBs act non-cooperatively by setting the partition prices equal to the additional value generated by their own datasets. Interestingly, such equilibrium survives even in the limit for $\beta \to 1$, i.e., both DBs have the same accuracy level. In such a case, both DBs charge a price equal to the additional value generated by the synergies γ. The equilibrium partitions' sizes offered by the DBs are:

$$d^*_1 = \frac{H(n)}{n(1 + M(n)(\gamma - \beta\alpha))}, \tag{11}$$

$$d^*_2 = d^*_1 \left(\frac{1 + M(n)\alpha}{1 + M(n)\beta\alpha} \right) \tag{12}$$

[20] See the proof of Propositions 1 and 2 in Appendix A.

[21] In equilibrium we show that partition prices are always positive. Thus, the DBs' individual rationality constraints are met.

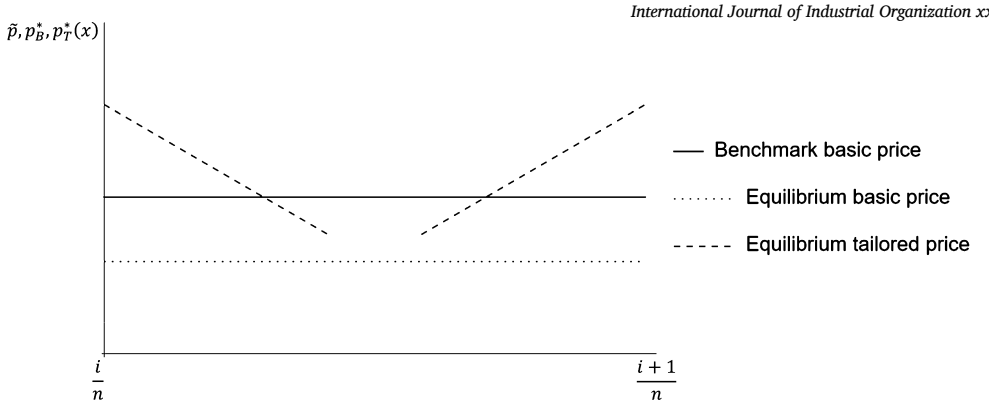

Fig. 3. Benchmark price, equilibrium basic price, and equilibrium tailored prices as a function of consumers' location. $n = 6, t = 5, \alpha = 0.5, \beta = 0.2, \gamma = 0.55$. Under these parameters, $\gamma^* \approx 0.57$, ensuring data sub-additivity.

where $0 < H(n) < M(n) < 1$ and $H'(n), M'(n) > 0, \forall n$.[22] This implies that both d_1^* and d_2^* are smaller than $\frac{1}{n}$, i.e., the DBs choose d^* to temper the *competition effect* of data by leaving some consumers unidentified. Moreover, as $\beta < 1$, we find that DB_2 offers larger partitions than DB_1: intuitively, DB_2 compensates its lower accuracy by providing data regarding more consumers.

The effects of α, β and γ on the equilibrum partition sizes are driven by the interplay between the *surplus extraction* and *competition effects*. DB_1's equilibrium partition is only influenced by the strength of synergies γ and its rival's accuracy $\beta\alpha$, and not by its own accuracy α. Indeed, as DB_1 offers smaller partitions than DB_2, in equilibrium a generic firm i identifies consumers in $\left[\frac{i}{n} - \frac{d_1^*}{2}, \frac{i}{n} + \frac{d_1^*}{2}\right]$ with accuracy γ, and consumers in $\left[\frac{i}{n} - \frac{d_2^*}{2}, \frac{i}{n} - \frac{d_1^*}{2}\right]$ and $\left[\frac{i}{n} + \frac{d_1^*}{2}, \frac{i}{n} + \frac{d_2^*}{2}\right]$ with accuracy $\beta\alpha$ (as shown in Fig. 1). An increase of γ, while enlarging both the competition and surplus extraction effects, has a stronger impact on the former due to the complementarity between firms' pricing strategies. To temper the competition effect, which negatively affects firms' profits, DB_1 reduces the equilibrium partition size. Instead, an increase of either α or β favors firms if they buy data from DB_2, as its accuracy increases. Then, DB_1 increases the partition size to decrease firms' profits should they buy data from DB_2. By doing so, it increases their willingness to pay for its data.

With regard to DB_2, its equilibrium partition size is also increasing in γ, as it tries to reduce the competition effect as well. Moreover, d_2^* is increasing in α and decreasing in β. Intuitively, all else equal, an increase in α implies a higher accuracy gap between the DBs' datasets, leading DB_2 to increase the partition size to compensate the lack in accuracy. Conversely, an increase of β reduces said gap, leading to the opposite effect.

Turning to the effects of data on firms, we find that equilibrium prices are equal to

$$p_{i,12}^{B*} = \frac{t}{n} - t\beta\alpha(d_2^* - d_1^*) - t\gamma d_1^* \tag{13}$$

and

$$p_{i,12}^{T}(x) = \begin{cases} p_{i-1,12}^{B^*} + 2tx - \frac{t}{n}(2i-1) & \text{for } x \in [\frac{i}{n} - \frac{d_2^*}{2}, \frac{i}{n}] \\ p_{i+1,12}^{B^*} - 2tx + \frac{t}{n}(2i+1) & \text{for } x \in [\frac{i}{n}, \frac{i}{n} + \frac{d_2^*}{2}], \end{cases} \tag{14}$$

with $p_{i,12}^{B*} = p_{i+1,12}^{B*} \forall i \in \{0, \ldots, n-1\}$. Indeed, as in equilibrium firms buy same-sized partitions, their basic prices are the same. In line with the previous literature (Montes et al., 2019; Bounie et al., 2021b; Abrardi et al., 2024), we find that the use of data for price discrimination intensifies competition between firms, leading to lower basic prices, as shown in (13). Moreover, tailored prices are higher than basic prices, as firms factor in the positional advantage they have over the identified consumers, as shown in (14). Although identified consumers are worse off than unidentified consumers, as they get higher prices, it does not imply that all identified consumers are worse off with respect to the benchmark model (the Salop model where data are absent). Indeed, as we show in Fig. 3, identified consumers who are located farther from the firms' locations enjoy a lower price than in the benchmark, as the *competition effect* of data dominates the *surplus extraction effect*.

As in equilibrium all firms purchase same-sized partitions, they obtain equal market shares of $\frac{1}{n}$. Then, as their equilibrium prices decrease with data, as described in (13) and (14), firms' equilibrium profits decrease with data.

A question naturally arises: as firms' profits decrease with data, why would they buy partitions from DBs in the first place? To answer this question, note that firms face a prisoner dilemma. Indeed, as all firms buy data, all of them price more aggressively, decreasing their equilibrium profits even before paying for data. Moreover, their profits decrease as data synergies increase, as the *competition effect* of data becomes stronger. However, not buying data would put them at a disadvantage with respect to their rivals,

[22] The expressions of $H(n), M(n)$ are provided in Appendix A.

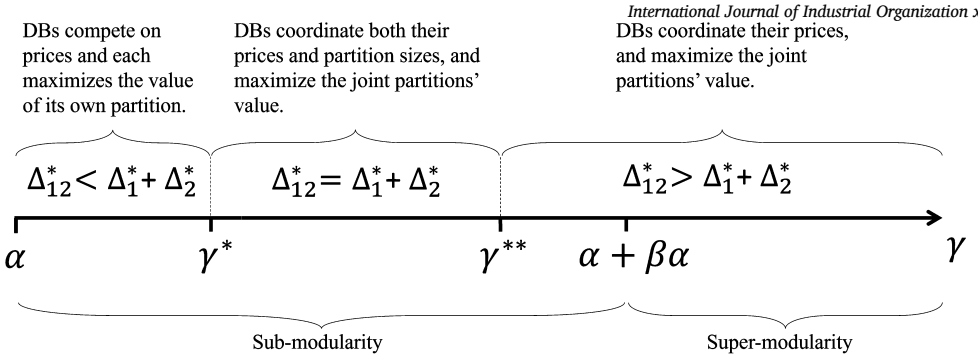

Fig. 4. DBs' equilibrium strategies as a function of γ.

resulting in lower profits. Thus, firms have a positive willingness to pay for data, leading them to buy both partitions. This result is in line with the previous literature on competing informed firms (Thisse and Vives, 1988; Montes et al., 2019; Bounie et al., 2021b).[23]

As already noted, our analysis shows that if DBs' datasets are sub-additive, they set their partitions' prices equal to the additional value generated by their own datasets. However, as the strength of synergies increases, both DBs increase their partition prices as they try to appropriate the additional value generated by the synergies. After a threshold $\gamma^* < \alpha + \beta\alpha$, the partition prices set under this pricing strategy would be too high, as firms would be better off by buying neither partition. The following proposition describes the DBs' equilibrium strategies when datasets are super-additive.

Proposition 2. *Datasets are super-additive iff $\gamma \geq \gamma^*$. If datasets are super-additive, in equilibrium, $\Delta_{i,k'}^* = \Delta_{k'}^*$, $w_{i,k}^* = w_k^*$ and $d_{i,k}^* = d_k^*$, $\forall i \in \{0, \ldots, n-1\}$. Both d_1^* and d_2^* are decreasing in γ, and $d_2^* \geq d_1^*$ $\forall \gamma \in [\gamma^*, 1]$. There exist γ^{**}, with $\gamma^* < \gamma^{**} < \alpha + \beta\alpha$ such that:*

(i) If $\gamma^* \leq \gamma \leq \gamma^{**}$ (sub-modular data), DBs set $d_2^* > d_1^*$ such that $\Delta_{12}^* = \Delta_1^* + \Delta_2^*$. There exist a unique equilibrium where $w_1^* = \Delta_1^*, w_2^* = \Delta_2^*$ and $w_1^* + w_2^* = \Delta_{12}^*$;

(ii) If $\gamma > \gamma^{**}$ (data can either be sub-modular or super-modular), DBs set $d_1^* = d_2^* = d^*$. Any pair (w_1^*, w_2^*) such that $w_k^* \geq \Delta_k, k \in \{1,2\}$ and $w_1^* + w_2^* = \Delta_{12}^*$ is a Nash equilibrium.

Proof. See Appendix A. ■

To better illustrate the results of Proposition 2, let us refer to Fig. 4, which shows the level of data additivity emerging in equilibrium, as a function of data modularity. If $\gamma < \gamma^*$, DBs act non-cooperatively and set prices equal to the additional value of their dataset, as described in Proposition 1. If $\gamma^* \leq \gamma \leq \gamma^{**}$ (point (i) of Proposition 2), datasets are still sub-modular. Nonetheless, the non-cooperative strategy described in Proposition 1 cannot be an equilibrium. Indeed, as datasets become super-additive, the non-cooperative strategy would lead to excessively high partition prices, and DBs cannot appropriate the additional value generated by the combined datasets. To get an intuition for this result, consider the following example. Suppose $\Delta_1^* = 4$, $\Delta_2^* = 3$ and $\Delta_{12}^* = 9$, which imply super-additive datasets as $\Delta_{12}^* > \Delta_1^* + \Delta_2^*$. If the DBs adopted the non-cooperative equilibrium strategy described for the sub-additive case, they would set prices equal to $w_1^* = \Delta_{12}^* - \Delta_2^* = 6$ and $w_2^* = \Delta_{12}^* - \Delta_1^* = 5$. This would imply that $\Delta_{12}^* < w_1^* + w_2^*$, which would result in firms not buying data. Thus, the DBs opt instead for a strategy that still induces firms to buy both partitions so that they can appropriate the positive synergies created by them. To this aim, DBs coordinate over two dimensions. First, they set the equilibrium partition sizes to make datasets super-additive, i.e., $\Delta_{12}^* = \Delta_1^* + \Delta_2^*$, despite the relatively weak synergies. This ensures that firms are (weakly) better off by buying from both DBs. Second, they set the equilibrium prices equal to the extra profits generated by the individual partitions, so that they can extract all the extra profits generated by the combination of the datasets (in our example, $w_1^* + w_2^* = \Delta_{12}^* = 9$).

Finally, if $\gamma > \gamma^{**}$ (point (ii) of Proposition 2), DBs do not coordinate the equilibrium partition sizes, as the value of the joint partitions is higher than the sum of the individual values of the partitions. In particular, to extract all of firms' willingness to pay, DBs must set partitions' prices such that the sum of the prices is equal to the value of the combined partitions. This could be achieved, for example, through algorithmic pricing, such that one DB immediately adjusts its price in response to a price change from the other DB. In this case, DBs can temper competition among themselves, and this makes them more effective in extracting profits from firms.

Interestingly, the number of firms in the downstream market affects the thresholds γ^* and γ^{**} identified in Proposition 2. Indeed, as we can see in Fig. 5, a higher number of firms implies that super-additivity can be achieved for a lower level of synergies. The intuition is that a higher number of firms entails fiercer competition downstream. Then, firms' willingness to pay for data increases as the threat of remaining uninformed is stronger.

A key result of Proposition 2 is that the coordinated sale between DBs takes place even when datasets are sub-modular. This marks a fundamental difference with existing results (see, e.g., Gu et al., 2022), and highlights that DBs, by properly choosing the size of

[23] Because of the prisoner's dilemma, firms in a dynamic setup might want to commit not to price discriminate to avoid the *competition effect* of data and achieve higher profits. We leave this issue to future research.

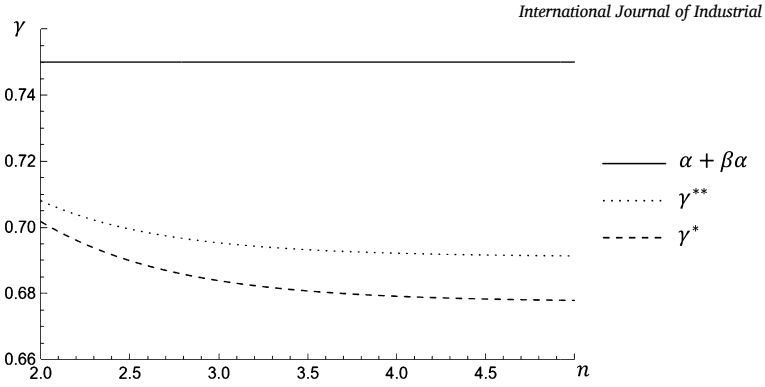

Fig. 5. Thresholds γ^* and γ^{**} and super-modularity threshold $\alpha+\beta\alpha$ as a function of n. $\alpha=0.5, \beta=0.5$.

the data partitions, can endogenously increase the value of the combined datasets above the value stemming from the technological synergies, lowering the synergy threshold after which they can coordinate.

6.2. Data accuracy and DBs' profits

The analysis conducted so far has highlighted how the DBs' equilibrium strategies depend on synergies between datasets, and how those synergies reverberate into the downstream market, affecting firm competition. Even though we consider data accuracies exogenous, in this Section we study how the levels of data accuracies α, β, and γ affect DBs' equilibrium profits. This allows us to provide an exploratory analysis on the incentives DB may have for data collection. To this aim, we focus on the sub-additive case, as in the super-additive one the coordination between DBs effectively allows them to avoid competition upstream.

Substituting the expressions of equilibrium data partitions (11) and (12) in the DBs' profit functions (8) and (9), respectively, we obtain:

$$\pi^*_{DB_1} = \frac{tH(n)^2}{2n}\frac{\gamma-\beta\alpha}{1+M(n)(\gamma-\beta\alpha)} \tag{15}$$

$$\pi^*_{DB_2} = \frac{tH(n)^2}{2n}\frac{(\gamma-\alpha)[1+M(n)(\gamma+\alpha-\beta\alpha)]+\beta\alpha M(n)^2(\gamma-\beta\alpha)^2}{[1+M(n)(\gamma-\beta\alpha)]^2(1+M(n)\beta\alpha)} \tag{16}$$

The following Proposition illustrates the effects of data accuracies on DBs' profits.

Proposition 3. *If data are sub-additive, Both DBs' profits are increasing in γ and decreasing in α. DB_1's profits are decreasing in β, whereas there exists a threshold $\bar{\bar{\beta}}$ such that, iff $\beta > \bar{\bar{\beta}}$, $\frac{d\pi^*_{DB_2}}{d\beta} < 0$.*

Proof. See Appendix A. ■

An increase in the level of synergies always increases both DBs' profits, because it increases firms' willingness to pay for the combined partitions. Moreover, an increase in α, all else equal, decreases DBs' profits. The intuition is that an increase in α increases the extra profits generated by the individual partitions Δ^*_1 and Δ^*_2, as they become more accurate. Given the same level of synergies, an increase in the value of individual datasets reduces firms' marginal value for the combined dataset and thus their willingness to pay for data.

Finally, Proposition 3 highlights the effect of β on DBs' profits. From (15), we find that DB_1's profits are always decreasing in β. The intuition is straightforward: an increase in β increases the individual value of DB_2's partitions, as they become more accurate. Thus, as long as the level of synergies remains constant, the additional value provided to firms by purchasing both datasets decreases, leading to lower DB_1's profits. The effect of β on DB_2's profit is more nuanced. On the one hand, an increase in β increases the probability of identifying consumers present only in DB_2's partition, which is larger than DB_1's one. On the other hand, an increase of β generates a strategic response from DB_1, which increases its equilibrium partition size (from (11)). This in turn intensifies the *competition effect* of data, reducing firms' profits and willingness to pay. If the vertical differentiation between the DBs is low, i.e., β is high enough, the competition effect dominates the effect of the improved accuracy, and an increase in β results in lower profits for DB_2.

The results of Proposition 3 suggest two interesting implications for DBs' incentives for data collection. First, it can be profitable for DB_2 to gather data attributes that are already available in DB_1's dataset, as DB_2 can still appropriate some value by selling information about consumers not present in DB_1's dataset. Second, DB_2 is better off when its accuracy is lower (but not excessively so) than DB_1's one. In other words, DB_2 would not want to match DB_1's higher accuracy, even if it would be costless to do so. Indeed, no vertical differentiation would intensify the competition between DBs to such an extent that it would erode any gain for

DB_2 for having a higher accuracy. This result is in line with that obtained by Belleflamme et al. (2020), who show that a monopolistic DB prefers selling partitions of different accuracies to two competing firms to temper competition between them.

Although modeling the data collection stage lies beyond the scope of this work, the results presented above, combined with those of Proposition 2, provide a basis for conjecturing about the potential implications of data collection costs. Indeed, using the results of Proposition 2, it is easy to show that DBs' profits always increase in γ, even in the super-additive case.[24] This observation highlights the presence of differing incentives for data collection depending on whether the collected data generate synergies or not. In particular, both DBs are incentivized to collect data that enhance dataset synergies, as the resulting improvement in accuracy enables them to extract greater profits from downstream firms. However, only DB_2 has an incentive to gather data that do not enhance synergies. Such data allow DB_2 to maintain a sufficient quality gap with DB_1, thereby softening competition in the DB market.

7. Extensions

7.1. Entry and welfare

In our baseline model, the number of downstream firms is exogenous. However, the data sale, by influencing the firms' profits, might influence their incentives to enter the market and, in turn, welfare. To explore the implications of DB competition on downstream entry and welfare, we assume that firms incur a fixed cost F if they enter the market. This cost can be interpreted as the cost of digitization, such as the investment needed to open an online retail shop. We assume sequential entry to avoid coordination problems and ignore integer constraints on n (Rhodes and Zhou, 2024). We thus add a Stage 0 to the timing described in Section 4.2, in which firms enter the market and pay the fixed cost F. As a benchmark we refer to the standard Salop (1979) model with endogenous entry, where entering firms make zero profits in equilibrium, resulting in $\widetilde{n} = \sqrt{\frac{t}{F}}$ and $\widetilde{CS} = \widetilde{TW} = v - \frac{5}{4}\sqrt{tF}$.

Firm entry in equilibrium is obtained from the free-entry condition, requiring that firms' profits, after paying for data and entry, are equal to zero. The following Proposition describes how the equilibrium level of entry is affected by the data sale.

Proposition 4. *The number n^* of entering firms in equilibrium is always lower than in the benchmark, i.e., $n^* < \widetilde{n}$ and it is decreasing in γ.*

Proof. See Appendix A. ■

In the limits for $\alpha = 0$ and $\gamma = 0$, data have no effect and our results converge to those obtained in the standard Salop model where data are absent. As $\alpha > 0$ and $\gamma > 0$, the sale of data by the DBs originates a *competition effect* so strong that it outweighs the *surplus extraction effect*, reducing firms' profits and, in turn, firm entry. Moreover, γ increases the value of the partitions offered by the DBs, leading to higher partition prices in equilibrium. Jointly taken, both the strong competition effect and the higher price of data reduce firms' profits and thus entry.

Proposition 4 has interesting implications for the super-additivity thresholds. Indeed, as an increase in the level of synergies γ results in lower entry, this in turn increases the thresholds γ^* and γ^{**}, as shown in Fig. 5. Then, the entry barrier effect lowers the scope for super-additivity.

The effect of the data sale on welfare is twofold. On the one hand, the *competition effect* induced by data leads to fiercer downstream competition and lower prices, benefiting consumers. On the other hand, the reduction in firm entry harms consumers. The surplus of consumers buying from firm i is defined as the integral of consumers' utility:

$$CS_i = \int_{\hat{x}_{i-1,i}}^{\hat{x}_{i,i+1}} U(x,i)dx. \tag{17}$$

Then, consumer surplus is $CS = \sum_i CS_i$. In particular, consumer surplus can be expressed as[25]

$$CS^* = u - \frac{5t}{4n} + \frac{nt}{2}((\gamma - \beta\alpha)d_1^* + \beta\alpha d_2^*). \tag{18}$$

The first two terms in (18) are the consumer surplus in the standard Salop (1979) model with an exogenous number of firms, while the third term is positive as $\gamma > \beta\alpha$. Thus, consumer surplus is higher in the presence of DBs if the number of firms is given. Let us define total welfare TW as the sum of consumer surplus CS, firm profits and the DBs' profits. We have

$$TW^* = CS^* + \sum_{i=0}^{n-1} \pi_i^* + \pi_{DB_1}^* + \pi_{DB_2}^*. \tag{19}$$

If the number of firms is given, total welfare is not affected by the DBs' presence, as it only transfers surplus from firms to consumers and DBs. However, as argued before, the reduction in firms' profits affects entry when the number of firms is endogenous. The following Proposition illustrates the conditions under which the overall effect is positive for consumer surplus and welfare.

[24] The level of profits in the superadditive case is also illustrated in Fig. 7 in the following section.

[25] Consumer surplus is computed in equilibrium, where all firms buy data from both DBs.

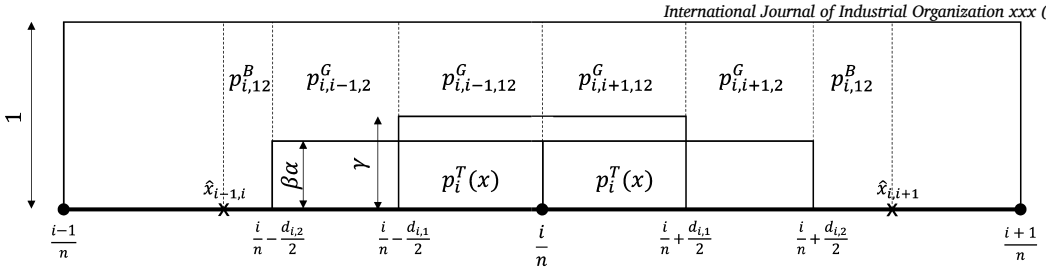

Fig. 6. Firm i's market share and prices when buying from both DB_1 and DB_2, assuming $d_{i,2} > d_{i,1}$.

Proposition 5. *In equilibrium, consumer surplus is decreasing in γ and increasing in β, and the opposite is true for total welfare. There exist two thresholds $\bar{\gamma} < \gamma^*$ and $\bar{\beta}$ such that $CS^* \geq \widetilde{CS}$ iff $\gamma \leq \bar{\gamma}$ and $\beta \geq \bar{\beta}$. For any level of α, β, and γ, $TW^* > \widetilde{TW}$.*

Proof. See Appendix A. ■

Proposition 5 highlights two novel results. First, the effect of the data sale on consumer surplus may be positive or negative, depending on the level of competition in the DB market and the level of synergies. Intuitively, an increase in β leads to lower data prices and higher entry, which in turn benefits consumers. This result is complementary to those obtained in recent studies on the effects of DBs on downstream competition. Abrardi et al. (2024) show that the presence of a monopolistic DB always leads to a decrease in consumer surplus due to the reduction of downstream entry. On the contrary, Bounie et al. (2021a) show that competition between DBs in the absence of entry dynamics always leads to higher consumer surplus. Our results bridge the gap between the aforementioned studies, by showing that the data sale has a more nuanced effect on consumer surplus. Depending on the level of DB competition, the use of data for price discrimination can either lead to consumer benefit or harm. However, if γ is high enough, the increase in the data partitions' prices stemming from the strong synergies always dominates the reduction of data partitions' prices given by the level of DB competition. Then, the higher partition prices reduce firm entry and always lead to consumer harm. Notably, the threshold after which consumers are always harmed is lower than the super-modularity threshold, implying that even sub-modular datasets can reduce consumer surplus.

Second, we find that total welfare always increases with respect to the standard Salop model. The data sale always reduces firms' profits, resulting in lower entry, which lowers the amount of profits dissipated in paying the entry cost F. In other words, the data sale partially solves the problem of excessive entry that is typical of the standard Salop model, leading to higher total welfare.

From a policy perspective, the key takeaway of Proposition 5 is that consumers can be harmed by the data sale even if data are sub-modular and DBs do not coordinate their sales. Indeed, relatively weak synergies are sufficient to reduce downstream entry and, in turn, consumer surplus. Moreover, when DBs do coordinate their data sales, the resulting monopolistic market outcome causes further consumer harm. Overall, by jointly considering the results of Propositions 2 and 5, we conclude that, by endogenizing downstream competition and thus the value of data, DBs coordinate their sale, causing consumer harm, even when data are sub-modular, provided that the overlap between datasets is sufficiently small.

7.2. Combining first- and third-degree price discrimination

In the baseline model, we assumed that the data partitions can either be perfectly informative with some probability, allowing firms to operate first-degree price discrimination, or be completely uninformative, inducing firms to adopt a uniform basic price. While this representation is useful to illustrate the key intuitions of our model, in practice the information provided by DBs might be more nuanced. For example, DBs might provide detailed information about some consumers, allowing firms to identify precisely their location, and only cursory information about others, allowing firms to identify their position over an interval but not pinpoint their exact location. We thus assume that DB_1 offers a partition d_i to firm i. With probability α, the firm is able to identify precisely the location of each consumer in d_i. Therefore, it operates first-degree price discrimination over consumers in d_i, while it charges a uniform basic price for consumers outside d_i. Conversely, with probability $1-\alpha$, the firm only learns which consumers are located in the segment d_i and which are not. In this case, it operates third-degree price discrimination, offering the basic price to consumers outside d_i, and a different uniform price to consumers in d_i.

This setup extends our baseline model as it allows us to consider simultaneously first- and third-degree price discrimination. In this respect, data are more informative than in our baseline setup, where data could be completely uninformative and firms had to adopt the basic price for all consumers.

We assume that third-degree price discrimination prices are set in Stage 4 of the game, simultaneously with first-degree price discrimination prices.

Fig. 6 provides a graphical representation of the model in this new setup. The firm can now offer two types of tailored prices to the consumers located inside the partitions. If their location is perfectly observed, they are offered the location-specific tailored prices $p^T_i(x)$ that perform first-degree price discrimination. Conversely, if firm i only learns which consumers belong to the data partition, it offers the basic price p^B_i to consumers outside the partition and uniform prices $p^G_{i,j,k}$ to the consumers inside the partition (the superscript G denotes the group price operating third-degree price discrimination), where $i \in \{0, \ldots, n-1\}$ is the firm charging the

price, $j \in \{i-1, i+1\}$ is firm i's direct rival on the arch where the price is offered, and $k \in \{1, 2, 12\}$ indicates whether the price is offered to consumers who are identified by either or both DBs' partitions.[26]

Similar to the location-specific tailored prices $p_i^T(x)$, firm i sets prices $p_{i,j,k}^G$ to match the direct rivals' basic prices in utility levels. In particular, they are set such that the consumer located farthest from firm i is indifferent between buying from firm i or firm j. This ensures that all other consumers to whom firm i offers $p_{i,j,k}^G$ prefer buying from firm i. For example, in the case illustrated in Fig. 6, firm i sets the price $p_{i,i+1,12}^G$ such that the consumer located in $\frac{i}{n} + \frac{d_{i,1}}{2}$ is indifferent between buying from firm i or firm $i+1$, leading to

$$p_{i,i+1,12}^G = p_{i+1}^B + \frac{t}{n}(2i+1) - 2t\left(\frac{i}{n} + \frac{d_{i,1}}{2}\right). \tag{20}$$

All other prices $p_{i,j,k}^G$ can be obtained analogously.

Assuming $d_{i,2} > d_{i,1}$, firm i's profits when buying from both DBs, prior to paying for data, thus become

$$\begin{aligned}\pi_{i,12}^G = \gamma\left(\int_{\frac{i}{n}-\frac{d_{i,1}}{2}}^{\frac{i}{n}+\frac{d_{i,1}}{2}} p_i^{\mathrm{T}}(x)\,dx\right) + (1-\gamma)\frac{d_{i,1}}{2}\left(p_{i,i+1,12}^{\mathrm{G}} + p_{i,i-1,12}^{\mathrm{G}}\right) + \\ + \beta\alpha\left(\int_{\frac{i}{n}-\frac{d_{i,2}}{2}}^{\frac{i}{n}-\frac{d_{i,1}}{2}} p_i^{\mathrm{T}}(x)\,dx + \int_{\frac{i}{n}+\frac{d_{i,1}}{2}}^{\frac{i}{n}+\frac{d_{i,2}}{2}} p_i^{\mathrm{T}}(x)\,dx\right) + (1-\beta\alpha)\left(\frac{d_{i,2}}{2} - \frac{d_{i,1}}{2}\right)\left(p_{i,i+1,2}^{\mathrm{G}} + p_{i,i-1,2}^{\mathrm{G}}\right) + \\ + p_{i,12}^{\mathrm{B}}\left(\hat{x}_{i,i+1} - \hat{x}_{i-1,i} - d_{i,2}\right),\end{aligned} \tag{21}$$

where the first line of Equation (21) refers to firm i's profits from consumers identified by both partitions, the second line refers to firm i's profits from consumers only identified by the largest partition, i.e., DB_2's partition in this example, and the third line refers to firm i's profits from unidentified consumers.

As the baseline model, we first find firms' equilibrium pricing strategies when buying or not buying data, and then identify DBs' equilibrium selling strategies and how they change depending on γ. The following Proposition shows the main results. Subscript G indicates equilibrium results in this extension of the model. We denote with γ_G^* and γ_G^{**} the thresholds levels of γ that characterize DBs' strategies in this setup.

Proposition 6. *If datasets allow both first-degree and third-degree price discrimination, in equilibrium, we have that $\gamma_G^{**} > \gamma^{**}$. Moreover: (i) $\gamma_G^{**} > \alpha + \beta\alpha$, and (ii) there exists a threshold $\tilde{\alpha}$ such that $\gamma_G^* < \alpha + \beta\alpha$ iff $\alpha \in (\tilde{\alpha}, 1)$.*

Proof. See Appendix A. ■

The key message of Proposition 6 is that the introduction of third-degree price discrimination makes super-additivity harder to achieve with respect to the baseline model. The main driving force behind the result of Proposition 6 is that data now allow identifying all consumers within the partition through first- or third-degree price discrimination. This implies that individual datasets are more valuable to firms, i.e. it is more difficult for DBs to induce firms to purchase both datasets. Synergies in this context constitute a weaker incentive for firms to purchase both datasets. This implies that the DBs' strategy of coordinating only on prices is now only feasible with super-modular datasets, i.e. $\gamma_G^{**} > \alpha + \beta\alpha$ (point (i) of Proposition 6). Instead, under specific conditions on α, super-additivity can still be achieved with sub-modular datasets (i.e., $\gamma_G^* < \alpha + \beta\alpha$, point (ii) of the Proposition), confirming our main result.

Interestingly, super-additivity can no longer be achieved with sub-modular datasets when α is low. Indeed, even if α is low, partitions provided by both DBs are still valuable to firms because they enable third-degree price discrimination. In this case, synergies provide a lower added value to firms. Then, firms are not willing to pay extra for the combined datasets if the latter are sub-modular.

To analyze the effects on entry, we again assume that entry entails a fixed cost F according to the timing presented in Section 7.1. The following Proposition describes how the data sale affects the equilibrium level of entry and welfare.

Proposition 7. *If datasets allow both first-degree and third-degree price discrimination, in equilibrium, $n_G^* < n^*$ and $TW_G^* > TW^*$.*

Proof. See Appendix A. ■

Even though super-additivity is harder to achieve in this framework, this does not imply that firms obtain higher profits. On the contrary, firms' profits are now lower, resulting in lower entry and higher welfare. Third-degree price discrimination lowers firms'

[26] For tractability, we assume that firm i can distinguish between consumers located on his left and on his right and offer them different prices $p_{i,j,k}^G$.

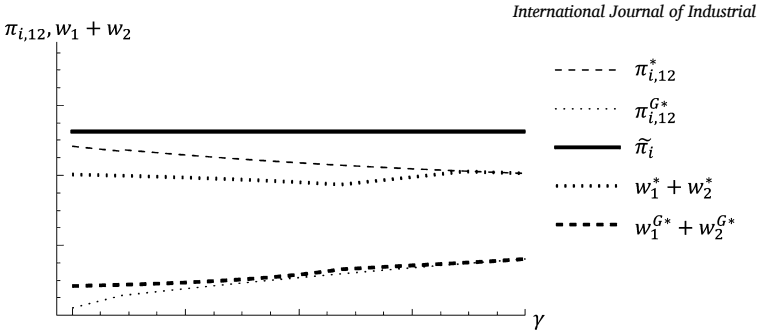

Fig. 7. Firms' profits prior to paying for data and sum of partition prices in the benchmark (standard Salop (1979)), baseline and extended model.

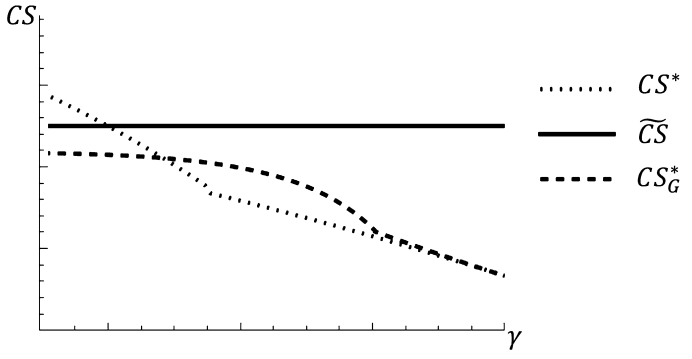

Fig. 8. Consumer surplus as a function of γ in the benchmark, baseline and extended model.

profits through two distinct channels. First, more informative data imply a stronger competition effect, leading firms to price more aggressively. Such an effect lowers firms' profits prior to paying for data, as we can see from Fig. 7. Second, the total value of data increases, as being uninformed and competing against informed firms constitutes a stronger disadvantage for the uninformed firms, as their rivals now also operate third-degree price discrimination. Thus, the total price of data increases with respect to the baseline model.

Due to the reduction of firm entry, lower profits are dissipated through the entry cost F, so welfare increases when datasets enable group pricing. In other words, the possibility to adopt group pricing reduces the inefficiency caused by the excessive entry that is typical of the Salop (1979). Nonetheless, the higher welfare might not benefit consumers. Indeed, the effect on consumer surplus is more nuanced, as we can see from Fig. 8. First, we find that consumer surplus is always lower than in the benchmark model because of the lower entry, regardless of γ. However, consumer surplus can be higher than in the baseline model for high enough values of γ. In fact, when data also allow third-degree price discrimination, strong data synergies lead DBs to increase the equilibrium partitions' sizes, which results in firms pricing more aggressively, benefiting consumers. Although this effect does not overcome the consumer harm caused by the reduction in entry, it weakens the negative effect of γ on consumer surplus.

8. Conclusions

With the growing centrality of consumer data in the digital economy, DBs have become key enablers of data-driven technologies. Their ability to transform data into valuable information allows them to influence downstream competition, with significant welfare implications. This work contributes to the expanding literature on the economic effects of DBs, by analyzing how DBs' competition and data synergies affect their strategies and, in turn, economic outcomes.

Our results highlight a novel channel through which DBs can temper competition among themselves, even if data synergies are weak. By strategically choosing the price and size of the partitions they sell to downstream firms, DBs can endogenously increase the value of data, enabling them to coordinate their sale and extract more profits from downstream firms. Thus, DBs can act strategically and achieve super-additivity in the value of data even when relatively weak synergies make datasets sub-modular, i.e., the predictive power of the combined datasets is lower than the sum of the predictive powers of individual datasets. We also find that, if downstream entry is taken into account, synergies between competing DBs reduce the number of downstream firms. This may cause consumer harm if competition between DBs is not strong enough. Moreover, if DBs can certify which consumers belong to the dataset and which not, firms can set different uniform prices inside and outside the data partition, in addition to tailored prices. In this situation, super-additivity becomes harder to achieve, although it can still be obtained with sub-modular datasets. The higher informative power of data intensifies downstream firms' competition, leading to a further reduction in entry.

From a policy perspective, our results imply that ensuring a level playing field in the DB market, i.e., decreasing the level of vertical differentiation in the DB market can have a positive outcome for consumers. Such a level playing field must be not only in terms

of similar data accuracy. Indeed, DBs' market power stems from having exclusive data on consumers, which in turn increases the synergies' strength if datasets are combined. This can be particularly harmful when the data synergies allow them to coordinate their sales, as DBs can then extract all available surplus from firms, in turn increasing the downstream market concentration. Although DBs are currently outside the scope of the recent data regulations, our results shed some light on DBs' huge potential of steering competition and access to data. Indeed, several Antitrust authorities are asking for higher transparency and awareness of how DBs operate, both upstream and downstream. In the upstream market, further investigation, both theoretical and empirical, is warranted about their data collection practices and the black box of their algorithms, exploring the endogenous nature of data quality. Indeed, our exploratory analysis on the DBs' incentives to collect data has highlighted opposing forces that may lead to counterintuitive outcomes. In the downstream market, another Antitrust issue might concern the adoption of data-sharing by firms to coordinate their strategies and thus mitigate the competition effect. We leave these issues for future research.

CRediT authorship contribution statement

Laura Abrardi: Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Carlo Cambini:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Flavio Pino:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Appendix. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijindorg.2025.103146.

Data availability

No data was used for the research described in the article.

References

Abrardi, L., Cambini, C., Congiu, R., Pino, F., 2024. User data and endogenous entry in online markets. J. Ind. Econ. 72, 1052–1088.
ACCC, 2023. Digital platform services inquiry–March 2024 report on data brokers: issues paper. Technical report. Australian Competition and Consumer Commission.
Aparicio, D., Metzman, Z., Rigobon, R., 2021. The pricing strategies of online grocery retailers. NBER working paper n. 28639.
Arrow, K.J., 1972. Economic Welfare and the Allocation of Resources for Invention. Springer.
Baik, Simon Anderson A., Larson, N., 2022. Price discrimination in the information age: prices, poaching, and privacy with personalized targeted discounts. Rev. Econ. Stud. 90 (5), 2085–2115.
Bajari, P., Chernozhukov, V., Hortaçsu, A., Suzuki, J., 2019. The impact of big data on firm performance: an empirical investigation. In: AEA Papers and Proceedings, vol. 109. American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 33–37.
Belleflamme, P., Lam, W.M.W., Vergote, W., 2020. Competitive imperfect price discrimination and market power. Mark. Sci. 39 (5), 996–1015.
Bergemann, D., Bonatti, A., 2011. Targeting in advertising markets: implications for offline versus online media. Rand J. Econ. 42 (3), 417–443.
Bergemann, D., Bonatti, A., 2019. Markets for information: an introduction. Annu. Rev. Econ. 11, 85–107.
Bergemann, D., Bonatti, A., Smolin, A., 2018. The design and price of information. Am. Econ. Rev. 108 (1), 1–48.
Bhargava, H.K., Dubus, A., Ronayne, D., Shekhar, S., 2024. The strategic value of data sharing in interdependent markets.
Biegel, B., Margulies, J., Maggi, G., Davis, C., 2018. The state of data 2018. Technical report. Winterberry Group.
Bounie, D., Dubus, A., Waelbroeck, P., 2021a. Competition and mergers with strategic data intermediaries. Available at SSRN.
Bounie, D., Dubus, A., Waelbroeck, P., 2021b. Selling strategic information in digital competitive markets. Rand J. Econ. 52 (2), 283–313.
Bounie, D., Dubus, A., Waelbroeck, P., 2022. Collecting and selling consumer information: selling mechanisms matter. Available at SSRN.
Braulin, F.C., 2023. The effects of personal information on competition: consumer privacy and partial price discrimination. Int. J. Ind. Organ. 87, 102923.
CEA, 2015. Big data and differential pricing. Technical report. Council of Economic Advisers.
Chen, Z., Choe, C., Matsushima, N., 2020. Competitive personalized pricing. Manag. Sci. 66 (9), 4003–4023.
Choe, C., Cong, J., Wang, C., 2023. Softening competition through unilateral sharing of customer data. Manag. Sci.
Christl, W., 2017. Corporate surveillance in everyday life. Technical report. Cracked Labs.
Crain, M., 2018. The limits of transparency: data brokers and commodification. New Media Soc. 20 (1), 88–104.
Delbono, F., Reggiani, C., Sandrini, L., 2024. Strategic data sales with partial segment profiling. Inf. Econ. Policy 68, 101102.
European Commission, 2024. Second report on the application of the general data protection regulation. Technical report. European Commission.
FTC, 2014. Data brokers: a call for transparency and accountability. Technical report. Federal Trade Commission, Washington, DC.
FTC, 2025. Ftc surveillance pricing 6(b) study: Research summaries a staff perspective. Technical report. Federal Trade Commission, Washington, DC.
Goldfarb, A., Tucker, C., 2019. Digital economics. J. Econ. Lit. 57 (1), 3–43.
Gu, Y., Madio, L., Reggiani, C., 2022. Data brokers co-opetition. Oxf. Econ. Pap. 74 (3), 820–839.
Haberer, B., Krämer, J., Schnurr, D., 2022. Do consumers benefit from selling their data? The economic effects of personal data brokers in digital markets. In: The Economic Effects of Personal Data Brokers in Digital Markets (March 9, 2022). TPRC, p. 46.
Hagiu, A., Wright, J., 2023. Data-enabled learning, network effects and competitive advantage. Rand J. Econ. 54, 638–667.
Hocuk, S., Martens, B., Prufer, P., Carballa Smichowski, B., Duch-Brown, N., 2022. Economies of scope in data aggregation: Evidence from health data. TILEC Discussion Paper no 020.
Ichihashi, S., 2021. Competing data intermediaries. Rand J. Econ. 52 (3), 515–537.
Iyer, G., Soberman, D., Villas-Boas, J.M., 2005. The targeting of advertising. Mark. Sci. 24 (3), 461–476.
Kirchner, L., 2020. Can algorithms violate fair housing laws? Markup.
Klein, T., Kurmangaliyeva, M., Prüfer, J., Prüfer, P., 2022. How important are user-generated data for search result quality?: Experimental evidence. TILEC Discussion Paper no 016.

Krämer, J., Colangelo, G., Richter, H., Schnurr, D., 2023. Data act: Towards a balanced EU data regulation. Technical report. Centre on Regulation in Europe.
Lee, G., Wright, J., 2023. Recommender systems and the value of user data. National University of Singapore Working Paper.
Leppälä, S., 2013. Arrow's paradox and markets for nonproprietary information. Technical report, Cardiff Economics Working Papers.
Leppälä, S., 2015. Economic analysis of knowledge: the history of thought and the central themes. J. Econ. Surv. 29 (2), 263–286.
Liu, Q., Serfes, K., 2004. Quality of information and oligopolistic price discrimination. J. Econ. Manag. Strategy 13 (4), 671–702.
Mishra, S., 2021. The dark industry of data brokers: need for regulation? Int. J. Law Inf. Technol. 29 (4), 395–410.
Montes, R., Sand-Zantman, W., Valletti, T., 2019. The value of personal information in online markets with endogenous privacy. Manag. Sci. 65 (3), 1342–1362.
Neumann, N., Tucker, C.E., Whitfield, T., 2019. Frontiers: how effective is third-party consumer profiling? Evidence from field studies. Mark. Sci. 38 (6), 918–926.
Ohm, P., 2010. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. 57, 1701–1777.
Rhodes, A., Zhou, J., 2024. Personalized pricing and competition. Am. Econ. Rev. 114 (7), 2141–2170.
Ruschemeier, H., 2022. Data brokers and European digital legislation. Eur. Data Prot. Law Rev. 9 (2023), 27–38.
Salop, S.C., 1979. Monopolistic competition with outside goods. Bell J. Econ., 141–156.
Shaffer, G., Zhang, Z.J., 1995. Competitive coupon targeting. Mark. Sci. 14 (4), 395–416.
Taylor, C., Wagman, L., 2014. Consumer privacy in oligopolistic markets: winners, losers, and welfare. Int. J. Ind. Organ. 34, 80–84.
Thisse, J.-F., Vives, X., 1988. On the strategic choice of spatial price policy. Am. Econ. Rev., 122–137.
Traub, A., McElwee, S., 2016. Bad Credit Shouldn't Block Employment: How to Make State Bans on Employment Credit Checks More Effective. Demos, Washington, DC.
Tuttle, B., 2013. Flight prices to get personal? airfares could vary depending on who is traveling. Time, March 5th, 2013.
Vickrey, W.S., 1964. Microstatics. Brace & World, Harcourt.
Villas-Boas, J.M., 2004. Consumer learning, brand loyalty, and competition. Mark. Sci. 23 (1), 134–145.
Yu, P., Zhang, J., 2024. Signaling targeting cost through list price. Manag. Sci.