Politecnico di Toriino

PhD programme in Management, Production and Design

Essays on Regulation and Competition in the Digital Markets

Raffaele Congiu

2022

Politecnico di Torino

Department of Management and Production Engineering

PhD programme in Management, Production and Design

XXXIV Cycle

S.S.D.: SECS-P/06

Essays on Regulation and Competition in the Digital Markets

Supervisor:

Carlo Cambini

Candidate:

Raffaele Congiu

Doctoral Examination Committee:

Prof. Marc Bourreau, Referee, Télécom Paris

Prof. Tommaso Duso, Referee, DIW Berlin

Prof. Luigi Benfratello, Politecnico di Torino

Prof. Federico Boffa, Libera Università di Bolzano

Prof. Laura Rondi, Politecnico di Torino

Coordinator: Arianna Alfieri

November 2018 - January 2022

Acknowledgements

I am indebted to my supervisor Carlo Cambini for his support and guidance over the years. His mentorship has shaped my approach to research.

I am thankful to the referees, Marc Bourreau and Tommaso Duso, for their precious insights and comments.

I thank the Economics Department of Télécom Paris for welcoming me during my visiting period.

This research has benefited greatly from the feedback that I received from professors and colleagues from the DIGEP department and from the participants at conferences and seminars I have presented at during my doctoral studies. I give special thanks to all co-authors of this dissertation: Laura Abrardi, Carlo Cambini and Flavio Pino for the first chapter; Lorien Sabatino and Geza Sapi for the second; Flavio Pino and Laura Rondi for the third.

Finally, I thank my friends and my family for their unyielding support.

Contents

List of Tables

Chapter 2

Chapter 3

List of Figures

Chapter 1

Chapter 2

Chapter 3

Declaration of Co-Authorship

The first chapter of this thesis is a joint work with Laura Abrardi, Carlo Cambini and Flavio Pino (Politecnico di Torino).

The second chapter is a joint work with Lorien Sabatino (Politecnico di Torino) and Geza Sapi (Düsseldorf Institute for Competition Economics (DICE), Heinrich Heine University of Düsseldorf).

The third chapter is a joint work with Flavio Pino and Laura Rondi (Politecnico di Torino).

Introduction

This thesis comprises three essays – a theoretical and two empirical ones – which study the dynamics of digital markets, focusing on regulation and competition. Each chapter analyses a single aspect of this multi-faceted topic: consumer data and competition in the first chapter; the impact of privacy regulation on website traffic and visitor behaviour in the second; how digital markets can affect traditional ones by looking at Airbnb's impact on the housing market in the third.

The first chapter investigates how a Data Broker (DB) can influence firm entry and downstream competition in oligopolistic markets by deciding how much data to sell and to whom. The effect of data on competition is the focus of two main strands of literature. The first one studies the impact of data exogenously available to firms. A more recent strand endogenises the information acquisition process, studying a monopolistic DB who sells data to a downstream duopoly. While the first strand highlights a pro-competitive effect of data, the second nuances it by showing that the DB can limit consumer surplus gains by reducing competition through exclusive data sales. This work contributes to the literature in two ways. First, by modelling an oligopoly market to analyse how the number of competing firms influences the DB's strategy and the market outcomes. Second, by endogenising firm entry.

The model consists of a circular city with entry costs à la Salop, where firms can enter the market and then acquire consumer data from a DB. Data are used by firms to identify consumers for price discrimination. The DB has information on all consumers, and she decides to which firms she sells data, the quantity sold to each one, and the price. The results show that the DB has the incentive to limit firm entry in the downstream market, as she benefits from the increased market concentration by extracting firms' profits through the price of data. Moreover, the DB has the incentive to under-serve the market by selling data to a subset of firms, so as to maximise their willingness to pay. Overall, both these effects lead to a reduction in downstream competition. The analysis shows that this reduction outweighs the pro-competitive effect of data highlighted by the previous literature. Consequently, consumer surplus is always lower in the presence of a monopolistic DB. These results are robust to the introduction of a privacy cost and to the reduction of the DB's bargaining power. Moreover, when taking into account the consumers' loss of privacy, the entry barrier effect is mitigated since data become less valuable to firms. As such, raising consumers' privacy awareness can be an effective lever to reduce the consumer harm induced by the DB.

The second chapter assesses the impact on website traffic and visitor behaviour of the introduction of the European Union's General Data Protection Regulation (GDPR). From its adoption, the GDPR has attracted considerable interest from researchers, policymakers and industry players across the globe, spurring a fast-growing literature that studies its impact on digital markets. So far, however, little attention has been devoted to understanding the legislation's effect on website's ability to attract visitors and on the way those users engage with website content. This work aims to fill this gap.

The analysis exploits the fact that the GDPR applies to EU residents – leaving the non-EU audience unaffected – to perform a difference-in-differences that relies on the geographic origin of website traffic. In particular, the treatment assignment identifies the traffic originated from EU countries, using US traffic as control group. Traffic data from about $5,000$ web domains in Europe and the US is used. The analysis documents an overall traffic reduction of approximately 15% in the long-run, and it finds a measurable reduction of user engagement with websites. These effects unfold fully with a delay, several months after the date of GDPR entry into force, following the issuance of the first large fine by the French Authority on Google. The overall traffic reduction is broken down into detailed acquisition channels. Traffic from direct visits, organic search, email, social media, display advertising and referrals dropped significantly, but paid search traffic – mainly Google search advertisement – was barely affected. The work finds evidence of an inverted U-shaped relationship between website size and traffic reduction due to the privacy regulation: the smallest and largest websites lost visitors, while medium ones were not affected or even gained from it. The results appear consistent with the view that users care about privacy and may avoid visiting a website in response to its data handling policy. The results also highlight how privacy regulation can impact market structure and may increase dependence on large advertising service providers. Enforcement matters as well: the effects were amplified considerably in the long-run, following the first significant fine issued eight months after full entry into force of the legislation.

The third chapter studies the impact of Airbnb's diffusion on house prices and rents in the Italian cities of Florence, Milan, Naples, Rome, and Turin. Airbnb's claim is to provide hosts with an additional source of income from unused capacity. Conversely, critics argue that – through the platform – landlords substitute from the long- to the short-term rental market, increasing rents and house prices. A growing number of empirical studies has recently started to inquire into the platform's impact on the housing market, with a recent strand of literature investigating the distributional effects of Airbnb's heterogeneous presence within cities. This work adds to both strands by investigating the impact of the platform at different levels – overall, across cities, and in the centre and suburbs – and by estimating the spillover effects of Airbnb presence in the city centre on the rents and house prices in the periphery.

The empirical strategy accounts for endogeneity and simultaneity problems. The analysis exploits an instrumental variable obtained from the interaction of an out-of-sample measure of tourist attraction that varies within cities (derived from Tripadvisor), and a measure of public awareness of Airbnb that varies over time (derived from Google searches). The analysis documents an increase in rents and, especially, in sale prices due to Airbnb's diffusion. Overall, an increase of 1 percentage point in Airbnb density rises house prices by 0.63%, translating to a 44.24 €/m^2 rise over the period of the analysis. However, the effect varies greatly across and within cities. Across cities, sale prices increase everywhere, from 162.31 €/m^2 in Milan to 19.37 €/m^2 in Rome. Rents are significantly affected in Florence and Naples, with effects that are sizeable when compared to price variations during the period of the analysis. The within-city effect is extremely heterogeneous, with some cities where it interests centre and suburbs and others where only the centre is affected. Whether the effect increases or reduces the gap between them changes on a by city basis, depending on the initial conditions of the two areas. Finally, the work finds evidence that the increase in Airbnb density in central areas has a negative effect on the property values in the suburbs. This is possibly due to the centre's increasing attractiveness at the expense of the suburbs following its increase in localised amenities. The results speak of an overarching effect, but also of differentiated impacts which require context-specific policies and evaluations.

CHAPTER 1

User Data and Endogenous Entry in Online Markets[*]

Laura Abrardi, Carlo Cambini, Raffaele Congiu, Flavio Pino[†]

This work investigates how the presence of a Data Broker (DB), who sells consumer information to downstream firms, affects firm entry and consumer surplus in an oligopoly market with horizontally differentiated goods, in which data allow firms to price discriminate. We show that the DB reduces firm entry by choosing the price and quantity of data and by selling data only to a subset of the entering firms. By doing so, the DB maximises firms' willingness to pay for data. Overall, the presence of the DB reduces both downstream competition and consumer surplus. Our results are robust to the introduction of a privacy cost and to alternative selling mechanisms entailing different degrees of DB's bargaining power.

[*] We thank Paul Belleflamme, Marc Bourreau, Antoine Dubus, Tommaso Duso, Johannes Johnen, Elisabetta Iossa, Tommaso Valletti, as well as participants to the 2022 Lidam seminar (Louvain) for useful comments.

[†] Politecnico di Torino, Department of Management, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy.

1. Introduction

Invisible hands move the market of data, but they might not be those of competition. Data Brokers (DBs) track consumers online, hoard massive amount of information and sell that intelligence in the form of targeted market segments based on the customer's needs. Though consumers can benefit from firms' targeted commercial offers, DBs might also have the power to affect market entry and steer competition simply by choosing to which firms (and to what extent) data are sold. This paper analyses a market where a DB sells consumer information to a number of horizontally differentiated downstream firms, which can use data for price discrimination. We highlight how the DB, by choosing the firms to which data are sold, and the price and quantity of data sold, can affect firm entry, firm profits and consumer surplus.

The advent of the digital economy has made personal data widely available. Once aggregated and processed, these data can be used to perform market research, customer base segmentation and targeted advertising. First degree price discrimination, once only a theoretical possibility, has become a reality.[1] However, collecting and processing data at a scale that makes it valuable requires unique resources and capabilities. The demand for such abilities has determined the growth of the DB sector, a highly concentrated industry whose revenue is estimated at USD 200 billion (FTC, 2014; Crain, 2018). DBs' business model compounds both online and offline sources, collecting data from commercial, government, and other publicly available sources – e.g., blogs, social media. Since they typically do not get their data directly from consumers, DBs are often away from the media's spotlight or people's awareness: yet, DBs are building intricate profiles with thousands of records on almost every household (FTC, 2014). Working in the background, DBs mostly engage in business-to-business relations, selling the processed information to downstream firms who want to reach specific consumers with targeted offers.

Given the huge potential to influence downstream competition, policymakers have often expressed concerns regarding the reach and the lack of transparency of this highly concentrated, and yet virtually unregulated industry. Recent literature (see, e.g., Montes et al., 2019) has pointed out how DBs have the incentive to increase some firms' market power by selling data selectively in downstream duopolistic markets. However, little is

[1] Mikians et al. (2012) show that individual consumer data such as geolocalization are used by firms to price discriminate them, with price differences of up to 166%. Similarly, Aparicio et al. (2021) show that the algorithms used by the leading online grocers in the U.S. personalise prices at the delivery zipcode level.

known on the strategies used by DBs when they serve markets populated by more than two competing firms, and how these strategies influence market entry and competition, firms' profits and consumer surplus.

The aim of this paper is to understand how a DB can influence firm entry and downstream competition in oligopolistic markets by deciding to whom and how much data to sell. We consider a circular city model with entry costs à la Salop (1979), where firms can enter the market and then acquire consumer data from a DB. Data are used by firms to identify consumers for price discrimination. The DB has information on all consumers, and she decides to which firms she sells data – making them informed – how much data she sells to each one (e.g., the full dataset or only a partition of it), and the price of data.

We find that the DB' optimal strategy entails a reduction of firm entry in the downstream market, relative to the benchmark case in which data are not available or are provided exogenously to the firms (as in Taylor and Wagman, 2014). Intuitively, a higher level of market concentration increases the overall profits of the market, which the DB can then extract through a higher price of data. In addition to this *entry barrier effect*, we also find that the DB influences the downstream market structure by selectively selling data to a subset of the entered firms. The possibility to compete having information that is precluded to rivals increases the firms' willingness to pay for data, and therefore the DB's profits. Overall, the DB lowers competition in the downstream market both by reducing entry, and by providing a competitive advantage to some of the firms. The reduction in competition ultimately harms consumers, who are all worse off when compared to a setting where data are not available.

We extend the basic model in three ways. First, we introduce a privacy cost for consumers when they receive a tailored offer (e.g., the annoyance of being contacted by somebody they have not disclosed their data to). When we take into account the consumers' loss of privacy, we find that the entry barrier effect is mitigated, since data become less valuable for firms. As such, raising consumers' awareness about privacy can be an effective lever to reduce the consumer harm induced by the DB.

Second, we show that the entry barrier effect is robust to alternative selling mechanisms adopted by the DB, namely the auction mechanism with or without reserve prices (see, e.g., Bounie et al., 2021) and Take It Or Leave It offers (as in Bergemann and Bonatti, 2019).[2] However, differently from the auction mechanism, under Take It Or

[2] The use of direct sales when selling data has been documented by the United States subcommittee on antitrust (Judiciary Committee, 2020).

Leave It offers the DB prefers to sell data to all entering firms. Interestingly, the DB's equilibrium strategy depends on the value of transportation costs relative to the entry cost. When transportation cost are low relative to the entry fixed cost, fewer firms enters the market, and the DB sells the whole dataset. Conversely, when transportation cost are high relative to the entry fixed cost, the market is less concentrated and the DB sells non-overlapping data partitions. Take It Or Leave It offers entail the highest consumer surplus among the analysed selling mechanisms.

Finally, we explore the possibility that the data sale occurs prior to firms' entry. This is the case, for instance, of emerging digital markets, where potential entrants anticipate the value of obtaining consumer data and thus make their entry decision after having obtained (or not obtained) data. Under this alternative timing, we find that the DB always maximises her entry barrier effect, regardless of the selling mechanism. This strategy allows her to increase concentration in the downstream market, leading to higher profits and higher consumer harm compared to the basic model.

The literature studying the impact of data on competition is growing. Firms can use consumer data to identify naive consumers (Johnen, 2020), or to distinguish between consumer groups with different price sensitivities (Colombo, 2018). de Cornière and Taylor (2020) provide a general framework in which data are a revenue-shifter, for a given level of consumers' utility. This framework usefully finds a wide range of applications in which data increase the quality of the information, but is ill suited for price discrimination in spatial competition settings where data provide information on the type of consumers (Armstrong and Vickers, 2001). When firms exogenously have data, the literature highlights a pro competitive effect, both under monopoly (Belleflamme and Vergote, 2016) and under competition.[3] As informed firms compete more fiercely, consumers benefit from lower prices. Although Taylor and Wagman (2014) show that the pro-competitive effect of data limits firm entry, this is due to the erosion of profits stemming from the intense competition, and not from the intervention of a DB who maximises the downstream surplus. A more recent strand of literature has endogenised the information acquisition process, either through firms' repeated interactions with consumers (Villas-Boas, 2004; Acquisti and Varian, 2005; Liu and Serfes, 2004; Bergemann and Bonatti, 2011; Hagiu and Wright, 2020) or by acquiring data from strategic actors (Bergemann and Bonatti,

[3] See for instance Thisse and Vives (1988), Shaffer and Zhang (1995), Bester and Petrakis (1996), C. R. Taylor (2003), Liu and Serfes (2004), Taylor and Wagman (2014), Shy and Stenbacka (2016) and Chen et al. (2020).

2015; de Cornière, 2016; Gu et al., 2019). In particular, Braulin and Valletti (2016), Montes et al. (2019) and Bounie et al. (2021) consider a monopolistic DB who sells data to a downstream duopoly through a series of auctions with negative externalities, as in Jehiel and Moldovanu (2000). These studies highlight how a DB can limit competition between two existing firms by selling data exclusively to one of them, thus extracting higher industry profits at the expense of consumer surplus. However, when three firms are present, Delbono et al. (2021) find that the DB always sells data to two or more firms – depending on the selling mechanism – and thus exclusive sales are never part of the equilibrium. A parallel stream of literature studies the role of competition between DBs on data collection. In particular, Ichihashi (2021), by studying a market with many data intermediaries and one downstream firm, shows that the non-rivalrous nature of data can lead to significant concentration in data markets.

We contribute to the existing literature in two ways. First, we extend the duopolistic setup to analyse how the number of competing firms in an oligopoly market influences the DB's strategy and the subsequent market outcomes. Second, we endogenise the number of firms present in the market by modelling their entry. This allows to highlight a novel effect of data, which we label as *entry barrier effect*, which emerges as a result of the DB's profit-maximising strategy. Our analysis shows that the reduction in competition given by the DB's entry barrier effect outweighs the pro-competitive effect of data, so that consumer surplus is ultimately reduced. To our knowledge, this is the first paper to highlight the entry barrier effect of the DB's behaviour and its potential anticompetitive nature.

From a policy perspective, a critical concern pertains to the concentration of the DBs' market and its effects on consumers. A key insight of previous literature on monopolistic DBs is that antitrust authorities should ban exclusive data deals to foster competition and protect consumers when the downstream market is a duopoly. However, our results suggest that in markets with more than two firms, the harm to competition stems from the entry barrier raised by a monopolistic DB. The negative effects of the entry barrier on consumers can be reduced by enforcing data sharing obligations with all firms or by intervening on the selling mechanism adopted by the DB. We find that the DB sells data to all firms if Take It Or Leave It offers (TIOLI) are used for the data sale, so that the TIOLI mechanism would be better for consumers than sales with auctions, especially in markets with high transportation cost relative to the entry fixed cost. While both these measures would be an effective tool to raise competition in the market, they might also involve a higher loss of consumer privacy.

The remainder of the paper is organised as follows. Section 2 presents the model, and Section 3 computes firms' equilibrium prices. Section 4 computes the DB's profits and her optimal strategy and discusses the consequent market outcomes. Section 5 analyses three model's extensions: introducing a privacy cost, reducing the DB's bargaining power, and allowing her to commit to data prices prior to firm entry. Section 6 concludes by discussing our results. All proofs are contained in the Appendix.

2. The Model

We consider a market where horizontally differentiated firms sell a product to a mass of consumers, whose preferences can be observed by a firm only if it purchases customer-specific data from a Data Broker (DB). For example, firms sell their products via e-commerce solutions, and the possibility of identifying the consumer through data acquired from a DB allows the firm to make personalised offers.

2.1. Consumers, Firms and the Data Broker

We consider a free-entry game with a market represented by a circular city of length 1 (Vickrey, 1964; Salop, 1979). Consumers are uniformly distributed on the circumference and normalised to 1, and their locations are indexed by $x \in [0,1)$ in counter-clockwise order. Let us denote with n the number of symmetric firms that enter the market, indexed by $i \in \{0, 1, 2, \dots, n-2, n-1\}$.[4] Their marginal cost of production is normalised to 0, while their entry in the market entails a cost F. We can think of F as the total costs incurred in the process of digitisation (see Anderson and Bedre-Defolie, 2021), such as the creation of an online retail shop. We assume that firms enter the market choosing equally spaced locations, so that the location of a generic firm i is indexed by $\frac{i}{n}$. Once firms enter the market, each consumer buys at most one unit of the product.

There is one Data Broker (DB) who has a dataset with the location of all consumers in the market. The DB can sell this information to firms that entered the market, allowing them to perform first-degree price discrimination on the identified consumers. The DB offers to each firm a data partition by setting up n auctions. Let us denote with $d_i \in [0,1]$

[4] As standard in the literature on markets with entry, we assume sequential entry to avoid coordination problems and ignore integer constraints on n. A similar approach has been recently adopted in Rhodes and Zhou (2021).

the data partition offered to firm i.[5] A partition d_i allows a firm to price discriminate on an arch of size d_i that contains firm i's location.[6] The partition set containing all partitions offered by the DB is $\mathbf{P} = (d_0, d_1, d_2, \ldots, d_{n-1})$. Once the auctions are concluded, we refer to the partition set containing all partitions sold in equilibrium as $\mathbf{P}^*$.

If a firm obtains a partition, it can offer location-specific tailored prices $p_i^{\mathrm{T}}(x)$ to the identified consumers and a basic price p_i^{B} to the others. Note that the number of consumers the firm serves through tailored prices depends on the amount of data it obtains.

2.2. Payoffs and Timing

When buying from firm i, a consumer located in x derives a net utility equal to:

$$U(x,i) = v - p_i^{\mathrm{T}}(x) - t * D(x,i)$$

if firm i has data on that consumer, or

$$U(x,i) = v - p_i^{\mathrm{B}} - t * D(x,i)$$

if it does not, where v is the gross utility, $p_i^{\mathrm{T}}(x) \geq 0$ is the tailored price of the product set by firm i to the identified consumer in position x, $p_i^{\mathrm{B}} \geq 0$ is the basic price set by firm i for the unidentified consumers, $t > 0$ is the transportation cost and $D(x,i)$ is the shortest arch between the consumer and firm i. A consumer in x buys from the firm i that maximises her utility $U(x,i)$. We assume that the market is fully covered: i.e., the gross utility is high enough that all consumers make a purchase. The location of an indifferent consumer between firms i and $i+1$ is $\widehat{x}_{i,i+1}$, i.e., $U(\widehat{x}_{i,i+1}, i) = U(\widehat{x}_{i,i+1}, i+1)$. A firm's profits can thus be defined as the integral of its prices over its market segment. Given that a firm offers a constant basic price to unidentified consumers, we can write its profits prior to paying for data as

$$\pi_i = \int_{\frac{i}{n}-\frac{d_i}{2}}^{\frac{i}{n}+\frac{d_i}{2}} p_i^{\mathrm{T}}(x)\, dx + p_i^{\mathrm{B}} \left(\widehat{x}_{i,i+1} - \widehat{x}_{i-1,i} - d_i\right)$$

where the first term on the right-hand side represents firm's profits over the identified consumers, while the second term represents its profits over the unidentified consumers. From this general expression we can see how the amount of data d_i influences firm i's

[5] By assuming public DB's offers, we rule out situations like secret contracting games as in Hart and Tirole (1988). In our model, firms are ex-ante identical and the DB's decision to sell data to any specific firm does not depend on the firm's identity.

[6] Through price discrimination, firms can extract more surplus from consumers who are close to their location. Since the DB's profits are directly proportional to firms' profits (gross of the price paid for data), her best strategy requires selling partitions containing firms' locations (see Bounie et al., 2021).

strategy. First, d_i determines the number of consumers the firm can offer a tailored price to. Second, d_i also influences the profits firm i makes from unidentified consumers, as a higher amount of data implies a smaller share of unidentified consumers, and thus less profits extracted through firm i's basic price. Finally, firm i's basic price, and in turn its profits, are influenced by its rivals' basic prices, which in turn depend on the amount of data they obtain. As such, firm i's prices, and thus its profits, depend on $\mathbf{P}$. We denote firm i's basic price and profits under a partition set as $p_i^{\mathrm{B}}(\mathbf{P})$ and $\pi_i(\mathbf{P})$ respectively.

Following Bounie et al. (2021), we assume that the DB sells data through a system of auctions with reserve prices. This assumption implies that the DB has all the bargaining power and can thus extract all surplus from firms (we relax this assumption in Section 5.2, where we assume alternative selling mechanisms, namely auctions without reserve prices and Take It Or Leave It offers). The DB chooses the partition set $\mathbf{P}$ and sets up n auctions. In each auction, all firms can participate and the DB sells a partition d_i that is particularly valuable to a specific firm, being centred on that firm's location. We denote a firm's profits when it wins its auction as $\pi_i^{\mathrm{W}}(\mathbf{P})$, while its profits when it loses are denoted as $\pi_i^{\mathrm{L}}\left(\mathbf{P}\right)$. While a generic firm i can bid in every auction, it is mostly interested in the one where d_i is sold. Thus, firm i valuation of d_i is not matched by any other firm. As such, firm i could bid lower than its true valuation and still obtain d_i: this would result in the DB not being able to extract all surplus. To avoid this scenario, the DB sets a reserve price w_i for each auction that is equal to firm i's true valuation of d_i. We define a firm's true valuation of data as the difference in firm's profits between winning or losing its specific auction under a partition set $\mathbf{P}$:

$$w_i = \pi_i^{\mathrm{W}}(\mathbf{P}) - \pi_i^{\mathrm{L}}(\mathbf{P})$$

The vector of reserve prices is denoted as $\mathbf{w} = (w_0, w_1,\ w_2, \ldots, w_{n-1})$. This selling mechanism implies that, while any firm could bid in any auction, the DB can tailor each auction to maximise the willingness to pay of some firms. As such, when the DB sets up an auction that awards a partition centred on firm i's location, we interchangeably refer to it as an offer made to firm i or as firm i's auction.

Finally, similar to Bounie et al. (2021), the DB declares the maximum number of auctions she is going to fulfil, k, which is common knowledge prior to firms' bidding. A fulfilled auction – or won auction, from the viewpoint of the firm – is one where the transaction takes place. Declaring the maximum rather than the actual number of fulfilled auctions is functional to minimise firms' profits if they lose their auction.[7] The DB fulfils a subset $\mathbf{J}$ of auctions after the firms have placed their bids. The reserve price w_i allows the DB to set the price of data equal to firm i's true valuation of data, as priorly described. As such, DB's profits can be written as the sum of firms' winning bids, which are equal to their reserve prices:

$$\pi_{\text{DB}}(\mathbf{P}, \mathbf{J}) = \sum_{i \in \mathbf{J}} w_i$$

We denote the cardinality of $\mathbf{J}$ as j: as such, $j \leq k$ is the number of fulfilled auctions.

The timing of the model is as follows:[8]

Stage 1. Firms enter the market and pay the fixed cost F.

Stage 2. The DB chooses a partition set $\mathbf{P}$, the reserve prices $\mathbf{w}$, and the maximum number of auctions she will fulfil k. All this information is common knowledge and offers are non-renegotiable.

Stage 3. Firms that entered the market individually and simultaneously bid in the auctions.

Stage 4: The DB observes the bids and chooses a subset $\mathbf{J}$ of auctions to fulfil. The winning firms receive their respective partitions and pay their price to the DB, corresponding to $w_i = \pi_i^{\text{W}}(\mathbf{P}) - \pi_i^{\text{L}}(\mathbf{P})$.

Stage 5. Firms set basic prices p_i^{B} for the anonymous consumers.

Stage 6. Firms set tailored prices ${p_i}^{\text{T}}(x)$ for the identified consumers if they have won an auction. Consumers purchase the product and profits are made.

[7] In a Hotelling setting with two firms, the DB's optimal strategy involves setting up two auctions and declaring that only one will be fulfilled (Montes et al., 2019; Bounie et al., 2021). This way, a firm knows that if it loses its auction, the other firm will win it. This strategy allows the DB to maximise firms' valuation of data, as firms are informed and competing against an uninformed rival if they win, and the opposite if they lose. However, declaring the number of fulfilled auctions is no longer optimal when moving to a circular city with $n \geq 3$, as firms always face two direct rivals. In this setting, the DB can maximise firms' evaluation of data only if she can change the number of fulfilled auctions depending on firms' behaviour. Consider a case where $n = 3$. A firm's profits are maximised when it is informed and competing against uninformed rivals, and thus when only its specific auction is fulfilled. On the other hand, a firm's profits are minimised when it is uninformed and competing against both informed rivals, resulting in two fulfilled auctions. As such, declaring the exact number of fulfilled auctions would be suboptimal for the DB, as she would not be able to simultaneously maximise firms' profits when they win and minimise them if they lose. By instead declaring the maximum number of fulfilled auctions, the DB can maximise firms' valuation of data. Sticking to the example with $n = 3$, the DB can declare that she will fulfil a maximum of two auctions. If all firms do not deviate from their equilibrium strategies, only one auction will be concluded, maximising the winning firm's profits. On the other hand, firms know that if they deviate, the DB can let both of their rivals win their respective auctions and thus minimise their profits.

[8] Stage 6 follows Stage 5 to ensure the existence of an equilibrium in pure strategies. See also Montes et al. (2019) for an analogous approach.

In the following sections we proceed by backward induction, identifying the firms' basic and tailored prices, and the DB's optimal strategy.

2.3. DB's Strategies

The DB can influence the degree of downstream competition by determining if and to what extent a firm and its rivals have access to consumer data. As already noted, she does so by deciding the maximum number of fulfilled auctions k, the subset of fulfilled auctions $\mathbf{J}$, the reserve prices $\mathbf{w}$ and the partition set $\mathbf{P}$. Although this would leave us with a conspicuous set of strategies, we can reduce them by eliminating some strategies that can never be part of an equilibrium, as stated by the following proposition.

Proposition 1. *Only two candidate equilibrium strategies are possible: i) the DB sells equally sized partitions to all the firms that entered, or ii) the DB sells equally sized partitions to half of the entered firms, alternating between informed and uninformed ones.*

Proof. See Appendix I. □

Proposition 1 states that the set of candidate equilibria includes only two types of strategies, namely either selling data to all entered firms, or selling data to every other firm. Intuitively, our circular city can be seen as a concatenation of Hotelling segments with symmetric firms located at their extremes. In equilibrium, the DB adopts the profits-maximising strategy in one of these Hotelling segments and replicates it on all other segments. By doing so, the DB can only have two viable strategies. First, she can sell equally sized partitions to both firms of each Hotelling segment. Replicated on all segments, this strategy implies that the DB sells equally sized partitions to all the firms that entered the market. We refer to this strategy as the *sale to all firms*, and we denote it with the subscript A. Second, the DB can sell data to one of the two firms of each Hotelling segment. Replicated on all segments, this strategy implies that the DB sells equally sized partitions to half of the firms, alternating those with data and those without. We refer to this strategy as the *sale to alternating firms*, denoting it with the subscript H. Notably, under both strategies a firm always faces direct rivals that obtain same-sized partitions: as such, in equilibrium the DB offers partitions that are centred on firms' locations. Suppose that firm i is offered a partition d_i: since both its direct rivals obtain equally sized partitions, in equilibrium the DB sells to firm i a partition such that it can identify consumer segments of size $\frac{d_i}{2}$ on each arch on which it competes.

Under the *sale to all firms*, the DB sets $d_i = d_{\mathrm{A}}\ \forall\ i$, thus offering a partition set $\mathbf{P_A} = (d_{\mathrm{A}}, d_{\mathrm{A}}, \ldots, d_{\mathrm{A}})$. Since the DB fulfils all auctions, $k = j = n$, and the partition set offered to firms is equal to the partition set resulting in equilibrium, $\mathbf{P_A} = \mathbf{P^*_A}$. To better explain the implications of this strategy, we focus on a generic firm i. If firm i wins its auction, its profits are $\pi_i^{\mathrm{W}}(\mathbf{P^*_A}) = \pi_i^{\mathrm{W}}(d_{\mathrm{A}}, d_{\mathrm{A}}, \ldots, d_{\mathrm{A}}, d_{\mathrm{A}}, d_{\mathrm{A}}, \ldots, d_{\mathrm{A}})$: that is, it is an informed firm competing against informed rivals. If firm i loses, it becomes the only uninformed firm in the market, with profits $\pi_i^{\mathrm{L}}(\mathbf{P^*_A}) = \pi_i^{\mathrm{L}}(d_{\mathrm{A}}, d_{\mathrm{A}},, \ldots, d_{\mathrm{A}}, 0, d_{\mathrm{A}}, \ldots, d_{\mathrm{A}})$. While this strategy maximises the number of paying firms, it does not maximise individual firms' willingness to pay. Previous literature on Hotelling settings (Thisse and Vives, 1988) has highlighted how data have two effects on firms' profits. An informed firm can extract more surplus from the identified consumers, increasing its profits. Moreover, an informed firm engages in price wars as it tries to poach the consumers of its rivals. These two effects are referred to by the literature as *surplus extraction effect* and *competition effect* respectively. The first one increases firms' profits, while the second one decreases them. In particular, when an informed firm faces informed rivals, the competition effect dominates the surplus extraction effect, and firms' profits decrease. As such, an individual firm's willingness to pay under the *sale to all firms* is lower than under the *sale to alternating firms.*

Under the *sale to alternating firms*, the equilibrium outcome entails $d_i = d_{\mathrm{H}}\ \forall\ i \in \{0, 2, 4, \ldots, n-2\}$, and $d_i = 0\ \forall\ i \in \{1, 3, 5, \ldots, n-1\}$. Therefore, the partition set relative to the equilibrium outcome is $\mathbf{P^*_H} = (d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0)$. Since the DB may decide not to fulfil some auctions, the partition outcome $\mathbf{P^*_H}$ may differ from the partition set offered to firms, denoted as $\mathbf{P_H}$. In particular, to achieve the partition outcome $\mathbf{P^*_H}$, the optimal partition set $\mathbf{P_H}$ offered by the DB under the *sale to alternating firms* is defined by the following proposition.

Proposition 2. *Under the sale to alternating firms, given the partition set outcome* $\mathbf{P^*_H} = (d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0)$*, the DB offers the partition set* $\mathbf{P_H} = (d_{\mathrm{H}}, 1, d_{\mathrm{H}}, 1, \ldots, d_{\mathrm{H}}, 1)$.

Proof. See Appendix II. □

Proposition 2 states that, under the *sale to alternating firms.*, the DB offers the whole dataset in all auctions that will not be fulfilled in equilibrium, and she offers d_{H} in all auctions that will be fulfilled. This strategy maximises the willingness to pay of the firms whose auctions will be fulfilled by the DB in equilibrium. The intuition is the following.

The DB is not constrained to fulfil all auctions, as she may fulfil only a subset of the auctions she sets up. In particular, in the *sale to alternating firms*, the DB can set up auctions for all firms, even though she means to fulfil only half of them. By doing so, the DB can use the auctions she will not fulfil as a threat to increase firms' willingness to pay for data, and consequently her profits. In fact, the DB's profits are equal to the difference between firms' profits when winning their auction and when losing it. If firm i is offered $d_i = d_{\mathrm{H}}$ and wins it auction, it competes against uninformed direct rivals (i.e., $i+1$ and $i-1$). This raises firm i's profits, as it becomes an informed firm competing against uninformed rivals. Conversely, if firm i loses its auction, the DB lets firm i's rivals win their respective partitions and sets such partitions to the full dataset. This minimises firm i's profits when losing, as it would be forced to compete without data against completely informed rivals. Overall, the DB's strategy of offering the full dataset in auctions that are not meant to be fulfilled hurts firm i the most if it loses its specific auction, increasing firm i's willingness to pay for data and, in turn, DB's profits.[9]

To maximise the threat posed on firms receiving d_{H}, the DB declares that she will fulfil at most $k = \frac{n}{2} + 1$ auctions. Under this strategy, if firm i loses its auction, the DB can make both its rivals win their respective auctions, since $k = \frac{n}{2} + 1$.

In the following sections we focus on subgame perfect Nash equilibria under these two strategies.

3. Equilibrium Prices

We proceed by backward induction, and find the equilibrium prices and firms' profits under the DB's strategies described in Section 2.3.

As a benchmark, we refer to the Salop (1979) model with marginal costs normalised to 0. In this setting, each firm sets a price ${p_i}^* = \frac{t}{n}$ and obtains a market share of $\frac{1}{n}$, resulting in profits ${\pi_i}^* = \frac{t}{n^2} - F$. The number of entering firms is $n^* = \sqrt{\frac{t}{F}}$, resulting in firms' prices ${p_i}^* = \sqrt{tF}$ and profits ${\pi_i}^* = 0$. Consumer surplus is $CS = v - \frac{5}{4}\sqrt{tF}$, which is also equal to total surplus (Taylor and Wagman, 2014).

In our setup, the indifferent consumers between firms $i-1$ and i, and between i and $i+1$, are:

$$\widehat{x}_{i-1,i} = \frac{2i-1}{2n} + \frac{p_i^{\mathrm{B}} - p_{i-1}^{\mathrm{B}}}{2t} \qquad \text{and} \qquad \widehat{x}_{i,i+1} = \frac{2i+1}{2n} + \frac{p_{i+1}^{\mathrm{B}} - p_i^{\mathrm{B}}}{2t} \tag{1}$$

[9] As already pointed out by Bounie et al. (2021; footnote 14), who obtain an analogous result in a duopoly setup, the threat of selling the full dataset to firm i's rivals is not renegotiation proof. In Section 5.2 we extend our model to other selling mechanisms in which the DB's strategy is renegotiation proof.

If firm i wins the auction, it obtains the data d_i and offers a tailored price $p_i^{\mathrm{T}}(x)$ to the identified consumers for each arch in which it competes, matching the competitor's offer in utility level and resulting in

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

Notice that tailored prices decrease as the rival's basic price decreases: as the competitive pressure rises, firms lower their tailored prices to match the rival's basic price. Depending on the amount of data d_i obtained, firm i can serve both identified and anonymous consumers on both arches. When firm i is offered d_i and wins the auction, its profits prior to paying for data are given by:

$$\pi_i^{\mathrm{W}}(\mathbf{P}) = \int_{\frac{i}{n}-\frac{d_i}{2}}^{\frac{i}{n}} p_i^{\mathrm{T}}(x)\,dx + \int_{\frac{i}{n}}^{\frac{i}{n}+\frac{d_i}{2}} p_i^{\mathrm{T}}(x)\,dx + p_i^{\mathrm{B}}(\mathbf{P})\left(\widehat{x}_{i,i+1} - \widehat{x}_{i-1,i} - d_i\right) - F \tag{3}$$

given the offered partition set $\mathbf{P} \in \{\mathbf{P_H}, \mathbf{P_A}\}$. The first two components of Equation (3) represent profits on the identified segment on the two sides of firm i and depend on the tailored price, while the third component represents profits on the anonymous segment and depends on the basic price. The profits on the identified segments are due to a surplus extraction effect: as firm i can identify consumers, it can offer them tailored prices to exactly match their willingness to pay for its product. Using the expression of the indifferent consumers from Equation (1) and of the tailored prices in Equation (2), we can rewrite the profits of the generic informed firm i in Equation (3) as

$$\begin{aligned}\pi_i^{\mathrm{W}}(\mathbf{P}) = \frac{d_i}{2n}\left(2t + np_{i-1}^{\mathrm{B}}(\mathbf{P}) + np_{i+1}^{\mathrm{B}}(\mathbf{P}) - ntd_i\right) \\ + p_i^{\mathrm{B}}(\mathbf{P})\left(\frac{n\left(p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P}) - 2p_i^{\mathrm{B}}(\mathbf{P})\right) + 2t}{2nt} - d_i\right) - F\end{aligned} \tag{4}$$

Conversely, if firm i loses the auction, it becomes uninformed obtaining profits

$$\pi_i^{\mathrm{L}}(\mathbf{P}) = p_i^{\mathrm{B}}(\mathbf{P})\left(\frac{n\left(p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P}) - 2p_i^{\mathrm{B}}(\mathbf{P})\right) + 2t}{2nt}\right) - F \tag{5}$$

given the offered partition set $\mathbf{P} \in \{\mathbf{P_H}, \mathbf{P_A}\}$. By taking the first-order condition of Equations (4) and (5) with respect to $p_i^{\mathrm{B}}(\mathbf{P})$, we obtain the firm's reaction function on basic prices:

$$p_{i(\mathrm{W})}^{\mathrm{B}}(\mathbf{P}) = \frac{t}{2n} - \frac{td_i}{2} + \frac{p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P})}{4} \tag{6}$$

$$p_{i(\mathrm{L})}^{\mathrm{B}}(\mathbf{P}) = \frac{t}{2n} + \frac{p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P})}{4} \tag{7}$$

The reaction function in Equation (6) is analogous to the reaction function of the standard Salop (1979) model, except for the term $-\frac{td_i}{2}$ in the expression of $p^{\mathrm{B}}_{i(\mathrm{W})}(\mathbf{P})$. The term $-\frac{td_i}{2}$ is related to the *competition effect* of data: as firm i acquires more data, the anonymous consumers it reaches are on average farther from its location, requiring the firm to lower its basic price.

3.1. Sale to Alternating Firms

Recall that, from Proposition 2, under the strategy of selling to alternating firms, a subset of firms are offered a positive partition of data d_{H}, whereas others are offered the full dataset – although their auctions will not be fulfilled.

Let us first analyse the equilibrium in the subgame in which the firms that are offered $d_i = d_H$ win their auction, while the firms that are offered the full dataset lose their auctions. In particular, let $i, i{+}2, i{+}4\ldots$ be the firms that obtain d_{H} and thus exhibit the reaction function expressed in Equation (6), while firms $i{-}1, i{+}1, i{+}3\ldots$ compete without data and thus present the reaction function expressed in (7). The system of reaction functions for all firms allows us to obtain the equilibrium basic prices and, by using (4), firm i's profits, as illustrated in the following proposition.

Proposition 3. *Under the sale to alternating firms, if firm i wins its auction and obtains* d_{H} *while its direct rivals lose their respective auctions, firm i's basic price is decreasing in* d_{H} *if* $d_{\mathrm{H}} < \frac{3}{2n}$*, and is zero otherwise:*

$$p_i^{\mathrm{B}*}(\mathbf{P}^*_{\mathbf{H}}) = \begin{cases} \frac{t}{n} - \frac{2}{3}td_{\mathrm{H}} & \textit{for} \;\; d_{\mathrm{H}} < \frac{3}{2n} \\ 0 & \textit{for} \;\; d_{\mathrm{H}} \geq \frac{3}{2n} \end{cases}$$

Firm i's profits follow an inverse U-shaped curve with respect to data for $d_{\mathrm{H}} < \frac{3}{2n}$*, and are constant otherwise:*

$$\pi_i^{\mathrm{W}*}(\mathbf{P}^*_{\mathbf{H}}) = \begin{cases} \frac{t}{n^2} + \frac{2d_{\mathrm{H}}t}{3n} - \frac{7td_{\mathrm{H}}^2}{18} - F & \textit{for} \;\; d_{\mathrm{H}} < \frac{3}{2n} \\ \frac{9t}{8n^2} - F & \textit{for} \;\; d_{\mathrm{H}} \geq \frac{3}{2n} \end{cases}$$

Moreover, firm i's profits are always higher than in the benchmark case.

Proof. See Appendix III. □

From the expression of firm i's basic price in Proposition 3, we observe that this price is positive only if $d_{\mathrm{H}} < \frac{3}{2n}$. When $d_{\mathrm{H}} < \frac{3}{2n}$, firm i's basic price is decreasing in d_{H}, while its profits follow an inverse U-shaped curve with respect to d_{H}. At first, the surplus extraction effect is stronger than the competition effect, increasing profits. However, the marginal surplus extraction is decreasing in d_{H} while the competition effect is linear, causing profits to peak and then decrease. After $d_{\mathrm{H}} \geq \frac{3}{2n}$, firm i's basic price reaches zero and the effect of incremental data vanishes, because firm i cannot poach additional consumers: those who are close to the uninformed rivals always prefer them over the informed firm due to their positional advantage. Note that firm's profits are negatively related to n. In fact, a higher number of firms intensifies competition, lowering basic prices, and at the same time it reduces firms' market segments.

We now focus on the subgame where firm i is offered d_{H} but loses its auction, while its direct rivals are offered $d_{i+1} = d_{i-1} = 1$ and win their auctions. Then, firm i becomes an uninformed firm competing against informed rivals. By offering the full dataset to firm i's rivals and choosing to fulfil their auctions, the DB minimises firm i's profits when it loses the auction.

In this scenario, firm i's profits are expressed by Equation (5), where $\mathbf{P} = \mathbf{P_H}$ and $p^{\mathrm{B}}_{i+1} = p^{\mathrm{B}}_{i-1} = 0$ because, being $d_{i+1} = d_{i-1} = 1 > \frac{3}{2n}$, they earn all of their profits through tailored prices. Firm i's equilibrium price and profits when losing the auction, given the offered partition set $\mathbf{P} = \mathbf{P_H}$ expressed in Proposition 2, are summarised in the following proposition.

Proposition 4. *Under the sale to alternating firms, if firm i loses its auction while its direct rivals obtain $d_{i+1} = d_{i-1} = 1$, firm i's basic price and profits are respectively $p^{\mathrm{B}*}_i(\mathbf{P_H}) = \frac{t}{2n}$ and $\pi^{\mathrm{L}*}_i(\mathbf{P_H}) = \frac{t}{4n^2} - F$.*

Proof. See Appendix IV. □

Proposition 4 highlights that firm i's profits when losing the auction do not depend on d_{H} and are always lower than its profits in the benchmark case. Moreover, by comparing the results of Propositions 3 and 4, we note that firm i's profits when losing the auction are always lower than its profits when it wins its auction.

It is worthwhile to note that firms' profits are positive even when they lose the auction and compete against fully informed rivals, due to their horizontal differentiation which ensures them a market share near their location. Therefore, uninformed firms can still make a profit. This feature has important implications for the number of entering firms, as we highlight in Section 4.1.

3.2. Sale to All Firms

We now consider the alternative DB's strategy of the *sale to all firms.* We first analyse the subgame in which a generic firm i wins its auction, and then we proceed to the subgame in which it loses it – given that all other firms win their respective auctions and receive data.

If the offered partition set is $\mathbf{P} = \mathbf{P_A}$ and all firms win their respective auction, firm i's profits are expressed by Equation (4), for all i, and all firms obtain a partition of size $d_i = d_\mathrm{A}$ centred on their respective location. Firms' reaction functions are expressed by (6), where $\mathbf{P} = \mathbf{P_A}$ and $d_i = d_\mathrm{A}$, for all i. The system of all reaction functions allows us to obtain the equilibrium basic prices and firms' profits, as reported in the following proposition.

Proposition 5. *Under the sale to all firms, if all firms win their respective auctions and obtain data d_A, firm i's basic price is*

$$p_i^{\mathrm{B}*}(\mathbf{P_A}) = \begin{cases} \frac{t}{n} - td_\mathrm{A} & \textit{for} \;\; d_\mathrm{A} < \frac{1}{n} \\ 0 & \textit{for} \;\; d_\mathrm{A} \geq \frac{1}{n} \end{cases}$$

for all i. Firm i's profits are

$$\pi_i^{\mathrm{W}*}(\mathbf{P_A}) = \begin{cases} \frac{t}{n^2} - \frac{td_\mathrm{A}^2}{2} - F & \textit{for} \;\; d_\mathrm{A} < \frac{1}{n} \\ \frac{t}{2n^2} - F & \textit{for} \;\; d_\mathrm{A} \geq \frac{1}{n} \end{cases}$$

Proof. See Appendix V. □

By observing the expressions of $p_i^{\mathrm{B}*}(\mathbf{P_A})$ and $\pi_i^{\mathrm{W}*}(\mathbf{P_A})$ in Proposition 5, we note that firm i's basic price and profits when all firms win their auction are both decreasing in d_A if $d_\mathrm{A} < \frac{1}{n}$, and zero otherwise.

Absent data, the indifferent consumer is in the middle of the arch between two firms (i.e., at a distance $\frac{1}{2n}$ from the firms' locations), and each firm sets an equal basic price. As the data d_A sold to the two firms increases, the location of the indifferent consumer remains the same due to firms' symmetry, but firms start price discriminating on ever more consumers. This causes the surplus extraction effect that increases profits. At the same time, however, the competition effect lowers all basic prices, and thus profits. Moreover, as the rival reduces its basic price, so does firm i with its tailored price, to match the rival's offer. This lowers the positive impact on profits of the surplus extraction effect. Although profits are affected by the two opposite forces of the competition and surplus extraction effect, the first is dominant due to the symmetry of firms' information, which intensifies their competition and produces a strong reduction of prices, thereby reducing profits.

However, above a threshold – namely, $d_{\mathrm{A}} \geq \frac{1}{n}$ – firms identify all consumers up to the indifferent one in the middle of the arch. Since now firm i only serves identified consumers through tailored prices, it has an incentive to set $p_i^{\mathrm{B}}(\mathbf{P_A})$ as low as possible to try expanding its market segment, as the basic price in this scenario only influences the indifferent consumer's location. After this threshold, additional data stop having an impact since they identify consumers who are too close to firms' rivals to be poached.

We now focus on the subgame in which firm i loses its auction, while its competitors obtain data. Although this subgame is off the equilibrium path, it is necessary to know firm i's payoff in this subgame to assess its willingness to pay for data. Since firm i does not obtain data, its profits are defined by Equation (5), while all other firms' profits can be expressed as in (4). From the system of all reaction functions, we obtain the following result.

Proposition 6. *Under the sale to all firms, if firm i loses its auction while all firms $i' \neq i$ win it and obtain $d_{\mathrm{A}} > 0$, firm i's basic price and profits are strictly decreasing in the quantity of data if $d_{\mathrm{A}} < \frac{3}{2n}$, and are constant otherwise.*

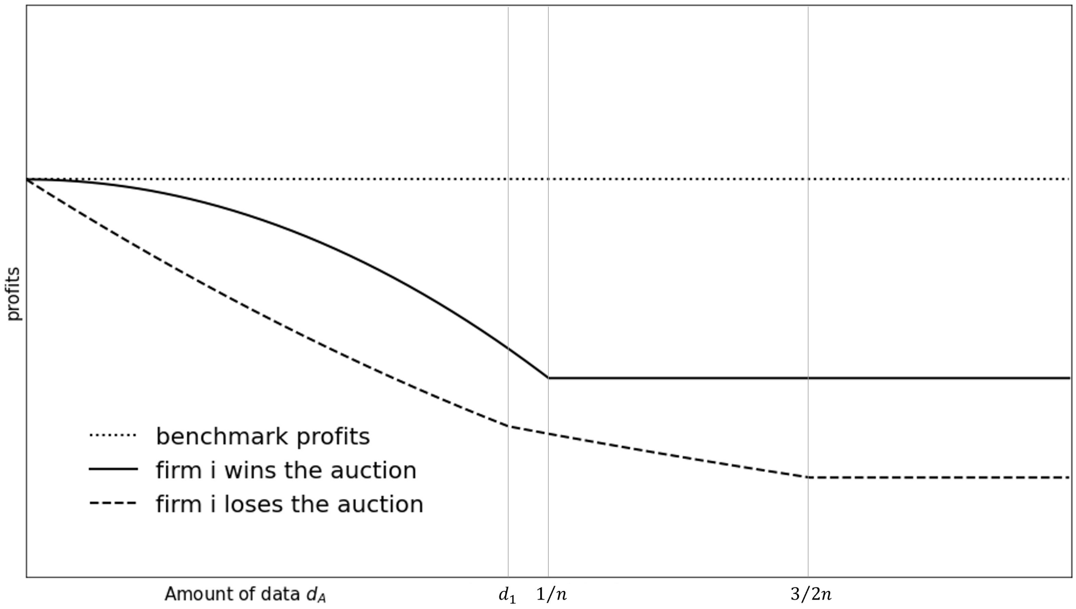
Proof. See Appendix VI. □

Being the only uninformed firm in the market, firm i is at a disadvantage vis-à-vis its rivals. First, firm i cannot adopt targeted pricing, thus limiting the profits it can extract from consumers. Second, the decrease in basic prices caused by the competition effect can be partially beneficial to informed firms: since they can lower their basic price more than the uninformed firm, they can expand their market segments. This second effect is stronger the closer the informed competitor is to the uninformed firm.

To gain further insights on this second effect, without loss of generality, let us focus on firm $i{+}1$: on one side, it competes against an uninformed firm (i.e., i), while on the other it competes against an informed one ($i{+}2$). In this situation, firm $i{+}1$ optimal strategy is to maximise the profits on the arch shared with firm i, on which it has an advantage, while sacrificing some profits on the arch shared with firm $i{+}2$. To do so, firm $i{+}1$ undercuts firm i's basic price, thus expanding its market share and increasing its profits on that arch. Firm $i{+}1$'s direct rival (i.e., $i{+}2$) can in turn increase its profits by slightly undercutting firm i+1's basic price: as such, firm i+2 expands its market segment on the arch it shares with firm i+1. In short, all firms expand their market segments on the arch closer to the uninformed firm, while they reduce their market segment on the other arch. When d_{A} is relatively low, all informed firms serve both identified and anonymous consumers, and thus set positive basic prices. After d_{A} reaches a first threshold, firm i+1 only serves identified consumers on the arch it shares with i+2, while it still serves anonymous consumers on the arch it shares with firm i.[10] Finally, after d_{A} reaches a second threshold (i.e., $d_{\mathrm{A}} \geq \frac{3}{2n}$), firm i+1 only serves identified consumers, and the effect of additional data becomes nil.

By comparing firm i's profits when winning or losing its auction, we find that firms are always better off when winning it by paying the DB's reservation price. Figure 1 provides a graphical representation of firms' profits functions in the two subgames.[11]

Figure 1. Firms' Profits in the *Sale to All Firms*



This figure shows firms' profits under the benchmark case (i.e., without data) and when winning or losing the auction prior to paying for data. $n = 4$ and $t = 20$.

As shown in Proposition 5, when all firms obtain data, their profits (represented by the solid line in Figure 1) are strictly decreasing for $d_{\mathrm{A}} \leq 1/n$, and constant otherwise. The threshold identifies the amount of data above which all consumers are identified in the market, and additional data have no impact on firms' strategy. The greater amount

[10] This threshold depends on the number of entering firms, and it is always lower than $\frac{1}{n}$. See Appendix VI for details.

[11] While we show profits functions for specific parameters values, we observe the same trends for all parameters values that grant market coverage (i.e., any value that does not violate the model's basic assumptions).

of data sold in the downstream market makes the competition effect outweigh the surplus extraction effect, resulting in firms being worse off under the presence of a DB. However, firms experience a prisoner's dilemma: while they would be better off not buying data, the threat of being uninformed and competing against informed rivals leads them to prefer participating in the DB's auctions.

4. DB's Equilibrium Profits

Having analysed the effect of data on firms, we can obtain DB's profits and maximise them with respect to d_{H} and d_{A} under the two strategies. We recall that the DB sets the auctions' reserve prices $\mathbf{w}$ equal to firms' willingness to pay for data: as such, she extracts all surplus from the firms that win the auctions. As a tie-breaker rule, we assume that if a firm is indifferent between winning or losing its auction, it prefers winning it. The DB's profits can be written as

$$\pi_{\mathrm{DB}}(\mathbf{P}, \mathbf{J}) = \sum_{i \in \mathbf{J}} w_i = \sum_{i \in \mathbf{J}} \pi_i^{\mathrm{W}}(\mathbf{P}) - \pi_i^{\mathrm{L}}(\mathbf{P}) \tag{8}$$

As such, firms' profits after winning their auction and paying for data are equal to

$$\pi_i^{\mathrm{W}}(\mathbf{P}) - w_i = \pi_i^{\mathrm{W}}(\mathbf{P}) - \left(\pi_i^{\mathrm{W}}(\mathbf{P}) - \pi_i^{\mathrm{L}}(\mathbf{P})\right) = \pi_i^{\mathrm{L}}(\mathbf{P})$$

That is, firms' remaining profits after paying for data are equal to their profits when losing their auction. Since firms enter the market as long as they make positive profits, the number of entering firms is given by the condition

$$\pi_i^{\mathrm{L}}(\mathbf{P}) = 0 \tag{9}$$

As a useful benchmark, we can refer to the standard Salop (1979) model (see Section 3) where, absent the DB, the number of entering firms is $\widetilde{n}^* = \sqrt{\frac{t}{F}}$. In the following sections we solve the game under the two strategies and compare the outcomes to assess the DB's preferred one.

4.1. Sale to Alternating Firms

When the DB opts for the *sale to alternating firms*, we can rewrite her profits as

$$\pi_{\mathrm{DB}}(\mathbf{P_H}, \mathbf{J}) = \sum_{i \in \mathbf{J}} w_i = \frac{n}{2}\left(\pi_i^{\mathrm{W}*}(\mathbf{P_H}) - \pi_i^{\mathrm{L}*}(\mathbf{P_H})\right) \tag{10}$$

The DB sets d_{H}^* so as to maximise firms' willingness to pay for data (10), for a given number of entering firms n. The solution of the DB's profit maximisation problem is expressed by the following proposition.

Proposition 7. *Under the sale to alternating firms, the DB offers $d_{\mathrm{H}}^{*} = \frac{6}{7n}$ in the auctions she wants to fulfil and $d_i = 1$ in the ones she does not want to fulfil.*

Proof. See Appendix VII. □

From Proposition 7, the amount of data sold in equilibrium by the DB to every other firm is $d_{\mathrm{H}}^{*} = \frac{6}{7n}$, which is lower than the amount of data $3/(2n)$ that would allow firms to identify all consumers on their market segment. This implies that the DB only sells data about high valuation consumers, located closer to the winning firms' positions, so that in equilibrium informed firms serve both identified and unidentified consumers. The DB adopts this strategy to temper downstream competition. In fact, letting informed firms identify all their consumers would result in price wars that would deplete their profits and, in turn, their willingness to pay for data.

4.2. Sale to All Firms

When the DB opts to sell data to all firms, her profits are:

$$\max_{d_{\mathrm{A}}} \pi_{\mathrm{DB}} = n\left(\pi_{\mathrm{i}}^{\mathrm{W}*}\left(\mathbf{P_A}\right) - \pi_{i}^{\mathrm{L}*}\left(\mathbf{P_A}\right)\right) \tag{11}$$

The amount of data d_A^{*} sold by the DB to each firm depends on the number of entering firms and is defined by the following proposition.

Proposition 8. *Let $\hat{n}$ be the number of entering firms such that the DB's profits when selling non overlapping partitions are equal to those when selling $d_A \geq \frac{3}{2n}$. Under the sale to all firms, the DB's strategy depends on the number of entering firms:*

- *If $n < \hat{n}$, the DB offers $d_{\mathrm{A}}^{*} = 1$ in all auctions and fulfils all of them.*
- *If $n \geq \hat{n}$, the DB offers non overlapping partitions in all the auctions and fulfils all of them.*

Proof. See Appendix VIII. □

As shown in the Appendix, the value of $\hat{n}$ is approximately equal to 3.34. The DB's strategy depends on the number of entering firms because firms' valuation of data depends on n. When n is sufficiently high, firms have small market shares. This implies that their average consumers are closer to their locations. Since data allows firms to extract more surplus from consumers closer to their locations, the DB opts to temper downstream competition by selling non-overlapping partitions. By doing so, the DB allows firms to extract more surplus from consumers, via the price of data. Conversely, when n is low, selling non-overlapping partitions would not be as effective, as consumers are on average farther from firms' locations. Thus, the DB offers the whole dataset to all firms: this strategy minimises firms' profits when losing their auction, as they risk to face completely informed rivals.

4.3. DB's Optimal Strategy

By comparing the DB's profits under the two possible strategies described in Propositions 7 and 8, we can identify the DB's optimal choice. Results are summarised in the following proposition.

Proposition 9. *In equilibrium, the DB sells data to alternating firms. The number of entering firms is* $n_{\mathrm{H}}^{*} = \frac{1}{2}\sqrt{\frac{t}{F}}$*, i.e.,* $n_{\mathrm{H}}^{*} = \frac{\widetilde{n}^{*}}{2}$*.*

Proof. See Appendix IX. □

The *sale to alternating firms* dominates the *sale to all firms*, implying that the DB prefers under-serving the market by excluding some firms from the data sale. In fact, selling data to alternating firms allows her to maximise firms' willingness to pay for data, as it increases the informed firm's profits when it competes against an uninformed rival while also maximising the threat posed to firms if they lose. While previous literature (e.g., Montes et al., 2019) advocated for oversight by policymakers of exclusive data arrangements in a duopoly setting, the result of Proposition 9 suggests that, in a less concentrated market, the DB under-serves the market even in the absence of exclusive deals, as she excludes some firms from the data sale.

Moreover, Proposition 9 highlights that the DB's optimal strategy reduces firms' entry relative to the benchmark case where data are absent. This entry barrier effect is due to the reduction of firms' profits, as they either pay for data (if they win) or face informed rivals (if they lose). Interestingly, the number of entering firms in equilibrium cannot be lower than $\frac{\widetilde{n}^{*}}{2}$, i.e., the entry deterrence caused by data is limited, as additional data stop reducing firms' profits after a threshold (due to the horizontal differentiation setting). This result expands the entry barrier effect of data identified by de Cornière and Taylor (2020) in a setting in which data affect the quality of the information held by firms. Our analysis shows that the entry barrier effect emerges also when data carry information on the consumers' preferences and can thus be used for price discrimination.

4.4. Consumer and Welfare Analysis

In the previous section we characterised the equilibrium, which sees the entry of n^*_{H} firms and their purchase of the partition set $\mathbf{P^*_H} = (d^*_{\mathrm{H}}, 0, ..., d^*_{\mathrm{H}}, 0)$. In this section we focus on the implications of this equilibrium on welfare and consumer surplus.

In our model, total welfare comprises consumer surplus, firm profits and the DB' profits. In particular, let us express it as

$$TW = CS + \sum_{i=0}^{n-1} \pi_i + \alpha \pi_{\mathrm{DB}} \tag{12}$$

where $\alpha \in [0,1]$ is the weight of DB's profits in the welfare function. The following proposition summarises the impact of the DB's equilibrium strategy on consumers surplus and welfare.

Proposition 10. *In equilibrium, consumer surplus is lower and, if α is sufficiently high, total welfare is higher than in the case in which consumer data are not available.*

Proof. See Appendix X. □

Proposition 10 compares the result of our model in terms of consumer surplus and welfare to the result of the standard Salop model. We find that the DB's entry barrier effect lowers consumer surplus. In fact, in our setup, consumer surplus under the *sale to alternating firms* can be expressed as

$$CS = u - \frac{5t}{4n} + \frac{\mathrm{nt} d^2_{\mathrm{H}}}{9} \tag{13}$$

The first two terms in Equation (13) are the consumer surplus in the standard Salop model: as more firms enter the market, consumers have lower transportation costs, and their surplus increases. The third term represents the effect of data on consumer surplus for a given number n of firms. A higher quantity of data intensifies competition between the entered firms and lowers basic prices, raising the surplus of (unidentified) consumers. Therefore, if the number of firms is given, the third term is positive and increasing in data, implying that the DB's presence has a positive impact on consumer surplus, as evidenced also by previous literature (Braulin and Valletti, 2016; Montes et al., 2019; Bounie et al., 2021).

However, our results highlight that, when firm entry is endogenous, an entry barrier effect of data arises. The limited entry hurts consumers: as fewer firms enter the market, the average transportation cost paid by consumers increases, more than offsetting the decrease in basic prices caused by the competition effect of data.

It should be noted that the reduction of consumer surplus stems from the effect of endogenous entry in the presence of a DB. In fact, Taylor and Wagman (2014) analyse a Salop model with firm entry where all firms can price discriminate on all consumers without the need of purchasing data, finding that consumer surplus is higher than in the standard Salop model. The presence of a monopolist DB who owns data and can sell them to firms can shape the downstream competition, resulting in consumer harm.

Proposition 10 also finds that total welfare is higher than in the benchmark case, although it is mostly appropriated by the DB. As a monopolist, the DB addresses the problem of excessive entry identified by Salop (1979), limiting the number of firms to the efficient level. By doing so, the DB maximises industry profits which she can subsequently extract. If the weight α of the DB's profit in the welfare function is sufficiently low (specifically, in the proof of Proposition 10 in the Appendix we show that it must be $\alpha \leq 0.84$), total welfare is lower than in the benchmark. This result shows how the increase in welfare is mainly driven by the increase of the DB's profits, causing redistributive concerns from a policymaking point of view.

5. Extensions

In this section, we extend our basic model along several directions. First, we show how the reduction in firms' entry and, in turn, in consumer surplus is robust to the introduction of a consumer privacy cost. Second, we introduce variations in the DB's bargaining power by exploring different selling mechanisms for the data sale. Third, we analyse a scenario where the DB can commit to the price of data before firms' entry decision. This alternative timing allows the DB to take into account the effect of data on the number of entering firms when choosing her strategy, and provides her with more bargaining power than in the basic model.

5.1. Introducing a Privacy Cost

In this section we assume that, when a consumer is offered a tailored price, she incurs a disutility $c > 0$ due to her loss of privacy, for example due to the annoyance at being price discriminated. Thus, when a consumer accepts a tailored offer, she obtains a utility

$$U(x,i) = v - p_i^{\mathrm{T}}(x) - t * D(x,i) - c$$

Focusing on consumers between firm i and i+1, we can express firm i's tailored price as

$$p_i^{\mathrm{T}}(x) = p_{i+1}^{\mathrm{B}} - 2tx + \frac{t}{n}(2i+1) - c$$

The privacy cost reduces the surplus that firms can extract through tailored prices. An informed firm offers tailored prices to a consumer located in x only if

$$p_i^{\mathrm{T}}(x) \geq p_i^{\mathrm{B}}(\mathbf{P}) \tag{14}$$

That is, only if the tailored price allows the firm to extract more surplus from that specific consumer than the basic price. Results under the *sale to alternating firms* in the presence of a privacy cost are summarised by the following proposition.

Proposition 11. *In equilibrium, if consumers incur a disutility $c > 0$ when they are offered a tailored price:*

a) *Consumer surplus is increasing in the privacy cost, i.e., $\frac{\partial CS}{\partial c} > 0$.*

b) *Total welfare is decreasing in the privacy cost, i.e., $\frac{\partial TW}{\partial c} < 0$.*

c) *DB's profits are decreasing in the privacy cost, i.e., $\frac{\partial \pi_{\mathrm{DB}}}{\partial c} < 0$.*

d) *If c is sufficiently high, firms offer their basic prices regardless of data, and we obtain the results of the standard Salop model.*

Proof. See Appendix XI. □

For low values of c and d, Inequality (14) is satisfied for all consumers who belong to firm i's market segment. Then, firm i offers tailored prices to all identified consumers. However, the reduction in surplus caused by the privacy cost lowers firm profits, if compared to the case where the privacy cost is absent.

For sufficiently high values of c and d_{H}, the tailored price for distant consumers does no longer cover the privacy cost, and the informed firm prefers offering them its basic price, even if it can identify them. In particular, an informed firm uses data as long as $d_{\mathrm{H}} \geq \frac{3}{2n} - \frac{3c}{2t}$. After this threshold, the informed firm prefers serving those consumers through basic prices. Note that, if $c \geq \frac{t}{n}$, informed firms prefer offering their basic price to all consumers and avoid using data.

We also find that the disutility c reduces the surplus extraction effect, but it does not influence the competition effect. In particular, when $c \geq \frac{2t}{3n}$, informed firms' profits are decreasing in data: even if Inequality (14) is satisfied for some consumers – i.e., a firm can extract more surplus from some consumers by offering them tailored prices instead of the basic price – the drop in its basic price results in overall lower profits.

Furthermore, the privacy cost reduces the DB's entry barrier effect. In fact, as already noted, when consumers face a privacy cost, a completely informed firm would still serve some consumers through its basic price. As a consequence, an uninformed firm faces milder competition, which results in higher profits and more firms entering the market. In a setup with privacy cost, the number of entering firms is

$$n_{\mathrm{H}}^{*} = \frac{t}{2\sqrt{tF} - c}$$

The number of entering firms increases with c: when $c \geq \frac{t}{n^*}$, no firms uses data, and the number of entering firms is equal to that of the standard Salop model, $n_{\mathrm{H}}^{*} = \sqrt{\frac{t}{F}}$.

The magnitude of the privacy cost also affects the DB's optimal strategy.

- For $c < \frac{2t}{3n_{\mathrm{H}}^{*}} = \frac{4}{5}\sqrt{tF}$, firms' profits first follow an inverse U-shaped curve and then become constant, similar to the ones shown in Figure 1: as such, the DB offers the optimal amount of data $d_{\mathrm{H}}^{*} = \frac{6}{7n_{\mathrm{H}}^{*}} - \frac{9c}{7t}$ to maximise firms' profits, and firms offer their tailored prices to all the identified consumers.
- For $\frac{4}{5}\sqrt{tF} = \frac{2t}{3n_{\mathrm{H}}^{*}} \leq c < \frac{t}{n_{\mathrm{H}}^{*}} = \sqrt{tF}$, firms' profits are decreasing with data, as the competition effect always outweighs the surplus extraction effect: as such, the DB offers $d_{\mathrm{H}}^{*} = 0$ in the auctions she wants to fulfil. Firms accept only to avoid facing informed rivals, which would result in lower profits.
- For $c \geq \frac{t}{n_{\mathrm{H}}^{*}} = \sqrt{tF}$, data has no value for firms since they prefer reaching all consumers through basic prices. In this situation, the DB sells no data.

Overall, we conclude that, when using data for price discrimination entails a privacy loss to consumers, consumer surplus increases via the reduction of the entry barrier effect. The entry of a higher number of firms allows consumers to buy products closer to their preferences, resulting in higher surplus. However, total welfare decreases, because the total costs of entry increases with the number of entering firms.

5.2. Decreasing DB's Bargaining Power: Alternative Selling Mechanisms

In the basic model, in line with the previous literature, we assume that the DB has all the bargaining power. In fact, she is able to charge the maximum price for data by threatening firms to sell massive amounts of data to their rivals through auctions with reserve prices (AR). In this section, we relax this assumption in two ways. First, we consider a case where a DB sets auctions without reserve prices (AU). In this case, firms can underbid, thus reducing the DB's profits. Second, we assume that the DB sells data through a Take It Or Leave It (TIOLI) mechanism. In this scenario, the DB offers a partition to each firm, and each firm individually and simultaneously accepts or declines the offer. Therefore, under the TIOLI mechanisms, and differently from auctions, the DB cannot subsequently decide not to fulfil an offer, thus eroding her bargaining power.

We find that the selling mechanism adopted does not affect the DB's strategy under the *sale to all firms*, so that the results of Proposition 8 hold also under AU and TIOLI. In fact, firms face the same choice under all selling mechanisms: if a firm accepts the DB's offer, all firms in the market are informed; if the firm declines it, it becomes the only uninformed one. Conversely, the selling mechanism affects the DB's strategy under the *sale to alternating firms*.

Let us focus on AU. When the DB cannot set reserve prices, a firm can win its auction simply by bidding above the valuations of the other firms, which are lower than its own owing to their distance. Then, selling different partitions to firms would be detrimental for the DB, as it would increase firms' underbidding. Indeed, the DB maximises her profits by offering same-sized partitions in all auctions, even in those that are not meant to be fulfilled. The absence of reserve prices ultimately reduces the DB's profit because she cannot simultaneously maximise firms' profits when winning and minimise them when losing.

Let us now focus on TIOLI. Under this mechanism, the DB has no ex-post control on the number of transactions she wants to conclude. In fact, differently from AR, the DB must fulfil all her offers under TIOLI. Then, under the sale to alternating firms under TIOLI, the DB offers the partition set $\mathbf{P_H^{TIOLI}} = \left(d_{\mathrm{H}}^{\mathrm{TIOLI}}, 0, \dots, 0, d_{\mathrm{H}}^{\mathrm{TIOLI}}, 0, \dots, d_{\mathrm{H}}^{\mathrm{TIOLI}}, 0\right)$, alternating the sale of same-sized data partitions and no data. To better understand the implications of this selling mechanism, suppose that firm i refuses the DB's offer. Under TIOLI, the DB cannot threat firm i to sell data to its direct rivals, and firm i would thus face uninformed rivals even when refusing the DB's offer.

The main effects of adopting the alternative selling mechanisms AU or TIOLI, which reduce the DB's bargaining power, are summarised in the following proposition.

Proposition 12. *Under AU and TIOLI, the quantity of data sold to downstream firms, the number of entering firms and consumer surplus in equilibrium are higher than under AR. While under AU the DB in equilibrium opts for the sale to alternating firms, under TIOLI she adopts the sale to all firms strategy.*

Proof. See Appendix XII. □

The reduction of the DB's bargaining power improves consumer surplus. Depending on the selling mechanism and on the number of entering firms, consumer surplus can either be lower or higher than in the absence of the DB. Under AR, the DB can always threaten firms to make them face completely informed rivals. Thus, the DB sets $d_{\mathrm{H}}^{\mathrm{AR}}$ so as to maximise firms' profits when they win. However, this is no longer the case when the DB's bargaining power is reduced.

Under AU, if a firm loses its auction, its rivals obtain partitions of the same size as the firm's one. The threat of being uninformed is thus reduced when compared to AR, where a losing firm always faces completely informed rivals. The DB's equilibrium strategy under AU involves the *sale to alternating firms* and the offer of $d_{\mathrm{H}}^{\mathrm{AU}^*} = \frac{4}{3n}$, which is larger than the partition sold under AR. Selling a larger partition allows the DB to increase the threat to firms when losing their auction, thus increasing their willingness to pay.

Under TIOLI, the *sale to alternating firms* would be suboptimal: the DB cannot properly threaten firms under this selling mechanism, and this would result in a lower willingness to pay. As a consequence, in equilibrium under TIOLI the DB opts for the *sale to all firms*, following the same strategy described in Section 4.2.

The change in the DB's strategy also spurs firms' entry. Under AU, a losing firm does not face completely informed rivals when losing, as they obtain $d_{\mathrm{H}}^{\mathrm{AU}^*} = \frac{4}{3n}$. Firms' profits when losing are thus higher than under AR, thus reducing the number of entering firms by $\frac{4}{9}$ with respect to the benchmark. Conversely, under TIOLI, the number of entering firms is reduced by a half when $n^* < \hat{n}$, and it is reduced by slightly less than $\frac{1}{4}$ when $n^* \geq \hat{n}$.

As shown in Section 4.3, consumer surplus increases with the partition size and with the number of entering firms. Since both are greater or equal under AU and TIOLI than under AR, consumer surplus increases when decreasing the DB's bargaining power. In particular, we find that consumer surplus is at its highest under TIOLI. When $n^* < \hat{n}$, the high amount of data sold to firms compensates the strong entry barrier effect: however, consumer surplus is still lower than in the benchmark. On the other hand, when $n^* \geq \hat{n}$, consumer surplus is higher than in the benchmark, as the quantity of data sold more than offsets the small reduction in firms' entry.

To sum up, we find that lowering the DB's bargaining power by adopting selling mechanisms based on TIOLI offers is an effective way to reduce the consumer harm caused by her presence. However, consumers are only better off when the DB is forced to sell through TIOLI and the downstream market is highly competitive.

5.3. Committing to the Price of Data: An Alternative Timing

In our baseline setup, firms enter the market in the initial stage, and then participate in auctions to acquire data. This framework is for example consistent with markets that are already established before the introduction of digital technologies and the possibility to price discriminate through data. However, such a timing may be less intuitive in the case of emerging digital markets, in which firms know already before entering that obtaining consumer data would give them an edge over competition. In this section, we explore the possibility that the DB sets up the auctions prior to firms' entry. As a consequence, firms make the entry decision only after observing the offer of data by the DB. In particular, the timing we analyse in this section is as follows:

> Stage 1. The DB chooses a partition set $\mathbf{P}$, the reserve prices $\mathbf{w}$, and the maximum number of auctions she will fulfil k. All this information is common knowledge.
>
> Stage 2. Firms individually and simultaneously bid in the auctions.
>
> Stage 3: The DB observes the bids and chooses a subset $\mathbf{J}$ of auctions to fulfil. The winning firms receive their respective partitions and pay their price to the DB, corresponding to $w_i = \pi_i^{\mathrm{W}}(\mathbf{P}) - \pi_i^{\mathrm{L}}(\mathbf{P})$.
>
> Stage 4. Firms enter the market and pay the fixed cost F.
>
> Stage 5. Firms set basic prices p_i^{B} for the anonymous consumers.
>
> Stage 6. Firms set tailored prices ${p_i}^{\mathrm{T}}(x)$ for the identified consumers if they have won an auction. Consumers purchase the product and profits are made.

Note that, in this setup, firms' equilibrium prices are defined by the same functions as in Section 3, as firms' price setting stage takes place in the final stages of the game, as in our baseline timing. However, the DB's strategy substantially departs from that of our basic model. In fact, in the basic model, the DB picks her strategy by taking the number of entering firms as given. Conversely, under this alternative setting, the DB explicitly anticipates the effect of the data sale on firms' entry. We analyse the DB's strategy under all the selling mechanisms presented in Section 5.2. The main outcomes of this setting are summarised in the following proposition.

Proposition 13. *If firms purchase data before entering the market, in equilibrium the DB adopts the following strategies:*

- *Under the auction with reserve prices, the DB opts for the sale to alternating firms, and offers d_H^* in the auctions she wants to fulfil and the full dataset in the ones she does not want to fulfil;*
- *Under the auction without reserve prices, the DB opts for the sale to alternating firms and offers the full dataset in all auctions;*
- *Under Take It Or Leave It offers, the DB opts for the sale to all firms and offers the full dataset to all entering firms.*

All these strategies maximise the DB's entry barrier effect, and the number of entering firms is $n^*_{\text{AR}} = n^*_{\text{AU}} = n^*_{\text{TIOLI}} = \frac{\widetilde{n}^*}{2}$. DB's profits are greater or equal, and consumer surplus is lower or equal than in the basic model.

Proof. See Appendix XIII. □

When the DB anticipates the effect of her strategy on firms' entry, we find that she always benefits from maximising the entry barrier effect: as competition in the downstream market is reduced, entering firms make higher profits, which the DB can then extract through the price of data. To better understand the implications of the different timing, let us focus on specific selling mechanisms.

Under the auction with reserve prices (AR), the DB's strategy maximises her profits for any given number of entering firms, as firms' profits when losing their auction do not depend on d_H. In equilibrium, firms' expected profits when entering the market are the same as in our basic model, leading to the same market outcomes.

Conversely, the alternative timing alters the DB's strategy under the auction without reserve prices (AU). While she still opts for the *sale to alternating firms*, as in our basic model, she instead offers $d_{\text{H}}^{\text{AU}^*} = 1$ (i.e., the whole dataset), as opposed to the amount of data $d_{\text{H}}^{\text{AU}^*} = \frac{4}{3n}$ offered under our baseline timing. By doing so, the DB minimises firms' profits when they lose, and in turn the number of entering firms. Although the DB cannot maximise her profits by maximising winning firms' profits, she can still do so by minimising competition in the downstream market.

Finally, under Take It Or Leave It (TIOLI), the DB offers the whole dataset to all entering firms, regardless of n. The aim of this strategy is again to minimise competition in the downstream market to extract higher profits, and it departs from the strategy adopted under our baseline timing, which depends on the number of entering firms.

To sum up, we find that if the data sale occurs before firms' entry, the DB always maximises her entry barrier effect. As her bargaining power is reduced, the DB floods the downstream market with data as a way to decrease firms' expected profits and, in turn, their entry. Her strategy ultimately harms consumers, who are always worse off than in the benchmark due to the increase of downstream market's concentration. Nonetheless, we find that consumer harm is minimised under TIOLI, consistently with the result obtained in Section 5.2.

6. Conclusions

With the steady growth of online services, DBs have become central players in the digital economy. Their ability to extract valuable information from consumers' data allows them to influence competition in retail markets, with important welfare implications. Our work contributes to the growing literature on the competitive effects of DBs by modelling an oligopoly market where the number of firms is endogenous.

We show that the presence of a DB reduces the entry of firms in the downstream market. The DB benefits from the increased concentration, as she can then extract firms' profits through the price of data. Previous literature on price discrimination in spatial competition settings has often highlighted a pro-competitive effect of data, as firms engage in price wars over the identified consumers. We show that, when entry endogenously depends on the DB's strategy, the entry barrier effect dominates the competition effect, leading to an overall decrease in competition in the market.

We also find that the DB has the incentive to under-serve the market by selling data to only a subset of firms. This result expands the insight developed in previous literature (Braulin and Valletti, 2016; Montes et al., 2019; Bounie et al., 2021), which advocates for a ban on exclusive data deals to benefit of consumers. We show that, when the number of firms is endogenous, the ban of exclusive deals should address single portions of the market. However, we also find that alternative mechanisms for the data sale (e.g., Take It Or Leave It offers) could induce the DB to avoid exclusive deals.

Overall, our results show that consumer surplus is lower in the presence of a monopolistic DB, while total welfare is mostly appropriated by the DB. As a consequence, if the weight of the DB's profits in the welfare function is sufficiently low, the presence of a DB is welfare decreasing.

The use of data by firms has implications not only for competition, but also in terms of privacy. We show that, if the use of data for price discrimination entails a privacy loss to consumers, the potential of the DB to raise entry barriers through data sales is reduced. As more firms enter the market, competition and consumer surplus increase. Thus, our results imply that an increase in consumers' privacy awareness can limit the consumer harm caused by the DB.

From a policymaking point of view, our results suggest that the presence of a DB that can manipulate the competitive dynamics by raising entry barriers is detrimental for consumers, despite the fact that the use of data intensifies competition between firms. However, we also find that consumer surplus can be raised by properly regulating the DB's selling mechanism. In particular, a competition authority could either mandate the sale of data to all entering firms, or enforce the use of direct sales (i.e., TIOLI offers). Such policies would effectively lower the DB's bargaining power, but would also lead to an increase in the amount of data sold. Therefore, the ensuing increase in competition would also be accompanied by a lower degree of consumer privacy.

Finally, we find that the DB's negative effect on welfare is stronger if firms purchase data before they decide to enter the market, as in this scenario the DB always chooses a strategy that minimises firms' entry, further reducing consumer surplus. Similarly to what we find in the basic model, reducing the DB's bargaining power would have positive effects on consumers. However, it would also lead the DB to flood the downstream market with data, as this strategy allows her to minimise the number of entering firms.

An important issue that remains to be addressed deals with the presence of competition in the collection of data at the DB's level. Indeed, competition between DBs would further limit the individual DB's bargaining power, possibly tempering their entry barrier effect. A careful analysis is needed to fully assess the implications of competition between DBs for entry in the downstream market and for consumers.

Acknowledgement: This work has been partially supported by Ministero dell'Istruzione, dell'Università e della Ricerca Award TESUN-83486178370409 finanziamento dipartimenti di eccellenza CAP. 1694 TIT. 232 ART. 6.

Bibliography

Acquisti, A., and Varian, H. R. (2005). Conditioning Prices on Purchase History. *Marketing Science*, *24*(3), 367–381. https://doi.org/10.1287/mksc.1040.0103

Anderson, S., and Bedre-Defolie, Ö. (2021). *Hybrid Platform Model* [mimeo, available on SSRN at 10.2139/ssrn.3867851].

Aparicio, D., Metzman, Z., and Rigobon, R. (2021). *The pricing strategies of online grocery retailers* [NBER working paper n. 28639].

Armstrong, M., and Vickers, J. (2001). Competitive Price Discrimination. *The RAND Journal of Economics*, *32*(4), 579–605. https://doi.org/10.2307/2696383

Belleflamme, P., and Vergote, W. (2016). Monopoly price discrimination and privacy: The hidden cost of hiding. *Economics Letters*, *149*, 141–144. https://doi.org/10.1016/j.econlet.2016.10.027

Bergemann, D., and Bonatti, A. (2011). Targeting in advertising markets: Implications for offline versus online media. *The RAND Journal of Economics*, *42*(3), 417–443. https://doi.org/10.1111/j.1756-2171.2011.00143.x

Bergemann, D., and Bonatti, A. (2015). Selling Cookies. *American Economic Journal: Microeconomics*, *7*(3), 259–294. https://doi.org/10.1257/mic.20140155

Bergemann, D., and Bonatti, A. (2019). Markets for Information: An Introduction. *Annual Review of Economics*, *11*, 85–107. https://doi.org/10.1146/annurev-economics-080315-015439

Bester, H., and Petrakis, E. (1996). Coupons and oligopolistic price discrimination. *International Journal of Industrial Organization*, *14*(2), 227–242. https://doi.org/10.1016/0167-7187(94)00469-2

Bounie, D., Dubus, A., and Waelbroeck, P. (2021). Selling strategic information in digital competitive markets. *The RAND Journal of Economics*, *52*(2), 283–313. https://doi.org/10.1111/1756-2171.12369

Braulin, F. C., and Valletti, T. (2016). Selling customer information to competing firms. *Economics Letters*, *149*, 10–14. https://doi.org/10.1016/j.econlet.2016.10.005

Chen, Z., Choe, C., and Matsushima, N. (2020). Competitive Personalized Pricing. *Management Science*, *66*(9), 4003–4023. https://doi.org/10.1287/mnsc.2019.3392

Colombo, S. (2018). Behavior- and characteristic-based price discrimination. *Journal of Economics & Management Strategy*, *27*(2), 237–250. https://doi.org/10.1111/jems.12244

Crain, M. (2018). The limits of transparency: Data brokers and commodification. *New Media & Society*, *20*(1), 88–104. https://doi.org/10.1177/1461444816657096

de Cornière, A., and Taylor, G. (2020). *Data and Competition: A General Framework with Applications to Mergers, Market Structure, and Privacy Policy* (SSRN Scholarly Paper No. ID 3547379). Social Science Research Network. Rochester, NY. Retrieved March 1, 2022, from https://papers.ssrn.com/abstract=3547379

de Cornière, A. (2016). Search Advertising. *American Economic Journal: Microeconomics*, *8*(3), 156–188. https://doi.org/10.1257/mic.20130138

Delbono, F., Reggiani, C., and Sandrini, L. (2021). *Strategic data sales to competing firms* (Technical Report JRC126568). JRC Digital Economy Working Paper. Seville, Spain. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/strategic-data-sales-competing-firms

FTC. (2014). *Data brokers: A call for transparency and accountability* (tech. rep.). Federal Trade Commission. Washington, DC.

Gu, Y., Madio, L., and Reggiani, C. (2019). *Exclusive Data, Price Manipulation and Market Leadership* (SSRN Scholarly Paper No. ID 3467988). Social Science Research Network. Rochester, NY. https://papers.ssrn.com/abstract=3467988.

Hagiu, A., and Wright, J. (2020). Data-enabled learning, network effects and competitive advantage. *Working Paper.*

Hart, O. D., and Tirole, J. (1988). Contract Renegotiation and Coasian Dynamics. *The Review of Economic Studies*, *55*(4), 509–540. https://doi.org/10.2307/2297403

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

Jehiel, P., and Moldovanu, B. (2000). Auctions with Downstream Interaction among Buyers. *The RAND Journal of Economics*, *31*(4), 768–791. https://doi.org/10.2307/2696358

Johnen, J. (2020). Dynamic competition in deceptive markets. *The RAND Journal of Economics*, *51*(2), 375–401. https://doi.org/10.1111/1756-2171.12318

Judiciary Committee. (2020). *Investigation of Competition in Digital Markets: Majority Staff Report and Recommendations* (tech. rep.). US House of Representatives. https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf

Liu, Q., and Serfes, K. (2004). Quality of Information and Oligopolistic Price Discrimination. *Journal of Economics & Management Strategy*, *13*(4), 671–702. https://doi.org/10.1111/j.1430-9134.2004.00028.x

Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. (2012). Detecting price and search discrimination on the internet. *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, 79–84. https://doi.org/10.1145/2390231.2390245

Montes, R., Sand-Zantman, W., and Valletti, T. (2019). The Value of Personal Information in Online Markets with Endogenous Privacy. *Management Science*, *65*(3), 955–1453. https://doi.org/10.1287/mnsc.2017.2989

Rhodes, A., and Zhou, J. (2021). Personalized Pricing and Privacy Choice. *Working Paper*. https://www.economics.utoronto.ca/index.php/index/research/downloadSeminarPaper/997875056

Salop, S. C. (1979). Monopolistic Competition with Outside Goods. *The Bell Journal of Economics*, *10*(1), 141–156. https://doi.org/10.2307/3003323

Searle, S. R. (1979). On inverting circulant matrices. *Linear Algebra and its Applications*, *25*, 77–89. https://doi.org/10.1016/0024-3795(79)90007-7

Shaffer, G., and Zhang, Z. J. (1995). Competitive Coupon Targeting. *Marketing Science*, *14*(4), 395–416. https://doi.org/10.1287/mksc.14.4.395

Shy, O., and Stenbacka, R. (2016). Customer Privacy and Competition. *Journal of Economics & Management Strategy*, *25*(3), 539–562. https://doi.org/10.1111/jems.12157

Taylor, C. R. (2003). Supplier Surfing: Competition and Consumer Behavior in Subscription Markets. *The RAND Journal of Economics*, *34*(2), 223–246. https://doi.org/10.2307/1593715

Taylor and Wagman, L. (2014). Consumer privacy in oligopolistic markets: Winners, losers, and welfare. *International Journal of Industrial Organization*, *34*, 80–84. https://doi.org/10.1016/j.ijindorg.2014.02.010

Thisse, J.-F., and Vives, X. (1988). On The Strategic Choice of Spatial Price Policy. *The American Economic Review*, *78*(1), 122–137.

Vickrey, W. S. (1964). *Microstatics*. Harcourt, Brace & World, Inc.

Villas-Boas, J. M. (2004). Price Cycles in Markets with Customer Recognition. *The RAND Journal of Economics*, *35*(3), 486–501. https://doi.org/10.2307/1593704

Appendix

I. Proof of Proposition 1

This proof proceeds in two steps. First, we demonstrate that the DB sells equally sized partitions to all even-indexed firms and to all odd-indexed firms. In particular, the DB can only offer partitions of size $\widehat{d}$ to even-indexed firms and of size $\widetilde{d}$ to odd-indexed firms. Thus, the DB only has four viable strategies:

1. Setting $\widehat{d} \neq \widetilde{d}, \widehat{d} > 0,\ \widetilde{d} > 0$. That is, she sells partitions of different sizes to all firms, alternating between the partitions' sizes.
2. Setting $\widehat{d} = \widetilde{d} > 0$. That is, she sells equally sized partitions to all firms.
3. Setting $\widehat{d} \neq \widetilde{d}, \widehat{d} = 0,\ \widetilde{d} > 0$ That is, she only sells data to odd-indexed firms, while she does not offer data to even-indexed firms
4. Setting $\widehat{d} = \widetilde{d} = 0$. That is, she does not sell data.

Second, we show that strategies 1. and 4. are always suboptimal for the DB, and thus that only strategies 2. and 3. can be the only Nash equilibria in pure strategies.

(I) The DB Sells Equally Sized Partitions to All Even-Indexed Firms and to All Odd-Indexed Firms

DB's profits are equal to the sum of the difference between firms' profits when they obtain their respective partition and when they do not obtain it:

$$\pi_{\text{DB}} = \sum_{i=0}^{n-1} \Delta\pi_i \tag{A.1}$$

where

$$\Delta\pi_0 = f\left(d_0, d_1, \ldots, d_{i-1}, d_i, \ldots, d_{n-1}\right)$$
$$\Delta\pi_1 = g\left(d_0, d_1, \ldots, d_{i-1}, d_i, \ldots, d_{n-1}\right)$$
$$\Delta\pi_2 = h\left(d_0, d_1, \ldots, d_{i-1}, d_i, \ldots, d_{n-1}\right)$$
$$\ldots$$

The different functions (e.g., $f, g, h \ldots$) derive from the fact that, while firms' profits depend on all firms' partitions, the way they do depends on the distribution of the partitions compared to the analysed firm's location.

Let us focus on $\Delta\pi_0$: the profits that the DB extracts from firm 0 depend on the partitions she sells to all firms. Not every partition influences $\Delta\pi_0$ in the same way. For example, the partition obtained by firm 1 (i.e., d_1) will have a different effect on $\Delta\pi_0$

than the partition obtained by firm 2 (i.e., d_2) has on $\Delta\pi_0$. However, the symmetry of the model allows us to draw conclusions regarding the effects of partitions obtained by firms that are equidistant from firm 0. Since firm 1 and firm $n{-}1$ are identical and both distant $\frac{1}{n}$ from firm 0, their partitions have the same effect on $\Delta\pi_0$. The same holds true for every pair of firms that are equidistant from firm 0, since firms are equally spaced on the circle. We thus have

$$\frac{\partial \Delta\pi_0}{\partial d_j} = \frac{\partial \Delta\pi_0}{\partial d_{n-j}} \qquad \forall\, j \in \left\{0, 1, 2, \dots, \frac{n}{2}\right\} \tag{A.2}$$

By symmetry, condition (A.2) can be applied to all i, obtaining

$$\frac{\partial \Delta\pi_i}{\partial d_j} = \frac{\partial \Delta\pi_i}{\partial d_{n-j}} \qquad \forall\, j \in \left\{0, 1, 2, \dots, \frac{n}{2}\right\},\ i \in \{0, 1, 2, \dots, n{-}1\} \tag{A.3}$$

Let us now focus on the relationship between $\Delta\pi_0$ and $\Delta\pi_1$. By the symmetry of firms, the effect of d_1 on $\Delta\pi_1$ is same that d_0 has on $\Delta\pi_0$. The same holds true when analysing the effect of equally distant firms. As an example, the effect of d_1 on $\Delta\pi_0$ is same of d_0 on $\Delta\pi_1$, as it is the effect that a direct rival's partition has on the analysed firm's profits difference. We can thus write

$$\frac{\partial \Delta\pi_i}{\partial d_{i+j}} = \frac{\partial \Delta\pi_k}{\partial d_{k+j}} \qquad \forall\, i, k \in \{0, 1, 2, \dots, n{-}1\},\ j \in \left\{0, 1, 2, \dots, \frac{n}{2}\right\} \tag{A.4}$$

We can now bring together (A.1), (A.3) and (A.4). The DB chooses the partition set $\mathbf{P} = (d_0, d_1, d_2, \dots, d_{n-1})$ to maximise the sum of $\Delta\pi_i$. Thus, at the equilibrium we have

$$\begin{aligned}
\frac{\partial \pi_{\text{DB}}}{\partial d_0} &= \frac{\partial \Delta\pi_0}{\partial d_0} + \frac{\partial \Delta\pi_1}{\partial d_0} + \frac{\partial \Delta\pi_2}{\partial d_0} + \dots + \frac{\partial \Delta\pi_{n-1}}{\partial d_0} = 0 \\
\frac{\partial \pi_{\text{DB}}}{\partial d_1} &= \frac{\partial \Delta\pi_0}{\partial d_1} + \frac{\partial \Delta\pi_1}{\partial d_1} + \frac{\partial \Delta\pi_2}{\partial d_1} + \dots + \frac{\partial \Delta\pi_{n-1}}{\partial d_1} = 0 \\
\frac{\partial \pi_{\text{DB}}}{\partial d_2} &= \frac{\partial \Delta\pi_0}{\partial d_2} + \frac{\partial \Delta\pi_1}{\partial d_2} + \frac{\partial \Delta\pi_2}{\partial d_2} + \dots + \frac{\partial \Delta\pi_{n-1}}{\partial d_2} = 0 \\
&\dots
\end{aligned}$$

From (A.4), we know that $\frac{\partial \Delta\pi_1}{\partial d_0} = \frac{\partial \Delta\pi_0}{\partial d_1}$, and the same can be applied to all the elements on the right side of the equation. We can thus rewrite $\frac{\partial \pi_{\text{DB}}}{\partial d_0}$ as

$$\frac{\partial \pi_{\text{DB}}}{\partial d_0} = \frac{\partial \Delta\pi_0}{\partial d_0} + \frac{\partial \Delta\pi_0}{\partial d_1} + \frac{\partial \Delta\pi_0}{\partial d_2} + \dots = \sum_{i=0}^{n-1} \frac{\partial \Delta\pi_0}{\partial d_i}$$

By applying (A.4), we can trace back every partial derivative of π_{DB} to the same form. In a general form, we obtain

$$\frac{\partial \pi_{\text{DB}}}{\partial d_k} = \sum_{i=0}^{n-1} \frac{\partial \Delta\pi_0}{\partial d_i} \qquad \forall\, k \in \{0, 1, 2, \dots, n{-}1\} \tag{A.5}$$

Equation (A.5) implies that the DB's profits are influenced in the same way by the partitions sold to any firm. Thus, the DB aims to maximise a given firm's difference in profits, and then she applies the same strategy to all other firms. We focus our analysis on a generic firm i. From (A.3), we know that $\Delta\pi_i$ is influenced in the same way by d_{i+1} and d_{i-1}. As such, in equilibrium the DB sets $d_{i+1} = d_{i-1}$. The same holds true for any pair d_{i+j}, d_{i-j}, as described in (A.3).The same reasoning can be applied when focusing on all other firms. For example, by looking at firm 1 we can conclude that $d_{1+j} = d_{1-j} \ \forall \ j \ \in \left\{0, 1, 2, \ldots, \frac{n}{2}\right\}$. By putting together all the equations, we find that the DB sells equally sized partitions to all even-indexed firms, which we denote as $\widehat{d}$, and equally sized partitions to all odd-indexed firms, which we denote as $\widetilde{d}$. Note how, under this strategy, a firm always faces direct rivals that obtain same sized partitions: as such, in equilibrium the DB offers partitions that are centred on firms' locations. Suppose that firm i is offered a partition $\widetilde{d}$: since both its direct rivals obtain $\widehat{d}$, in equilibrium the DB sells firm i a partition such that it can identify consumer segments of size $\frac{\widetilde{d}}{2}$ on each arc on which it competes.

(II) the DB Either Sells Equally Sized Partitions to All Firms or Equally Sized Partitions to Alternating Firms

Step (I) leaves the DB with four possible strategies, as priorly described. We now want to demonstrate that strategies 1. and 4. are suboptimal for the DB. First, we can discard strategy 4.: since in our model the DB does not sustain any costs, her minimum profits are 0. As such, a strategy where the DB sets $\widehat{d} = \widetilde{d} = 0$, which results in her profits being 0, can never dominate any other strategy.

We move on to strategy 1., where the DB sets $\widehat{d} \neq \widetilde{d}, \widehat{d} > 0, \widetilde{d} > 0$. We show that this strategy is always dominated by strategy 2., where the DB sets $\widehat{d} = \widetilde{d} > 0$. To do so, we solve the model under strategy 1. The DB offers a partition set $\mathbf{P} = \left(\widetilde{d}, \widehat{d}, \widetilde{d}, \ldots, \widehat{d}\right)$. Without loss of generality, we focus on a generic firm i, to which the DB offers a partition $\widetilde{d}$. The indifferent consumers between firms i, i+1 and i−1 can be obtained by equating utility levels, and they are:

$$\widehat{x}_{i-1,i} = \frac{2i-1}{2n} + \frac{p_i^{\mathrm{B}} - p_{i-1}^{\mathrm{B}}}{2t} \qquad \text{and} \qquad \widehat{x}_{i,i+1} = \frac{2i+1}{2n} + \frac{p_{i+1}^{\mathrm{B}} - p_i^{\mathrm{B}}}{2t} \tag{A.6}$$

Firm i offers a tailored price $p_i^{\mathrm{T}}(x)$ to the identified consumers, matching the competitor's offer in utility level. It sets a tailored price for each arc where it competes, resulting in

$$p_i^{\mathrm{T}}(x) = \begin{cases} p_{i-1}^{\mathrm{B}} + 2tx - \frac{t}{n}(2i-1) & \text{for } \; x \in [\frac{i}{n} - \frac{d_i}{2}, \frac{i}{n}] \\ p_{i+1}^{\mathrm{B}} - 2tx + \frac{t}{n}(2i+1) & \text{for } \; x \in [\frac{i}{n}, \frac{i}{n} + \frac{d_i}{2}] \end{cases} \tag{A.7}$$

Firm i's profits are thus given by:

$$\pi_i^{\mathrm{W}}(\mathbf{P}) = \int_{\frac{i}{n}-\frac{\widetilde{d}}{2}}^{\frac{i}{n}} p_i^{\mathrm{T}}(x)\,dx + \int_{\frac{i}{n}}^{\frac{i}{n}+\frac{\widetilde{d}}{2}} p_i^{\mathrm{T}}(x)\,dx + p_i^{\mathrm{B}}(\mathbf{P})\left(\widehat{x}_{i,i+1} - \widehat{x}_{i-1,i} - \widetilde{d}\right) - F \tag{A.8}$$

Using the expression of the indifferent consumers from (A.6) and of the tailored prices in (A.7), we can rewrite the profits of the generic informed firm i in (A.8) as

$$\begin{aligned} \pi_i^{\mathrm{W}}(\mathbf{P}) = {} & \frac{\widetilde{d}}{2n}\left(2t + \mathrm{np}_{i-1}^{\mathrm{B}}(\mathbf{P}) + \mathrm{np}_{i+1}^{\mathrm{B}}(\mathbf{P}) - nt\widetilde{d}\right) \\ & + p_i^{\mathrm{B}}(\mathbf{P})\left(\frac{n\left(p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P}) - 2p_i^{\mathrm{B}}(\mathbf{P})\right) + 2t}{2nt} - \widetilde{d}\right) - F \end{aligned} \tag{A.9}$$

Similarly, the profits of its rival i+1 firm are

$$\begin{aligned} \pi_{i+1}^{\mathrm{W}}(\mathbf{P}) = {} & \frac{\widehat{d}}{2n}\left(2t + np_i^{\mathrm{B}}(\mathbf{P}) + np_{i+2}^{\mathrm{B}}(\mathbf{P}) - nt\widehat{d}\right) \\ & + p_{i+1}^{\mathrm{B}}(\mathbf{P})\left(\frac{n\left(p_i^{\mathrm{B}}(\mathbf{P}) + p_{i+2}^{\mathrm{B}}(\mathbf{P}) - 2p_{i+1}^{\mathrm{B}}(\mathbf{P})\right) + 2t}{2nt} - \widehat{d}\right) - F \end{aligned} \tag{A.10}$$

By taking the first-order condition of (A.9) with respect to $p_i^{\mathrm{B}}(\mathbf{P})$ and of (A.10) with respect to $p_{i+1}^{\mathrm{B}}(\mathbf{P})$, we obtain firms' reaction function on basic prices

$$\begin{gathered} p_i^{\mathrm{B}}(\mathbf{P}) = \frac{t}{2n} - \frac{t\widetilde{d}}{2} + \frac{p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P})}{4} \\ \text{and} \\ p_{i+1}^{\mathrm{B}}(\mathbf{P}) = \frac{t}{2n} - \frac{t\widehat{d}}{2} + \frac{p_i^{\mathrm{B}}(\mathbf{P}) + p_{i+2}^{\mathrm{B}}(\mathbf{P})}{4} \end{gathered} \tag{A.11}$$

The system of equations (A.11) for all $i = 0, \ldots, n-1$ allows us to obtain the equilibrium basic prices and, by replacing them in (A.9), firm i's profits. In matrix form we have $\mathbf{A} * \mathbf{p} = \mathbf{b}$, where $\mathbf{p}$ is the price vector, and $\mathbf{b}$ is the known terms vector. Assuming that the DB offers $\widetilde{d}$ to even indexed firms, we obtain

$$\begin{bmatrix} 4 & -1 & \dots & 0 & 0 & 0 & \dots & -1 \\ -1 & 4 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 4 & -1 & 0 & \dots & 0 \\ 0 & 0 & \dots & -1 & 4 & -1 & \dots & 0 \\ 0 & 0 & \dots & 0 & -1 & 4 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & \dots & 0 & 0 & 0 & \dots & 4 \end{bmatrix} * \begin{bmatrix} p_0^{\mathrm{B}}(\mathbf{P}) \\ p_1^{\mathrm{B}}(\mathbf{P}) \\ \dots \\ p_{i-1}^{\mathrm{B}}(\mathbf{P}) \\ p_i^{\mathrm{B}}(\mathbf{P}) \\ p_{i+1}^{\mathrm{B}}(\mathbf{P}) \\ \dots \\ p_{n-1}^{\mathrm{B}}(\mathbf{P}) \end{bmatrix} = \begin{bmatrix} \frac{2t}{n} - 2t\widetilde{d} \\ \frac{2t}{n} - 2t\widehat{d} \\ \dots \\ \frac{2t}{n} - 2t\widehat{d} \\ \frac{2t}{n} - 2t\widetilde{d} \\ \frac{2t}{n} - 2t\widehat{d} \\ \dots \\ \frac{2t}{n} - 2t\widehat{d} \end{bmatrix}$$

Matrix $\mathbf{A}$ is circulant, tridiagonal and symmetric. The inverse of this type of matrix has been computed by Searle (1979). We obtain

$$A^{-1} = \begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{n-1} & a_0 & \dots & a_{n-2} \\ \dots & \dots & \dots & \dots \\ a_1 & a_2 & \dots & a_0 \end{bmatrix}$$

where, in our specific case, $a_j = -\frac{1}{2\sqrt{3}} * \left(\frac{(2+\sqrt{3})^j}{1-(2+\sqrt{3})^n} - \frac{(2-\sqrt{3})^j}{1-(2-\sqrt{3})^n} \right)$. A property of this type of matrix is that $a_j = a_{n-j}\ \forall j \neq 0, \frac{n}{2}$ if n is even. Moreover, in our particular case, $\sum_{j=0}^{n-1} a_j = \frac{1}{2}$. We can now write $\mathbf{p} = \mathbf{A^{-1}} * \mathbf{b}$. We obtain

$$\begin{bmatrix} p_0 \\ p_1 \\ \dots \\ p_{n-1} \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{n-1} & a_0 & \dots & a_{n-2} \\ \dots & \dots & \dots & \dots \\ a_1 & a_2 & \dots & a_0 \end{bmatrix} * \begin{bmatrix} \frac{2t}{n} - 2t\widetilde{d} \\ \frac{2t}{n} - 2t\widehat{d} \\ \dots \\ \frac{2t}{n} - 2t\widehat{d} \end{bmatrix}$$

Thus, we can write

$$p_i^{\mathrm{B}} = \left(\frac{2t}{n} * \sum_{j=0}^{n-1} a_j \right) - 2t \sum_{j=0}^{\frac{n-2}{2}} \widetilde{d} a_{2j} - 2t \sum_{j=0}^{\frac{n-2}{2}} \widehat{d} a_{2j+1}$$

Since $\sum_{j=0}^{n-1} a_j = \frac{1}{2}$, we can simplify and obtain

$$p_i^{\mathrm{B}} = \frac{t}{n} - 2t \sum_{j=0}^{\frac{n-2}{2}} \widetilde{d} a_{2j} - 2t \sum_{j=0}^{\frac{n-2}{2}} \widehat{d} a_{2j+1}$$

Due to the symmetry properties of the coefficients a_j, we also obtain a similar form for p_{i-1}^{B} and p_{i-1}^{B}:

$$p_{i-1}^{\mathrm{B}} = p_{i+1}^{\mathrm{B}} = \frac{t}{n} - 2t\sum_{j=0}^{\frac{n-2}{2}} \widehat{d}a_{2j} - 2t\sum_{j=0}^{\frac{n-2}{2}} \widetilde{d}a_{2j+1}$$

We find that in our case $\sum_{j=0}^{\frac{n-2}{2}} a_{2j} = \frac{1}{3}$ and $\sum_{j=0}^{\frac{n-2}{2}} a_{2j+1} = \frac{1}{6}$. Thus, we can rewrite basic prices as

$$p_i^{\mathrm{B}} = \frac{t}{n} - \frac{2}{3}t\widetilde{d} - \frac{1}{3}t\widehat{d} \qquad \text{and} \qquad p_{i-1}^{\mathrm{B}} = p_{i+1}^{\mathrm{B}} = \frac{t}{n} - \frac{2}{3}t\widehat{d} - \frac{1}{3}t\widetilde{d} \tag{A.12}$$

By replacing the basic prices from (A.12) in firms' profits functions (A.9) and (A.10), we obtain

$$\pi_i^{\mathrm{W}}(\mathbf{P}) = \frac{t}{9n^2}\left(9 - 2n\left(n\widetilde{\widehat{d}\widetilde{d}} + \widetilde{d}\left(\frac{7}{4}n\widetilde{d} - 3\right) - n\widehat{d}^2 + 3\widehat{d}\right)\right) - F \tag{A.13}$$

$$\pi_{i-1}^{\mathrm{W}}(\mathbf{P}) = \pi_i^{\mathrm{W}}(\mathbf{P}) = \frac{t}{9n^2}\left(9 - 2n\left(n\widetilde{\widehat{d}\widetilde{d}} + \widehat{d}\left(\frac{7}{4}n\widehat{d} - 3\right) - n\widetilde{d}^2 + 3\widetilde{d}\right)\right) - F \tag{A.14}$$

We now compute firms' profits when they do not obtain their partition. Suppose that firm i does not obtain its partition: as such, in equilibrium $d_i = 0$. By imposing it in (A.9), we obtain that firm i's profits are

$$\pi_i^{\mathrm{L}}(\mathbf{P}) = p_i^{\mathrm{B}}(\mathbf{P})\left(\frac{n\left(p_{i+1}^{\mathrm{B}}(\mathbf{P}) + p_{i-1}^{\mathrm{B}}(\mathbf{P}) - 2p_i^{\mathrm{B}}(\mathbf{P})\right) + 2t}{2nt}\right) - F \tag{A.15}$$

We can again compute firms' basic prices by solving the n-equations system. The only difference from the already analysed subgame is that firm i's known term has $d_i = 0$ instead of $d_i = \widetilde{d}$. As such, we can compute the new basic prices by simply subtracting $\widetilde{d}a_{i-j}$ from the basic prices p_j^{B} computed in (A.12). Thus, we obtain

$$p_i^{\mathrm{B}} = \frac{t}{n} - 2t\widetilde{d}\left(\frac{1}{3} - a_0\right) - \frac{1}{3}t\widehat{d} \quad \text{and} \quad p_{i-1}^{\mathrm{B}} = p_{i+1}^{\mathrm{B}} = \frac{t}{n} - \frac{2}{3}t\widehat{d} - 2t\widetilde{d}\left(\frac{1}{6} - a_1\right) \tag{A.16}$$

By replacing the basic prices of (A.16) in (A.15), we obtain

$$\pi_i^{\mathrm{L}}(\mathbf{P}) = \frac{t}{9n^2}\left(6a_0n\widetilde{d} - 2n\widetilde{d} - n\widehat{d} + 3\right)\left(-6a_0n\widetilde{d} + 6a_1n\widetilde{d} + n\widetilde{d} - n\widehat{d} + 3\right) - F \tag{A.17}$$

Following the same procedure, we obtain firm i+1's profits in the subgame where it does not obtain data:

$$\pi_{i+1}^{\mathrm{L}}(\mathbf{P}) = \frac{t}{9n^2}\left(6a_0n\widehat{d} - 2n\widehat{d} - n\widetilde{d} + 3\right)\left(-6a_0n\widehat{d} + 6a_1n\widehat{d} + n\widehat{d} - n\widetilde{d} + 3\right) - F \tag{A.18}$$

Finally, we compute DB's profits. We can write them as

$$\pi_{\mathrm{DB}} = \frac{n}{2}\left(\pi_i^{\mathrm{W}}(\mathbf{P}) - \pi_i^{\mathrm{L}}(\mathbf{P})\right) + \frac{n}{2}\left(\pi_{i+1}^{\mathrm{W}}(\mathbf{P}) - \pi_{i+1}^{\mathrm{L}}(\mathbf{P})\right) \tag{A.19}$$

Replacing firms' profits from (A.17) and (A.18) and simplifying, we obtain

$$\begin{aligned}\pi_{\mathrm{DB}} = \frac{t}{3}\Big(& 6na_0^2\left(\widetilde{d}^2 + \widehat{d}^2\right) - 6na_0a_1\left(\widetilde{d}^2 + \widehat{d}^2\right) - 3na_0\left(\widetilde{d}^2 + \widehat{d}^2\right) \\ & +2na_1\left(\widetilde{d}^2 + \widehat{d}^2 + \widetilde{d}\widehat{d}\right) - 3a_1\left(\widetilde{d} + \widehat{d}\right) - \frac{n}{4}\left(\widetilde{d}^2 + \widehat{d}^2\right) - n\widetilde{d}\widehat{d} + \frac{3}{2}\widetilde{d} + \frac{3}{2}\widehat{d}\Big) \qquad \text{(A.20)}\end{aligned}$$

By computing FOCs of (A.20) for both $\widetilde{d}$ and $\widehat{d}$, we find that both partitions have the same effect on DB's profits; to maximise them, the DB would set $\widetilde{d} = \widehat{d}$. However, since $\widetilde{d} \neq \widehat{d}$ by hypothesis, we find that setting $\widetilde{d} \neq \widehat{d}$ is suboptimal for the DB. On the other hand, this case does not test the corner solution where either $\widetilde{d} = 0$ or $\widehat{d} = 0$. Thus, we are left with two cases: $\widetilde{d} = \widehat{d}$ or $\widetilde{d} \neq \widehat{d}$, $\widetilde{d} = 0$ or $\widehat{d} = 0$, which are the strategies 2. and 3.

II. Proof of Proposition 2

DB's profits are equal to the sum of the reserve prices w_i, which in turn are set equal to firms' difference in profits between winning or losing their specific auction. We can thus write

$$\pi_{\mathrm{DB}}(\mathbf{P_H}, \mathbf{J}) = \sum_{i \in \mathbf{J}} w_i$$

with

$$w_i = \pi_i^{\mathrm{W}}(\mathbf{P_H}) - \pi_i^{\mathrm{L}}(\mathbf{P_H})$$

From the proof of Proposition 1, we know that under the *sale to alternating firms* the DB only fulfils half of the auctions: as such, the subset $\mathbf{J}$ is given and has a cardinality $j = \frac{n}{2}$. We refer to the partition set sold in equilibrium as $\mathbf{P}^*_{\mathbf{H}}$, while the partition set offered to firms is $\mathbf{P_H}$. The DB can also influence firms' willingness to pay through the auctions she does not fulfil. This is due to the fact if a firm loses its specific auction, other firms can win theirs, and by the DB's ability to claim the maximum number of auctions k she is going to fulfil. First, we focus on firms' profits when they win their respective auctions, which are equal to those at the equilibrium. Therefore, these profits do not depend neither on the auctions that are not fulfilled nor on the maximum number of auctions that will be fulfilled, since in equilibrium the DB will fulfil the subset $\mathbf{J}$ which is given under the *sale to alternating firms*. On the other hand, firms' profits when losing their respective auctions are influenced by k and, in turn, by the partitions offered by the DB in the auctions she does not want to fulfil. Suppose that the DB offers a partition set $\mathbf{P_H} = (d_{\mathrm{H}}, d, \dots, d, d_H, d, \dots, d_{\mathrm{H}}, d)$. At the equilibrium, the DB only fulfils the auctions where she offers d_{H}, resulting in $\mathbf{P}^*_{\mathbf{H}} = (d_{\mathrm{H}}, 0, \dots, 0, d_{\mathrm{H}}, 0, \dots, d_{\mathrm{H}}, 0)$. Suppose that the DB offers d_{H} to firm i, and that she claims $k = \frac{n}{2}$. Then, if firm i loses, the DB could fulfil one of the auctions where she offers d. In particular, she could fulfil the auctions of one of firm i's direct rivals, $i{+}1$ and $i{-}1$. This would reduce firm i's profits when losing its auction, as it would result in it being uninformed while facing an informed rival. As such, the DB wants to set $\mathbf{P_H}$ and k to minimise firms' profits when they lose their respective auctions in which she offers them d_{H}. Intuitively, a firm's profits are minimised when it is uninformed and competing against direct rivals who obtain all consumer data, as already observed by Bounie et al. (2021) in a duopoly setting. Thus, the DB sets $\mathbf{P_H} = (d_{\mathrm{H}}, 1, \ldots, 1, d_H, 1, \ldots, d_{\mathrm{H}}, 1)$ and $k = \frac{n}{2} + 1$. This way, if firm i loses its auction, the DB can fulfil both auctions of its direct rivals $i{+}1$ and $i{-}1$. This minimises firm i's profits when losing its auction, and in turn maximises its willingness to pay.

III. Proof of Proposition 3

We organise this proof by computing equilibrium prices and profits under two cases. First, when winning firms serve both unidentified and identified firms: focusing on firm i, this condition holds as long as

$$\frac{i}{n} + \frac{d_{\mathrm{H}}}{2} < \widehat{x}_{i,i+1} \tag{A.21}$$

that is, as long as winning firms cannot identify the indifferent consumers. When condition (A.21) holds, winning firms set basic prices greater than 0, as they still serve some unidentified consumers. When condition (A.21) is no longer satisfied, winning firms only serve identified consumers and we fall in the second case. Using the non-negative price constraint (see, e.g., Montes et al., 2019 and Bounie et al., 2021), we assume that winning firms set their basic prices equal to 0, and additional data do not longer influence firms' decisions as they do not allow to conquer any new consumers.

(I) Winning Firms Serve Both Identified and Unidentified Consumers

When condition (A.21) holds, we need to solve an n-equations systems to compute firms' equilibrium prices, where the equations alternate between

$$p_i^{\mathrm{B}}(\mathbf{P_H}) = \frac{t}{2n} - \frac{td_{\mathrm{H}}}{2} + \frac{p_{i+1}^{\mathrm{B}}(\mathbf{P_H}) + p_{i-1}^{\mathrm{B}}(\mathbf{P_H})}{4} \quad \text{and} \quad p_{i+1}^{\mathrm{B}}(\mathbf{P_H}) = \frac{t}{2n} + \frac{p_{i+2}^{\mathrm{B}}(\mathbf{P_H}) + p_i^{\mathrm{B}}(\mathbf{P_H})}{4}$$

In matrix form we have $\mathbf{A} * \mathbf{p} = \mathbf{b}$, where $\mathbf{p}$ is the price vector, and $\mathbf{b}$ is the known terms vector. Assuming that the DB offers d_{H} to firms $0, 2, \dots, i{-}2, i, .., n{-}2$, we obtain

$$\begin{bmatrix} 4 & -1 & \dots & 0 & 0 & 0 & \dots & -1 \\ -1 & 4 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 4 & -1 & 0 & \dots & 0 \\ 0 & 0 & \dots & -1 & 4 & -1 & \dots & 0 \\ 0 & 0 & \dots & 0 & -1 & 4 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & \dots & 0 & 0 & 0 & \dots & 4 \end{bmatrix} * \begin{bmatrix} p_0^{\mathrm{B}}(\mathbf{P_H}) \\ p_1^{\mathrm{B}}(\mathbf{P_H}) \\ \dots \\ p_{i-1}^{\mathrm{B}}(\mathbf{P_H}) \\ p_i^{\mathrm{B}}(\mathbf{P_H}) \\ p_{i+1}^{\mathrm{B}}(\mathbf{P_H}) \\ \dots \\ p_{n-1}^{\mathrm{B}}(\mathbf{P_H}) \end{bmatrix} = \begin{bmatrix} \frac{2t}{n} - 2td_{\mathrm{H}} \\ \frac{2t}{n} \\ \dots \\ \frac{2t}{n} \\ \frac{2t}{n} - 2td_{\mathrm{H}} \\ \frac{2t}{n} \\ \dots \\ \frac{2t}{n} \end{bmatrix}$$

Matrix $\mathbf{A}$ is circulant, tridiagonal and symmetric. The inverse of this type of matrix has been computed by Searle (1979). We obtain

$$A^{-1} = \begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{n-1} & a_0 & \dots & a_{n-2} \\ \dots & \dots & \dots & \dots \\ a_1 & a_2 & \dots & a_0 \end{bmatrix}$$

Where, in our specific case, $a_j = -\frac{1}{2\sqrt{3}} * \left(\frac{\left(2+\sqrt{3}\right)^j}{1-\left(2+\sqrt{3}\right)^n} - \frac{\left(2-\sqrt{3}\right)^j}{1-\left(2-\sqrt{3}\right)^n} \right)$. It is worth noting that $a_j > a_{j+1} \quad \forall j \in \{0, \frac{n}{2}{-}1\}$. Another property of this type of matrix is that $a_j = a_{n-j}\ \forall j$. Moreover, in our particular case, $\sum_{j=0}^{n-1} a_j = \frac{1}{2}$. We can now compute the vector of prices p as $\mathbf{p} = \mathbf{A^{-1}} * \mathbf{b}$. We can thus write

$$\begin{bmatrix} p_0^{\mathrm{B}}(\mathbf{P_H}) \\ p_1^{\mathrm{B}}(\mathbf{P_H}) \\ \dots \\ p_{i-1}^{\mathrm{B}}(\mathbf{P_H}) \\ p_i^{\mathrm{B}}(\mathbf{P_H}) \\ p_{i+1}^{\mathrm{B}}(\mathbf{P_H}) \\ \dots \\ p_{n-1}^{\mathrm{B}}(\mathbf{P_H}) \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & \dots & a_{i-1} & a_i & a_{i+1} & \dots & a_{n-1} \\ a_{n-1} & a_0 & \dots & a_{i-2} & a_{i-1} & a_i & \dots & a_{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n-i+1} & a_{n-i+2} & \dots & a_0 & a_1 & a_2 & \dots & a_{n-i} \\ a_{n-i} & a_{n-i+1} & \dots & a_{n-1} & a_0 & a_1 & \dots & a_{n-i-1} \\ a_{n-i-1} & a_{n-i} & \dots & a_{n-2} & a_{n-1} & a_0 & \dots & a_{n-i-2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_1 & a_2 & \dots & a_i & a_{i+1} & a_{i+2} & \dots & a_0 \end{bmatrix} * \begin{bmatrix} \frac{2t}{n} - 2td_{\mathrm{H}} \\ \frac{2t}{n} \\ \dots \\ \frac{2t}{n} \\ \frac{2t}{n} - 2td_{\mathrm{H}} \\ \frac{2t}{n} \\ \dots \\ \frac{2t}{n} \end{bmatrix}$$

We first focus on p_i^{B}. We have

$$p_i^{\mathrm{B}}(\mathbf{P_H}) = \left(\frac{2t}{n} * \sum_{j=0}^{n-1} a_j \right) - 2td_{\mathrm{H}} \left(a_{n-i} + a_{n-i+2} + \dots + a_0 + \dots + a_{n-i-2} \right) \tag{A.22}$$

Using the properties described before, we can rewrite (A.22) as

$$p_{\mathrm{i}}^{\mathrm{B}*}\left(\mathbf{P_H}\right)=\frac{t}{n}-2td_{\mathrm{H}}\left(a_0+a_{\frac{n}{2}}+2\sum_{j=1}^{\frac{n}{4}-1}a_{2j}\right)=\frac{t}{n}-\frac{2}{3}td_{\mathrm{H}} \tag{A.23}$$

Following the same method, prices of uninformed firms are

$$p_{i+1}^{\mathrm{B}*}\left(\mathbf{P_H}\right)=p_{i-1\ \mathrm{H}}^{\mathrm{B}*}=\frac{t}{n}-\frac{1}{3}td_{\mathrm{H}} \tag{A.24}$$

We obtain indifferent consumers' locations by replacing (A.23) and (A.24) in (1), obtaining

$$\widehat{x}_{i-1,i}=\frac{2i-1}{2n}-\frac{d_{\mathrm{H}}}{6}\quad\text{and}\quad\widehat{x}_{i,i+1}=\frac{2i+1}{2n}+\frac{d_{\mathrm{H}}}{6} \tag{A.25}$$

We can compute firm i's profits by replacing (A.23), (A.24) and (A.25) in firm i's profits function, obtaining

$$\pi_i^{\mathrm{W}*}\left(\mathbf{P_H}\right)=\frac{t}{n^2}+\frac{2d_{\mathrm{H}}t}{3n}-\frac{7td_{\mathrm{H}}^2}{18}-F$$

In the same way we can compute firm i+1's profits, leading to

$$\pi_{i+1}^{\mathrm{L}*}\left(\mathbf{P_H}\right)=\frac{t}{n^2}-\frac{2d_{\mathrm{H}}t}{3n}+\frac{\mathrm{td}_H^2}{9}-F$$

(II) Firms Only Serve Identified Consumers

We now focus on the case where winning firms only serve identified consumers. This happens when

$$\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}\geq\widehat{x}_{i,i+1}=\frac{2i+1}{2n}+\frac{d_{\mathrm{H}}}{6} \tag{A.26}$$

Solving (A.26), we obtain that firms only serve identified consumers when

$$d_{\mathrm{H}}\geq\frac{3}{2n} \tag{A.27}$$

While (A.27) holds, firm i sets its basic price $p_{\mathrm{i\ H}}^{\mathrm{B}*}=0$, and we can rewrite its profits function as

$$\pi_i^{\mathrm{W}}\left(\mathbf{P_H}\right)=\int_{\widehat{x}_{i-1,i}}^{\frac{i}{n}}p_i^{\mathrm{T}}(x)\,dx+\int_{\frac{i}{n}}^{\widehat{x}_{i,i+1}}p_i^{\mathrm{T}}(x)\,dx-F$$

While its rival's profits are

$$\pi_{i+1}^{\mathrm{L}}\left(\mathbf{P_H}\right)=p_{i+1}^{\mathrm{B}}\left(\mathbf{P_H}\right)\left(\frac{n\left(p_{i+2}^{\mathrm{B}}\left(\mathbf{P_H}\right)+p_i^{\mathrm{B}}\left(\mathbf{P_H}\right)-2p_{i+1}^{\mathrm{B}}\left(\mathbf{P_H}\right)\right)+2t}{2nt}\right)-F \tag{A.28}$$

Since $p_{i+2}^{\mathrm{B}}\left(\mathbf{P_H}\right)=p_i^{\mathrm{B}}\left(\mathbf{P_H}\right)=0$, We can derive $p_{i+1}^{\mathrm{B}}\left(\mathbf{P_H}\right)$ by taking the FOCs of (A.28), obtaining $p_{i+1}^{\mathrm{B}*}\left(\mathbf{P_H}\right)=\frac{t}{2n}$. The same reasoning can be applied to firm i−1 due to symmetry. By replacing the basic prices in the profits functions, we obtain

$$\pi_i^{\mathrm{W}*}\left(\mathbf{P_H}\right)=\frac{9t}{8n^2}-F \quad \text{and} \quad \pi_{i+1}^{\mathrm{L}*}\left(\mathbf{P_H}\right)=\frac{t}{4n^2}-F$$

IV. Proof of Proposition 4

When firm i loses its auction, both its rivals obtain the whole dataset (i.e., $d_{i+1}=d_{i-1}=1$). This subgame is specular to the case where firm i wins the auction and $d_i \geq \frac{3}{2n}$. As such, firm i's basic price and profits are equal to firm $i+1$'s ones in Step (II) of the proof of Proposition 3, leading to

$$p_i^{\mathrm{B}*}\left(\mathbf{P_H}\right)=\frac{t}{2n} \quad \text{and} \quad \pi_i^{\mathrm{L}*}\left(\mathbf{P_H}\right)=\frac{t}{4n^2}-F$$

V. Proof of Proposition 5

This proof separately examines two cases. The first is the case in which firm i serves both identified and non-identified consumers. The second is the one in which firm i serves only identified consumers.

(I) Firm i Serves Both Identified and Unidentified Consumers

Let us solve the n-equations system in its matrix form, which is

$$\begin{bmatrix} 4 & -1 & 0 & 0 & \dots & 0 & -1 \\ -1 & 4 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 4 & -1 & \dots & 0 & 0 \\ 0 & 0 & -1 & 4 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 4 & -1 \\ -1 & 0 & 0 & 0 & \dots & -1 & 4 \end{bmatrix} * \begin{bmatrix} p_0^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_1^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_2^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_3^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ \dots \\ p_{n-2}^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_{n-1}^{\mathrm{B}}\left(\mathbf{P_A}\right) \end{bmatrix} = \begin{bmatrix} \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \dots \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \end{bmatrix}$$

We can compute the vector of prices as $\mathbf{p}=\mathbf{A^{-1}}*\mathbf{b}$, obtaining

$$\begin{bmatrix} p_0^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_1^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ \dots \\ p_{i-1}^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_i^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ p_{i+1}^{\mathrm{B}}\left(\mathbf{P_A}\right) \\ \dots \\ p_{n-1}^{\mathrm{B}}\left(\mathbf{P_A}\right) \end{bmatrix} = \begin{bmatrix} a_0 & a_1 & \dots & a_{i-1} & a_i & a_{i+1} & \dots & a_{n-1} \\ a_{n-1} & a_0 & \dots & a_{i-2} & a_{i-1} & a_i & \dots & a_{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n-i+1} & a_{n-i+2} & \dots & a_0 & a_1 & a_2 & \dots & a_{n-i} \\ a_{n-i} & a_{n-i+1} & \dots & a_{n-1} & a_0 & a_1 & \dots & a_{n-i-1} \\ a_{n-i-1} & a_{n-i} & \dots & a_{n-2} & a_{n-1} & a_0 & \dots & a_{n-i-2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_1 & a_2 & \dots & a_i & a_{i+1} & a_{i+2} & \dots & a_0 \end{bmatrix} * \begin{bmatrix} \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \dots \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \frac{2t}{n}-2td_{\mathrm{A}} \\ \dots \\ \frac{2t}{n}-2td_{\mathrm{A}} \end{bmatrix}$$

We can thus write firm i's basic price as $p_i^{\mathrm{B}}(\mathbf{P_A}) = (\frac{2t}{n} - 2td_{\mathrm{A}}) * \sum_{j=0}^{n-1} a_j$, obtaining $p_i^{\mathrm{B}*}(\mathbf{P_A}) = \frac{t}{n} - td_{\mathrm{A}}$. By replacing the basic prices in firm i's profits function, we obtain

$$\pi_i^{\mathrm{W}*}(\mathbf{P_A}) = \frac{t}{n^2} - \frac{td_{\mathrm{A}}^2}{2} - F$$

(II) Firm i Only Serves Identified Consumers

Let us apply the non-negative price constraint. Since $p_i^{\mathrm{B}*}(\mathbf{P_A}) = \frac{t}{n} - td_{\mathrm{A}}$, we conclude that firm i only serves identified consumers when $d_{\mathrm{A}} \geq \frac{1}{n}$. This corresponds to firm i identifying the consumer at the centre of the arch between itself and its rivals. Since all firms have the same amount of data, all of them set their basic prices equal to 0. By replacing these basic prices in firm i's profits, we obtain $\pi_i^{\mathrm{W}*}(\mathbf{P_A}) = \frac{t}{2n^2} - F$.

VI. Proof of Proposition 6

We organise this proof by examining three cases, depending on the firms' ability to identify consumers. In the first case, all informed firms serve at least some unidentified consumers. In the second, all informed firms except firm i's direct rivals only serve identified consumers. In the third, all informed firms only serve identified consumers.

(I) All Informed Firms Serve Both Identified and Unidentified Consumers

When firm i loses its auction under the *sale to all firms*, it becomes the only uninformed firm in the market. We can rewrite the vector of prices as in the proof of Proposition 5, except that the i-th component of vector $\mathbf{b}$ is $\frac{2t}{n}$ instead of $\frac{2t}{n} - 2td_{\mathrm{A}}$, as firm i is uninformed. By using the a_j coefficients' properties, we can write the basic prices as

$$p_i^{\mathrm{B}*}(\mathbf{P_A}) = \frac{t}{n} - td_{\mathrm{A}} + 2td_{\mathrm{A}}a_0 \quad \text{and} \quad p_{i-j}^{\mathrm{B}*}(\mathbf{P_A}) = p_{i+j}^{\mathrm{B}*}(\mathbf{P_A}) = \frac{t}{n} - td_{\mathrm{A}} + 2td_{\mathrm{A}}a_j \qquad (\text{A.29})$$

We recall that $a_j > a_{j+1} \forall j \in \{0, \frac{n}{2} - 1\}$. As such, firm i sets the highest basic price, and each firm's basic price decreases with its distance from firm i. By replacing (A.29) in firm i's profits function, we obtain

$$\pi_i^{\mathrm{L}*}(\mathbf{P_A}) = \left(\frac{t}{n} - td_{\mathrm{A}} + 2td_{\mathrm{A}}a_o\right)\left(2d_{\mathrm{A}}(a_1 - a_0) + \frac{1}{n}\right) - F \qquad (\text{A.30})$$

(II) All Informed Firms Except Firm i's Direct Rivals Only Serve Identified Consumers

As above, the fact that firm i is the only uninformed firm in the market creates an asymmetry, which causes firms' basic prices to decrease with their distance from firm

i. This also implies that the indifferent consumers' positions are skewed toward firm i's location. Without loss of generality, let us focus on firms $i{-}2$ and $i{-}1$. The former sets a lower basic price than the latter, as it is more distant from firm i. As such, the indifferent consumer placed between them $\widehat{x}_{i-2,i-1}$ is located closer to firm $i{-}1$ instead of at the centre of the arch like in the proof of Proposition 5. By replacing the basic prices, the indifferent consumer is located in $\widehat{x}_{i-2,i-1} = \frac{2i-3}{2n} + d_{\mathrm{A}}(a_1 - a_2)$. Firm $i{-}2$ can identify consumers up to $\frac{i-2}{n} + d_{\mathrm{A}}$. As such, we derive that if $d_{\mathrm{A}} \geq \frac{1}{2n(\frac{1}{2}+a_1-a_2)}$, then firm $i{-}2$ only serves identified consumers and sets its basic price equal to 0. For simplicity, we refer to this threshold as d_1. Moreover, as $(a_j - a_{j+1})$ decreases with j, all other informed firms except $i{+}1$ and $i{-}1$ also set their basic prices equal to 0. Without loss of generality, we focus on firms $i{-}1$ and i, as the model is symmetric with respect to firm i. Under these conditions, firm $i{-}1$ identifies all consumers on the arch it shares with firm $i{-}2$, while it still serves some unidentified consumers on the arch it shares with firm i. We can write firms' profits functions as

$$\pi^{\mathrm{W}}_{i-1}\left(\mathbf{P_A}\right) = \int_{\widehat{x}_{i-2,i-1}}^{\frac{i-1}{n}} p^{\mathrm{T}}_{i-1,i-2}(x)\,dx + \int_{\frac{i-1}{n}}^{\frac{i-1}{n}+d_{\mathrm{A}}} p^{\mathrm{T}}_{i-1,i}(x)\,dx + p^{\mathrm{B}}_{i-1}\left(\mathbf{P_A}\right)\left(\widehat{x}_{i-1,i} - \frac{i-1}{n} - d_{\mathrm{A}}\right) - F \quad \text{(A.31)}$$

The first term on the right-hand side represents firm $i{-}1$'s profits on the arch it shares with firm $i{-}2$. The second represents the profits it extracts from the identified consumers on the arch it shares with firm i. The third represents the profits firm $i{-}1$ makes on the unidentified consumers on the arch it shares with firm i. Firm i's profits are instead

$$\pi^{\mathrm{L}}_{i}\left(\mathbf{P_A}\right) = 2p^{\mathrm{B}}_{i}\left(\mathbf{P_A}\right)\left(\frac{i}{n} - \widehat{x}_{i-1,i}\right) - F \quad \text{(A.32)}$$

as the arches on which it competes are symmetric. We now replace the tailored prices and the indifferent consumers' location with the formulas provided in Equation (1) and (2). By computing FOCs of (A.31) and (A.32) we obtain the equilibrium basic prices, which are

$$p^{\mathrm{B}*}_{i-1}\left(\mathbf{P_A}\right) = \frac{t\left(3-2nd_{\mathrm{A}}\right)}{5n} \quad \text{and} \quad p^{\mathrm{B}*}_{i}\left(\mathbf{P_A}\right) = \frac{t\left(4-nd_{\mathrm{A}}\right)}{5n} \quad \text{(A.33)}$$

Replacing the basic prices from (A.33) into firm i's profits function, we obtain

$$\pi^{\mathrm{L}*}_{i}\left(\mathbf{P_A}\right) = \frac{t(nd_{\mathrm{A}}-4)^2}{25n^2} - F \quad \text{(A.34)}$$

(III) All Informed Firms Only Serve Identified Consumers

Finally, we focus on the case where also firms $i-1$ and $i+1$ only serve identified consumers. With respect to firm i, this subgame is the same as the one analysed in the proof of Proposition 4 where an uninformed firm competes against completely informed rivals. As such, we find the same solution: when $d_{\text{A}} \geq \frac{3}{2n}$, all informed firms set their basic prices equal to 0, leading to

$$p_i^{\text{B}*}\left(\mathbf{P_A}\right) = \frac{t}{2n} \quad \text{and} \quad \pi_i^{\text{L}*}\left(\mathbf{P_A}\right) = \frac{t}{4n^2} - F \tag{A.35}$$

To sum up, we put together results from the three cases when the DB opts for the *sale to all firms* and firm i loses its auction. Firm i's basic price, as shown in (A.29), (A.33) and (A.35) is

$$p_i^{\text{B}*}\left(\mathbf{P_A}\right) = \begin{cases} \frac{t}{n} - td_{\text{A}} + 2td_{\text{A}}a_0 & \text{for } \ d_{\text{A}} < d_1 \\ \frac{t(4-nd_{\text{A}})}{5n} & \text{for } \ d_1 \leq d_{\text{A}} < \frac{3}{2n} \\ \frac{t}{2n} & \text{for } \ d_{\text{A}} \geq \frac{3}{2n} \end{cases}$$

while its profits, as shown in (A.30), (A.34) and (A.35) are

$$\pi_i^{\text{L}}\left(\mathbf{P_A}\right) = \begin{cases} \left(\frac{t}{n} - td_{\text{A}} + 2td_{\text{A}}a_o\right)\left(2d_{\text{A}}\left(a_1 - a_0\right) + \frac{1}{n}\right) - F & \text{for } \ d_{\text{A}} < d_1 \\ \frac{t(nd_{\text{A}}-4)^2}{25n^2} - F & \text{for } \ d_1 \leq d_{\text{A}} < \frac{3}{2n} \\ \frac{t}{4n^2} - F & \text{for } \ d_{\text{A}} \geq \frac{3}{2n} \end{cases}$$

VII. Proof of Proposition 7

The DB solves the problem

$$\max_{d_{\text{H}}} \pi_{\text{DB}} = \frac{n}{2}\left(\pi_i^{\text{W}*}\left(\mathbf{P_H}\right) - \pi_i^{\text{L}*}\left(\mathbf{P_H}\right)\right) \tag{A.36}$$

where

$$\pi_i^{\text{W}*}\left(\mathbf{P_H}\right) = \begin{cases} \frac{t}{n^2} + \frac{2d_{\text{H}}t}{3n} - \frac{7td_{\text{H}}^2}{18} - F & \text{for } \ d_{\text{H}} < \frac{3}{2n} \\ \frac{9t}{8n^2} - F & \text{for } \ d_{\text{H}} \geq \frac{3}{2n} \end{cases} \tag{A.37}$$

$$\pi_i^{\text{L}*}\left(\mathbf{P_H}\right) = \frac{t}{4n^2} - F \quad \text{for } \ 0 \leq d_{\text{H}} \leq 1$$

Thus, we can rewrite DB's profits by replacing (A.37) in (A.36), obtaining

$$\max_{d_{\text{H}}} \pi_{\text{DB}} = \begin{cases} \frac{n}{2}\left(\frac{3t}{4n^2} + \frac{2d_{\text{H}}t}{3n} - \frac{7td_{\text{H}}^2}{18}\right) & \text{for } \ d_{\text{H}} < \frac{3}{2n} \\ \frac{n}{2}\left(\frac{7t}{8n^2}\right) & \text{for } \ d_{\text{H}} \geq \frac{3}{2n} \end{cases}$$

When the DB sets $d_{\mathrm{H}} < \frac{3}{2n}$, she opts for $d^*_{\mathrm{H}} = \frac{6}{7n}$ and derives profits $\pi^*_{\mathrm{DB}} = \frac{29t}{56n}$. When she sets $d_{\mathrm{H}} \geq \frac{3}{2n}$, then her profits are constant and equal to $\pi^*_{\mathrm{DB}} = \frac{7t}{16n}$. By directly comparing the two results, we find that the DB maximises her profits by setting $d^*_{\mathrm{H}} = \frac{6}{7n}$.

Finally, the number of entering firms would be such that their profits after paying for entry and data are 0. We obtain the number of entering firms by solving

$$\pi_i^{\mathrm{L}*}\left(\mathbf{P_H}\right) = \frac{t}{4n^2} - F = 0$$

which results in

$$n^*_{\mathrm{H}} = \frac{1}{2}\sqrt{\frac{t}{F}} \tag{A.38}$$

VIII. Proof of Proposition 8

By using the expressions for $\pi_i^{\mathrm{W}*}(\mathbf{P_A})$ and $\pi_i^{\mathrm{L}*}(\mathbf{P_A})$ obtained in the proof of Proposition 5 and 6, we can express the DB's profits as

$$\max_{d_{\mathrm{A}}} \pi_{\mathrm{DB}} = \begin{cases} n\left(\frac{t}{n^2} - \frac{td_{\mathrm{A}}^2}{2} - \left(\frac{t}{n} - td_{\mathrm{A}} + 2td_{\mathrm{A}}a_o\right)\left(2d_{\mathrm{A}}\left(a_1 - a_0\right) + \frac{1}{n}\right)\right) & \text{for } \; d_{\mathrm{A}} < d_1 \\ n\left(\frac{t}{n^2} - \frac{td_{\mathrm{A}}^2}{2} - \frac{t(nd_{\mathrm{A}}-4)^2}{25n^2}\right) & \text{for } \; d_1 \leq d_{\mathrm{A}} < \frac{1}{n} \\ n\left(\frac{t}{2n^2} - \frac{t(nd_{\mathrm{A}}-4)^2}{25n^2}\right) & \text{for } \; \frac{1}{n} \leq d_{\mathrm{A}} < \frac{3}{2n} \\ n\left(\frac{t}{2n^2} - \frac{t}{4n^2}\right) & \text{for } \; d_{\mathrm{A}} \geq \frac{3}{2n} \end{cases}$$

First, we prove that the second and third part of DB's profits are always suboptimal. By computing FOC of the second part with respect to d_A, we find that it is monotonically decreasing in d_A over its domain. Thus, the DB would always prefer the first part to the second one. By computing FOC of the third part with respect to d_A, we find that it is monotonically increasing in d_A over its domain. Thus, the DB would always prefer the fourth part to the third one.

To assess the DB's equilibrium strategy, we maximise her profits with respect to d_A. If the DB sets $d_A < d_1$, her profits are maximised for

$$d^*_A = \frac{1 - 2a_1}{n\left(-8a_0^2 + a_0\left(8a_1 + 4\right) - 4a_1 + 1\right)} \tag{A.39}$$

By substituting (A.39) in DB's profits, we obtain

$$\pi_{\mathrm{DB}} = \frac{t}{2}(1 - 2a_1)d^*_A \tag{A.40}$$

Instead, if the DB sets $d_A \geq \frac{3}{2n}$, her profits are equal to

$$\pi_{\mathrm{DB}} = \frac{t}{4n} \tag{A.41}$$

We now compare (A.40) and (A.41) to assess the DB's equilibrium strategy. The DB sets $d_A = d_A^*$ if

$$\frac{(1-2a_1)^2}{(-8a_0^2 + a_0(8a_1+4) - 4a_1 + 1)} \geq \frac{1}{2} \tag{A.42}$$

While we are not able to find an explicit solution to (A.42), we find that the inequality is satisfied for $n \geq \hat{n} \approx 3.34$. Thus, when $n < \hat{n}$, the DB sets $d_A \geq \frac{3}{2n}$; when $n \geq \hat{n}$, the DB sets $d_A = d_A^*$.

Having found the DB's equilibrium strategy as a function of n, we now proceed to the firms' entry stage.

If $n < \hat{n}$, the number of entering firms is given by solving

$$\pi_i^{\mathrm{L}*}(\mathbf{P_A}) = \frac{t}{4n^2} - F = 0$$

leading to $n_A^* = \frac{1}{2}\sqrt{\frac{t}{F}}$.

If instead $n \geq \hat{n}$, the number of entering firms is given by solving

$$\pi_i^{\mathrm{L}*}(\mathbf{P_A}) = \left(\frac{t}{n} - td_{\mathrm{A}}^* + 2td_{\mathrm{A}}^* a_o\right)\left(2d_{\mathrm{A}}^*(a_1 - a_0) + \frac{1}{n}\right) = 0 \tag{A.43}$$

To isolate all the terms that depend exponentially from n, it is useful to rewrite d_{A}^* as

$$d_{\mathrm{A}}^* = \frac{\alpha(n)}{n} \tag{A.44}$$

where

$$\alpha(n) = \frac{1-2a_1}{-8a_0^2 + a_0(8a_1+4) - 4a_1 + 1}$$

By replacing (A.44) in (A.43), we can rewrite it as

$$(1 - \alpha(n) + 2a_0\alpha(n))\,(2(a_1 - a_0\alpha(n) + 1) = \frac{F}{t}n^2 \tag{A.45}$$

We refer to the left-side of the equation as $A(n)$. By studying $A(n)$, we find that it is monotonically decreasing in n and quickly approaches an asymptote:

$$\lim_{n\to\infty} A(n) = \frac{36\sqrt{3} - 99}{1644\sqrt{3} - 2915}$$

To find the number of entering firms, we approximate $A(n)$ with

$$A(n) \approx \frac{1}{n^3} + \frac{36\sqrt{3} - 99}{1644\sqrt{3} - 2915} \tag{A.46}$$

This approximation overestimates the true value of $A(n)$ by less than 1% over its domain (i.e., $n \geq 2$). We recall however that in our basic model the DB chooses her strategy given n: as such, this approximation does not affect the DB's strategy, and it is only done to estimate the possible effect of the DB's strategies on firm entry.

By replacing (A.46) in (A.45), we obtain

$$\frac{1}{n^3} + \frac{36\sqrt{3} - 99}{1644\sqrt{3} - 2915} - \frac{F}{t} n^2 = 0 \tag{A.47}$$

To find an explicit solution to (A.47), we use the Newton-Raphson approximation method. This method starts by providing a first guess of the solution, denoted as x_0, and iteratively improving the approximation:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

By looking at (A.47), we see that the solution is close to $\frac{3}{4}\sqrt{\frac{t}{F}}$. Thus, by posing $x_0 = \frac{3}{4}\sqrt{\frac{t}{F}}$ we obtain

$$n_{\text{A}}^* \approx \sqrt{\frac{t}{F}} \frac{4096\left(1644\sqrt{3} - 2915\right)\frac{F}{t} + 243\left(1708\sqrt{3} - 3091\right)\sqrt{\frac{t}{F}}}{8\left(1644\sqrt{3} - 2915\right)\left(512\frac{F}{t} + 81\sqrt{\frac{t}{F}}\right)} \tag{A.48}$$

By analysing (A.48), we find that the number of entering firms when the DB sets $d_{\text{A}} = d_{\text{A}}^*$ is slightly above $\frac{3}{4}\sqrt{\frac{t}{F}}$.

IX. Proof of Proposition 9

We directly compare results from the proof of Proposition 7 and 8. Under the *sale to alternating firms,* the DB obtains $\pi_{\text{DB}}^*(\mathbf{P_H}) = \frac{29t}{56n}$.

Under the *sale to all firms*, when $n < \hat{n}$, the DB obtains $\pi_{\text{DB}}^*(\mathbf{P_A}) = \frac{t}{4n}$. By comparing profits with the *sale to alternating firms*, we find that the DB prefers it over the *sale to all firms.*

When $n \geq \hat{n}$, DB's profits are equal to

$$\pi_{\text{DB}}^*(\mathbf{P_A}) = \frac{t}{2}(1 - 2a_1)d_A^*$$

where

$$d_A^* = \frac{1 - 2a_1}{n\left(-8a_0^2 + a_0\left(8a_1 + 4\right) - 4a_1 + 1\right)}$$

The DB opts for the *sale to alternating firms* if

$$\frac{29}{56} > \frac{1}{2}\frac{(1 - 2a_1)^2}{-8a_0^2 + a_0\left(8a_1 + 4\right) - 4a_1 + 1}$$

Which is always satisfied for $n \geq 2$. We can thus conclude that the DB always opts for the *sale to alternating firms*, regardless of n.

X. Proof of Proposition 10

Consumer utility after purchasing the product from a firm is equal to

$$U(x,i)=v-p_i(x)-t\left(x-\frac{i}{n}\right)$$

where $p_{i,i+1}(x)$ can either be $p_i^{\mathrm{B}*}(\mathbf{P_H})$ if firm i does not identify the consumer located in x or $p_i^{\mathrm{T}}(x)$ if it identifies it. Due to the model's symmetry, we can derive the consumer surplus by first computing it on the arch between firms i and i+1, and then by simply multiplying this result by n.

In equilibrium, the DB sets $d_{\mathrm{H}}^{*}=\frac{6}{7n_{\mathrm{H}}^{*}}$, as concluded in Proposition 7. Suppose that the DB offers d_{H}^{*} to firm i: as such, firm i serves both identified and unidentified consumers, and it sets a basic price equal to $p_i^{\mathrm{B}*}(\mathbf{P_H})=\frac{t}{n}-\frac{2}{3}td_{\mathrm{H}}$, while its rivals set a basic price $p_{i+1}^{\mathrm{B}*}(\mathbf{P_H})=p_{i-1}^{\mathrm{B}*}(\mathbf{P_H})=\frac{t}{n}-\frac{1}{3}td_{\mathrm{H}}$. Firm i's tailored price is $p_{i,i+1}^{\mathrm{T}*}(x)=p_{i+1}^{\mathrm{B}*}(\mathbf{P_H})-2tx+\frac{t}{n}(2i+1)$. Finally, the indifferent consumer is located in $\widehat{x}_{i,i+1}=\frac{2i+1}{2n}+\frac{d_{\mathrm{H}}}{6}$. All these results have been previously obtained in the proof of Proposition 3. We can now compute consumer surplus on the arch between firms i and i+1. Consumers located between $\frac{i}{n}$ and $\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}$ buy from firm i and pay the tailored price; consumers between $\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}$ and $\widehat{x}_{i,i+1}$ buy from firm i and pay the basic price; consumers between $\widehat{x}_{i,i+1}$ and $\frac{i+1}{n}$ buy from firm i+1 and pay its basic price. The sum of these terms must then be multiplied by n to obtain the overall consumer surplus. Formally, we have

$$CS=n\left(\int_{\frac{i}{n}}^{\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}} v-p_{i,i+1}^{\mathrm{T}*}(x)-t\left(x-\frac{i}{n}\right)dx+\int_{\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}}^{\widehat{x}_{i,i+1}} v-p_i^{\mathrm{B}*}(\mathbf{P_H})-t\left(x-\frac{i}{n}\right)dx\right.$$
$$\left.+\int_{\widehat{x}_{i,i+1}}^{\frac{i+1}{n}} v-p_{i+1}^{\mathrm{B}*}(\mathbf{P_H})-t\left(\frac{i+1}{n}-x\right)dx\right) \tag{A.49}$$

By replacing the prices and the indifferent consumer's location in (A.49), we obtain

$$CS=u-\frac{5t}{4n}+\frac{\mathrm{nt}d_{\mathrm{H}}^{2}}{9} \tag{A.50}$$

Since $d_{\mathrm{H}}^{*}=\frac{6}{7n_{\mathrm{H}}^{*}}$ and $n_{\mathrm{H}}^{*}=\frac{1}{2}\sqrt{\frac{t}{F}}$, we can further rewrite (A.50) as

$$\mathrm{CS}=u-\frac{229}{98}\sqrt{tF} \tag{A.51}$$

Consumer surplus where data are absent is equal to $\mathrm{CS}^{\mathrm{BENCH}}=u-\frac{5t}{\widetilde{n}^{*}}$ with $\widetilde{n}^{*}=\sqrt{\frac{t}{F}}$. This results in $\mathrm{CS}^{\mathrm{BENCH}}=u-\frac{5}{4}\sqrt{tF}$, which is always higher than the consumer surplus reported in (A.51) (i.e., when the DB is present).

We now compute total welfare. DB profits, as shown in the proof of Proposition 7, are equal to

$$\pi_{\mathrm{DB}}^{*} = \frac{29t}{56n_{\mathrm{H}}^{*}} = \frac{29}{28}\sqrt{tF} \tag{A.52}$$

We recall how firms that win their auction are left with $\pi_i^{\mathrm{L}}\left(\mathbf{P_H}\right) = \frac{t}{4n_{\mathrm{H}}^{*2}} - F$ after paying for data, while firms that do not win their auctions obtain $\pi_{i+1}^{\mathrm{L}*}\left(\mathbf{P_H}\right) = \frac{t}{n_{\mathrm{H}}^{*2}} - \frac{2d_{\mathrm{H}}^{*}t}{3n_{\mathrm{H}}^{*}} + \frac{\mathrm{td}_{H}^{*2}}{9} - F$. We can thus rewrite total firms' profits as

$$\pi_{\text{firms}}^{*} = \sum_{i=0}^{n-1} \pi_i^{\mathrm{L}}\left(\mathbf{P_H}\right) = \frac{n_{\mathrm{H}}^{*}}{2}\left(\frac{t}{4n_{\mathrm{H}}^{*2}} - F\right) + \frac{n_{\mathrm{H}}^{*}}{2}\left(\frac{t}{n_{\mathrm{H}}^{*2}} - \frac{2d_{\mathrm{H}}^{*}t}{3n_{\mathrm{H}}^{*}} + \frac{td_{H}^{*2}}{9} - F\right) \tag{A.53}$$

By replacing $d_{\mathrm{H}}^{*} = \frac{6}{7n_{\mathrm{H}}^{*}}$ and $n_{\mathrm{H}}^{*} = \frac{1}{2}\sqrt{\frac{t}{F}}$, in (A.53), we obtain

$$\pi_{\text{firms}}^{*} = \frac{43}{196}\sqrt{tF} \tag{A.54}$$

Finally, we compute total welfare as the sum of (A.51), (A.52) and (A.54). We obtain

$$TW = CS + \pi_{\mathrm{DB}}^{*} + \pi_{\text{firms}}^{*} = u - \frac{229}{98}\sqrt{tF} + \frac{29}{28}\sqrt{tF} + \frac{43}{196}\sqrt{tF} = u - \frac{53}{49}\sqrt{tF} \tag{A.55}$$

When data are absent, total welfare is equal to consumer surplus since firms make profits equal to 0. As such, we find that total welfare increases under the DB's presence. On the other hand, most of the total welfare is appropriated by the DB. To show this, we compare total welfare under the DB's absence and presence, and we weight her profits by a coefficient α. The aim is to show for which weight α total welfare in the presence of a DB is lower than in the benchmark case. We can write

$$u - \frac{5}{4}\sqrt{tF} > u - \frac{229}{98}\sqrt{tF} + \frac{43}{196}\sqrt{tF} + \alpha\frac{29}{28}\sqrt{tF} \tag{A.56}$$

from which we obtain

$$\alpha^{*} < \frac{170}{203} \tag{A.57}$$

Thus, as long as (A.57) holds, total welfare is lower than under the benchmark case.

XI. Proof of Proposition 11

We introduce a disutility $c > 0$ that consumers incur when firms offer them tailored prices. We analyse this extension under the *sale to alternating firms*, which we have proven being optimal for the DB. First, we compute firms' profits when winning or losing their respective auctions. Then, we assess the DB's optimal strategy and her profits. Finally, we compute consumer surplus and total welfare.

(I) Equilibrium Prices

We focus on firm i, assuming that the DB offers it $d_i = d_{\mathrm{H}}$. Due to the added disutility, firm i sets tailored prices equal to

$$p_i^{\mathrm{T}}(x) = \begin{cases} p_{i-1}^{\mathrm{B}} + 2tx - \frac{t}{n}(2i-1) - c & \text{for } \; x \in [\frac{i}{n} - \frac{d_H}{2}, \frac{i}{n}] \\ p_{i+1}^{\mathrm{B}} - 2tx + \frac{t}{n}(2i+1) - c & \text{for } \; x \in [\frac{i}{n}, \frac{i}{n} + \frac{d_H}{2}] \end{cases} \tag{A.58}$$

By substituting (A.58) in (4) we obtain

$$\begin{aligned} \pi_i^{\mathrm{W}}(\mathbf{P_H}) = {} & \frac{d_{\mathrm{H}}}{2n}\left(2t + np_{i-1}^{\mathrm{B}}(\mathbf{P_H}) + np_{i+1}^{\mathrm{B}}(\mathbf{P_H}) - ntd_{\mathrm{H}} - 2nc\right) \\ & + p_i^{\mathrm{B}}(\mathbf{P_H})\left(\frac{n\left(p_{i+1}^{\mathrm{B}}(\mathbf{P_H}) + p_{i-1}^{\mathrm{B}}(\mathbf{P_H}) - 2p_i^{\mathrm{B}}(\mathbf{P_H})\right) + 2t}{2nt} - d_{\mathrm{H}}\right) - F \end{aligned} \tag{A.59}$$

While its rivals' profits are

$$\pi_{i+1}^{\mathrm{L}}(\mathbf{P_H}) = p_{i+1}^{\mathrm{B}}(\mathbf{P_H})\left(\frac{n\left(p_{i+2}^{\mathrm{B}}(\mathbf{P_H}) + p_i^{\mathrm{B}}(\mathbf{P_H}) - 2p_{i+1}^{\mathrm{B}}(\mathbf{P_H})\right) + 2t}{2nt}\right) - F \tag{A.60}$$

Comparing the profits functions with Equation (4) and (5), we can see how the disutility does not affect firm i or firm i+1's basic prices. As such, we obtain the same basic prices as in Section 3.1:

$$p_i^{\mathrm{B}*}(\mathbf{P_H}) = \frac{t}{n} - \frac{2}{3}td_{\mathrm{H}} \quad \text{and} \quad p_{i+1}^{\mathrm{B}*}(\mathbf{P_H}) = \frac{t}{n} - \frac{1}{3}td_{\mathrm{H}} \tag{A.61}$$

By substituting (A.61) in (A.59) and (A.60), we obtain

$$\begin{gathered} \pi_i^{\mathrm{W}*}(\mathbf{P_H}) = \frac{t}{n^2} + d_{\mathrm{H}}\left(\frac{2t}{3n} - c\right) - \frac{7}{18}td_{\mathrm{H}}^2 - F \\ \text{and} \\ \pi_{i+1}^{\mathrm{L}*}(\mathbf{P_H}) = \frac{t}{n^2} - \frac{2td_{\mathrm{H}}}{3n} + \frac{1}{9}td_{\mathrm{H}}^2 - F \end{gathered} \tag{A.62}$$

These profits functions hold as long as firm i benefits from offering tailored prices instead of basic prices to all identified consumers. Given the model's symmetry, we focus on the arch between firm i and i+1 without loss of generality. Thus, the profits functions hold as long as

$$p_i^{\mathrm{B}*}(\mathbf{P_H}) \leq p_i^{\mathrm{T}}(x)$$

We rewrite it as

$$\frac{t}{n} - \frac{2}{3}td_{\mathrm{H}} \leq \frac{t}{n} - \frac{1}{3}td_{\mathrm{H}} - 2tx + \frac{t}{n}(2i+1) - c \tag{A.63}$$

The inequality (A.63) must hold for any consumer identified by firm i. We recall that, on the arch between firm i and firm i+1, firm i identifies consumers up to the location

$\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}$. If (A.63) holds for the farthest identified consumer, then it also holds for all the other identified consumers, as tailored prices allow to extract more surplus from closer consumers (while the surplus extraction through basic prices is independent from the consumers' locations). By setting $x=\frac{i}{n}+\frac{d_{\mathrm{H}}}{2}$ and solving with respect to d_{H}, we find

$$d_{\mathrm{H}} \leq \frac{3}{2n}-\frac{3c}{2t} \tag{A.64}$$

Thus, as long as (A.64) holds, firm i offers tailored prices to all identified consumers. After this threshold, firm i offers its basic price to the newly identified consumers, as doing so grants it higher profits. From (A.64) we can also see how, if $c \geq \frac{t}{n}$, then firm i would never opt for the use of tailored prices. As such, we can conclude that, if $c \geq \frac{t}{n}$, no firm would buy data.

We now consider the case in which (A.64) does not hold. In this case, firm i only offers its tailored prices to consumers located between $\frac{i}{n}-\frac{3}{4n}+\frac{3c}{4t}$ and $\frac{i}{n}+\frac{3}{4n}-\frac{3c}{4t}$. We can thus rewrite its profits as

$$\begin{aligned}\pi_i^{\mathrm{W}}(\mathbf{P_H}) = \int_{\frac{i}{n}-\frac{3}{4n}+\frac{3c}{4t}}^{\frac{i}{n}} p_i^{\mathrm{T}}(x)\,dx + \int_{\frac{i}{n}}^{\frac{i}{n}+\frac{3}{4n}-\frac{3c}{4t}} p_i^{\mathrm{T}}(x)\,dx \\ + p_i^{\mathrm{B}}(\mathbf{P_H})\left(\widehat{x}_{i,i+1}-\widehat{x}_{i-1,i}-\left(\frac{3}{2n}-\frac{3c}{2t}\right)\right)-F\end{aligned}$$

which can be rewritten as

$$\begin{aligned}\pi_i^{\mathrm{W}}(\mathbf{P_H}) = \frac{3(cn-t)}{8tn^2}\left(cn-2t-2np_{i+1}^{\mathrm{B}}(\mathbf{P_H})-2np_{i-1}^{\mathrm{B}}(\mathbf{P_H})\right) \\ + p_i^{\mathrm{B}}(\mathbf{P_H})\left(\frac{n\left(p_{i+1}^{\mathrm{B}}(\mathbf{P_H})+p_{i-1}^{\mathrm{B}}(\mathbf{P_H})-2p_i^{\mathrm{B}}(\mathbf{P_H})+3c\right)-t}{2nt}\right)-F\end{aligned} \tag{A.65}$$

On the other hand, its rivals' profits maintain the same form presented before. We compute FOCs of (A.60) and (A.65) with respect to the basic prices, obtaining

$$\begin{aligned}\frac{\partial \pi_i^{\mathrm{W}}(\mathbf{P_H})}{\partial p_i^{\mathrm{B}}(\mathbf{P_H})} &= 4p_i^{\mathrm{B}}(\mathbf{P_H})-p_{i+1}^{\mathrm{B}}(\mathbf{P_H})-p_{i-1}^{\mathrm{B}}(\mathbf{P_H})-3c+\frac{t}{n}=0 \\ \frac{\partial \pi_{i+1}^{\mathrm{W}}(\mathbf{P_H})}{\partial p_{i+1}^{\mathrm{B}}(\mathbf{P_H})} &= 4p_{i+1}^{\mathrm{B}}(\mathbf{P_H})-p_i^{\mathrm{B}}(\mathbf{P_H})-p_{i+2}^{\mathrm{B}}(\mathbf{P_H})-\frac{2t}{n}=0\end{aligned} \tag{A.66}$$

We solve the n-equations system in its matrix form, and obtain

$$\begin{aligned}p_i^{\mathrm{B}*}(\mathbf{P_H}) = \left(3c-\frac{t}{n}\right)\left(a_0+a_{\frac{n}{2}}+2\sum_{j=1}^{\frac{n}{4}-1}a_{2j}\right)+\frac{2t}{n}\left(\frac{1}{2}-a_0-a_{\frac{n}{2}}-2\sum_{j=1}^{\frac{n}{4}-1}a_{2j}\right) \\ = \frac{1}{3}\left(3c-\frac{t}{n}\right)+\frac{1}{6}\frac{2t}{n}=c\end{aligned} \tag{A.67}$$

$$p_{i+1}^{\mathrm{B}*}\left(\mathbf{P_H}\right)=p_{i-1}^{\mathrm{B}*}\left(\mathbf{P_H}\right)=\left(3c-\frac{t}{n}\right)\left(\frac{1}{2}-a_0-a_{\frac{n}{2}}-2\sum_{j=1}^{\frac{n}{4}-1}a_{2j}\right)$$

$$+\frac{2t}{n}\left(a_0+a_{\frac{n}{2}}+2\sum_{j=1}^{\frac{n}{4}-1}a_{2j}\right)=\frac{1}{6}\left(3c-\frac{t}{n}\right)+\frac{1}{3}\frac{2t}{n}=\frac{t}{2n}+\frac{c}{2} \qquad \text{(A.68)}$$

By substituting (A.67) and (A.68) in the profits functions, we obtain

$$\pi_i^{\mathrm{W}*}\left(\mathbf{P_H}\right)=\frac{9t^2-6cnt+5c^2n^2}{8tn^2}-F \quad \text{and} \quad \pi_{i+1}^{\mathrm{L}*}\left(\mathbf{P_H}\right)=\frac{(t+cn)^2}{4tn^2}-F$$

(II) DB's Profits

Like in the basic model, DB's profits are equal to the sum of the difference in firms' profits between winning and losing their auction. When a firm wins, it obtains $d_i = d_{\mathrm{H}}$; when it loses, both its direct rivals obtain the whole dataset. We also recall that, if $c \geq \frac{t}{n}$, then no firms would buy data, and the DB's profits would thus be 0. DB's profits are equal to

$$\max_{d_{\mathrm{H}}}\pi_{\mathrm{DB}}=\begin{cases}\frac{n}{2}\left(\frac{t}{n^2}+d_{\mathrm{H}}\left(\frac{2t}{3n}-c\right)-\frac{7}{18}td_{\mathrm{H}}^2-\frac{(t+cn)^2}{4tn^2}\right) & \text{for } d_{\mathrm{H}}\leq\frac{3}{2n}-\frac{3c}{2t}\\ \frac{n}{2}\left(\frac{9t^2-6cnt+5c^2n^2}{8tn^2}-\frac{(t+cn)^2}{4tn^2}\right) & \text{for } d_{\mathrm{H}}>\frac{3}{2n}-\frac{3c}{2t}\end{cases} \qquad \text{(A.69)}$$

We start from the case where $d_{\mathrm{H}} \leq \frac{3}{2n}-\frac{3c}{2t}$. We can rewrite DB's profits as

$$\pi_{\mathrm{DB}}=\frac{3t}{8n}+\frac{d_{\mathrm{H}}t}{3}-\frac{c}{4}-\frac{c^2n}{8t}-\frac{\mathrm{cn}d_{\mathrm{H}}}{2}-\frac{7ntd_{\mathrm{H}}^2}{36} \qquad \text{(A.70)}$$

By computing the FOC of (A.70) with respect to d_{H}, we obtain

$$d_{\mathrm{H}}^*=\frac{6}{7n}-\frac{9c}{7t} \qquad \text{(A.71)}$$

The number of entering firms is instead given by equating to 0 firms' profits when losing their auction. Since a losing firm faces completely informed rivals, we obtain the number of entering firms by solving

$$\pi_i^{\mathrm{L}*}=\frac{(t+cn)^2}{4tn^2}-F=0$$

from which we obtain

$$n_{\mathrm{H}}^*=\frac{t}{2\sqrt{tF}-c} \qquad \text{(A.72)}$$

By replacing (A.72) in the inequality $c \geq \frac{t}{n}$, we obtain that, if $c \geq \sqrt{tF}$, no firm would buy data and the DB's profits would be 0. We can rewrite the optimal partition size set by the DB, (A.71), as

$$d_{\mathrm{H}}^{*} = \frac{12\sqrt{tF}}{7t} - \frac{15c}{7t} \tag{A.73}$$

The partition size d_{H}^{*} is greater than 0 only if $c < \frac{4}{5}\sqrt{tF}$. After this threshold, the DB is better off not offering data in the auctions she wants to conclude, while still threatening firms to sell the whole dataset to their rivals. The reason is the following: when $c \geq \frac{4}{5}\sqrt{tF}$, firm i's profits when winning its auction become monotonically decreasing in d_{H}. The disutility in using tailored prices is so high that, while tailored prices could still extract more surplus on individual consumers than the basic prices, the overall effect on profits is negative due to the competition effect. As such, the DB is better off offering a partition of size 0. Thus, when $d_{\mathrm{H}} \leq \frac{3}{2n} - \frac{3c}{2t}$, DB's profits are

$$\pi_{\mathrm{DB}}^{*} = \begin{cases} \frac{9\sqrt{tF}}{14} - \frac{39c}{28} + \frac{11tF}{14\left(2\sqrt{tF}-c\right)} & \text{for} \quad c < \frac{4}{5}\sqrt{tF} \\ \frac{\left(\sqrt{tF}-c\right)\left(3\sqrt{tF}-c\right)}{2\left(2\sqrt{tF}-c\right)} & \text{for} \quad \frac{4}{5}\sqrt{tF} \leq c < \sqrt{tF} \end{cases} \tag{A.74}$$

We refer to the first part as $\pi_{\mathrm{DB}}^{c_{\mathrm{low}}*}$, and to the second as $\pi_{\mathrm{DB}}^{c_{\mathrm{high}}*}$

We now move to the case where $d_{\mathrm{H}} > \frac{3}{2n} - \frac{3c}{2t}$. The number of entering firms is still $n_{\mathrm{H}}^{*} = \frac{t}{2\sqrt{tF}-c}$. Thus, we can rewrite DB's profits as

$$\pi_{\mathrm{DB}}^{*} = \frac{5c^2 - 12c\sqrt{tF} + 7tF}{4\left(2\sqrt{tF} - c\right)} \tag{A.75}$$

By comparing (A.74) and (A.75), we find that DB's profits are always maximised when $d_{\mathrm{H}} \leq \frac{3}{2n} - \frac{3c}{2t}$ for any $0 \leq c < \sqrt{tF}$, $t > F > 0$. Moreover, DB's profits are decreasing in c.

(III) Consumer Surplus and Total Welfare

We start by computing consumer surplus when $c < \frac{4}{5}\sqrt{tF}$. First, we note that firms' basic prices are not influenced by the disutility c. Second, as shown in Step (I), firms internalise the disutility by subtracting c from their tailored prices. As such, the direct net effect on consumers is 0: while consumers incur it, they are compensated with a reduced price. As such, the computation of consumer surplus under this scenario is equal to the one shown in the proof of Proposition 10, and we obtain

$$CS = u - \frac{5t}{4n_{\mathrm{H}}^{*}} + \frac{4}{9}n_{\mathrm{H}}^{*}d_{\mathrm{H}}^{*2}t \tag{A.76}$$

However, both the number of entering firms and the amount of data sold depend on c: by substituting (A.72) and (A.73) in (A.76) we find

$$CS^{c_{\text{low}}} = u - \frac{5}{4}\left(2\sqrt{tF} - c\right) + \frac{\left(5c - 4\sqrt{tF}\right)^2}{98\sqrt{tF} - 49c} \tag{A.77}$$

By computing the FOC with respect to c, we find that (A.77) is increasing in c. When $\frac{4}{5}\sqrt{tF} \leq c < \sqrt{tF}$, at the equilibrium no firm has data, and the general formula for consumer surplus is the same as in the standard Salop model: however, the number of entering firms is still influenced by the DB's presence and the disutility, as they affect firms' willingness to pay for data. Thus, we obtain

$$CS^{c_{\text{high}}} = u - \frac{5t}{4n_{\text{H}}^*} = u - \frac{5}{4}\left(2\sqrt{tF} - c\right) \tag{A.78}$$

which is also increasing in c.

Finally, we need to compute total firms' profits to assess the effect of privacy loss on total welfare. First, winning firms are left with zero profits after paying the entry cost F, as the DB extracts all the extra profits they can make. Thus, total firms' profits are only composed of the profits of firms that do not obtain data. When $c < \frac{4}{5}\sqrt{tF}$, the DB sells partitions of size $d_{\text{H}}^* = \frac{12\sqrt{tF}}{7t} - \frac{15c}{7t}$, and total firms' profits are

$$\pi_{\text{firms}}^{c_{\text{low}}*} = \frac{n_{\text{H}}^*}{2}\pi_{i+1}^{\text{L}*}\left(\mathbf{P_H}\right) = \frac{n_{\text{H}}^*}{2}\left(\frac{t}{n_{\text{H}}^{*2}} - \frac{2td_{\text{H}}^*}{3n_{\text{H}}^*} + \frac{1}{9}td_{\text{H}}^{*2} - F\right)$$

which we can rewrite as

$$\pi_{\text{firms}}^{c_{\text{low}}*} = \frac{t}{4\sqrt{tF} - 2c}\left(\frac{\left(7\left(2\sqrt{tF} - c\right) + \left(5c - 4\sqrt{tF}\right)\right)^2}{49t} - F\right) \tag{A.79}$$

On the other hand, when $\frac{4}{5}\sqrt{tF} \leq c < \sqrt{tF}$, at the equilibrium the DB does not sell data. Firms' profits are thus equal to the ones in the standard Salop model, although the number of entering firms still depends on the disutility c. We find

$$\pi_{\text{firms}}^{c_{\text{high}}*} = \frac{n_{\text{H}}^*}{2}\left(\frac{t}{n_{\text{H}}^{*2}} - F\right) = \frac{c^2 - 4c\sqrt{tF} + 3tF}{4\sqrt{tF} - 2c} \tag{A.80}$$

We can now compute total welfare. When $c < \frac{4}{5}\sqrt{tF}$, total welfare is given by the sum of (A.74), (A.77) and (A.79), obtaining

$$TW^{c_{\text{low}}} = \frac{9\sqrt{tF}}{14} - \frac{39c}{28} + \frac{11tF}{14\left(2\sqrt{tF}-c\right)} + u - \frac{5}{4}\left(2\sqrt{tF}-c\right) + \frac{\left(5c-4\sqrt{tF}\right)^2}{98\sqrt{tF}-49c} + \frac{t}{4\sqrt{tF}-2c}\left(\frac{\left(7\left(2\sqrt{tF}-c\right)+\left(5c-4\sqrt{tF}\right)\right)^2}{49t} - F\right)$$

When $\frac{4}{5}\sqrt{tF} \leq c < \sqrt{tF}$, total welfare is the sum of (A.74), (A.78) and (A.80), obtaining

$$TW^{c_{\text{high}}} = \frac{\left(\sqrt{tF}-c\right)\left(3\sqrt{tF}-c\right)}{2\left(2\sqrt{tF}-c\right)} + u - \frac{5}{4}\left(2\sqrt{tF}-c\right) + \frac{c^2-4c\sqrt{tF}+3tF}{4\sqrt{tF}-2c}$$

By computing the FOC with respect to c under both scenarios, we find that total welfare is always decreasing in c.

XII. Proof of Proposition 12

To analyse the effects of decreasing the DB's bargaining power, we search for the DB's equilibrium strategies under the auction without reserve prices (suffix AU) and Take It Or Leave It (suffix TIOLI) mechanisms. We recall how the *sale to all firms* leads to the same result under all selling mechanisms, so that the results of proofs of Proposition 5, Proposition 6 and Proposition 8 are still valid under these selling mechanisms. Instead, we need to solve the model again under the *sale to alternating firms.*

(I) Auction Without Reserve Prices

Under the *sale to alternating firms*, the partition set sold is $\mathbf{P}^*_{\mathbf{H}} = (d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0)$. However, as discussed in Section 2.3, the DB can offer a partition set $\mathbf{P_H}$ which is different from $\mathbf{P}^*_{\mathbf{H}}$, as she can decide to not fulfil some of the auctions she sets up. When the DB can set reserve prices, she offers a partition set $\mathbf{P_H} = (d_{\mathrm{H}}, 1, d_{\mathrm{H}}, 1, \ldots, d_{\mathrm{H}}, 1)$ and then only fulfils the auctions where she offers d_{H}, as shown in the proof of Proposition 2. Instead suppose that the DB offers a generic partition set $\mathbf{P_H} = (d_{\mathrm{H}}, d_1, d_{\mathrm{H}}, d_3, \ldots, d_{\mathrm{H}}, d_{n-1})$. First, due to the result on model's symmetry obtained in the proof of Proposition 1, we can conclude that the DB would set $d_1 = d_3 = \ldots = d_{n-1} = d$. Second, since we know that in equilibrium the DB sells data to half of the firms, the cardinality of the subset $\mathbf{J}$ of fulfilled auctions is $j = \frac{n}{2}$. Without loss of generality, we focus on firms 0 and 1. If firm 0 wins its auction, all other firms that are offered d_{H} win, as they are identical to it and have access to the same set of information. On the other hand, if firm 0 loses, all the firms that are offered d_{H} lose. The same reasoning can be applied to firm 1 and the other firms that are offered d. Thus, firms' willingness to pay for data are equal to

$$\pi_0^{\mathrm{W}}\left(d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0\right)-\pi_0^{\mathrm{L}}(0, d, 0, d, \ldots, 0, d)$$

and

$$\pi_1^{\mathrm{W}}(0, d, 0, d, \ldots, 0, d)-\pi_1^{\mathrm{L}}\left(d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0\right)$$

We want to show that $d=d_{\mathrm{H}}$ in equilibrium. Suppose that $d>d_{\mathrm{H}}$ (the opposite case is solved similarly): then it is straightforward to show that

$$\begin{aligned}\pi_1^{\mathrm{W}}(0, d, 0, d, \ldots, 0, d)-\pi_1^{\mathrm{L}}(d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0 >& \\ &\pi_0^{\mathrm{W}}\left(d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0\right)-\pi_0^{\mathrm{L}}(0, d, 0, d, \ldots, 0, d)\end{aligned}$$

That is, firm 1's willingness to pay is higher than firm 0's one. Since there are no reserve prices, firm 1 can win its auction by offering firm 0's willingness to pay plus ε, where ε is an arbitrary small number. As such, the DB chooses d as low as possible to maximise firm 0's willingness to pay and, in turn, firm 1's: that is, she chooses $d=d_{\mathrm{H}}$.

Under AU, the DB offers a partition set $\mathbf{P_H}=\left(d_{\mathrm{H}}, d_{\mathrm{H}}, d_{\mathrm{H}}, d_{\mathrm{H}}, \ldots, d_{\mathrm{H}}, d_{\mathrm{H}}\right)$, and the partition set sold in equilibrium is $\mathbf{P}_{\mathbf{H}}^{*}=\left(d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0\right)$. Since all firms are symmetric and are offered same-sized partitions, the DB is indifferent between fulfilling the auctions of even-indexed or odd-indexed firms. This change from the auction with reserve prices (AR) has implications on firms' expected profits when they lose their auction. Under AR, a firm knew that, if it tried to deviate, the DB would fulfil its direct rivals' auctions. On the other hand, under AU, an even-indexed firm knows that, if it loses its auction, the DB will fulfil all the odd-indexed firms' auctions. Without loss of generality, we can focus on firm 0. If firm 0 wins, its profits are

$$\pi_0^{\mathrm{W\ AU}}\left(d_{\mathrm{H}}, 0, d_{\mathrm{H}}, 0, \ldots, d_{\mathrm{H}}, 0\right)$$

If it loses, they are

$$\pi_0^{\mathrm{L\ AU}}\left(0, d_{\mathrm{H}}, 0, d_{\mathrm{H}}, \ldots, 0, d_H\right)$$

Firm 0's profits when winning have already been computed in the proof of Proposition 3. We find

$$\pi_0^{\mathrm{W\ AU}*}\left(\mathbf{P_H}\right)=\begin{cases}\frac{t}{n^2}+\frac{2 d_{\mathrm{H}} t}{3 n}-\frac{7 t d_{\mathrm{H}}^2}{18}-F & \text{for } d_{\mathrm{H}}<\frac{3}{2 n} \\ \frac{9 t}{8 n^2}-F & \text{for } d_{\mathrm{H}} \geq \frac{3}{2 n}\end{cases} \tag{A.81}$$

In a similar way, firm 0's profits when losing are equal to firm $i{+}1$'s profits derived in the proof of Proposition 3. We find

$$\pi_0^{\mathrm{L\ AU}*}(\mathbf{P_H}) = \begin{cases} \frac{t}{n^2} - \frac{2d_{\mathrm{H}}t}{3n} + \frac{td_{\mathrm{H}}^2}{9} - F & \text{for } d_{\mathrm{H}} < \frac{3}{2n} \\ \frac{t}{4n^2} - F & \text{for } d_{\mathrm{H}} \geq \frac{3}{2n} \end{cases} \tag{A.82}$$

Thus, we can compute DB's profits as

$$\max_{d_{\mathrm{H}}} \pi_{\mathrm{DB}}^{\mathrm{AU}} = \begin{cases} \frac{n}{2}\left(\frac{t}{n^2} + \frac{2d_{\mathrm{H}}t}{3n} - \frac{7td_{\mathrm{H}}^2}{18} - F - \left(\frac{t}{n^2} - \frac{2d_{\mathrm{H}}t}{3n} + \frac{\mathrm{t}d_H^2}{9} - F\right)\right) & \text{for } d_{\mathrm{H}} < \frac{3}{2n} \\ \frac{n}{2}\left(\frac{9t}{8n^2} - F - \left(\frac{t}{4n^2} - F\right)\right) & \text{for } d_{\mathrm{H}} \geq \frac{3}{2n} \end{cases}$$

which we can rewrite as

$$\max_{d_{\mathrm{H}}} \pi_{\mathrm{DB}}^{\mathrm{AU}} = \begin{cases} \frac{d_{\mathrm{H}}t(8-3d_{\mathrm{H}}n)}{12} & \text{for } d_{\mathrm{H}} < \frac{3}{2n} \\ \frac{7t}{16n} & \text{for } d_{\mathrm{H}} \geq \frac{3}{2n} \end{cases} \tag{A.83}$$

When $d_{\mathrm{H}} < \frac{3}{2n}$, we obtain DB's profits by equating to 0 the FOC of the first part of (A.83) with respect to d_{H}, obtaining d_{H}^*. Since $\frac{4t}{9n} > \frac{7t}{16n}$ is always satisfied, we find that the DB's equilibrium strategy under the *sale to alternating firms* implies $d_{\mathrm{H}}^* = \frac{4}{3n}$.

We now compare the DB's profits under the *sale to all firms* and the *sale to alternating firms*, following the same approach as in the proof of Proposition 9. Under the *sale to alternating firms,* the DB obtains $\pi_{\mathrm{DB}}^{\mathrm{AU}^*}(\mathbf{P_H}) = \frac{4t}{9n}$.

Under the *sale to all firms*, when $n < \hat{n}$, the DB obtains $\pi_{\mathrm{DB}}^{\mathrm{AU}^*}(\mathbf{P_A}) = \frac{t}{4n}$. By comparing profits with the *sale to alternating firms*, we find that the DB prefers the latter.

When $n \geq \hat{n}$, DB's profits are equal to

$$\pi_{\mathrm{DB}}^*(\mathbf{P_A}) = \frac{t}{2}(1-2a_1)d_A^*$$

where

$$d_A^* = \frac{1-2a_1}{n\left(-8a_0^2 + a_0\left(8a_1+4\right) - 4a_1 + 1\right)}$$

The DB opts for the *sale to alternating firms* if

$$\frac{4}{9} > \frac{1}{2}\frac{(1-2a_1)^2}{-8a_0^2 + a_0\left(8a_1+4\right) - 4a_1 + 1}$$

which is always satisfied for $n \geq 2$. To sum up, the DB always opts for the *sale to alternating firms*, regardless of n.

We now derive the number of entering firms, which is is given by binding the first part of the piecewise function (A.82), so that

$$\frac{t}{n^2} - \frac{2d_{\mathrm{H}}^{*}t}{3n} + \frac{td_{\mathrm{H}}^{*\,2}}{9} - F = 0 \tag{A.84}$$

Solving (A.84) with respect to n gives us

$$n_{\mathrm{H}}^{*} = \frac{9\sqrt{tF} - 3d_{\mathrm{H}}^{*}t}{9F - td_{\mathrm{H}}^{*\,2}} = \frac{5}{9}\sqrt{\frac{t}{F}} \tag{A.85}$$

leading to DB's profits being

$$\pi_{\mathrm{DB}}^{\mathrm{AU}*} = \frac{4}{5}\sqrt{tF}$$

To compute consumer surplus, we can use the general formula for the *sale to alternating firms* provided in the proof of Proposition 10. In our case, we obtain

$$CS^{\mathrm{AU}} = u - \frac{5t}{4n_{\mathrm{H}}^{*}} + \frac{n_{\mathrm{H}}^{*}td_{\mathrm{H}}^{*2}}{9} \tag{A.86}$$

where $d_{\mathrm{H}}^{*} = \frac{4}{3n_{\mathrm{H}}^{*}}$ and $n_{\mathrm{H}}^{*} = \frac{5}{9}\sqrt{\frac{t}{F}}$. By replacing these values, we can rewrite (A.86) as

$$CS^{\mathrm{AU}} = u - \frac{341}{180}\sqrt{tF}$$

(II) Take It Or Leave It Offers

Under TIOLI, the DB cannot offer a partition set different from the one that results in equilibrium, as she has to fulfil all the offers she makes if firms are willing to pay the corresponding price. In the proof of Proposition 1, we concluded that the DB offers same-sized partitions to every other firm, i.e., $\mathbf{P} = \left(\widetilde{d}, \widehat{d}, \widetilde{d}, \ldots, \widehat{d}\right)$. We proved that $\widetilde{d}$ and $\widehat{d}$ could not be different: however, our demonstration did not rule out the corner solution where either $\widetilde{d}$ or $\widehat{d}$ were equal to 0. This is because, under both AU and AR, firms that at the equilibrium do not obtain data could have had data offered to them in auctions that the DB did not fulfil, and these offers influence firms' willingness to pay. This is not the case under TIOLI. As already shown in the proofs of Proposition 3 and 5, firms are always better off when they obtain data. As such, under this mechanism, only those firms to which the DB does not offer data will have no data at the equilibrium. Thus, under TIOLI, the analysis shown in the proof of Proposition 1 allows us to also discard the corner solution, leaving only the *sale to all firms* as a viable strategy. As already derived in the proof of Proposition 8, the DB's strategy depends on the number of entering firms.

If $n < \hat{n}$, the DB offers the whole dataset to all firms. The number of entering firms is $n^*_{\mathrm{A}} = \frac{1}{2}\sqrt{\frac{t}{F}}$, and her profits are

$$\pi^{\mathrm{TIOLI}*}_{\mathrm{DB}} = \frac{1}{2}\sqrt{tF}$$

Regarding consumer surplus, we first note that all firms set basic prices equal to 0, as they identify all the consumers they serve. As such, indifferent consumers are located at the centre of each arch, which is split in two identical halves. We can thus write consumer surplus as

$$CS^{\mathrm{TIOLI}} = 2n^*_{\mathrm{A}}\left(\int_{\frac{i}{n^*_{\mathrm{A}}}}^{\frac{2i+1}{2n^*_{\mathrm{A}}}} v - p^{\mathrm{T}*}_i(x) - t\left(x - \frac{i}{n^*_{\mathrm{A}}}\right)dx\right) \tag{A.87}$$

where

$$p^{\mathrm{T}}_i(x) = -2tx + \frac{t}{n^*_{\mathrm{A}}}(2i+1) \tag{A.88}$$

By substituting (A.88) in (A.87) and solving, we obtain

$$CS^{\mathrm{TIOLI}} = u - \frac{3t}{4n^*_{\mathrm{A}}} = u - \frac{3}{2}\sqrt{tF}$$

Instead, when $n \geq \hat{n}$, the DB sets

$$d^*_A = \frac{1 - 2a_1}{n\left(-8a_0^2 + a_0\left(8a_1 + 4\right) - 4a_1 + 1\right)}$$

In the proof of Proposition 8, we have approximated the number of entering firms as

$$n^*_{\mathrm{A}} \approx \sqrt{\frac{t}{F}}\frac{4096\left(1644\sqrt{3} - 2915\right)\frac{F}{t} + 243\left(1708\sqrt{3} - 3091\right)\sqrt{\frac{t}{F}}}{8\left(1644\sqrt{3} - 2915\right)\left(512\frac{F}{t} + 81\sqrt{\frac{t}{F}}\right)}$$

DB's profits are equal to

$$\pi^{\mathrm{TIOLI}}_{DB} = \frac{t}{2}(1 - 2a_1)d^*_A$$

Note how all firms offer equal basic prices and have equal market shares. As such, we can compute consumer surplus as that gained on an arch between a firm's location and its closest indifferent consumer, multiplied by $2n^*_{\mathrm{A}}$. We can thus write

$$\begin{aligned} CS = 2n^*_{\mathrm{A}}\Bigg(&\int_{\frac{i}{n}}^{\frac{i}{n}+\frac{d^*_{\mathrm{A}}}{2}} v - p^{\mathrm{T}*}_{i,i+1}(x) - t\left(x - \frac{i}{n}\right)dx + \\ &\int_{\frac{i}{n}+\frac{d^*_{\mathrm{A}}}{2}}^{\widehat{x}_{i,i+1}} v - p^{\mathrm{B}*}_i\left(\mathbf{P_A}\right) - t\left(x - \frac{i}{n}\right)dx\Bigg) \end{aligned} \tag{A.89}$$

Under the *sale to all firms*, firm i's basic price is $p^{\mathrm{B}*}_i\left(\mathbf{P_A}\right) = \frac{t}{n^*_{\mathrm{A}}} - td^*_{\mathrm{A}}$, the indifferent consumer is located in $\widehat{x}_{i,i+1} = \frac{i}{n^*_{\mathrm{A}}} + \frac{1}{2n^*_{\mathrm{A}}}$ and firm i's tailored price is $p^{\mathrm{T}*}_{i,i+1}(x) = p^{\mathrm{B}*}_{i+1} -$

$2tx + \frac{t}{n_{\text{A}}^*}(2i+1)$. By replacing these values in (A.89), we can rewrite consumer surplus as

$$CS = v - \frac{5t}{4n_{\text{A}}^*} + 2n_{\text{A}}^* t {d_{\text{A}}^*}^2$$

As we can see, consumer surplus is increasing in n_{A}^*: thus, our overestimation of the number of entering firms in (A.48) is an upper bound on consumer surplus. To properly compare results, we do a second approximation by underestimating $A(n)$: this in turn underestimates the number of entering firms, and gives us a lower bound on consumer surplus. We find that

$$A(n) \approx \frac{1}{n^3} + \frac{53}{100} \tag{A.90}$$

always underestimates $A(n)$. Through the Newton-Raphson approximation method, we find that the lower bound for the number of entering firms is

$$n_{\text{A low}}^* \approx \frac{102400\frac{F}{t} + 11799\sqrt{\frac{t}{F}}}{200\left(\frac{512}{\frac{t}{F}^{\frac{3}{2}}} + 81\right)} \tag{A.91}$$

By comparing DB's profits, consumer surplus and the quantity of data sold under the different selling mechanisms, we conclude that the first decreases while the others increase, as the DB's bargaining power is lowered. Moreover, consumer surplus is higher than in the benchmark only under TIOLI when $n \geq \hat{n}$, even when underestimating n as in (A.91), while it is lower in all the other scenarios.

XIII. Proof of Proposition 13

This proof is organised in five parts. First, we assess the *sale to all firms*, as its outcomes do not vary when changing the selling mechanism. Then, we individually assess the DB's strategy in equilibrium under the three selling mechanisms: auctions with reserve prices (AR), auctions without reserve prices (AU) and Take It Or Leave It offers (TIOLI). Finally, we compare our findings. We recall that, since the DB anticipates the effect of her data sale on firm entry, the DB first anticipates the number of entering firms with respect to her strategy, and then she chooses the amount of data she sells so as to maximise her profits. Moreover, the number of entering firms is given by posing firms' profits after paying for entry and data equal to 0.

(I) Sale to All Firms

The proof for the *sale to all firms* proceeds in two steps. First, we show that the

DB either sets $d_\mathrm{A} < d_1$ or $d_\mathrm{A} \geq \frac{3}{2n}$, where $d_1 = \frac{1}{2n(\frac{1}{2}+a_1-a_2)}$. Then, we show that she maximises her profits by setting $d_\mathrm{A} \geq \frac{3}{2n}$.

(I.1) The DB Either Sets $d_\mathrm{A} < d_1$ or $d_\mathrm{A} \geq \frac{3}{2n}$

By using the expressions for $\pi_i^{\mathrm{W}*}(\mathbf{P_A})$ and $\pi_i^{\mathrm{L}*}(\mathbf{P_A})$ obtained in the proof of Proposition 5 and 6, we can express the DB's constrained maximisation problem as

$$\max_{d_\mathrm{A}} \pi_\mathrm{DB} = \begin{cases} n\left(\frac{t}{n^2} - \frac{td_\mathrm{A}^2}{2} - \left(\frac{t}{n} - td_\mathrm{A} + 2td_\mathrm{A}a_o\right)\left(2d_\mathrm{A}\left(a_1 - a_0\right) + \frac{1}{n}\right)\right) & \text{for } d_\mathrm{A} < d_1 \\ n\left(\frac{t}{n^2} - \frac{td_\mathrm{A}^2}{2} - \frac{t(nd_\mathrm{A}-4)^2}{25n^2}\right) & \text{for } d_1 \leq d_\mathrm{A} < \frac{1}{n} \\ n\left(\frac{t}{2n^2} - \frac{t(nd_\mathrm{A}-4)^2}{25n^2}\right) & \text{for } \frac{1}{n} \leq d_\mathrm{A} < \frac{3}{2n} \\ n\left(\frac{t}{2n^2} - \frac{t}{4n^2}\right) & \text{for } d_\mathrm{A} \geq \frac{3}{2n} \end{cases}$$

subject to the following constraint:

$$\text{s.t.} \quad \begin{cases} \left(\frac{t}{n} - td_\mathrm{A} + 2td_\mathrm{A}a_o\right)\left(2d_\mathrm{A}\left(a_1 - a_0\right) + \frac{1}{n}\right) - F \geq 0 & \text{for } d_\mathrm{A} < d_1 \\ \frac{t(nd_\mathrm{A}-4)^2}{25n^2} - F \geq 0 & \text{for } d_1 \leq d_\mathrm{A} < \frac{3}{2n} \\ \frac{t}{4n^2} - F \geq 0 & \text{for } d_\mathrm{A} \geq \frac{3}{2n} \end{cases}$$

where the constraint is equal to firms' profits after paying for the entry cost and data. We focus on the second, third and fourth part of the DB's profits function: in particular, we want to show that, if the DB sets $d_\mathrm{A} \geq d_1$, then she sets $d_\mathrm{A} \geq \frac{3}{2n}$. To do so, we first note from Figure 1 how the constraint regarding entering firms is decreasing in d_A, and it becomes fixed for $d_\mathrm{A} \geq \frac{3}{2n}$. Moreover, DB's profits are inversely proportional to the number of entering firms, while they are directly proportional to firms' willingness to pay for data. Thus, if a given d_A grants a higher willingness to pay by firms and a smaller number of entering firms than another one, then the former strategy dominates the latter.

The number of entering firms is minimised when $d_\mathrm{A} \geq \frac{3}{2n}$. As such, strategies where $d_1 \leq d_\mathrm{A} < \frac{1}{n}$ or $\frac{1}{n} \leq d_\mathrm{A} < \frac{3}{2n}$ can only be optimal if they allow the DB to extract more surplus from individual firms than the strategy where $d_\mathrm{A} \geq \frac{3}{2n}$.

Suppose that the number of entering firms n is given: we can rewrite DB's profits when $d_1 \leq d_\mathrm{A} < \frac{1}{n}$ as

$$\pi_\mathrm{DB}\left(d_1 \leq d_\mathrm{A} < \frac{1}{n}\right) = \frac{9t}{25n} + \frac{8d_\mathrm{A}t}{25} - \frac{27d_\mathrm{A}^2tn}{50}$$

Which is a downward facing parabolic function with its maximum in $d_\mathrm{A} = \frac{8}{27n}$. Since $\frac{8}{27n} < d_1 \; \forall \; n$, we find that this function is decreasing in d_A over its domain. As such, if the DB sets $d_1 \leq d_{\mathrm{A}} < \frac{1}{n}$, she opts for $d^*_{\mathrm{A}} = d_1 = \frac{1}{2n(\frac{1}{2}+a_1-a_2)}$. On the other hand, when the DB sets $d_{\mathrm{A}} \geq \frac{3}{2n}$, her profits are

$$\pi_{\mathrm{DB}}\left(d_{\mathrm{A}} \geq \frac{3}{2n}\right) = \frac{t}{4n}$$

We want to show that

$$\pi_{\mathrm{DB}}\left(d_{\mathrm{A}} \geq \frac{3}{2n}\right) > \pi_{\mathrm{DB}}\left(d_1\right) \;\; \forall \; n$$

We can rewrite the inequality as

$$1 > \frac{36}{25} + \frac{32nd_1}{25} - \frac{108d_1^2n^2}{50} \tag{A.92}$$

The right-hand side of (A.92) is monotonically increasing in n, and

$$\lim_{n\to\infty} \frac{36}{25} + \frac{32nd_1}{25} - \frac{108d_1^2n^2}{50} \approx 0.85 < 1$$

Thus, we can conclude that the DB would opt for $d_{\mathrm{A}} \geq \frac{3}{2n}$ instead of $d_1 \leq d_{\mathrm{A}} < \frac{1}{n}$, as it allows her to better extract surplus from firms and to further hinder their entry. We can apply the same reasoning for the case where $\frac{1}{n} \leq d_{\mathrm{A}} < \frac{3}{2n}$. From Figure 1 we can see how firms' profits when winning are the same when $\frac{1}{n} \leq d_{\mathrm{A}} < \frac{3}{2n}$ or when $d_{\mathrm{A}} \geq \frac{3}{2n}$, while their profits when they lose are minimised when $d_{\mathrm{A}} \geq \frac{3}{2n}$. Thus, setting $d_{\mathrm{A}} \geq \frac{3}{2n}$ enables the DB to extract more surplus from firms. Since this amount of data also minimises the number of entering firms, we can conclude that the set of strategies where $\frac{1}{n} \leq d_{\mathrm{A}} < \frac{3}{2n}$ is also dominated by setting $d_{\mathrm{A}} \geq \frac{3}{2n}$. By putting together the two results, we conclude that if the DB sets $d_{\mathrm{A}} \geq d_1$, then she opts for $d_{\mathrm{A}} \geq \frac{3}{2n}$.

(I.2) DB's Profits Are Maximised When Setting $d_{\mathrm{A}} \geq \frac{3}{2n}$

For simplicity, we refer to the DB setting $d_{\mathrm{A}} \geq \frac{3}{2n}$ as d_{high}, while we refer to the DB setting $d_{\mathrm{A}} < d_1$ as d_{low}. When the DB sets $d_{\mathrm{A}} \geq \frac{3}{2n}$, she maximises

$$\pi_{\mathrm{DB}}(\mathbf{P_A}^{d_{\mathrm{high}}}) = \frac{t}{4n} \tag{A.93}$$

under the constraint

$$\pi_i{}^{\mathrm{L}}(\mathbf{P_A}^{d_{\mathrm{high}}}) = \frac{t}{4n^2} - F \geq 0 \tag{A.94}$$

Firms enter as long as (A.94) holds. By binding the constraint – i.e., setting $\pi_i{}^{\mathrm{L}}(\mathbf{P_A}^{d_{\mathrm{high}}}) = 0$ – we obtain

$$n^*_{d_{\mathrm{high}}} = \frac{1}{2}\sqrt{\frac{t}{F}} \tag{A.95}$$

By replacing (A.95) in (A.93) we obtain

$$\pi_{\mathrm{DB}}(\mathbf{P_A}^{d_{\mathrm{high}}}) = \frac{1}{2}\sqrt{tF} \tag{A.96}$$

The DB maximises

$$\pi_{\mathrm{DB}}(\mathbf{P_A}^{d_{\mathrm{low}}}) = n\left(\frac{t}{n^2} - td_{\mathrm{low}}^2 - \left(\frac{t}{n} - td_{\mathrm{low}} + td_{\mathrm{low}}a_o\right) * \left(2d_{\mathrm{low}}\left(a_1 - a_0\right) + \frac{1}{n}\right)\right)$$

which can be rewritten as

$$\pi_{\mathrm{DB}}(\mathbf{P_A}^{d_{\mathrm{low}}}) = t\left(d_{\mathrm{low}}\left(1 - 2a_1\right) - n\frac{d_{\mathrm{low}}^2}{2}\left(1 + 4\left(1 - 2a_0\right)\left(a_0 - a_1\right)\right)\right) \tag{A.97}$$

under the constraint

$$\pi_i{}^{\mathrm{L}}(\mathbf{P_A}^{d_{\mathrm{low}}}) = \left(\frac{t}{n} - td_{\mathrm{low}} + 2td_{\mathrm{low}}a_o\right)\left(2d_{\mathrm{low}}\left(a_1 - a_0\right) + \frac{1}{n}\right) - F \geq 0 \tag{A.98}$$

We want to show that

$$\pi_{\mathrm{DB}}(\mathbf{P_A}^{d_{\mathrm{low}}}) < \pi_{\mathrm{DB}}(\mathbf{P_A}^{d_{\mathrm{high}}}) \tag{A.99}$$

for all relevant values of d_{low}, t and F. In particular, we recall that $0 \leq d_{\mathrm{low}} < \frac{1}{n}$, $t > 0$, $F > 0$, $t > F$. The last condition is needed so that at least one firm enters the market (as such, $n \geq 1$). It is useful to express $\frac{F}{t} = k$, with $0 < k < 1$. We can rewrite (A.99) as

$$2d_{\mathrm{low}}\left(1 - 2a_1\right) - nd_{\mathrm{low}}^2\left(1 + 4\left(1 - 2a_0\right)\left(a_0 - a_1\right)\right) < \sqrt{k} \tag{A.100}$$

To solve (A.100), we bind the constraint (A.98), find the number of entering firms and replace it in (A.97). However, we recall that a_0 and a_1 are exponential functions in n, and as such they heavily complicate this process. To find an explicit solution, we operate a series of round ups on the left-hand side of (A.100). By showing that the left side is smaller than the right side even after the round ups, we prove that also the original inequality holds.

First, we compute the number of entering firms when $d = d_{\mathrm{low}}$. Since firms enter as long as their outside option is no lower than 0, the number of entering firms is given by binding the constraint (A.98):

$$\pi_i{}^{\mathrm{L}}(\mathbf{P_A}^{d_{\mathrm{low}}}) = \left(\frac{t}{n} - td_{\mathrm{low}} + 2td_{\mathrm{low}}a_o\right)\left(2d_{\mathrm{low}}\left(a_1 - a_0\right) + \frac{1}{n}\right) - F = 0 \tag{A.101}$$

By replacing the explicit forms of a_0 and a_1 in (A.101), we can rewrite it as

$$\frac{1}{n^2} - \frac{2d_{\mathrm{low}}}{\sqrt{3}n}f(n) + \frac{1}{3}d_{\mathrm{low}}^2 f(n)^2 - \frac{F}{t} = 0 \tag{A.102}$$

where

$$f(n) = \frac{\left(\sqrt{3} - 1\right)\left(2 + \sqrt{3}\right)^n + \left(1 + \sqrt{3}\right)\left(2 - \sqrt{3}\right)^n - 2\sqrt{3}}{\left(2 + \sqrt{3}\right)^n + \left(2 - \sqrt{3}\right)^n - 2} \tag{A.103}$$

Our objective would be to solve (A.102) with respect to n, to obtain $n(d_{\mathrm{low}})$. However, (A.102) depends on n both linearly and exponentially, which increases the complexity of finding an explicit form of $n(d)$. On the other hand, the equality is a second-order polynomial in d_{low}: as such, we can easily obtain $d(n)$. By solving (A.102) with respect to d we obtain

$$d_1^*(n) = \frac{\sqrt{3}\left(n\sqrt{k}+1\right)}{nf(n)} \quad \text{and} \quad d_2^*(n) = -\frac{\sqrt{3}\left(1-n\sqrt{k}\right)}{nf(n)} \tag{A.104}$$

where $k = \frac{F}{t}$. From Salop (1979) we know that, if a DB is absent, the number of entering firms is $n = \sqrt{\frac{t}{F}}$. As such, our solution must satisfy $d_{\text{low}}\left(\sqrt{\frac{t}{F}}\right) = 0$, which gives us $d_{\text{low}} = d_2^*(n)$. Having found $d_{\text{low}}(n)$, we need to invert the function to obtain $n(d_{\text{low}})$. However, the presence of the exponential function $f(n)$ still poses problems when searching for an explicit solution. As such, we round $f(n)$ to find an explicit form of $n(d_{\text{low}})$. We recall that we want to round up $\pi_{\text{DB}}(\mathbf{P_A}^{d_{\text{low}}})$, which is inversely proportional to n. As such, we need to round down $n(d_{\text{low}})$, which requires rounding up $f(n)$. Through a study of the function, we find that $0.6197\frac{1.0489n-1.0566}{0.7806n-0.4757}$ is always higher than $f(n)$ $\forall$ $n \geq 1$, while it closely approximates its trend. As such, we approximate (A.103) as

$$f(n) \approx 0.6197 * \frac{1.0489n - 1.0566}{0.7806n - 0.4757} \tag{A.105}$$

Replacing (A.105) in the correct solution from (A.104), we obtain

$$d_{\text{low}}(n) = -\frac{\sqrt{3}\left(1-n\sqrt{k}\right)(0.7806n-0.4757)}{n*0.6197(1.0489n-1.0566)} \tag{A.106}$$

We can now replace (A.106) in (A.102), obtaining

$$n^2\left(0.6197 + 1.0489d_{\text{low}} + 0.7806\sqrt{3k}\right)$$
$$- n\left(0.6197 * 1.0566d_{\text{low}} + 0.7806\sqrt{3} + 0.4757\sqrt{3k}\right) + 0.4757\sqrt{3} = 0$$

Being a second-order polynomial in n, we find two explicit solutions:

$$n\left(d_{\text{low}}\right) = \frac{0.66d_{\text{low}} + 0.4757\sqrt{3k} + 0.7806\sqrt{3}}{1.3d + 1.56\sqrt{3k}}$$
$$\pm \frac{1.31\sqrt{-0.277\sqrt{3}\left(2.6d_{\text{low}} + 3.12\sqrt{3k}\right) + \left(0.5d_{\text{low}} + 0.363\sqrt{3k} - 0.597\sqrt{3}\right)^2}}{2.6d_{\text{low}} + 1.56\sqrt{3k}}$$

Since we know that $n(0) = \sqrt{\frac{t}{F}} = \sqrt{\frac{1}{k}}$, we can conclude that the correct solution is the one with the plus sign.

We have obtained $n\left(d_{\text{low}}\right)$ rounded down, which in turn rounds up $\pi_{\text{DB}}(\mathbf{P_A}^{d_{\text{low}}})$. Next, we aim to round the exponential terms present in the left side of (A.100) to increase it. We first focus on $(1-2a_1)$. This function is monotonically increasing in n, and its limit is

$$\lim_{n\to\infty}(1-2a_1)=2-\frac{2}{\sqrt{3}}$$

The term $(1-2a_1)$ has a positive effect on the left side of (A.100); as such, by rounding up $(1-2a_1)$ we are also rounding up the left side of (A.100). Thus, we approximate

$$(1-2a_1)\approx 2-\frac{2}{\sqrt{3}} \tag{A.107}$$

Next, we focus on $(1+4(1-2a_0)(a_0-a_1))$. This function is monotonically increasing in n, and has a negative effect on the left side of (A.100); since we want to round up that side, we need to round down $(1+4(1-2a_0)(a_0-a_1))$. Through a study of the function, we find that the function $1.36\frac{8n-1}{8n+2}$ is always lower than $(1+4(1-2a_0)(a_0-a_1))\ \ \forall\ n\geq 1$, while it closely approximates its trend. As such, we approximate

$$(1+4(1-2a_0)(a_0-a_1))\approx 1.36\frac{8n-1}{8n+2} \tag{A.108}$$

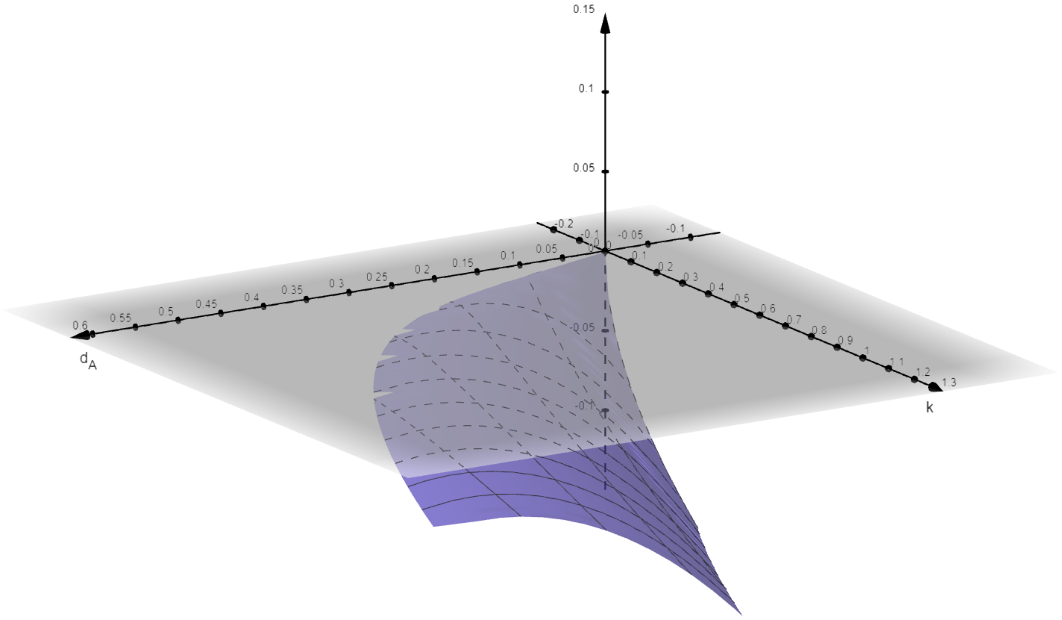
By replacing (A.107) and (A.108) in (A.100) and setting $n=n(d_{\text{low}})$ we obtain

$$\left(2-\frac{2}{\sqrt{3}}\right)2d_{\text{low}}-n(d_{\text{low}})\,d_{\text{low}}^2\left(1.36\frac{8n(d_{\text{low}})-1}{8n(d_{\text{low}})+2}\right)-\sqrt{k}<0 \tag{A.109}$$

The inequality (A.109) only depends on d_{low} and k, with $0<d_{\text{low}}<\frac{1}{n(d_{\text{low}})}$, and $0<k<1$. As such, we can plot the left-hand function for all the proper couples (d_{low},k). As we can see in Figure A.1, the inequality always holds: thus, the DB's strategy is the same as the one adopted in the basic model when $n<\hat{n}$: the DB sets $d_{\text{A}}\geq\frac{3}{2n_{\text{A}}^*}$ with $n_{\text{A}}^*=\frac{1}{2}\sqrt{\frac{t}{F}}$, obtaining $\pi_{\text{DB}}(\mathbf{P_A}^{d_{\text{high}}})=\frac{1}{2}\sqrt{tF}$, while consumer surplus is equal to

$$CS=u-\frac{3t}{4n_{\text{A}}^*}=u-\frac{3}{2}\sqrt{tF}$$

FIGURE A.1. Difference in DB's Profits Between Selling d_{high} and d_{low}



This figure shows the left term of inequality (A.109) for $0 < d_{\text{low}} < \frac{1}{2n(d_{\text{low}})}$ and $0 < k < 1$.

(II) Auction With Reserve Prices

As priorly stated, the difference between the basic model and this extension is that the DB's strategy is directly influenced by the number of entering firms, as she anticipates the effect of the data sale on it. However, under the auction with reserve prices (AR), the number of entering firms is constant, as a losing firm always faces completely informed rivals. Thus, we find that the DB adopts the strategy already described in the proof of Proposition 7. By comparing DB's profits under the *sale to all firms* and the *sale to alternating firms*, we find that the DB still opts for the latter, leading to the market outcomes already described in the proofs of Proposition 9 and 10.

(III) Auction Without Reserve Prices

Under AU, as already shown in the proof of Proposition 12, the DB offers same sized partitions in all the auctions. Without loss of generality, we focus our analysis on firm 0.

Firm 0's profits when winning have already been computed in the proof of Proposition 3. We find

$$\pi_0^{\text{W AU}*}(\mathbf{P_H}) = \begin{cases} \frac{t}{n^2} + \frac{2d_{\text{H}}t}{3n} - \frac{7td_{\text{H}}^2}{18} - F & \text{for } d_{\text{H}} < \frac{3}{2n} \\ \frac{9t}{8n^2} - F & \text{for } d_{\text{H}} \geq \frac{3}{2n} \end{cases} \tag{A.110}$$

In a similar way, firm 0's profits when losing are equal to firm $i{+}1$'s profits derived in the proof of Proposition 3. We find

$$\pi_0^{\mathrm{L\ AU}*}\left(\mathbf{P_H}\right)=\begin{cases}\frac{t}{n^2}-\frac{2d_{\mathrm{H}}t}{3n}+\frac{td_{\mathrm{H}}^2}{9}-F & \text{for } \ d_{\mathrm{H}}<\frac{3}{2n}\\ \frac{t}{4n^2}-F & \text{for } \ d_{\mathrm{H}}\geq\frac{3}{2n}\end{cases}\tag{A.111}$$

Thus, we can compute DB's profits as

$$\max_{d_{\mathrm{H}}}\pi_{\mathrm{DB}}^{\mathrm{AU}}=\begin{cases}\frac{n}{2}\left(\frac{t}{n^2}+\frac{2d_{\mathrm{H}}t}{3n}-\frac{7td_{\mathrm{H}}^2}{18}-F-\left(\frac{t}{n^2}-\frac{2d_{\mathrm{H}}t}{3n}+\frac{\mathrm{td}_{H}^2}{9}-F\right)\right) & \text{for } \ d_{\mathrm{H}}<\frac{3}{2n}\\ \frac{n}{2}\left(\frac{9t}{8n^2}-F-\left(\frac{t}{4n^2}-F\right)\right) & \text{for } \ d_{\mathrm{H}}\geq\frac{3}{2n}\end{cases}$$

which we can rewrite as

$$\max_{d_{\mathrm{H}}}\pi_{\mathrm{DB}}^{\mathrm{AU}}=\begin{cases}\frac{d_{\mathrm{H}}t(8-3d_{\mathrm{H}}n)}{12} & \text{for } \ d_{\mathrm{H}}<\frac{3}{2n}\\ \frac{7t}{16n} & \text{for } \ d_{\mathrm{H}}\geq\frac{3}{2n}\end{cases}\tag{A.112}$$

When $d_{\mathrm{H}}<\frac{3}{2n}$, we obtain the number of entering firms by binding the first part of the piecewise function (A.111), so that

$$\frac{t}{n^2}-\frac{2d_{\mathrm{H}}t}{3n}+\frac{\mathrm{td}_{H}^2}{9}-F=0\tag{A.113}$$

Solving (A.113) with respect to n gives us

$$n_{\mathrm{H}}^*=\frac{9\sqrt{tF}-3d_{\mathrm{H}}t}{9F-td_{\mathrm{H}}^2}\tag{A.114}$$

When $d_{\mathrm{H}}\geq\frac{3}{2n}$, the number of entering firms is constant and given by binding the second part of the piecewise function (A.111), so that

$$\frac{t}{4n^2}-F=0\tag{A.115}$$

Solving (A.115) with respect to n, we obtain

$$n_{\mathrm{H}}^*=\frac{1}{2}\sqrt{\frac{t}{F}}\tag{A.116}$$

By substituting (A.114) and (A.116) in (A.112), we can rewrite DB's profits as

$$\max_{d_{\mathrm{H}}}\pi_{\mathrm{DB}}^{\mathrm{AU}}=\begin{cases}\frac{d_{\mathrm{H}}t\left(72F+td_{\mathrm{H}}^2-27d_{\mathrm{H}}\sqrt{tF}\right)}{108F-12td_{\mathrm{H}}^2} & \text{for } \ d_{\mathrm{H}}<\frac{3}{2n_{\mathrm{H}}^*}\\ \frac{7}{8}\sqrt{tF} & \text{for } \ d_{\mathrm{H}}\geq\frac{3}{2n_{\mathrm{H}}^*}\end{cases}\tag{A.117}$$

By computing FOCs of (A.117) for $d_{\mathrm{H}}<\frac{3}{2n_{\mathrm{H}}^*}$ with respect to d_{H}, we find that DB's profits are monotonically increasing in d_{H}. As such, the DB opts for $d_{\mathrm{H}}^*\geq\frac{3}{2n_{\mathrm{H}}^*}$ and obtains profits

$$\pi_{\mathrm{DB}}^{\mathrm{AU}*}=\frac{7}{8}\sqrt{tF}\tag{A.118}$$

As profits in (A.118) are higher than under the *sale to all firms*, in equilibrium the DB opts for the *sale to alternating firms*. To compute consumer surplus, we can use the general formula for the *sale to alternating firms* provided in the proof of Proposition 10. In our case we obtain

$$CS^{\mathrm{AU}} = u - \frac{5t}{4n^*_{\mathrm{H}}} + \frac{n^*_{\mathrm{H}} t d^{*2}_{\mathrm{H}}}{9} \tag{A.119}$$

where $n^*_{\mathrm{H}} = \frac{1}{2}\sqrt{\frac{t}{F}}$ and $d^*_{\mathrm{H}} = \frac{3}{2n^*_{\mathrm{H}}}$, as it is the limit case after which data exhaust their marginal effect. We can rewrite (A.119) as

$$CS^{\mathrm{AU}} = u - \frac{t}{n^*_{\mathrm{H}}} = u - 2\sqrt{tF}$$

(IV) Take It Or Leave It Offers

As already described in the proof of Proposition 12, the *sale to alternating firms* is always suboptimal under TIOLI, as the DB cannot properly threat firms when they refuse to buy data. Thus, the DB opts fort the *sale to all firms*, leading to the market outcomes already described in the first part of this proof.

(V) Comparison

By comparing the market outcomes under the three selling mechanisms, we find that DB's profits decrease and consumer surplus increases as the DB's bargaining power is reduced. However, consumer surplus is always lower than in the benchmark case, and the number of entering firms is always minimised.

CHAPTER 2

The Impact of Privacy Regulation on Web Traffic: Evidence From the GDPR[*]

Raffaele Congiu[†], Lorien Sabatino[†], Geza Sapi[‡]

We use traffic data from around 5,000 web domains in Europe and United States to investigate the effect of the European Union's General Data Protection Regulation (GDPR) on website visits and user behaviour. We document an overall traffic reduction of approximately 15% in the long-run and find a measurable reduction of engagement with websites. Traffic from direct visits, organic search, email marketing, social media links, display advertising and referrals dropped significantly, but paid search traffic – mainly Google search ads – was barely affected. We observe an inverted U-shaped relationship between website size and traffic reduction due to privacy regulation: the smallest and largest websites lost visitors, while medium ones were less affected. Our results appear consistent with the view that users care about privacy and may defer visits in response to website data handling policies. Privacy regulation can impact market structure and may increase dependence on large advertising service providers. Enforcement matters as well: the effects were amplified considerably in the long-run, following the first significant fine issued eight months after the entry into force of the GDPR.

[*] We thank Similarweb (www.similarweb.com) for providing us the data used in this work. The views expressed in this article are solely those of the authors and may not, under any circumstances, be regarded as representing an official position of organizations the authors are affiliated with.

[†] Politecnico di Torino, Department of Management, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy.

[‡] Düsseldorf Institute for Competition Economics (DICE), Heinrich Heine University of Düsseldorf.

1. Introduction

The rise of the internet has enabled new products and services, revolutionising the ways we work and interact. In the United States (US), as in other mature economies, the growth of the internet economy exceeds that of other sectors by orders of magnitude (Interactive Advertising Bureau, 2021). The shift of private and commercial activities to the internet goes hand in hand with increased attention to online security, privacy and the protection of data.[1] In recent years, policy makers around the world reacted to increased data protection concerns by putting in place a wave of regulations with the aim of protecting online and offline privacy, thereby limiting the ways companies collect and use data.[2] Typically, the goal of such privacy regulations is to empower consumers by giving them additional control over how firms gather and use their personal information.

One of the most far-reaching regulatory interventions to boost privacy is without doubt the General Data Protection Regulation (GDPR) of the European Union (EU). The GDPR entered into force in May 2018 with the intention of protecting personal privacy across all domains, both online and offline. The regulation mandates data protection "*by design and by default*" and prescribes several principles and tools for managing data. It requires data controllers to inform individuals about the collection and use of data. Firms collecting personal data must obtain the users' informed opt-in consent to such practices. The regulation assigns liability to firms handling data and – for the first time – imposes significant fines on privacy breaches up to 4% of global turnover.

While the GDPR applies both online and offline to any entity handling personal information, no other economic domain is expected to be as heavily affected by it as the internet ecosystem. Today's internet relies heavily on the collection and use of visitor data. Websites attract visitors by offering a broad spectrum of content ranging from news, music and entertainment, to commerce and other services. More often than not, websites monetise at least in part by tracking their visitors and providing targeted advertisements to individuals or homogeneous groups of users. Digital stars such as Google and Facebook grew to titans of the internet by harvesting user data from millions of websites,

[1] In Europe, in 2011 around 40% of survey respondents were concerned about their behaviour being recorded through the internet when browsing, downloading files, and accessing content online (European Commission, 2011), in 2015 less than a quarter of respondents reported to trust online businesses to protect their personal data (European Commission, 2015). Growing online privacy concerns likely followed the general trend in internet use (Ourworldindata.org, 2021), and likely reach back to the early 2000s (Auxier et al., 2019).

[2] At the time of writing this article, the Australian Privacy Act of 1988 is under review (Coos, 2020), and legislation strengthening privacy is under way in Canada and the United States (Government of Canada, 2020; IAPP, 2021).

connecting users' browsing paths, aggregating these signals into user profiles and offering advertising services to third parties to target consumers based on their specific interests and characteristics.[3]

Since its introduction, the GDPR has attracted considerable interest from researchers, policy-makers, and industry players across the globe.[4] So far, less attention has been devoted to understanding the consequences of the GDPR on website's ability to attract internet users and the altered ability and likelihood of those users to engage with website content. Our research aims to fill this gap. We investigate the effect of the GDPR on online traffic and user behaviour in the European Union (EU).

We use data on traffic – broken down by channel – for about $5,000$ websites in Europe and the United States. We exploit the fact that the GDPR applies to European users, leaving the non-European audience unaffected. We apply a difference-in-differences (DiD) analysis that exploits the geographic origin of website traffic. In particular, our treatment assignment identifies the traffic originated from EU countries, leaving US traffic in the control group. This implies that multinational websites having visitors from both the EU and abroad are treated only for the portion of traffic coming from EU countries.

Overall, we find that the GDPR led to a reduction in web traffic, and website visits decrease by approximately 15% in the long-run. This effect unfolded fully with a delay, several months after the hard date of the GDPR entry into force, and following the issuance of the first large fine. We break down the overall traffic reduction into detailed traffic acquisition channels. In the short-run, direct website traffic and visits triggered by email marketing messages reduced by 4.5% and 7% respectively. In the long-run – coinciding with the first large fine – traffic reduction amplified significantly. Email and display advertising traffic collapsed by 35% and 29%, respectively. Visits from referrals (such as links on 3rd party websites) and from social media reduced too, together with direct (-9.5%) and organic search visits (7%). Strikingly, against this backdrop we find that website traffic from paid search – mainly Google search advertisement – was barely affected.

The GDPR was much anticipated and was feared to affect predominantly small businesses negatively (Kottasová, 2018; BBC, 2018). We in turn find an inverted U-shaped

[3] The tracker networks of Google and Facebook have been found embedded into respectively 80% and 37% of the most popular websites (Hu et al., 2020).

[4] The first research articles assessing the impact of the GDPR came from the computer science domain, and revolved around measuring websites' technical compliance with the regulation (Nouwens et al., 2020; Utz et al., 2019; Matte et al., 2020; Sørensen and Kosta, 2019).

relationship between website size (measured in visits) and the traffic reduction due to the privacy regulation. Small websites lost traffic, but so did the largest ones, while medium websites remained unaffected and may have even grown. On the intensive margin, we observe a significant deterioration of user engagement with websites following the GDPR, both in the short and in the long term. We estimate a significant reduction in average visit duration and the number of web pages visited, as well as a measurable increase in website bounce rate – the share of website visitors that leave almost immediately after arriving on a website.

This paper makes several contributions to the fast-growing literature on the impact of the GDPR in the digital market. First, we investigate how the GDPR affected web traffic and user engagement, which measure the extensive and intensive margins of internet consumption. Second, we are able to explore both short- and long-run effects of the GDPR. This is important, as most of the effects unfold several months after the enactment of the regulation. Third, contrary to other related studies (Peukert et al., 2020; Goldberg et al., 2019), our empirical strategy is based on the geographical source of website traffic, rather than on the country domain.[5] In such a framework, the treatment assignment targets precisely the European audience that falls under the scope of the GDPR. Finally, we assess empirically the differentiated effects of privacy regulation based on website size (Campbell et al., 2015; Dimakopoulos and Sudaric, 2018; Sabatino and Sapi, 2019). To the best of our knowledge, we are the first in providing evidence of an inverted U-shaped relationship between website size and privacy regulation effects.

All in all, we document and measure some anticipated and less anticipated effects of the GDPR. While the effects we emphasise here indicate an overall loss of traffic via most channels and less interaction with websites, this in itself does not imply that the GDPR is a harmful regulation. On the contrary: the decline of direct website traffic – the largest source of website visits – may be the result of users' conscious choice, after the increased prominence of privacy policies and websites' pop-ups to obtain consent to these policies.

The remainder of the paper is organised as follows. Section 2 presents an overview of the related literature and our contribution. Section 3 describes the main provisions of the GDPR. Section 4 introduces the data used in our empirical analysis. Section 5 discusses our empirical strategy. Section 6 reports our main findings. Section 7 provides robustness checks. Section 8 concludes.

[5] A country domain is for example *.de* for Germany, and *.fr* for France in a website's URL.

2. Literature Review

Our research draws from the growing body of the economic literature on data and privacy regulation, surveyed recently and extensively by Acquisti et al. (2016). This literature emphasises the trade-off between consumer protection from data exploitation and market efficiency. In their seminal paper, Goldfarb and Tucker (2011) study the effectiveness of advertising in the EU after the implementation of the 2002 EU ePrivacy Directive, finding that the policy change reduced advertising effectiveness. Campbell et al. (2015) argue that privacy regulation imposes costs on all firms, but disproportionately on smaller ones. On the contrary, focusing on the 2009 European E-privacy Directive, Sabatino and Sapi (2019) find that mainly large firms were negatively affected, while small firms experienced no significant negative effects. Our results partially support both views: while we observe the greater traffic reduction for small websites, the largest ones lose traffic as well, while medium ones remain unaffected.

Our work adds to the recent line of empirical research studying the impact of the GDPR in digital markets. Gal and Aviv (2020) argue that, under certain market conditions, the GDPR may reduce welfare by increasing data market concentration and limiting synergies. Johnson et al. (2020) find a 15% reduction in website operators' use of web technology providers following the GDPR, although this drop was reabsorbed within a few months. Moreover, they find that small providers were more strongly affected, and that Google and Facebook's market shares increased as a consequence of websites substituting from smaller to larger technology vendors. Peukert et al. (2020) report a decrease in third-party trackers and cookies and an increase in first-party trackers following the introduction of the GDPR. They identify a significant increase in the prominence of the leading firm, Google, at the expense of competitors.

In the same strand, Libert et al. (2018) observe significantly less third-party cookies set without consent on a sample of EU news websites in the period immediately before and after the introduction of the legislation, but Sørensen and Kosta (2019) do not find conclusive evidence of a reduction in the number of third party requests. Still in this strand, Aridor et al. (2020) investigate the effect of the GDPR on firms' ability to track users and generate revenues through online advertising. Their findings highlight a reduction in the number of consumers, but also an increase in the effectiveness of tracking on the remaining users. We contribute to this literature by providing evidence of the impact of the GDPR on website traffic – a measure of consumption in the digital domain – and user engagement.

Two recent papers are closely related to our research. Goldberg et al. (2019) examine the GDPR's impact on website pageviews (and revenue) for over a thousand websites, using data for 32 weeks around the GDPR's enactment from Adobe's website analytics platform. The study documents a reduction of approximately 12% in both EU user website pageviews and website e-commerce revenue after the GDPR's enforcement deadline. The reduction is larger for traffic originating from email and display advertisement. The authors find no evidence that consent interfaces would dissuade users from browsing sites.

Our paper departs from Goldberg et al. (2019) in several respects. First, our data includes more than $6,500$ country-domain pairs. The large number of domains in our dataset allows us to analyse heterogeneous effects by domain size category (in terms of visits). This allows us to identify significantly different effects of the privacy regulation by website size. Second, we analyse a longer time period, with our data encompassing 95 weeks, from the 5th of January 2018 to the 25th of October 2019. This is an important addition, because we document larger effects in the long-run, possibly linked to the issuance of the first large GDPR fine in early 2019. Finally, our identification strategy differs from that of Goldberg et al. (2019). The authors take as preferred control group the same set of sites in the year before GDPR enactment. Since we observe traffic by source country, we can assign treatment based on the geographic source of traffic, rather than a domain's location.[6] Despite these differences, some of our main findings are remarkably close to those of Goldberg et al. (2019). Most notably, we also observe an around 10% reduction in website visits, and a particularly large reduction of traffic originating from email marketing and display advertising.

Another closely related paper developed independently from ours is Schmitt et al. (2021). The authors analyse the impact of the GDPR on web traffic and user engagement using the same data provider as we do in this article. In line with our findings, Schmitt et al. (2021) also report an increasing long term negative effect of the GDPR on web traffic. Our study differs from Schmitt et al. (2021) in several ways. First, we exploit additional disaggregation of the data to observe not only the country of traffic origin but also the channel that generate traffic. This is important, as we find that some paid traffic

[6] Since the GDPR applies to EU residents, regardless where the website they visit is located, our identification strategy is fully in line with the logic of the GDPR.

acquisition channels are particularly adversely affected. Second, we find significant long-run effects, consistent with the step up of enforcement and the issuance of the first large fine by the French regulator in January 2019. Finally, we are the first to provide evidence on a possible inverted U-shaped relationship between the traffic reduction induced by the privacy regulation and website size, as measured by the number of visits before the GDPR.

There is a small but growing body of empirical literature analysing the broader economic impact of the GDPR. Jia et al. (2021) study the effect of the regulation on venture capital investment in new technology firms, finding a reduction in the number of monthly deals in the EU compared to the US. Zhuo et al. (2021) study the effect of the GDPR on interconnections among network operators. The authors report that the legislation had no impact on the number and type of agreements at the internet layer and may have had a minor effect on the entry and the number of network customers. Through an experiment with a large telecom provider, Godinho de Matos and Adjerid (2021) find evidence suggesting that more informed final user consent to data processing mandated by the GDPR can have positive effects to both consumers and firms.

3. Institutional Background

The General Data Protection Regulation (GDPR) was passed in April 2016 and came into effect on the 25th of May 2018 across the EU. The regulation is labelled "general" because it applies to any firm handling data, offline and online likewise. The regulation was well anticipated, and the market had a two-year period between enactment and entry-into-force to prepare. The GDPR aims to strengthen and harmonise the legislation concerning privacy, data collection and processing within the European Union. It reshapes the way personal data of EU residents are collected and processed, providing new rights to access and control these data. These improved user rights relate mainly to access, rectification, erasure, and portability of user data. The GDPR places the burden of justifying data processing with lawful motivation upon data controllers, who are obliged to obtain explicit and informed consent from data subjects in order to handle those data. The legislation imposes obligations on organisations operating anywhere in the world as long as they collect or process EU residents' data. Furthermore, it imposes large fines for infractions, up to €20 million or 4% of worldwide annual revenues, whichever is higher.[7]

[7] The global scope of the legislation and the severity of the fines are key differences compared to Directive 95/46/EC, which the GDPR replaced.

The GDPR can impact website traffic by reducing the intensity and effectiveness of online advertising in (at least) three ways. First, by interfering with tracking technologies such as cookies, trackers, fingerprinting, beacons – which are used to gather data on users' activities throughout the Internet.[8] Typically, such tracking happens without the user being aware of it (The Economist, 2014). The GDPR affects this mechanism by limiting data processing to well-specified cases.[9] Moreover, a website using web technologies to identify and target users can do so only after obtaining explicit and informed consent.[10] Restrictions in data collection and processing can reduce both the effectiveness and the intensity of advertising (Goldfarb and Tucker, 2011). The intensity can be reduced because of higher costs incurred by firms to perform advertising or because some forms of it become unlawful. For example, email advertising may decrease because email lists might have been redacted as a result of the GDPR since email addresses qualify as data protected by the regulation. Therefore, firms cannot send unsolicited emails to any address they collected. On the other hand, the effectiveness of advertisement may decrease because, as fewer users consent to data collection and processing, advertisement becomes less tailored to the specific user.

The second way through which the GDPR can impact online advertising is by affecting the perceived legal risks of website operators. Under the GDPR, websites are jointly responsible with third-party providers. As a consequence, both may reduce advertisement to avoid the risk of sanctions. Peukert et al. (2020) find evidence that fear of legal repercussions led to an increase in concentration in the website technology vendor market, as website operators perceived top players as more trustworthy.

Finally, the GDPR can impact advertising by affecting online user behaviour. From its approval to its enforcement, the legislation has been widely discussed, bringing privacy

[8] Through the collection and processing of these data, the operator – either the website or a third-party web technology vendor – can infer user preferences: websites can exchange and pool visitor identifiers and visit logs. By combining these logs and linking them to individual visitors, the online advertising industry is able to create detailed user profiles with visit histories, interests and purchases. This information can then be used for targeting advertisements to the user when she visits an ad-supported webpage.

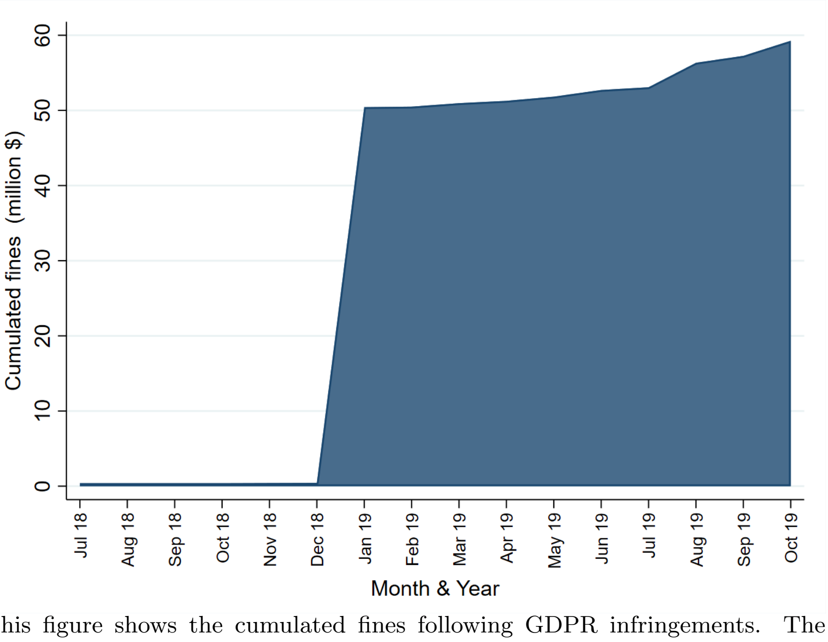
[9] Specifically, the fulfilment of a legal obligation or a contract, the protection of vital interests of a person, reasons of public interest, firm or a third party's legitimate interest and, finally, consumer consent. Whether online tracking does or does not fall into these motivations is debated. At the very least however, the GDPR increased the risk for firms to wrongly classify their practices into these bins, since they are liable and are subject to fines.

[10] The GDPR is not the first regulation to introduce the prerequisite of user explicit consent (opt-in) for data collection. Already under the revised 2009 ePrivacy Directive, consent was needed to place cookies on a users' device. Thus, the key difference of the GDPR has been to extend the scope of the territorial applicability, to sensibly increase fines in case of infringements, and to adopt a technology-neutral approach.

concerns to the forefront of public debate.[11] The boom in online pop-ups asking for consent to place cookies on the user's device has been hard to miss. European internet users will also long remember the storm of emails from various websites asking to opt-in to various mailing lists before and shortly after the enactment of the GDPR. The greater awareness, in turn, can influence the degree to which users actively try to protect their privacy. User can impact advertising by avoiding websites that are recognised as more intrusive, choosing not to opt-in to data handling policies, or increasingly using technologies that limit tracking and advertisement.

From a legal perspective, there has been little enforcement following the entry into force of the GDPR. In the second half of 2018, only nine fines for GDPR infringements were imposed around Europe, for a total of only €440,000. The first significant fine – the third largest to date – came on the 21st of January 2019. On that occasion, the French regulator imposed a €50 million penalty on Google "*for lack of transparency, inadequate information and lack of valid consent regarding the ads personalization*" (CNIL, 2019). That fine was perceived by some commentators as the end of an unofficial grace period, with The Economist going as far as calling it "*the start of a war*" (The Economist, 2019).

Figure 1. Cumulated Fines Under the GDPR



This figure shows the cumulated fines following GDPR infringements. The jump in January 2019 refers to the financial penalty imposed by the French privacy authority (CNIL) against Google. Source: GDPR Enforcement Tracker at https://www.enforcementtracker.com/.

[11] A search on Google Trends on the term "*GDPR*" illustrates this well: there was a large surge in English language web-searches on this keyword around May 2018, the entry into force of the regulation. https://trends.google.com/trends/explore?date=2018-01-01%202019-12-31&geo=GB&q=gdpr.

Figure 1 shows the cumulated fines from July 2018 to October 2019. The jump in January 2019 coincides with the French Google penalty. In retrospect, the following months do not show an increase in regulatory enforcement on a comparable level with the expectations of the time. However, the prominence of the first large fine – issued to a company as visible as Google – might have influenced firms' compliance decisions.

4. Data

4.1. Data Description and Summary Statistics

Our main source of data is Similarweb, a web analytics service provider.[12] It tracks information on website traffic volume – i.e. the total number of visits for each domain – as well as several measures of user online behaviour. The provider records daily data along the following key metrics:

- **Visits**: Total visits to a domain via all traffic channels. A visit is the access of a user to one or more pages of a website. All subsequent page views belong to the same visit, until the visitor is inactive for 30 minutes.[13]
- **Average visit duration**: The average visit length measured in seconds, which is calculated as the time elapsed between the first and last page view of a visit.
- **Average pages per visit**: The average number of pages viewed per visit.
- **Bounce rate**: The share of users who close the website after landing on it without loading other pages.

Visits is our main outcome of interest. We refer to the other three outcomes as engagement measures, as they provide information on the way users interact with the website.

A website is defined by its domain, so we use both terms interchangeably.[14] A domain can receive traffic from different countries and through different *channels*. For instance, a visitor from the UK can land on an e-commerce website registered anywhere in the world through an advertisement seen on a social platform. Our data allow us to observe the visits by country of origin of the visitor, as well as the traffic channel through which the website was reached. In particular, Similarweb categorises traffic into the following channels:[15]

[12] www.similarweb.com

[13] We use the terms visits and traffic interchangeably.

[14] This means that we treat google.com and google.fr as different domains. While a domain like sites.google.com may host several websites, they constitute a single domain for our analysis.

[15] Note that these are the same traffic channels Google Analytics tracks. They can be therefore seen as the industry standard classification.

- **Direct**: Website visits by users directly typing the website address into the URL bar in a browser. Direct traffic also covers clicking on a bookmark or on a link from outside the browser (but not in emails).
- **Display Advertising**: Traffic from users clicking on an advertising banner or video advertisement shown on a third-party website (except social networks).
- **Email**: Website visits following the user clicking on a link provided by a web-based mail client. For example, clicks in marketing emails on links pointing to the website.
- **Organic search**: Traffic originating from the organic (i.e. non-paid) results in an internet search engine, such as Google or Bing.
- **Paid search**: Traffic originating from clicks on paid advertisements on the result page of a search engine, such as Google or Bing.
- **Referral**: Traffic generated by third-party links on most other websites than social networks and search engines. This includes paid links in blogs and other affiliates, and free traffic such as media coverage.
- **Social**: Traffic coming from social media websites such as Facebook or YouTube, either through users posting links for free or advertisement.

Finally, for every domain Similarweb also reports the share of traffic originating from each country at a monthly level, along with domain's traffic rank in that country.

We select the largest websites active in the US and in major European countries. In particular, we observe the top 1,000 domains in five EU countries: France, Germany, Italy, Netherlands and United Kingdom, together with the top 2,000 US domains in terms of visits. Since domains are ranked by traffic volume within each country, we collect 7,000 domain-country pairs. However, as some domains belong to more than one country's top thousand (e.g. amazon.com), we have data for 5,300 unique domains. Of these, we keep the 4,957 websites that were active both before and after the GDPR.

Traffic data is further disaggregated by channel, so that the cross-section is defined by the triple domain-country-channel and comes in daily frequency.[16] Our data span two years from the 1st of November 2017 to the 31st of October 2019, covering about seven months before and 17 months after the GDPR's implementation. The panel is unbalanced, as some domain-country couples have missing data for various periods, with

[16] An observation in our dataset for example corresponds to the number of daily visits to bbc.com in the United Kingdom via clicks on organic search results.

no clear pattern.[17] Since traffic data show a high degree of daily volatility, we aggregate the observations at the weekly level. We define a week as a Friday-to-Friday period, as the GDPR entered into force on Friday, 25th of May 2018. The panel covers 94 weeks, from the 5th of January 2017 to the 25th of October 2019. The final dataset covers around 4 million observations at the channel-domain-country-week level. Summary statistics are provided in Table 1.

TABLE 1. Summary Statistics

	Mean	**SD**	**Min**	**Max**	**Observations**
Overall					
Traffic	315,112	9,973,771	0	1,970,498,007	4,338,812
Avg. Visit Duration (seconds)	394.45	876.99	0	86,235	3,844,734
Avg. Pages Visited (N)	6.40	9.54	0.10	1,263.59	3,984,617
Bounce Rate	0.45	0.24	0.00	1.00	3,839,368
Traffic by channel					
Direct	1,485,277	26,116,612	17	1,970,498,007	619,999
Display Advertising	13,611.48	92,249.68	0	6,042,370	619,760
Email	55,130.18	614,207.90	0	83,178,035	619,959
Organic Search	443,680.79	3,275,475	0	225,132,942	619,998
Paid Search	21,493.86	191,691.46	0	21,070,592	619,106
Referrals	103,117.49	618,041.18	0	28,195,818	619,995
Social	82,916.88	896,998.53	0	51,305,963	619,995
Traffic by size					
Small	46,461.67	99,237.95	0	5,154,861	1,956,409
Medium	146,163.58	330,465.12	0	9,629,733	1,241,459
Large	317,105.24	1,457,990	0	40,895,073	986,016
Giant	5,048,721	52,419,507	0	1,970,498,007	154,928

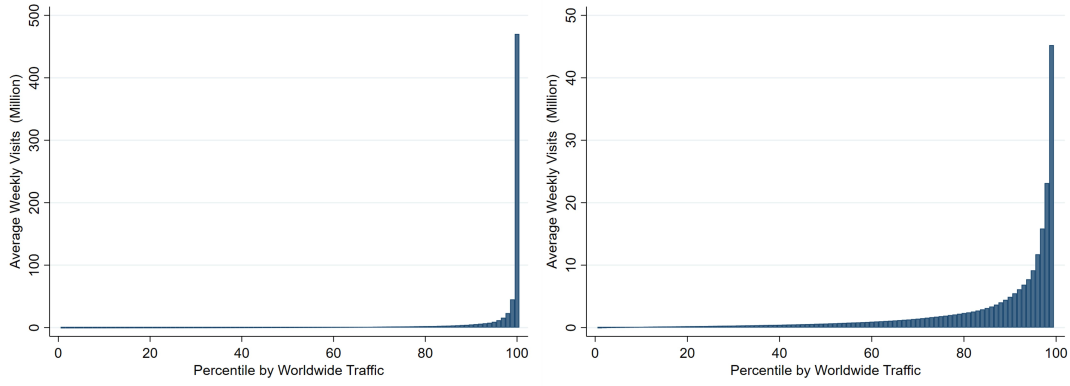
This table shows summary statistics for the main variables used in the empirical analysis. Source: authors' elaboration of Similarweb data.

We consider our list of websites very extensive. The websites in our dataset cover a very significant share of websites typical users in the selected countries ever visited. Website traffic appears highly concentrated, as reflected in our data: Figure 2 shows the average weekly traffic by percentile of the distribution of domain's worldwide traffic before the GDPR.[18] The traffic for every percentile is shown on the left panel, while the right panel removes the top 1st percentile to avoid flattening the scale. It is evident that the distribution is highly skewed towards the top: the handful of most visited websites attract the vast majority of internet visits. Therefore, selecting the top 1,000 domain by country (2,000 for the US) ensures that our sample covers a high share of internet traffic in these countries and can be regarded as representative.

[17] Missing observations do not happen systematically over time or across websites.

[18] We obtain this distribution by taking each domain's weekly worldwide traffic (i.e., not limited to the six countries in our analysis) and averaging it across the time periods that precede the introduction of the GDPR. Then, for each percentile, we calculate the average weekly traffic depicted in Figure 2.

Figure 2. Concentration in the Digital Market

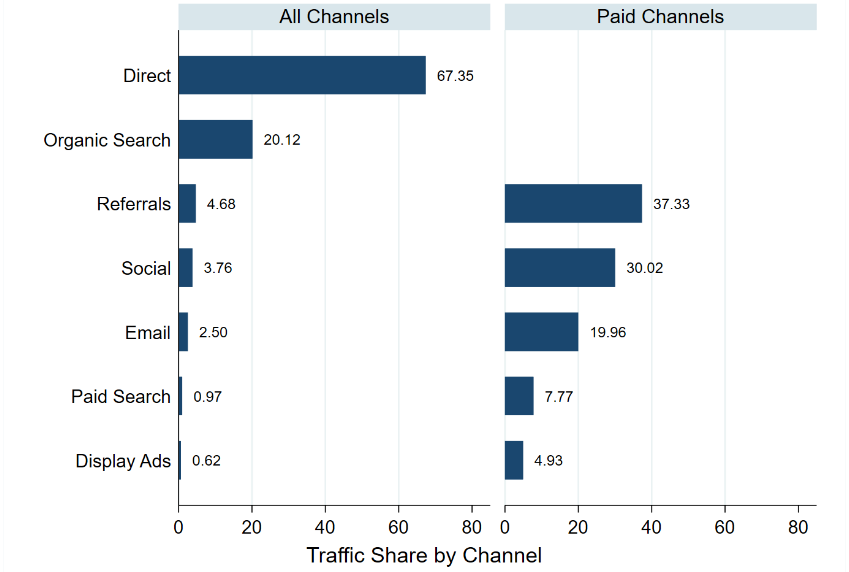


This figure shows average visits (in thousands) for each percentile of the distribution of worldwide traffic before the GDPR. On the left, all percentiles are listed. On the right panel, the top 1st percentile is excluded. Source: Similarweb data.

Our data by channel allows us to distinguish between paid and unpaid traffic. The former generates either a remuneration to a third party (e.g. advertisement fees for Google) or is the result of an internal marketing strategy (e.g. email campaign, where email addresses may be purchased). The latter does not induce any remuneration to a third-party website, such as the traffic generated by directly typing a domain's URL into the browser. In this view, we consider Direct traffic and Organic Search as unpaid traffic, while the other channels are labelled as paid.

Figure 3 shows the traffic distributed by channel. The left panel reports the contribution of each channel on total website traffic. Most of the traffic comes from the Direct channel, amounting to about 67% of total website traffic. Organic search makes up for another 20%. Taken together, these unpaid channels amount to almost 90% of overall traffic. The remaining 12% of overall website traffic can be regarded as *paid.* The right panel of Figure 3 shows the channel shares for paid traffic only. While referrals and social networks constitute the bulk of paid traffic, each channel has a non-negligible share, with email amounting to 20% of website paid traffic.

FIGURE 3. Distribution of Web Traffic Across Channels



This figure shows the traffic share by channel. On the left, all traffic is considered when determining each channel's share. On the right, only paid traffic is considered. Source: Similarweb data.

4.2. Recording Bias

A potential concern affecting our data is that the GDPR may have had an impact on the ability of our data provider to measure website traffic and usage. Goldberg et al. (2019) refer to this as the *recording bias.* For example, if users increasingly opt-out of any form of data sharing with third-parties, websites may have been less able to report data about these metrics. We took several steps to investigate to what extent the recording bias may arise on our data.

First, we explicitly inquired with Similarweb about a potential bias in measurement associated with the GDPR. Similarweb confirmed in writing that this was not the case.

Second, according to publicly available information and Similarweb's own account, the data provider combines data from several sources, including directly from websites, internet providers, public data sources as well as an anonymised panel of users through browser extensions.[19] Much of these data are not individually identifiable, and therefore largely fall outside the GDPR.[20] No user-level data is necessary for any of the analysis we do. Similarweb also declares publicly to not rely on any personally identifiable data in its

[19] https://licreativetechnologies.com/seo/how-to-track-website-traffic-similarweb/, retrieved on the 1st of February 2022.

[20] This even applies to such seemingly sensitive items as gender, which we do not use in our analysis: https://support.similarweb.com/hc/en-us/articles/360001253797-Website-Demographic-Data, retrieved on the 1st of February 2022.

collection and elaboration process, again, largely rendering it immune to measurement issues due to the GDPR.[21] In particular, it explains that: "*We employ a multi-step verification process to ensure data collected is devoid of any Personally Identifiable Information (PII) [...] Behavioral data is shared anonymously and aggregated at the site- and app-level rather than the user-level [...] Data is never used for advertising or targeting, and we don't use "cookies" to collect behavioral data.*"

Third, we also gather additional information on Similarweb's user panel, which is partly based on tracking users' browsing behaviour through its own browser extension.[22] Similarweb makes users of the web extension aware of being tracked, and in exchange it provides statistics on the websites they visit. In February 2022, the add-on had more than $800,000$ users in the Chrome web store alone, not counting users of other browsers. Given that we focus on the top few thousand websites in the largest European countries and the United States, it appears to us that solely based on such a panel Similarweb would be able to capture the metrics we analyse in great detail. As the Chrome extension was already available on the 10th of August 2017, it appears unlikely that users of such a browser extension would have revoked their consent to be tracked due to the GDPR.[23]

Finally, Similarweb is a widely audited industry-standard source of website traffic information, providing data to several companies engaged deeply in web-traffic measurement, such as Adobe, Google and The Economist, as well as to researchers (Calzada and Gil, 2020; Lu et al., 2020; Schmitt et al., 2021). Schmitt et al. (2021) compare Similarweb's data with a German data provider on web audience (AGOF) which was likely not affected by GDPR, finding no significant deviation between the two sources.

Overall, we conclude that our data are fit for purpose and unlikely to suffer from the recording bias. We now turn to the results of our main analysis.

5. Empirical Model

Our main interest is understanding how the GDPR affected website traffic as well as user behaviour on the internet, in the short and in the long term. Our empirical strategy

[21] https://support.similarweb.com/hc/en-us/articles/360001631538-Similarweb-Data-Methodology, retrieved on the 1st of February 2022.

[22] See for example https://chrome.google.com/webstore/detail/similarweb-traffic-rank-w/hoklmmgfnpapgjgcpechhaamimifchmp, retrieved on the 2nd of February 2022.

[23] We found the first historic version of the web page from the *Wayback Machine*, https://web.archive.org/web/20170810122245/https://chrome.google.com/webstore/detail/similarweb-traffic-rank-w/hoklmmgfnpapgjgcpechhaamimifchmp

applies a difference-in-differences (DiD) approach to compare our outcomes of interest in EU countries with the US, before and after the introduction of the GDPR.

Our data are unique as they allow us to identify the geographic origin of a website's audience, as opposed to the country of registration of the domain. Our treatment assignment reflects the main provision of the GDPR, since it is defined as the traffic coming from one of the EU countries that fall under the GDPR. Therefore, our definition of treatment and control groups is very precise: it does not derive from the domain of the observed website, but rather on the geographical source of traffic for the specific website. As a consequence, each website can be partially treated if its traffic comes from both the US and EU countries. Another important dimension that we exploit refers to the different types of traffic channel. That is, for a specific geographical source of traffic, we observe the channel through which the website is reached. This further increases the sources of variation in our analysis, and allows us to investigate potential heterogeneous effects of the GDPR on specific traffic channels.

Our baseline model takes the following form:

$$y_{i,k,c,t} = \beta_0 + \beta_1 Post_t + \beta_2 EU_c + \beta_3 Post_t \times EU_c + \alpha_i + \kappa_k + \epsilon_{i,k,c,t} \tag{1}$$

where $y_{i,k,c,t}$ is one of our outcomes of interest[24] for website i, from channel k and country c at time (weekly date) t.[25] *Post* is an indicator identifying the period from the introduction of the GDPR onward (the 25th of May, 2018). *EU* is a dummy identifying whether the country from which website traffic comes is an EU member state. α and κ are website and channel fixed effects respectively. Finally, ϵ is a mean-zero error term.

The main coefficient of interest in Equation (1) is β_3, which captures the average causal impact of the GDPR on website outcomes across channels. For example, taking website visits as the dependent variable, the estimated coefficient measures the average (percentage) difference in traffic coming from the EU and the US induced by the GDPR.[26] Since we also want to assess the main channels through which the GDPR affects web traffic, we further interact *Post*, *EU* $Post \times EU$ with a full set of dummies identifying

[24] Our dependent variables are *visits*, *average visit duration*, *pages visited*, and *bounce rate*. For the first three variables we apply the following monotonic increasing transformation $\ln(x+1)$, so as to include null values in the estimation. Since the bounce rate takes values between zero and one, we do not apply any transformation to it.

[25] As explained in Section 4, the traffic acquisition channels are *Direct*, *Display Advertising*, *Paid* and *Organic Search*, *Referrals*, *Social* and *Email*.

[26] The logic is analogous when the depend variable is one of our engagement measures.

traffic channel k. In this setting, we estimate differentiated causal effects of the GDPR for each traffic channel.

Equation (1) does not allow us to disentangle short and long term impacts of the GDPR. In particular, if website operators anticipated weak enforcement of the regulation during the subsequent months following the GDPR's introduction, then we might expect to see little effect during such a period. In this case, β_3 would underestimate the true effect of the regulation. For this reason we also run the following alternative model

$$y_{i,k,c,t} = \beta_0 + \beta_1^S Post_t^S + \beta_1^L Post_t^L +$$

$$+\beta_2 EU_c + \beta_3^S Post_t^S \times EU_c + \beta_3^L Post_t^L \times EU_c + \alpha_i + \kappa_k + \epsilon_{i,k,c,t} \tag{2}$$

where $Post^S$ identifies the period between the 25th of May 2018 and the week before the 21st of January 2019, that is the period between the entry into force of the GDPR and the French fine on Google, while $Post^L$ is active from the end of January 2019 onward. Thus, in this specification the coefficient β_3^S identifies the short-run causal impact of the GDPR, while β_3^L identifies the long-run effect on the outcome variables. Additional interactions of $Post^S$, $Post^L$, EU, $Post^S \times EU$, and $Post^L \times EU$ with channel-specific dummies allow us to identify diverse effects across traffic channels both in the short and in the long term.

Finally, to understand the dynamics behind the estimated parameters of interest in both (1) and (2), we interact our treatment (EU) dummy with time specific dummies, leading to the following equation:

$$y_{i,k,c,t} = \beta_0 + \sum_t \beta_{1,t} Time_t \times EU_c + \alpha_i + \tau_t + \gamma_c + \kappa_k + \epsilon_{i,k,c,t} \tag{3}$$

where τ and γ denote time and country fixed effects respectively. We also estimate Equation (3) by traffic channel in order to assess potential differential dynamics across those channels.

5.1. Parallel Trends Test

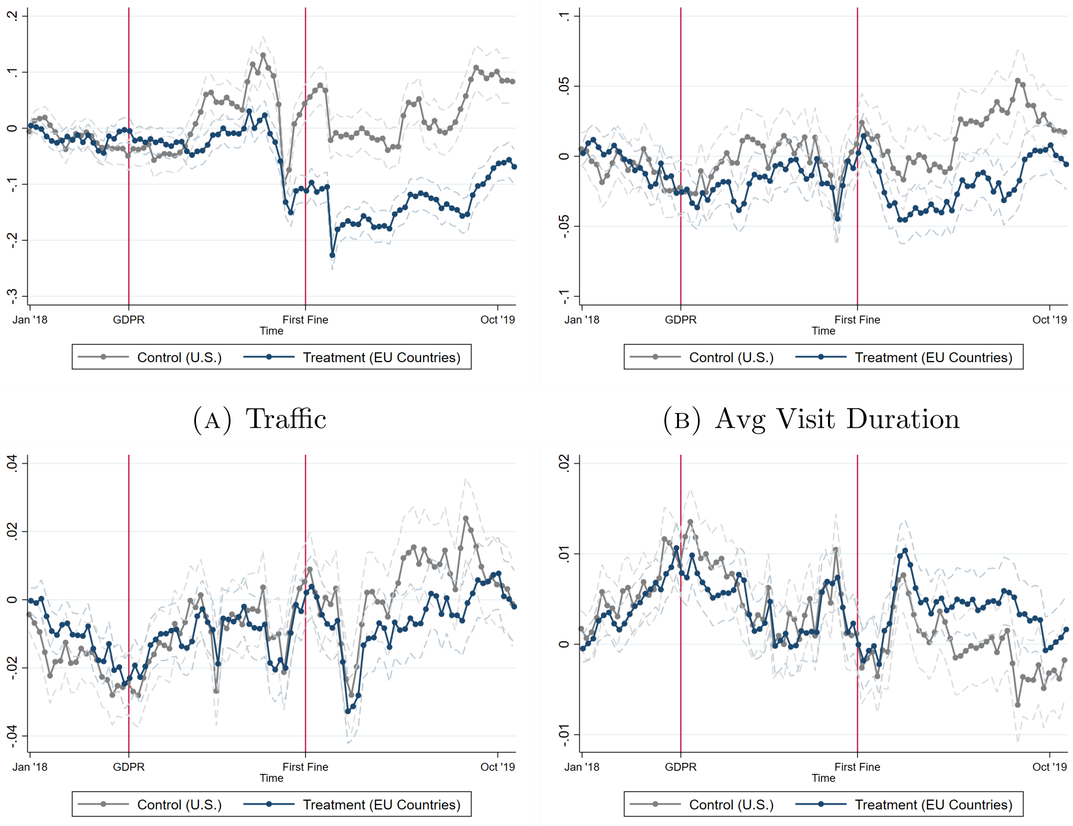
Equations (1) and (2) identify the causal impact of the GDPR on web traffic and user engagement if US online behaviour provides a valid counterfactual for EU countries. The identifying assumption is that, absent the GDPR, EU and US traffic would have had a similar pattern. We test such an assumption by checking for parallel trends before GDPR enactment. Conditional on fixed effects, we should observe a similar trend in the dependent variables between treatment and control pre-GDPR.

We start by estimating the following equation separately for the US and EU:

$$y_{i,k,c,t} = \beta_0 + \tau_t + \alpha_i + \kappa_k + \epsilon_{i,k,c,t}. \tag{4}$$

We then plot the estimated coefficients associated to the time dummies τ_t for both the US and the EU in Figure 4, together with the corresponding 95% confidence interval. In doing so, we exclude the first-period dummy to avoid multicollinearity. Hence, each coefficient represents a (percentage) variation of the dependent variable normalised to the first week.

FIGURE 4. Parallel Trends: Graphical Evidence



(A) Traffic

(B) Avg Visit Duration

(C) Avg Pages Visited

(D) Bounce Rate

This figure shows the OLS coefficients of the time dummies τ_t of Equation (4) for both the US (gray line) and the EU (blue line), together with the corresponding 95% confidence interval. The first-period dummy is excluded to avoid multicollinearity. Robust standard errors are clustered by domain.

Each panel corresponds to one of our variables of interest, namely web traffic (Panel A), average visit duration (Panel B), average pages visited (Panel C), and bounce rate (Panel D). It is reassuring to observe that, before the GDPR, the US and the EU behave similarly for all dependent variables, suggesting that the US provides a valid counterfactual for European online behaviour. What is more, US and EU follow a similar path after the GDPR, reacting very similarly to seasonal shocks.[27] The picture also shows a significant reduction in European traffic with respect to the US, both in the short and in the long term, with a similar pattern emerging also for average visit duration. On the contrary, as for average pages visited and bounce rate, we only observe a long-run separation of the two trends. These preliminary pieces of evidence suggest that (i) the US and Europe behaved similarly before the GDPR, and (ii) the regulation negatively affected web traffic and user engagement in European countries.

We also test formally for parallel linear trends before the GDPR between our treatment and control groups by estimating the following equation:

$$y_{i,k,c,t} = \beta_0 + \gamma_1 Trend_{EU} + \gamma_2 Trend_{US} + \alpha_i + \kappa_k + \epsilon_{i,k,c,t}, \tag{5}$$

where $Trend_{EU}$ and $Trend_{US}$ are European-specific and US-specific linear trends respectively. We estimate Equation (5) only for the period before the GDPR. Ideally, one would like to observe similar estimates for γ_1 and γ_2, implying no differential trends between treatment and control before the shock.

Table 2 shows the estimated coefficients of Equation (5) for each variable of interest. Overall, we observe similar trends between the US and Europe, particularly for the engagement measures (Columns 2-4). In fact, when we test for the equality of the two coefficients of interest, we cannot reject the null hypothesis for average visit duration, average pages visited and bounce rate. On the contrary, we do reject that the two trends are equal for web traffic at the 5% level (Column 1). However, the difference between the two coefficients is fairly small (0.0016), most likely determined by the minor jump few weeks before the GDPR that we observe in Figure 4 Panel (A).

[27] For instance, looking at Figure 4 Panel (A), we observe a drop in web traffic few weeks before and after Google's fine both in the EU and US. The same happens for the engagement metrics, where the two curves react in parallel over time.

Table 2. Parallel Trends Estimates

VARIABLES	(1) Traffic	(2) Avg Visit Duration	(3) Avg Pages Visited	(4) Bounce Rate
Trend$_{EU}$	-0.0004	-0.0016***	-0.0011***	0.0005***
	(0.0006)	(0.0003)	(0.0002)	(0.0001)
Trend$_{US}$	-0.0020***	-0.0010**	-0.0011***	0.0005***
	(0.0007)	(0.0005)	(0.0002)	(0.0001)
H_0: Trend$_{EU}$ =Trend$_{US}$				
F-test	4.40	0.88	0.05	0.09
Observations	906,899	807,279	834,344	806,740
R-squared	0.719	0.422	0.601	0.497

Presented are OLS estimated coefficients of Equation (5). The dependent variable is the $\ln(x+1)$, where x is the number of visits to domain i from channel k and country c, at time (weekly date) t. The Post dummy takes value 1 from GDPR introduction onward, while EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

In summary, our parallel trends tests suggest that US traffic provides a valid control group for European traffic. When we analyse non-parametrically both trends, we find no systematic deviations before the GDPR. When we run a formal test of equal linear trends, we reject the null hypothesis of equal trends only for traffic. However, both the magnitude of the difference in trends and the F-test are small, suggesting the absence of a significant divergence in trends between treatment and control group.

6. Results

We start with the results on website visits, looking first at the overall impact of the GDPR on traffic and then analysing effects by traffic channel and website size category. We then discuss the results for visitor engagement.

6.1. Website Traffic: Average Effect

Table 3 reports *Ordinary Least Square* (OLS) estimated coefficients for different specifications of Equation (1) when the dependent variable is the natural logarithm of website traffic (plus one). The first column collects estimates from our baseline model. The coefficient associated to *Post* is positive and statistically significant, suggesting a general expansion in web traffic over time. At the opposite end, the coefficient associated to *EU* is strongly negative, implying that European countries experience lower traffic compared to the US. The coefficient associated to $Post \times EU$ captures the impact of the GDPR on web traffic. The coefficient is negative and strongly significant, and points to a reduction

Table 3. Diff-in-Diff on Web Traffic

VARIABLES	(1) Traffic	(2) Traffic	(3) Traffic	(4) Traffic	(5) Traffic	(6) Traffic	(7) Traffic
Post × EU	-0.104***	-0.104***	-0.104***	-0.104***	-0.104***	-0.106***	-0.064***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
Post	0.034***						
	(0.011)						
Europe	-2.190***	-2.190***					
	(0.044)	(0.044)					
Domain FE	X	X	X	X			
Channel FE	X	X	X				
Time FE		X	X				
Country FE			X				
Channel-Time FE				X	X	X	X
Country-Channel FE				X	X		
Domain-Channel FE					X		
Domain-Channel-Country FE						X	X
Domain-Time FE							X
Observations	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812
R-squared	0.694	0.694	0.700	0.706	0.928	0.942	0.956

Presented are OLS estimated coefficients of Equation (1). The dependent variable is the $\ln(x+1)$, where x is the number of visits to domain i from channel k and country c, at time (weekly date) t. The Post dummy takes value 1 from GDPR introduction onward, while EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

in web traffic of about 10%.[28] Hence, following the introduction of the GDPR, European website traffic experienced a significant reduction compared to US traffic.

Columns (2)-(6) report estimated coefficients of Equation (1) as we add different sets of fixed effects. It is interesting to observe that $Post \times EU$ coefficient barely moves, which increases our confidence in the baseline specification. Finally, in Column (7) we also control non-parametrically for website-specific time trends by including domain-time fixed effects. This specification identifies the main parameter of interest by exploiting time-variation across country of origin within each domain, which implies that only domains receiving traffic from more than one country contribute to the estimation of $Post \times EU$. That is, the estimated coefficient captures the causal impact of the GDPR only for websites with multinational traffic. The coefficient is still negative and strongly significant, although smaller in magnitude. Since multinational domains are typically larger websites, at this point we might expect potential heterogeneous effects of the GDPR on web traffic based on website size (Section 6.3).

To identify long and short term effects of the GDPR on web traffic, we iterate the same battery of regressions on Equation (2), where we split the $Post$ dummy into $Post^S$ and $Post^L$, which identify short- and long-run periods from the GDPR introduction respectively. In particular, $Post^S$ is active from the GDPR's enactment (25th of May 2018) until the week before the French fine imposed on Google – the so-called "*start of*

[28] This is fully in line with the estimated traffic drop reported by Goldberg et al. (2019).

a war" by data protection authorities against privacy violators (The Economist, 2019) – which happened on the 21st of January 2019. On the other hand, $Post^L$ takes value one from that date onward. Hence, the interactions $Post^S \times EU$ and $Post^L \times EU$ identify the causal impact of the GDPR on web traffic in the short and in the long term respectively.

Table 4 collects estimated coefficients of Equation (2). From Column (1) we observe that both $Post^S$ and $Post^L$ are positive and significant, with the latter being larger in magnitude. The coefficients associated with $Post^S \times EU$ and $Post^L \times EU$ are both negative and significant. Moreover, they are statistically different from each other. The estimated $Post^S \times EU$ coefficient implies a reduction of 4% of web traffic induced by the policy change in the eight months after the introduction of the GDPR, while the $Post^L \times EU$ estimated coefficient suggests a much larger effect in the long-run of -15.7%. Again, estimates barely move when we add different sets of fixed effects (Columns 2-6). When we exploit only multinational website heterogeneity, the coefficients follow a similar path, although both short and long term effects are lower in magnitude. Thus, results from Table 4 suggest that most of the negative effect on web traffic materialises in the long-run, after January 2019. As a consequence, the timing of enforcement might have played a major role in spurring compliance with the law and, therefore, on the variation in website traffic.

What are the dynamics behind the aforementioned results? One possibility is that DiD estimates may hide potential time-varying treatment effects. That is, the impact of the GDPR on web traffic increases over time, implying a divergent path in the post-GDPR period between EU and US traffic. Another possibility is the existence of a transitory period that may coincide with the lack of enforcement and compliance characterising a *grace period* in the months right after GDPR enactment. In such a case, a potential impact on web traffic may take some time to materialise. We investigate this issue by estimating Equation (3), in which we interact the EU identifier with time-specific dummies after controlling for domain, channel, time, and country fixed effects.

Table 4. Diff-in-Diff on Web Traffic in the Short- and Long-Run

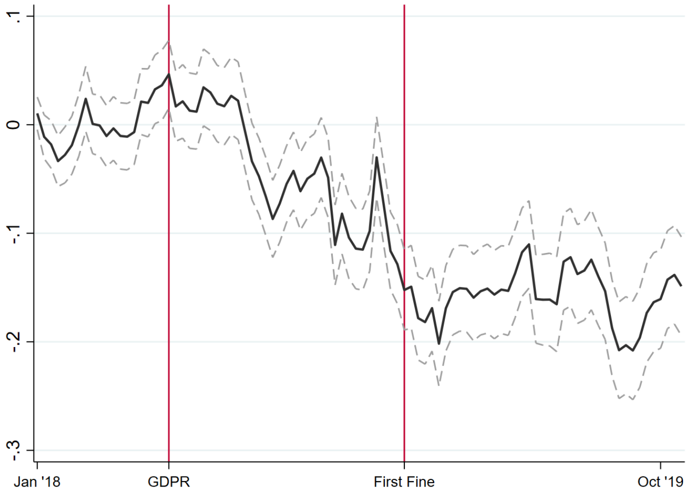
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Traffic	Traffic	Traffic	Traffic	Traffic	Traffic	Traffic
Post^S × EU	-0.040***	-0.040***	-0.040***	-0.040***	-0.040***	-0.042***	-0.027**
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)
Post^L × EU	-0.157***	-0.157***	-0.157***	-0.157***	-0.157***	-0.159***	-0.094***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)
Post^S	0.026***						
	(0.010)						
Post^L	0.042***						
	(0.014)						
EU	-2.190***	-2.190***					
	(0.044)	(0.044)					
Domain FE	X	X	X	X			
Channel FE	X	X	X				
Time FE		X	X				
Country FE			X				
Channel-Time FE				X	X	X	X
Country-Channel FE				X	X		
Domain-Channel FE					X		
Domain-Channel-Country FE						X	X
Domain-Time FE							X
Observations	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812
R-squared	0.694	0.694	0.700	0.706	0.928	0.942	0.956

Presented are OLS estimated coefficients of Equation (2). The dependent variable is the $\ln(x+1)$, where x is the number of visits to domain i from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from GDPR introduction to the week before the fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Figure 5 displays the estimated coefficients (hard black line) and 95% confidence intervals (dashed lines) from the dynamic DiD of Equation (3). Red vertical lines identify the entry into force of the GDPR and the end of the grace period identified by the French financial penalty imposed to Google. First, we notice that pre-GDPR estimated coefficients are fairly constant, and almost all of them are not statistically different from zero. This is in line with our parallel trends test, as it suggests no diversified trends between treatment and control groups before the policy change.

Second, we observe a negative trend in EU web traffic *vis-à-vis* the US starting after the 25th of May 2018 – the date of GDPR enactment – which stops after Google's fine. Third, estimated coefficients stabilise in the long-run, implying an average 15% reduction in web traffic compared to the pre-GDPR period. Hence, the dynamic DiD highlights that the downward shift in the EU web traffic does not materialise right after the introduction of the GDPR. Consistently with the idea of a strict relation between enforcement and compliance, the negative impact of the GDPR on EU web traffic arises eight months later the regulation's entry into force.

Figure 5. Web Traffic: Dynamic Effect



This figure shows the $\hat{\beta}_{1,t}$ coefficients from Equation (3) with the associated 95% confidence interval. The first-period dummy is excluded to avoid multicollinearity. The dependent variable is the $\ln(x+1)$, where x is the number of visits to domain i from channel k and country c, at time (weekly date) t. Robust standard errors are clustered by domain.

In summary, our DiD estimates suggest a negative effect of the GDPR on web traffic. The effect is large and economically significant, implying an average 15% reduction in long-run web traffic. However, this effect does not materialise immediately after the entry into force of the GDPR, but unfolds fully the following year. The dynamic DiD highlights that the transition started during the grace period and stopped a few weeks after the end of the regulatory holiday period, emphasising the role of enforcement on website compliance with GDPR rules.

6.2. Analysis by Traffic Acquisition Channel

How did the GDPR affect website traffic by various traffic channels? If the effect by such channel differs, is the loss of a specific channel counterbalanced by an increase in traffic in another channel? We answer these questions by interacting $Post^S$,$Post^L$, $Post^S \times EU$, and $Post^L \times EU$ from Equation (2) with channel-specific dummies in order to identify diversified effects across traffic channels both in the short and in the long term.

Table 5 collects estimated coefficients $Post^S \times EU$ and $Post^L \times EU$ for each traffic channel. First, we notice that the short-run effects are smaller in magnitude compared to long-run ones for every channel. We observe a short-run negative impact of the GDPR in selected channels, including Direct, Email, Referrals, and Social. The largest coefficient is for Email traffic, suggesting a reduction of about 8% in this channel. A smaller but

Table 5. Heterogeneous Effects on Web Traffic by Traffic Channel

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Traffic	Traffic	Traffic	Traffic	Traffic	Traffic
Post^S × EU × Direct	-0.039***	-0.039***	-0.038***	-0.040***	-0.041***	-0.026**
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.013)
Post^S × EU × Display Advertising	-0.053**	-0.053**	-0.052**	-0.055**	-0.058**	-0.043*
	(0.026)	(0.026)	(0.026)	(0.025)	(0.025)	(0.024)
Post^S × EU × Email	-0.079***	-0.079***	-0.079***	-0.075***	-0.077***	-0.062***
	(0.017)	(0.017)	(0.017)	(0.016)	(0.016)	(0.017)
Post^S × EU × Organic Search	-0.010	-0.010	-0.010	-0.007	-0.009	0.006
	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)	(0.013)
Post^S × EU × Paid Search	-0.029	-0.028	-0.028	-0.038	-0.040	-0.026
	(0.030)	(0.030)	(0.030)	(0.030)	(0.030)	(0.029)
Post^S × EU × Referrals	-0.028*	-0.028*	-0.028*	-0.028*	-0.029*	-0.014
	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.017)
Post^S × EU × Social	-0.042***	-0.042***	-0.042***	-0.040***	-0.041***	-0.026*
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
Post^L × EU × Direct	-0.088***	-0.088***	-0.088***	-0.088***	-0.089***	-0.023
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.019)
Post^L × EU × Display Advertising	-0.381***	-0.381***	-0.380***	-0.381***	-0.386***	-0.320***
	(0.037)	(0.037)	(0.037)	(0.036)	(0.036)	(0.034)
Post^L × EU × Email	-0.291***	-0.291***	-0.291***	-0.287***	-0.288***	-0.223***
	(0.025)	(0.025)	(0.025)	(0.024)	(0.024)	(0.025)
Post^L × EU × Organic Search	-0.073***	-0.073***	-0.073***	-0.071***	-0.072***	-0.007
	(0.016)	(0.016)	(0.016)	(0.015)	(0.015)	(0.019)
Post^L × EU × Paid Search	-0.068*	-0.068*	-0.067*	-0.075*	-0.079**	-0.013
	(0.040)	(0.040)	(0.040)	(0.039)	(0.039)	(0.039)
Post^L × EU × Referrals	-0.079***	-0.079***	-0.079***	-0.080***	-0.081***	-0.016
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.024)
Post^L × EU × Social	-0.123***	-0.123***	-0.123***	-0.120***	-0.121***	-0.055***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.020)
Domain FE	X	X	X			
Channel FE	X					
Channel-Time FE		X	X	X	X	X
Country-Channel FE			X	X		
Domain-Channel FE				X		
Domain-Channel-Country FE					X	X
Domain-Time FE						X
Observations	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812	4,338,812
R-squared	0.696	0.696	0.707	0.928	0.942	0.956

Presented are OLS estimated coefficients of Equation (2) interacted with channel-specific dummies. Channels are categorised in Section 4, and identify the way through which the website is reached. The dependent variable is the $\ln(x+1)$, where x is the number of visits to domain i from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from GDPR introduction to the week before the fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

strongly significant effect concerns Direct traffic, where the coefficient implies a reduction of about 4%. Hence, also unpaid traffic channels have been affected by GDPR restriction, and in the same direction as paid channels.[29]

[29] This is consistent with the view that different traffic acquisition channels are complements, rather than substitutes.

Most of the negative effect of the GDPR on web traffic materialises in the long-run, specifically after the grace period, at the beginning of 2019. This is true for each traffic channel, as shown by the channel-specific $Post^L \times EU$ coefficients in Table 5. The impact on the Email and Display advertising channels is particularly severe, as the coefficients imply a reduction in traffic induced by the GDPR of about 35% and 29% respectively.[30] Although these coefficient estimates are rather large for email and display advertising traffic, they also reflect the relatively small amount of traffic coming from those specific channels. In fact, on average less than 3% of website traffic comes from Email and Display Advertising (Figure 3), implying that even a small absolute reduction in traffic may account for a large variation in percentage points.

We also observe a significant reduction in traffic coming from other paid channels, such as Referrals and Social. Only the coefficient associated to Paid Search is not statistically significant at 5%, neither in the short nor in the long term.[31] What is more, unpaid channels are negatively affected too. Our results suggest a reduction in Direct traffic of about 9%, while traffic from Organic Search experience a 7% reduction. This is somewhat surprising because we would not expect these channels to rely on targeted messages to the same extent as paid advertising does. The effect likely comes through increased user awareness about privacy: the GDPR triggered an increased use of various information banners where websites seek user consent to data handling upon entry. A not insignificant group of users likely turn away from websites upon being presented these pop-ups.

We find that both paid and unpaid website traffic channels are negatively affected by the GDPR in the long-run. Paid channels are more severely affected, although also Organic Search and Direct traffic reduce. Finally, the coefficients in Column (6) of Table 5 suggest that the impact of the GDPR on traffic is significantly smaller for multinational websites. In the long-run only paid channels such as Email, Display Advertising, and Social are negatively affected, with the magnitude being significantly lower compared to the ones in Columns (1)-(5). Hence, potential heterogeneity may arise depending on firm size, paving the way to our dedicated analysis in Section 6.3.

[30] The strong negative effect on email traffic is in line with industry testimonies about the "*death of email marketing*" (Harris, 2018).

[31] This is as expected. Personal data is not used much in paid search. As explained by a recent report of the UK competition authority: "*search ads rely only in a limited way on personalisation, rather they are primarily targeted to match key search terms entered on search engines (ie the 'search query'), which typically provides most of the information needed to serve a relevant ad*", https://assets.publishing.service.gov.uk/media/61b86aee8fa8f5037ffaa347/Appendix_I_-_Considering_the_impacts_of_Apples_ATT.pdf, retrieved on the 1st of February 2022.

6.3. Analysis by Website Size

We now turn to the analysis of the effect of the GDPR on traffic by website size category. Figure 2 shows average visits (in thousands) for each percentile of the distribution of worldwide traffic before the GDPR. It highlights how website traffic is distributed in a highly heterogeneous way, with the presence of a sizeable group (59%) of small websites that however account only for a small share of total traffic (6%), a modest group (30%) of medium ones (13% of traffic), few (10%) large websites (23% of traffic) and a handful (1%) of giant websites that account for the largest share of traffic (57%). Due to the significant heterogeneity in website size, we analyse whether the effects of the GDPR differ for domains of different size. We take into account the high skewness of the traffic distribution when classifying websites by size. We split domains in four categories according to the percentile of worldwide traffic they fall in: small ($1 \leq p \leq 59$), medium ($60 \leq p \leq 89$), large ($90 \leq p \leq 99$), giant ($p = 100$)

Table 6 reports estimated coefficients for two different specifications of Equation 2 by website size. We observe that the short-run effect ($Post^S \times EU$) is negative and statistically significant only for small websites, implying a traffic reduction of about 14%. The long term effect ($Post^L \times EU$) is negative and statistically significant for all websites but medium ones. The coefficient varies strongly by website size, implying a reduction of website traffic of about 41% for small websites, and 7% and 16% for large and very large websites, respectively. These results are robust to different specifications, with coefficients remaining practically unchanged when adding interacted fixed effects.

Overall, our results point to an inverted U-shaped long-run relationship between website size and traffic change due to privacy regulation: the smallest and the largest websites registered a significant drop in visits, while medium ones lost fewer visitors. This is an important result that nuances policy discussions preceding the GDPR, where commentators warned about small firms being particularly severely affected by privacy regulation (Cherry, 2017). The European Commission called the allegation that "*GDPR is overwhelming for small businesses*" a "*myth*" (European Commission, 2019). Our results show that small websites were indeed hit particularly hard by the GDPR. However, so were large websites as well. Medium websites in turn remained largely spared from associated traffic loss.

Table 6. Diff-in-Diff on Web Traffic by Website Size

	Small		Medium		Large		Giants	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Traffic	Traffic	Traffic	Traffic	Traffic	Traffic	Traffic	Traffic
$\text{Post}^S \times$ EU	-0.142***	-0.143***	0.024	0.024	-0.008	-0.009	-0.046	-0.046
	(0.025)	(0.025)	(0.016)	(0.016)	(0.016)	(0.016)	(0.031)	(0.031)
$\text{Post}^L \times$ EU	-0.394***	-0.394***	-0.038	-0.038	-0.073***	-0.073***	-0.155***	-0.155***
	(0.035)	(0.035)	(0.024)	(0.024)	(0.024)	(0.024)	(0.047)	(0.047)
Domain FE	X		X		X		X	
Channel FE	X		X		X		X	
Time FE	X		X		X		X	
Country FE	X		X		X		X	
Channel-Time FE		X		X		X		X
Country-Channel FE		X		X		X		X
Domain-Channel FE		X		X		X		X
Observations	1,956,409	1,956,409	1,241,459	1,241,459	986,016	986,016	154,928	154,928
R-Squared	0.658	0.925	0.704	0.936	0.725	0.917	0.819	0.909

Presented are OLS estimated coefficients of Equation (2). The dependent variable is the $\ln(x+1)$, where x is the number of visits to domain i from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from GDPR introduction to the week before the fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

6.4. User Engagement

Results from Tables 3-4 suggest a reduction in website traffic due to the GDPR's enforcement. That is, websites receive less visits from users. However, the restrictions imposed by the GDPR may affect also user behaviour once the visitor has effectively reached the website. Hence, the policy change can affect both the extensive (website visits) and intensive margin (website engagement) of online behaviour. For instance, a less effective display ad may drive the user to a dull website for which she has no particular interest. In such a situation, we can expect the user to quickly shift to another activity, either offline or online.

We investigate whether the GDPR has affected also the intensive margin of online behaviour by estimating Equation 2 on three metrics of user engagement, namely average visit duration, average pages visited, and bounce rate.[32] Table 7 collects OLS coefficients from such regressions. Focusing on average visit duration (first column) we observe a significant reduction both in the short- and in the long-run of 1.7% and 3.5% respectively. A similar pattern emerges for the average number of pages visited within the website, with estimates pointing to a reduction of 0.7% and 1.5% in the short and long term respectively. As for the bounce rate, results point toward a reduction in engagement only in the long-run of 0.006 percentage points. Thus, the GDPR has a negative impact on both intensive and extensive margins. Consistently with previous results on web traffic,

[32] For a description of the three variables see Section 4.

the adverse effect on user engagement induced by the GDPR materialises mostly in the long-run.

Table 7. Diff-in-Diff on User Engagement

VARIABLES	(1) Avg Visit Duration	(2) Avg Pages Visited	(3) Bounce Rate
$\text{Post}^S \times \text{EU}$	-0.017***	-0.007**	0.000
	(0.006)	(0.003)	(0.001)
$\text{Post}^L \times \text{EU}$	-0.035***	-0.015***	0.006***
	(0.007)	(0.005)	(0.002)
Channel-Time FE	X	X	X
Country-Channel FE	X	X	X
Domain-Channel FE	X	X	X
Observations	3,844,720	3,984,614	3,839,355
R-squared	0.550	0.721	0.697

Presented are OLS estimated coefficients of Equation (2). In the first two columns, the dependent variable is the $\ln(x + 1)$, where x is the average visit duration measured in seconds (Column 1) and the number of pages visited (Column 2), while in Column (3) the dependent variable is the bounce rate on domain i, for traffic coming from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from GDPR introduction to the week before the fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Are the results driven by specific channels? Table 8 suggests that that is the case. When we disentangle heterogeneous effects by traffic channel, we find that the negative impact of the GDPR on user engagement comes from those users reaching the website mainly through email, although also engagement from social networks and organic search traffic is significantly impacted both in the short and long term. In the long-run, average visit duration decreases mainly for paid traffic channels, including Email, Referrals, Paid Search, and Social, thus indicating online advertising being less effective also in terms of engagement.

All in all, this battery of results highlights another adverse effect in addition to the drop in web traffic. User engagement is negatively affected, and particularly following the effective enforcement of the GDPR after the issuance of the first large fine. The policy change affected disproportionally paid channels, although we find also evidence of a negative effect from organic search traffic engagement, pointing toward potential complementarities across channels also in the intensive margin.

Table 8. Heterogeneous Effects on User Engagement by Traffic Channel

VARIABLES	(1) Avg Visit Duration	(2) Avg Pages Visited	(3) Bounce Rate
Post^S × EU × Direct	-0.003	-0.002	-0.002*
	(0.005)	(0.004)	(0.001)
Post^S × EU × Display Advertising	0.009	-0.001	-0.002
	(0.022)	(0.009)	(0.004)
Post^S × EU × Email	-0.051***	-0.026***	0.009***
	(0.012)	(0.006)	(0.002)
Post^S × EU × Organic Search	-0.010*	-0.010**	0.003
	(0.006)	(0.004)	(0.002)
Post^S × EU × Paid Search	-0.042*	0.003	-0.004
	(0.022)	(0.010)	(0.004)
Post^S × EU × Referrals	-0.001	0.005	-0.006***
	(0.008)	(0.005)	(0.002)
Post^S × EU × Social	-0.025***	-0.013**	0.002
	(0.010)	(0.005)	(0.002)
Post^L × EU × Direct	-0.002	0.002	0.001
	(0.007)	(0.006)	(0.002)
Post^L × EU × Display Advertising	0.010	0.000	0.003
	(0.024)	(0.010)	(0.004)
Post^L × EU × Email	-0.090***	-0.057***	0.022***
	(0.015)	(0.008)	(0.003)
Post^L × EU × Organic Search	-0.013*	-0.008	0.005**
	(0.007)	(0.005)	(0.002)
Post^L × EU × Paid Search	-0.079***	-0.019*	-0.001
	(0.023)	(0.011)	(0.005)
Post^L × EU × Referrals	-0.028***	-0.004	-0.000
	(0.010)	(0.007)	(0.003)
Post^L × EU × Social	-0.052***	-0.020***	0.007***
	(0.012)	(0.007)	(0.002)
Channel-Time FE	X	X	X
Country-Channel FE	X	X	X
Domain-Channel FE	X	X	X
Observations	3,844,720	3,984,614	3,839,355
R-squared	0.550	0.721	0.697

Presented are OLS estimated coefficients of Equation (2) interacted with channel-specific dummies. Channels are categorised in Section 4, and they identify the way through which the website is reached. In the first two columns, the dependent variable is the $\ln(x+1)$, where x is the average visit duration measured in seconds (Column 1) and the number of pages visited (Column 2), while in Column 3 the dependent variable is the bounce rate on domain i, for traffic coming from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from GDPR introduction to the week before the fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

7. Robustness Checks

We run two series of robustness checks that consider the presence of potential spillovers. The first one deals with the presence of multinational traffic, which might generate spillovers between treatment and control if multinational domains apply a uniform privacy policy to both US and EU visitors. The second robustness check assesses the presence of differential effects for websites located in the EU compared to those in the US, which may also suggest potential spillovers. This might happen if EU websites comply with the GDPR also when serving US visitors, or if US websites do not comply with the GDPR at all. Both robustness checks confirm our main results, namely a reduction in web traffic and user engagement following the entry into force of the GDPR, and in particular in the long-run.

7.1. Excluding Multinational Traffic

In Equations (1) and (2) the treatment assignment covers traffic originating from the EU. Since a website can receive visitors from more than one country, this implies that multinational websites are treated only for the portion of traffic coming from Europe. This is in line with the main provision of the GDPR, which aims to protect all EU residents regardless of the website's actual location. However, an implicit assumption for the identification of the causal impact of the GDPR on web traffic is that multinational domains comply with the GDPR only for EU visitors. That is, websites are assumed to apply a different privacy policy for EU visitors and to visits coming from the US. Although we do informally observe websites treating visitors from the US and the EU differently, this may not always be the case.[33] It may be possible that websites with a high share of EU traffic comply *tout court*, while domains with most of the traffic coming from the US may prefer not to comply with the limitations imposed by the GDPR. In the extreme case, websites may decide to block access to EU visitors altogether.[34]

We test for this issue by focusing only on websites that receive most of the traffic from one of the countries of our analysis. In particular, we focus on domains with more than 85% of their global traffic pre-GDPR coming from either one of the EU countries in our analysis (i.e., France, Germany, Italy, Netherlands, and the UK) or the US.[35] Moreover, we select only the traffic originated from the country that meets the 85% threshold. In this way, we clean the dataset from multinational domains (such as facebook.com), as well as from country traffic that cannot be clearly allocated between treatment and control group.

[33] We casually surfed several multinational websites using a VPN service that allows us to manually set our IP address as US or European. A large number of websites provide a different experience tailored to the perceived country of origin.

[34] There is some evidence that US news websites blocked EU users right after the enactment of GDPR, because they were not able to comply with the new data protection rules in the short term. See https://digiday.com/media/u-s-sites-continue-block-european-visitors-post-gdpr/.

[35] Despite the focus on specific EU countries, Similarweb provides metric for global traffic for each domain. Given a domain's global traffic, we can derive the share of traffic coming from each of the observed countries at each point in time and for each domain.

TABLE 9. Robustness Check - No Multinational Traffic

VARIABLES	(1) Traffic	(2) Avg Visit Duration	(3) Avg Pages Visited	(4) Bounce Rate
Post^S × EU	-0.037**	-0.032***	-0.015***	0.004**
	(0.017)	(0.009)	(0.005)	(0.002)
Post^L × EU	-0.179***	-0.053***	-0.022***	0.010***
	(0.024)	(0.011)	(0.007)	(0.002)
Channel-Time FE	X	X	X	X
Country-Channel FE	X	X	X	X
Domain-Channel FE	X	X	X	X
Observations	1,864,121	1,658,465	1,710,556	1,649,193
R-squared	0.938	0.574	0.739	0.731

Presented are OLS estimated coefficients of Equation (2). The dependent variable is the $\ln(x + 1)$, where x is the number of visits (Column 1), the average visit duration measured in seconds (Column 2), and the number of pages visited (Column 3), while in Column (4), the dependent variable is the bounce rate on domain i, for traffic coming from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from the GDPR introduction to the week before the fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table 9 collects the estimated coefficients of Equation (2) in this setting. By looking at website traffic (Column 1), we find that both the short and long term (negative) effects of the GDPR remain intact. The estimated coefficients are very much in line with our main results (Table 4 Column 5), suggesting an increasing long-run effect, with the magnitudes also being similar. The short-run effect remains largely unchanged (0.3% difference), while the long-run effect increases slightly by 2%. This suggests that increasing the precision of the treatment assignment rises the estimated impact, as one would expect. However, it also highlights that the potential bias is fairly small, which in turn increases our confidence in the research design.

A similar argument applies for the engagement measures. A comparison with Table 7 shows that, if anything, cleaning from multinational traffic increases the estimated coefficients. In this case, the difference between the two estimates is significantly larger. For instance, the long-run impact on average visit duration goes from -3.5% to -5.3%, which implies a 50% increase in magnitude. Moreover, while results from Table 7 Column (3) do not suggest any significant short term effect on the bounce rate, we now observe a short term increase of about 0.4 percentage points.

In conclusion, the results from this robustness check confirm our main findings, namely a negative impact of the GDPR on web traffic and user engagement. Although the bias from multinational traffic does not significantly affect estimates on web traffic, it somewhat underestimates the impact on user engagement metrics, implying that results from Table 7 are likely even conservative.

7.2. US vs. EU Websites

Another potential source of concern might arise if websites apply a uniform privacy policy based on their home country rather than the users' location. This may happen if EU-based websites comply with the GDPR *tout court*, or if US-based domains fail to comply with the GDPR for the EU visitors, or both. We deal with this issue by differentiating US- and EU-based websites through their domains. In particular, we categorise the ".com" URLs as US-based domains, and ".fr" and ".de" as French and German, respectively, and proceed for the other EU-countries in our dataset analogously. We then estimate Equation (2) separately for US- and EU-based domains. In this way, we check for the presence of spillovers between treatment and control traffic within US- or EU-based domains.

Table 10 collects estimated coefficients from such an experiment. In Columns (1)-(4) only EU-based domains are used. The estimates are qualitatively the same as in our main analysis. The long-run impact on web traffic is however larger, suggesting (not surprisingly) that EU-based websites might have complied more rigorously with the GDPR. The same holds for user engagement measures, as the long-run effects are generally higher than the ones in Table 7, with the difference being sizeable only for average visit duration. Even when we focus on US-based domains (Columns 5-8) the results are qualitatively stable, implying long-run increasing effects of the GDPR. Estimates are slightly lower in magnitude compared with our main results, but the difference is not particularly large. Hence, also EU traffic for US-based websites have been affected by the GDPR.

Overall, our robustness checks confirm the results in the main analysis, pointing to a reduction in web traffic and user engagement induced by the GDPR, particularly in the long-run. They also validate our empirical design, since we do not detect large biases induced by spillovers between treatment and control.

Table 10. Robustness Check - Assessing Potential Spillovers

	EU Domains Only				US Domains Only			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Traffic	Avg Visit Duration	Avg Pages Visited	Bounce Rate	Traffic	Avg Visit Duration	Avg Pages Visited	Bounce Rate
Post^S × EU	-0.040*	-0.015	-0.012**	0.002	-0.036***	-0.013	-0.003	-0.002
	(0.024)	(0.010)	(0.006)	(0.002)	(0.014)	(0.009)	(0.005)	(0.002)
Post^L × EU	-0.246***	-0.053***	-0.017**	0.006**	-0.128***	-0.025**	-0.015**	0.006**
	(0.032)	(0.013)	(0.007)	(0.003)	(0.021)	(0.011)	(0.007)	(0.002)
Channel-Time FE	X	X	X	X	X	X	X	X
Country-Channel FE	X	X	X	X	X	X	X	X
Domain-Channel FE	X	X	X	X	X	X	X	X
Observations	2,026,044	1,760,007	1,830,579	1,758,138	2,312,768	2,084,713	2,154,035	2,081,217
R-squared	0.925	0.528	0.703	0.693	0.929	0.569	0.736	0.701

Presented are OLS estimated coefficients of Equation (2). The dependent variable is the $\ln(x+1)$, where x is the number of visits (Columns 1 and 5), the average visit duration measured in seconds (Columns 2 and 6), and the number of pages visited (Columns 3 and 7), while in Columns (4) and (8), the dependent variable is the bounce rate for domain i, for traffic coming from channel k and country c, at time (weekly date) t. Post^S is a dummy taking value 1 from the GDPR introduction to the week before the first large fine imposed by the French privacy authority to Google (21st of January 2109), while Post^L activates from that date onward. EU is an indicator that identifies traffic generated from European countries. Robust standard errors clustered by domain in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

8. Conclusions

In this paper we investigated the effect of the European Union's General Data Protection Regulation on websites' ability to attract visitors and on users' interaction with websites. We document an overall average traffic reduction of 15% in the long-run following the GDPR's implementation, and a measurable reduction of user engagement with websites. Traffic from direct visits, organic search, email marketing and social media links, display advertising and referrals dropped significantly after the GDPR. Email marketing and display advertising experienced a large, near 30% reduction of traffic in the long-run.

Our results carry relevance for broader economic policy. They appear consistent with the view that users care about privacy and actively opt out of visits as a result of better information about website data handling policies. Paid search traffic – mainly Google search advertisement – was barely affected. Privacy regulations can therefore impact market structure and may increase dependence on large advertising service providers. Our results also indicate that enforcement matters. The effects were amplified considerably following the first large fine issued eight months after full entry into force of the regulation.

We furthermore document an inverted U-shaped relationship between website size and the change in traffic due to privacy regulation. While the smallest and largest websites lost visits, medium websites remained largely unaffected and may have even gained from the GDPR. This confirms pre-GDPR worries about a potential adverse effect of privacy regulation on small firms, but nuance it by finding a negative impact also on large firms.

While there is a positive correlation between traffic and revenues in the internet economy, our results are not directly translatable into welfare effects of the regulation. First, digital platforms may better monetise from the remaining users, and this in turn may offset the revenue loss from lower visitors. Second, we do not measure consumer privacy benefits, nor can we reliably convert lost visits into costs. Overall however, it appears to us that additional consumer benefits may easily outweigh the implied losses of website traffic. Web traffic is probably the domain where consumer reaction to privacy regulation is the largest as some website visits may be deferred. And this loss of visits may not be equivalent to economic harm. For example, the decline of direct website visits may be the result of users' choice to refrain from interacting with privacy-intrusive websites. Since the GDPR enabled such informed choice by mandating that websites obtain informed consent to data policies, even lost website visits may imply net welfare gains.

Overall, our analysis demonstrates that privacy regulation had a measurable impact on the online economy. We take a step towards better understanding how these effects are distributed across different players on various levels of the value chain. However, better understanding the impact mechanism behind these results – i.e., what is driving the reduction in traffic and user engagement – remains an important task for further research.

Acknowledgement: This work has been partially supported by Ministero dell'Istruzione, dell'Università e della Ricerca Award TESUN-83486178370409 finanziamento dipartimenti di eccellenza CAP. 1694 TIT. 232 ART. 6.

Bibliography

Acquisti, A., Taylor, C., and Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, *54*(2), 442–492. https://doi.org/10.1257/jel.54.2.442

Aridor, G., Che, Y.-K., and Salz, T. (2020). *The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR* (Working Paper No. 26900). National Bureau of Economic Research. https://doi.org/10.3386/w26900

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., and Turner, E. (2019). Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. Retrieved January 28, 2022, from https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

BBC. (2018). GDPR: Data protection overhaul hits small businesses. *BBC News*. Retrieved January 28, 2022, from https://www.bbc.com/news/uk-wales-44199910

Calzada, J., and Gil, R. (2020). What Do News Aggregators Do? Evidence from Google News in Spain and Germany. *Marketing Science*, *39*(1), 134–167. https://doi.org/10.1287/mksc.2019.1150

Campbell, J., Goldfarb, A., and Tucker, C. (2015). Privacy Regulation and Market Structure. *Journal of Economics & Management Strategy*, *24*(1), 47–73. https://doi.org/10.1111/jems.12079

Cherry, M. (2017). *Small businesses at risk of being overwhelmed by data protection burden.* https://www.euractiv.com/section/data-protection/opinion/small-businesses-at-risk-of-being-overwhelmed-by-data-protection-burden/

CNIL. (2019). The CNIL's restricted committee imposes a financial penalty of 50 Million euros against GOOGLE LLC — CNIL. Retrieved November 4, 2021, from https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc

Coos, A. (2020). Australian Government Kicks off Privacy Act Review. Retrieved January 28, 2022, from https://www.endpointprotector.com/blog/australian-government-kicks-off-privacy-act-review

Dimakopoulos, P. D., and Sudaric, S. (2018). Privacy and platform competition. *International Journal of Industrial Organization*, *61*, 686–713. https://doi.org/10.1016/j.ijindorg.2018.01.003

European Commission. (2011). EU: Attitudes on Data Protection and Electronic Identity in the European Union. Retrieved January 28, 2022, from https://joinup.ec.europa.eu/collection/eidentity-and-esignature/document/eu-attitudes-data-protection-and-electronic-identity-european-union

European Commission. (2015). *Special eurobarometer 431: Data protection* (tech. rep.). Directorate-General for Communication. Brussels, Belgium. Retrieved January 28, 2022, from https://data.europa.eu/data/datasets/s2075_83_1_431_eng?locale=en

European Commission. (2019). Mythbusting: General data protection regulation - fact sheet. Retrieved January 28, 2022, from https://ec.europa.eu/info/sites/default/files/100124_gdpr_factsheet_mythbusting.pdf

Gal, M. S., and Aviv, O. (2020). The Competitive Effects of the GDPR. *Journal of Competition Law & Economics*, *00*(00), 1–43. https://doi.org/10.1093/joclec/nhaa012

Godinho de Matos, M., and Adjerid, I. (2021). Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider. *Management Science*. https://doi.org/10.1287/mnsc.2021.4054

Goldberg, S., Johnson, G., and Shriver, S. (2019). *Regulating Privacy Online: An Economic Evaluation of the GDPR* (SSRN Scholarly Paper No. ID 3421731). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3421731

Goldfarb, A., and Tucker, C. E. (2011). Privacy Regulation and Online Advertising. *Management Science*, *57*(1), 57–71. https://doi.org/10.1287/mnsc.1100.1246

Government of Canada. (2020). Fact Sheet: Digital Charter Implementation Act, 2020. Retrieved January 28, 2022, from https://www.ic.gc.ca/eic/site/062.nsf/eng/00119.html

Harris, W. (2018). Why marketers need to stop saying "Email marketing is dead". Retrieved January 28, 2022, from https://www.campaignmonitor.com/blog/email-marketing/why-marketers-need-to-stop-saying-email-marketing-is-dead/

Hu, X., de Tangil, G. S., and Sastry, N. (2020). Multi-country Study of Third Party Trackers from Real Browser Histories. *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 70–86. https://doi.org/10.1109/EuroSP48549.2020.00013

IAPP. (2021). The california privacy rights act of 2020. Retrieved December 12, 2021, from https://iapp.org/resources/article/the-california-privacy-rights-act-of-2020/

Interactive Advertising Bureau. (2021). Study Finds Internet Economy Grew Seven Times Faster Than Total U.S. Economy, Created Over 7 Million Jobs in the Last Four Years. Retrieved January 28, 2022, from https://www.iab.com/news/study-finds-internet-economy-grew-seven-times-faster/

Jia, J., Jin, G. Z., and Wagman, L. (2021). The Short-Run Effects of the General Data Protection Regulation on Technology Venture Investment. *Marketing Science*, *40*(4), 661–684. https://doi.org/10.1287/mksc.2020.1271

Johnson, G. A., Shriver, S. K., and Du, S. (2020). Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry? *Marketing Science*, *39*(1), 33–51. https://doi.org/10.1287/mksc.2019.1198

Kottasová, I. (2018). These companies are getting killed by GDPR. Retrieved January 28, 2022, from https://money.cnn.com/2018/05/11/technology/gdpr-tech-companies-losers/index.html

Libert, T., Graves, L., and Nielsen, R. K. (2018). Changes in Third-Party Content on European News Websites after GDPR. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-08/Changes%20in%20Third-Party%20Content%20on%20European%20News%20Websites%20after%20GDPR_0.pdf

Lu, S., Wang, X. (, and Bendle, N. (2020). Does Piracy Create Online Word of Mouth? An Empirical Analysis in the Movie Industry. *Management Science*, *66*(5), 2140–2162. https://doi.org/10.1287/mnsc.2019.3298

Matte, C., Bielova, N., and Santos, C. (2020). Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. *arXiv:1911.09964 [cs]*. Retrieved January 28, 2022, from http://arxiv.org/abs/1911.09964

Nouwens, M., Liccardi, I., Veale, M., Karger, D., and Kagal, L. (2020). Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. *arXiv:2001.02479 [cs]*. https://doi.org/10.1145/3313831.3376321

Ourworldindata.org. (2021). Share of the population using the Internet. https://ourworldindata.org/grapher/share-of-individuals-using-the-internet?tab=chart

Peukert, C., Bechtold, S., Batikas, M., and Kretschmer, T. (2020). *European Privacy Law and Global Markets for Data* (Working Paper No. 01/2020). ETH Zürich. Zürich, Switzerland. https://doi.org/10.3929/ethz-b-000406601

Sabatino, L., and Sapi, G. (2019). *Online privacy and market structure: Theory and evidence* (Working Paper No. 308). DICE Discussion Paper. Retrieved January 28, 2022, from https://www.econstor.eu/handle/10419/192877

Schmitt, J., Miller, K. M., and Skiera, B. (2021). The Impact of Privacy Laws on Online User Behavior. *arXiv:2101.11366 [econ, q-fin]*. Retrieved March 15, 2022, from http://arxiv.org/abs/2101.11366

Sørensen, J., and Kosta, S. (2019). Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. *The World Wide Web Conference*, 1590–1600. https://doi.org/10.1145/3308558.3313524

The Economist. (2014). Getting to know you. *The Economist*. Retrieved October 11, 2021, from https://www.economist.com/special-report/2014/09/11/getting-to-know-you

The Economist. (2019). The French fine against Google is the start of a war. *The Economist*. Retrieved November 4, 2021, from https://www.economist.com/business/2019/01/24/the-french-fine-against-google-is-the-start-of-a-war

Utz, C., Degeling, M., Fahl, S., Schaub, F., and Holz, T. (2019). (Un)informed Consent: Studying GDPR Consent Notices in the Field. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 973–990. https://doi.org/10.1145/3319535.3354212

Zhuo, R., Huffaker, B., claffy, k., and Greenstein, S. (2021). The impact of the General Data Protection Regulation on internet interconnection. *Telecommunications Policy*, *45*(2), 102083. https://doi.org/10.1016/j.telpol.2020.102083

CHAPTER 3

The Uneven Effect of Airbnb on the Housing Market. Evidence Across and Within Italian Cities[*]

Raffaele Congiu, Flavio Pino, Laura Rondi[†]

We investigate whether the diffusion of Airbnb may exacerbate the differences between central and suburban areas by affecting residential property values and rents. We leverage on the heterogeneity across five Italian cities – Florence, Milan, Naples, Rome, and Turin – that differ for touristic intensity, housing costs, and disparity between centre and periphery. To relate our analysis to existing literature, we first quantify the overall and the city-specific effects of Airbnb intensity. Then we estimate whether the impact is constant across city centre and suburbs and, finally, what is the effect of Airbnb's diffusion in the centre on the periphery. We find that Airbnb growth increases rents and, more significantly, sale prices. Overall, a 1 pp increase in listing density leads to an increase of 44.24 €/m^2 in house prices between 2014 and 2019. However, the impact differs both across cities and within cities (more in the centre). Finally, Airbnb diffusion in the centre is related to a price decrease in the suburbs, in line with the idea that home-sharing Airbnb might contribute to reinforce inequality by raising the attractiveness of the city centre. Overall, the high heterogeneity of the effects suggests that Airbnb's diffusion benefits some parts of the city while leaving others behind.

[*] The authors gratefully acknowledge the real estate portal *Idealista* and the Italian register of the real estate market *OMI* (*Agenzia delle Entrate - Territorio - Osservatorio del mercato immobiliare*) for kindly providing data on the Italian rental market. For useful comments and suggestions, the authors thank Laura Abrardi, Elena Cefis, Gilles Duranton, Lorien Sabatino, Raffaele Scuderi, Huaxin Wang-Lu, and all participants to the XIX SIEPI Workshop, the 62nd SIE Conference and the 9th iCare Conference.

[†] Politecnico di Torino, Department of Management, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy.

1. Introduction

Airbnb is probably the most well-known face of the sharing economy. The platform's claim is to provide hosts with an additional source of income from what was previously unused capacity, and to present guests with an affordable and personal accommodation experience. While this appears economically efficient, the diffusion of the platform has been criticised by residents and local administrators, who complain that Airbnb benefits landlords and tourists at the expense of local renters by raising rental rates. Not surprisingly, a growing number of academic studies has inquired into the positive and negative externalities of short-term rental platforms and provided empirical evidence of their impact on the real estate market (Horn and Merante, 2017; Garcia-López et al., 2020; Barron et al., 2020; Duso et al., 2020; Ayouba et al., 2020; Koster et al., 2021; Franco and Santos, 2021). Recently, the literature has started to question the distributional impact of the heterogeneous presence of Airbnb in cities highly exposed to visitor flows, postulating a spillover effect of Airbnb entry in one part of the city on prices in other parts of the city (Calder-Wang, 2021). This effect may result from an endogenous increase of the attractiveness of one part at the expense of the other (Almagro and Domínguez-Iino, 2021; Xu and Xu, 2021).

Our paper embraces this view and, after estimating the impact of Airbnb density on house prices and rents overall and at the city level for five Italian cities, it digs into within-city effects and, ultimately, it estimates the impact of Airbnb presence in the city centre on the property values and rental rates in the periphery. Due to its tourist vocation, Italy has been particularly affected by Airbnb's diffusion, though its pressure is not uniformly distributed across towns. We are motivated to this research question by the inherent variety characterising the five cities we study – Rome, Florence, Milan, Turin, and Naples – and by the spatial dimension of the inequality between their central and peripheral areas, which itself differs city by city.

The five cities vary from each other in terms of size, tourist inflow and seasonality, business vocation, topographic constraints and economic performance. Florence is the epitome of the arts town, with a relatively small and highly protected centre. Rome is both a touristic and a political attraction, being the capital of Italy, while Milan is the economic and financial centre of Italy, but over the years its fashion industry has more and more allured millions of visitors (even surpassing Florence numerically). Both have large peripheries where the living conditions may be harsh. Naples is especially alluring in the summer as it is located on the seaside, while it also attracts tourists (albeit less than previous cities) due to the artistic value of its (very congested) city centre. Finally, Turin is an industrial city, headquarter of a motor vehicle company (FIAT) that has deeply affected the suburban areas as well as the demographics, but is also close to ski resorts in the Alps and home to an admired and consistent baroque city centre. Although average income is higher in Milan and in Rome and lower in Naples and Turin, the highest income inequality is in Milan (second highest, Rome) and the lowest in Florence.

All in all, the five cities aptly represent the heterogeneity of the housing market and of the tourist industry in the Italian cities interested by the Airbnb phenomenon. To grasp what this heterogeneity implies, it is worth noting that although the number of listings has more than tripled in all cities, Rome has three times the number of visitors of Florence, which in turn triples the visitors of Turin and Naples. However, from 2014 to 2019, the number of tourists has grown faster in Naples and in Rome than in Florence (ISTAT, 2016, 2020). At the same time, the average listing density of Airbnb has increased more in Milan – where property values have skyrocketed – and in Naples – where they declined – than in Florence and Rome.

Digging into this disparate picture, our paper provides evidence of Airbnb's impact on house prices and rents in this representative sample of towns. We set the scene by estimating the effect of Airbnb on the cities altogether. We then explore the heterogeneity of the housing market by estimating how the impact differs across towns. Next, we turn to within-city effects. First, we estimate whether Airbnb only affects the city centre or if its impact also extends to the periphery. Second, we investigate whether the increase in Airbnb's penetration in the city centre might also influence the periphery through spillovers and cross effects.

To address these questions, we collected data from a variety of sources. Daily data on individual Airbnb listings were obtained from AirDNA, a large provider of short-term rental analytics. Rent and sale prices were provided by Idealista, a major online real estate portal. Our unit of observation is the zone (i.e., the neighbourhood into which Idealista divides the housing market of each city) and the time dimension is the year-quarter, from 2014 to 2019.

Estimation of these impacts raises potential endogeneity concerns that derive from omitted variable bias as well as from simultaneity problems due to shocks to rents and sale prices that could affect the decision to list an apartment on the platform but are unrelated to Airbnb. We address these concerns by including a large set of control variables, zone-specific fixed effects and spatial-time interacted effects to account for city-specific seasonality and within-city housing market trends. Moreover, we implement an instrumental variable approach (Barron et al., 2020) that exploits the interaction of an out-of-sample measure of tourist attraction that varies within cities (derived from Tripadvisor) with a measure of public awareness of Airbnb that varies over time (derived from Google searches). We control for identification threats that may come from the high correlation between tourist attraction and centrality by including centre- and suburbs-specific year-level fixed effects and zone-level time-varying controls associated with urban revival processes.

We find that Airbnb's diffusion has determined an increase of both rents and house prices, where the latter appears more affected. The impact differs by city and, within each city, between the centre and the suburban areas. Overall, our analysis shows that, on average, an increase of 1 percentage point in Airbnb density leads to an increase in sale prices of 44.38 €/m^2 over the period of the analysis, but the impact differs across cities, from an average increase of 162.31 €/m^2 in Milan to 72.04 €/m^2 in Florence and only 19.37 €/m^2 in Rome. The effect on rent is more significant in Florence and Naples, with an estimated increase of 19 and 37 €cent/m^2 respectively, which, compared to rents' variation over time, implies that Airbnb is responsible for 10% and 65% of the rent increases in the two cities. When we focus on within-city effects, the picture confirms to be extremely heterogeneous. We find that the effect of Airbnb density on house value is positive and significant both in the city centre and in the suburbs in Florence, Milan and Rome, but only in the city centre of Naples and Turin. However, quantitatively the effect differs greatly, as the increases in Milan and Rome's city centres are much higher than those in the suburbs, whereas in Florence the price increase is higher in the periphery. Our findings from the analysis on spillover effects, running from Airbnb's growing presence in the centre to house prices and rents in the suburbs, shows that the increase in density in central areas has a negative effect on the property values in the suburbs. The evidence is stronger for Milan and Rome – which have the largest gap between neighbourhoods with the highest and lowest average income – and weaker for Turin, but the direction of the effect is the same. Overall, this suggests that the presence of Airbnb in the centre and the related boost to the supply of localised amenities may reinforce the core's attractiveness at the expense of the peripheries. On the opposite side, we find that in Florence the presence of Airbnb has a stronger positive effect on the suburbs than on the centre's prices, seemingly leading prices to converge. As Florence is the smallest of the analysed cities and its centre has the highest listing density, this could result from the fact that the suburbs are becoming more desirable, as they are relatively close to the city centre.

Considering that in the starting year, the wedge between the prices in the centre and in the suburbs was smaller in Florence and larger in Milan and Rome, these findings suggest that the Airbnb presence has a reinforcing effect on the housing market, by widening large gaps and reducing small ones. Therefore, our analysis hints that the impact of Airbnb on within-city inequality depends on the initial conditions of the different areas. The urban policies and the calls for regulation of home sharing activity by individual cities should take into account these initial conditions and to what extent central and suburban areas are adequately equipped to deal with short-term rental platforms and, possibly, to benefit from them.

We contribute to the literature in several ways. First, although Italy has a strong touristic vocation, this is the first study estimating the impact of Airbnb on the Italian housing market, to the best of our knowledge. Second, by focusing on five cities with heterogeneous housing markets and socio-economic characteristics, we provide evidence of an overarching effect but also of the need for bespoke policies. Third, by disentangling Airbnb's effect within the cities, we show the unevenness of Airbnb's effect between city centres and suburbs and the potential spillovers between them, in line with the findings of the recent literature that studies the distributional impact of home-sharing platforms and the mechanisms behind the reinforcement of residential sorting. Overall, our findings can help to better understand whether Airbnb's diffusion is benefiting some parts of the city while leaving other neighbourhoods behind.

The paper is organised as follows. In Section 2 we review the relevant literature from which we derive our conceptual framework. In Section 3 we describe the geographical scope of our study and briefly characterise the heterogeneity of the five cities. In Section 4 we describe the data and, in Section 5, the empirical strategy. In Section 6 we present the results and in Section 7 we conclude.

2. Related Literature and Conceptual Framework

2.1. Empirical Evidence

Our work contributes to a growing literature that studies the economic impact of short-term rentals on the housing market. Sheppard and Udell (2016) study Airbnb's impact on the value of New York City's residential property. They argue that Airbnb's diffusion can increase property value – e.g., by offering new income streams to house owners, thus reducing the cost of ownership; increasing the demand for space due to a growth in local tourist population; raising the quality of a neighbourhood due to the local economic impact of tourists – or decreasing it, when the presence of tourists generates negative externalities to residents. They find that doubling the number of Airbnb listings in a 300m radius around a property leads to a 6% to 11% increase in its value. Horn and Merante (2017) inquire into the short-term effects on rents of Airbnb's penetration in Boston. They find that an increase in density of Airbnb listings equal to a standard deviation rises rents by 0.4% and reduces the number of units offered for rent by 5.9%. The increase in rental rates reaches 3.1% for neighbourhoods in the top decile of Airbnb density. Barron et al. (2020) study the impact of Airbnb on house prices and rents in the USA. They find that a 1% increase in Airbnb listings leads to a 0.018% increase in rents and 0.026% in sale prices. They underline how the impact decreases with the share of owner-occupiers, suggesting that the price increase is driven by a reallocation of supply from the long-term to the short-term rental market. Koster et al. (2021) investigates the impact of the platform on Los Angeles County. They exploit the exogenous shock offered by the introduction of a legislation restricting short-term rentals which affected only some cities of the County. They find a significant reduction in rents and house prices of about 2% following the introduction of the legislation.

A branch of this literature assesses Airbnb's impact on European housing markets. Ayouba et al. (2020) investigate whether Airbnb listings affect rental prices in eight French cities. They find that an increase in the number of Airbnb listings is linked to a raise in rental prices for some cities. A one percent increase in Airbnb density in a given neighbourhood leads to a 0.5 percent increase in rents in Pairs, which is the town affected the most. However, when focusing on commercial listings, they find that the impact more than doubles to 1.2 percent. In some cities, Airbnb impact surprisingly increases with the share of home-occupiers, and it decreases with hotel density. Garcia-López et al. (2020) study the impact of Airbnb's diffusion in Barcelona's housing market. They find that

Airbnb had a significant effect on housing rents and sale prices in Barcelona, especially in the city's most touristic areas, where they attribute to Airbnb's presence a 7% increase in rents and a 17% and 14% increase in transaction and posted prices. They impute this effect to the reduction in the long-term supply of housing units. Duso et al. (2020) assess Airbnb's impact on rental prices in Berlin. They exploit the exogenous shock to the market caused by the enforcement of a law that limits short-term rentals to identify the impact on rents. They find that an additional Airbnb listing increases by at least seven cents the average monthly rents per square meter. Finally, Franco and Santos (2021) investigate Airbnb's impact on house prices and rents in Portugal. They find that one percentage point increase in Airbnb density results in a 3.7% increase in house prices, while they find no evidence of an effect on rents. The impact on sale prices is greater in city centres and tourist areas. In line with the literature, we find that Airbnb's presence increases rent and, more significantly, sale prices. The effect varies greatly both across and within cities, suggesting that Airbnb's effect strongly depends on the cities' inherent characteristics. Table 1 provides a summarised view of the empirical methods adopted by the reviewed literature.

Table 1. The Literature on the Impact of Airbnb on House Prices and Rents

Author	**Objective**	**Method**	**Dependent variable**	**Independent variable**	**Control variables**
Sheppard and Udell (2016)	Impact of Airbnb on house prices in New York City.	Fixed-effect model (hedonic) and diff-in-diff. SE clustered at the census tract level.	Sale price of a given house. Data from 2003 to 2015	Number of active (i.e., first review) listings in a 300m radius from the property sold. Alternative measures: price, capacity, bedroom, beds, reviews. 12 points in time during 2015-2016.	Information about the house being sold; areas of interest (e.g., parks, cemeteries, airports, subway entrances); tax lot; census tract level information on education, racial and ethnic demographics, employment measures; crimes by precinct. Year of sale and neighbourhood fixed effects. Not all data is at census-tract level.
Horn and Merante (2017)	Impact of Airbnb on asking rents and on the number of houses available for rent in Boston.	Fixed-effect model. Asking rents used with a 1-month lag with respect to the Airbnb density measure to minimise the risk of reverse causality. SE clustered at the census tract level.	Asking rent of a given house at a given month. Data for the six months from 08-2015 to 01-2016.	Density of Airbnb in a census tract in the previous month: number of listings over number of housing units.	Number of beds and bathrooms, square footage. Number of newly built rental units in a given tract. Population, housing units, crime level, building permits and restaurant licenses issuances at the tract level. Census tract and month fixed effects.
Barron et al. (2020)	Impact of Airbnb on house prices and rents in the USA.	IV is the interaction of Google Trends global search index with a measure of how touristy a zipcode is in 2010 (proxied by establishments in the accommodation and food industries). SE clustered at the zipcode level.	Median sale price of houses at zipcode-month level. Median long-term rental price of houses at zipcode-month level.	Number of active listings in each zipcode (active starting from host join date). Data from 2008 to 2016.	Zipcode level 5-year estimates of income level, population, education, employment rate, owner-occupancy rate. 1-year estimates of housing vacancy rates at the metropolitan area (CBSA). Zipcode fixed effects, CBSA time varying effects, correlated with number of listings.
Koster et al. (2021)	Impact of Airbnb on house prices and rents in L.A. County	Spatial regression discontinuity design which compares changes in prices across municipality borders after localised bans to Airbnb. Diff-in-diff to estimate effect on rents. SE clustered at the census block level.	Transaction price and monthly rent of a given house. Data from 2014 to 2018.	Density of Airbnb listings in a 200m radius from the property. Monthly data of 15 points in time from October 2014 to September 2018.	Property and neighbourhood characteristics, location controls (distance to beach and to the Home-Sharing Ordinance municipality border). Census block and area-month fixed effects.
Ayouba et al. (2020)	Impact of Airbnb on asking rents in eight French cities.	Hedonic regression allowing for heteroscedasticity and spatial error autocorrelation of unknown forms. Distinction between nonprofessional and professional renters and on all tenancy agreements and only new ones. B-spline functions for some controls. Lagged variables to limit endogeneity.	Asking monthly rent of a given apartment at a given year for 2014-2015.	Density of Airbnb listings in a given neighbourhood for a given year. A differentiation is made between professional and non-professional hosts.	Structural characteristics of dwellings, accessibility to jobs and services, socioeconomic context, environmental quality around housing. Time fixed effects.

Author	**Objective**	**Method**	**Dependent variable**	**Independent variable**	**Control variables**
Garcia-López et al. (2020)	Impact of Airbnb on housing market in Barcelona.	Fixed-effect models (hedonic). IV is the interaction of Google Trends local search index with a measure of how touristy a zipcode is (obtained from Tripadvisor reviews). SE clustered at the neighbourhood level.	Residual resulting from a hedonic regression of log rents or prices on time dummies and unit characteristics, from 2007 to 2017.	Number of Airbnb listings. 21 points in time, from 2015 to 2018.	Neighbourhood and time fixed effects. Neighbourhood level time trends and demographic effects (i.e., average age, population density, average household occupancy rate, unemployment rate, relative income, and percentage of foreign residents).
Duso et al. (2020)	Impact of Airbnb on asking rents in Berlin.	IV using an exogenous shock: introduction of ban on the use of apartments as short-term rentals. They consider only “entire home” dwelling types because only these are affected by the law. SE clustered at the zipcode level.	Asking monthly rent per m2 of a given apartment for 2013-2018.	Number of active listings in a 250m radius from the property sold. Monthly data from 2015 to 2018.	Neighbourhood characteristics: number of restaurants, bus stops, supermarkets, level of noise, air quality, age of buildings. Apartment characteristics: size, number of rooms, parking lot. Linear and quadratic monthly trend and zipcodes fixed effects.
Franco and Santos (2021)	Impact of Airbnb on house prices and rents in Portugal.	IV is the interaction of Google Trends global search index with a measure of how touristy a municipality or parish is pre-Airbnb expansion (proxied by Airbnb density). Diff-in-diff to compare results. SE clustered at the municipality or parish level.	Average housing rent and transaction price in a given municipality or parish at a given quarter. Data from 2010 to 2016.	Density of Airbnb listings in a given municipality or parish in a given quarter. Data from September 2016.	Census data on socio-demographic characteristics, number of dwellings, population. Further data on area and building characteristics at the parish level for the cities of Porto and Lisbon. List of time-invariant amenities. Year-quarter fixed effects.

2.2. The Effect of Short-Term Rental Platforms on the Housing Market

A few studies have enquired into the transmission mechanism through which short-term rental platforms can impact the housing market, highlighting how the effect can go in opposite directions. This section summarises the main contributions.

First and foremost, short-term rental platforms have reduced many of the frictions that were present in the short-term rental market, both on a transactional and on a trust level (Einav et al., 2016), driving some homeowners to switch from the long-term to the short-term rental market. Since, especially in the short term, housing supply is inelastic, this substitution leads to a price increase in the former market and to a reduction in the latter, as observed by Horn and Merante (2017) and Zervas et al. (2017). The magnitude of the substitution effect depends on many factors. Short-term rental prices are usually higher than long-term prices, and they often elude or are subject to a more lenient revenue taxation. On the one hand, owners can be attracted by the fewer restrictions given by short-term contracts, especially so in jurisdictions with strong tenant protection laws. On the other hand, owners could prefer long-term rentals due to risk aversion (e.g., due to fear of property depreciation caused by impolite short-term renters) or to reduce effort costs required by managing a short-term rental. In the long run, the quantity of houses that can supply short- and long-term rentals should increase, reducing the impact of home-sharing on the supply side. However, the increase in supply may not happen as it depends on a variety of factors such as land availability and building regulations, as documented by Gyourko and Molloy (2015).

Second, short-term rental platforms can increase the attractiveness of previously less interesting city areas: this effect, documented for the US by Farronato and Fradkin (2018) and Coles et al. (2018), can drive both long-term and short-term prices upwards due to a general demand increase. Relatedly, harsh increases in tourist presence may lower the attractiveness of an area for local residents, as pointed out by Filippas and Horton (2018) in a study on New York City.

Finally, short-term rental platforms' effects on rental prices also reflect on sale prices. House value can be measured by the present value of all future revenues and costs, including possible incomes from renting (Poterba, 1984). As such, any change in the rental market is reflected on the sale market with a higher magnitude. Moreover, since short-term rental platforms allow the host to rent unused capacity, this additional source of possible future income should drive up sale prices even further.

Although the literature identifies various and, to some extent, discordant effects, both theoretical and empirical studies suggest that the predominant one is the substitution effect. Indeed, our results show a general increase in both rent and sale prices. Moreover, these effects are stronger in the city centre, where Airbnb's presence has raised the most during the analysed time period.

2.3. The Welfare Redistribution Effect of Short-Term Rental Platform Within Cities

A few recent studies have focused on how the effect of short-term rental platforms can actively shape the welfare distribution within a city, leading to heterogeneous effects on different areas and demographic groups as documented for Amsterdam by Almagro and Domínguez-Iino (2021). First, the entry of short-term rental platforms can actively shape touristic flows and, in turn, the distribution of amenities within the city. As such, amenities in touristic areas start focusing on satisfying the tourists' taste, which are misaligned with that of residents. In the end, the damage that short-term rental platforms do to residents is two-fold: their average rent raises, and the amenities shift farther from their taste as they try to accommodate tourists. Both of these effects are directly proportional to the magnitude of touristic flows, and thus are more concerning in more touristically attractive areas such as the city centres.

A similar analysis on the heterogeneous effect of short-term rental platforms has also been carried out in New York by Calder-Wang (2021). Her results show how the entry of short-term rental platforms causes a substantial welfare loss for renters, as the increase in rent prices more than offsets the additional gains from short-term renting. Moreover, this loss is particularly high in rich neighbourhoods, mostly inhabited by high-income and well-educated white people. This is due to the higher penetration of short-term rentals platforms, and the fact that high-income renters' preferences are more aligned with those of tourists.

Koster et al. (2021) take a quasi-experimental approach on the subject, analysing the differences of short-term rental platforms' effect in cities in Los Angeles County where Home Sharing Ordinances (HSOs) are in place. Their results suggest that HSOs reduces short-term rental platforms' supply and, in turn, housing prices and rents. This effect is particularly strong in areas with high touristic attractiveness. Moreover, they argue how short-term rental platforms redistribute welfare from long-term renters to homeowners as more supply is located in the platforms, and how HSOs can reduce this welfare transfer.

Finally, short-term rental platforms can also have a role in the distribution of capital investments within a city. As documented for Chicago by Xu and Xu (2021), the entry of shot-term rental platforms increases the number of residential renovation projects, as landlords are more inclined towards renovating to better compete on the short-term rental market. Moreover, they observe how the elasticity of renovation projects with respect to short-term rental platforms' penetration increases within declining neighbourhoods; that is, the investment responses have been stronger in non-gentrified neighbourhoods, likely due to lower investment costs.

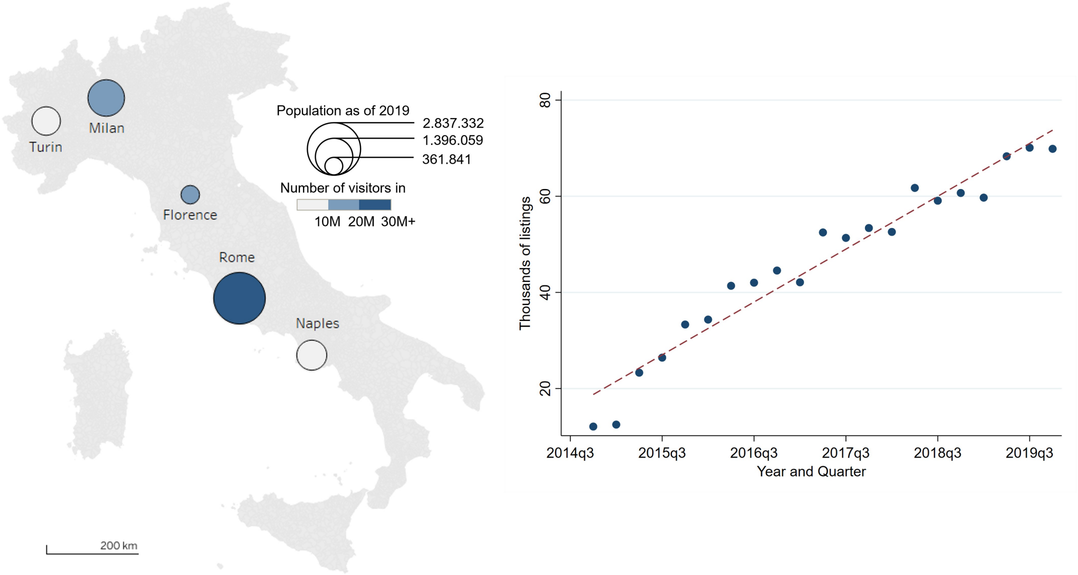
We contribute to this strand of literature by investigating whether an increase of Airbnb's presence in the city centre affects prices in that city's suburbs. Indeed, we find a significant negative effect in most cities, hinting that the presence of Airbnb in the centre and the related boost to the supply of localised amenities may actually reinforce its attractiveness at the expense of the peripheries. We suggest that the distributional impact of Airbnb within cities depends on the initial conditions of the areas, indicating that urban policy calls for regulation should take them into account.

3. Geographical Scope of the Analysis

In Figure 1, the left pane shows the location of the five cities, their population and the number of visitors in 2019, while the right pane reports the overall number of listings per quarter in the five cities, testifying to the explosive growth of the platform in the country.

Our choice of cities – Florence, Milan, Naples, Rome, and Turin – is driven by their importance in the economic and political life of the country as well as their symbolic, worldwide fame as touristic attractions and their inherent variety. In this section we describe how they differ in terms of Airbnb density, housing market indicators, demographic and economic conditions. Interestingly, these differences amplify when we focus on the comparison between the city centre and the peripheries. Indeed, one purpose of this paper is to start from the heterogeneity of the initial conditions to investigate how the impact of Airbnb's presence changes with the specific characteristics of each city. We then proceed to further disentangle the effect of Airbnb within cities. More precisely, we divide each city into the city centre, which typically attracts the large majority of tourists, and the suburbs, which might be only residually affected by house sharing – when the centre is saturated or through spillovers from the centre.

Figure 1. The Five Cities



Left pane: Location, number of residents and visitor as of 2019. Right pane: Total number of listings (the dashed line is the linear fit). Sources: AirDNA and ISTAT.

Table 2 summarises information on rents and sales prices as well as Airbnb's listing density (the number of listings divided by the number of dwellings) in each city and overall, at the beginning and the end of the sample period (i.e., the last quarters of 2014 and 2019). Airbnb's presence has grown overall, both in Florence – where it increased from 1.20% to 5.84% – and in Turin, where it was almost non-existent in 2014. However, the evolution of the housing market indicators is flat in most towns, probably as a consequence of the financial crisis of 2008. This suggests that a potential role of Airbnb's presence might be to slow down, or reverse, the descent of prices, at least in some cities.

Rome – the capital – is by far the most populous and most visited city in the country, with almost 3 million citizens and over 33 million visitors in 2019. Its rents and sale prices are about average with respect to our sample, but property values decreased, on average, over the period. Milan is the second most populous and third most visited city in Italy as well as the economic and financial capital. It is a highly productive and dynamic city, as shown by the highest and growing average income. Not surprisingly, Milan exhibits the highest sale prices and rental rates. Florence is significantly smaller than the previous two cities but ranks first in terms of tourist density and second as Italy's most visited town, after Rome. Its high attractiveness together with its limited geographical extension result in the highest average Airbnb density in our sample, together with high house prices and rents, the highest after Milan. Finally, Naples, a touristic seaside town rich of archaeological sites (e.g., Pompei) and Turin, an industrial town (headquarter of the

Table 2. Housing Market Characteristics by City at the Beginning and the End of the Period

	2014q4				2014	
	Monthly Rent	Sale Price	Airbnb Density	Store Density	Revenue	Tourists/capita
Average	12.11	3,338.79	0.41%	12.45	24,954	7.52
Florence	12.90	3,504.76	1.20%	8.64	23,624	22.82
Milan	13.65	3,557.11	0.29%	17.48	30,156	7.73
Naples	9.28	2,759.34	0.15%	15.28	19,880	2.96
Rome	12.63	3,633.47	0.43%	9.87	24,577	8.26
Turin	7.90	1,945.49	0.11%	9.83	22,542	3.39
	2019q4				2019	
	Monthly Rent	Sale Price	Airbnb Density	Store Density	Revenue	Tourists/capita
Average	13.47	3,147.18	2.32%	12.98	26,019	9.75
Florence	14.73	3,827.65	5.84%	8.81	24,444	30.28
Milan	17.71	3,891.87	2.11%	18.90	32,330	8.92
Naples	9.85	2,231.41	2.23%	15.59	19,757	4.00
Rome	12.65	3,076.66	1.87%	10.01	25,262	11.11
Turin	7.57	1,584.09	0.74%	10.15	23,793	4.27

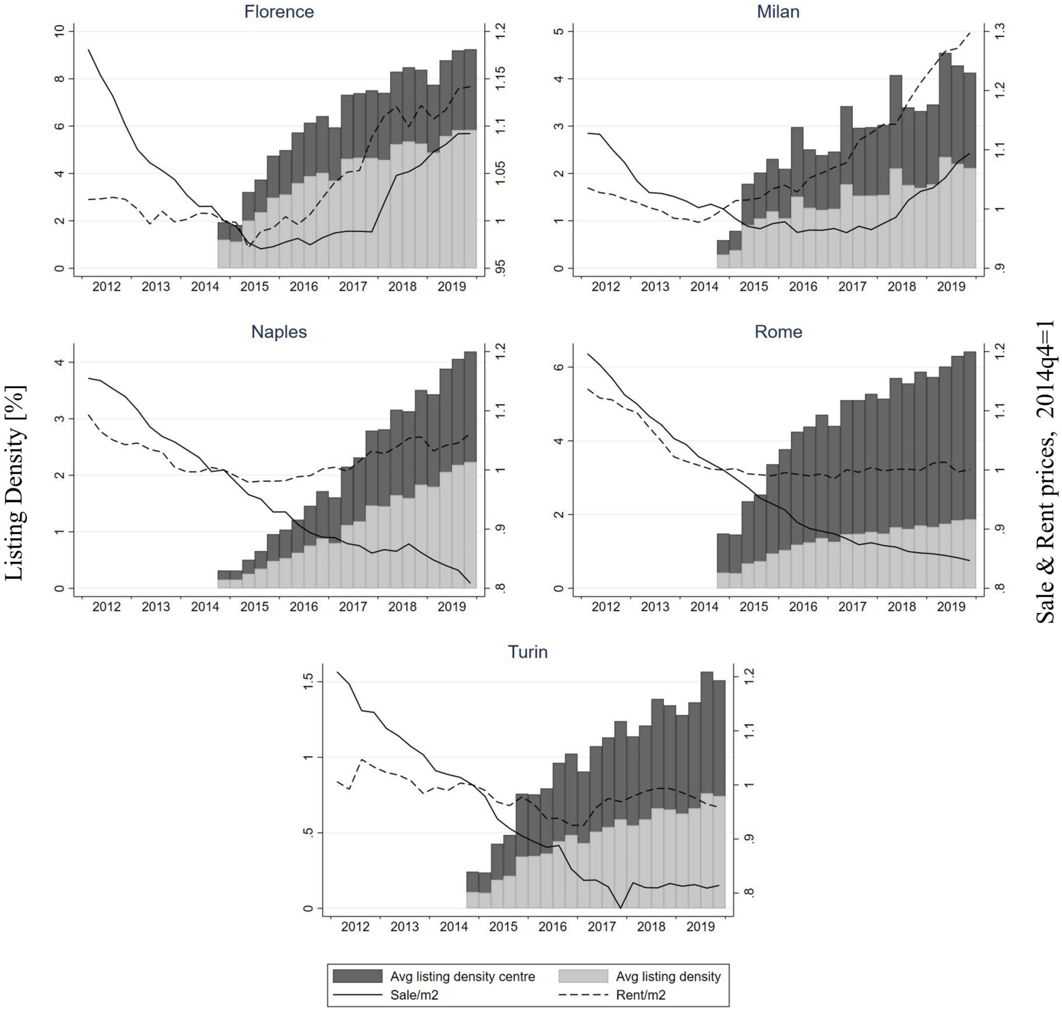
This table shows average values for monthly rent, sale price, listing and store density – overall and by city – for the last quarter of 2014 and 2019. Average revenue and tourists per capita at the yearly level are shown for year 2014 and 2019. Rent and sale prices are expressed in euros per square meter. Sources: AirDNA Idealista, ISTAT and OMI.

motor vehicle company FIAT, now Stellantis, i.e., FIAT, Chrysler and PSA), complete our analysis. These cities attract relatively less tourists but still rank tenth and twelfth as most visited cities in Italy, respectively. Compared to the previous cities, house prices, rents and Airbnb's presence are sensibly lower, especially in Turin (property values are 40.7% of Milan's), as well as per capita income, especially in Naples (61% of Milan's).

Figure 2 shows the evolution over time of average rents, sale prices and Airbnb density for each city. Rents and sale prices are shown from the first quarter of 2012 and are normalised to the last quarter of 2014. Listing density is shown from the last quarter of 2014 as the average of all zones (light grey) as the average of zones located in the city centre (dark grey).[1]

[1] To classify whether a zone belongs to the centre or the periphery, we adopt the definitions of OMI, the Italian register of the real estate market of the Internal Revenue Service. See the data section for further details.

Figure 2. Airbnb Density, Rents and House Prices From 2012 to 2019 (Average Data by Zone)



This figure shows the evolution over time of average rents, sale prices and Airbnb density for each city. Rents and sale prices are shown from the first quarter of 2012 and are normalised to the last quarter of 2014. Average listing density is shown from the last quarter of 2014 for both the average zone (light grey) and for the city centre (dark grey). Sources: AirDNA, Idealista, OMI.

The five cities exhibit different trends of house prices and rents. Starting from 2014, when the home sharing phenomenon was dawning, Florence and Milan experienced a steady increase in both rents and sales prices. In the other cities, sale prices have been falling over time, although in Rome and especially in Turin the descent reduced its speed at the end of 2017. The evolution of rents is similar in Rome, Naples and Turin, as they first declined until around 2015 and afterwards stabilised or slightly reverted the trend. Airbnb density also varies greatly between cities. Florence approaches densities of 30% in the top decile. Rome and Milan reach peaks of about 15% and 10% respectively, while Turin stays below 3% even in top decile zones.

3.1. A Closer Look Within the Five Cities: Centre vs. Suburbs

All cities exhibit, to a great extent, large differences between the city centre and the peripheries under many aspects. Table 3 reports the average values of the main variables in the city centre and in the suburbs. In Appendix I we also provide maps for individual cities showing, for each zone, Airbnb's density and a measure of tourist attraction derived from Trip Advisor (defined *tourist attraction score* throughout the paper, as explained in Section 5.1.1).

Table 3. Airbnb Presence and Housing Market Characteristics Within City

2014q4	Suburbs				Center			
	Monthly Rent	Sale Price	Airbnb Density	Store Density	Monthly Rent	Sale Price	Airbnb Density	Store Density
Average	11.09	2817.53	0.10%	6.97	14.18	4392.28	1.05%	23.51
Florence	12.41	3193.87	0.25%	3.77	13.28	3746.57	1.94%	12.44
Milan	12.24	2764.23	0.16%	8.50	17.08	5487.60	0.59%	39.34
Naples	8.19	2300.23	0.02%	8.04	10.58	3310.27	0.31%	23.96
Rome	11.49	3116.82	0.06%	6.46	15.92	5114.53	1.49%	19.64
Turin	7.45	1778.04	0.04%	6.08	8.79	2280.38	0.24%	17.35
2019q4	Suburbs				Center			
	Monthly Rent	Sale Price	Airbnb Density	Store Density	Monthly Rent	Sale Price	Airbnb Density	Store Density
Average	12.22	2487.49	0.70%	7.35	16.02	4480.45	5.58%	24.36
Florence	14.01	3372.26	1.46%	3.85	15.29	4181.84	9.25%	12.66
Milan	15.93	2709.56	1.29%	9.31	22.04	6770.54	4.13%	42.24
Naples	8.16	1759.23	0.60%	8.23	11.88	2798.04	4.19%	24.42
Rome	11.43	2595.42	0.28%	6.68	16.14	4456.20	6.42%	19.57
Turin	7.07	1321.08	0.36%	6.27	8.58	2110.10	1.51%	17.91

This table shows average values for monthly rent, sale price, listing and store density – in the suburbs and in the city centre – for the last quarter of 2014 and 2019. Sources: AirDNA Idealista and OMI.

First, as expected, the density of Airbnb listings is much higher in the centre than in the suburbs in all cities, but particularly so in Florence and Rome. This is in line with the tourist attractions distribution, which is heavily skewed towards the city centre, as we can see from the maps in Appendix I. The gap between centre and suburbs was modest at the beginning of the period, but then, as the home sharing phenomenon set off, it has become remarkable. Second, house prices and rents are also very different, especially in Milan and Rome whereas in Florence the gap is smaller. Turin's home market confirms to be the most stagnant, with the lowest prices in both years and in both areas, and the gap seems to have widened from 2014 to 2019. Third, when we focus on the demographic characteristics described by the time invariant variables (Appendix

Table A.1), we find that the share of graduated and employed people is much higher in the centre in all cities. The share of elderly residents is slightly higher in the city centre, except in Milan and Turin, whereas the share of houses occupied by owners (more than 60% on average) is slightly higher in the suburbs, except in Naples. According to Barron et al. (2020), the impact of Airbnb presence on house prices and rents is lower where the share of owner occupancy is higher. Finally, to compare average income levels, we rely on data by the Ministry of Economy and Finance (MEF) for the poorest and the richest postal code (zip code) neighbourhoods in 2019 (Appendix Table A.2), as the information about income distribution at the Idealista zone level was unavailable (MEF, 2021). Assuming that the poorest zip code is in the periphery and the richest in the centre, we find that income inequality within city is highest in Milan (where the ratio "highest to lowest income" is as high as 5.3) and Rome (4.2) and lowest in Florence (1.97), whereas Naples and Turin stand in the middle (3.5). More in general, it is worth noting that the heterogeneity across cities is much higher amongst the "rich" zip codes (the ratio between Milan and Florence is 2.48) than amongst "poor" zip codes (the Florence to Naples ratio is 1.65), thus suggesting that suburban areas in Italy tend to share similar disadvantaged conditions.

4. Data

4.1. Rent and Sale Prices Data and Neighbourhood Definition

Our source of rent and sale prices is Idealista, a major online real estate portal operating in the Italian market (Idealista, 2021). Idealista divides each city into zones, that is, geographical areas sharing common characteristics. We will use the term zone and neighbourhood interchangeably. Idealista data cover 287 zones from the first quarter of 2012 to the first quarter of 2020, and the number of zones per city varies significantly according to each city's characteristics. For each zone, Idealista provides an estimate of the monthly rental rates and the transaction prices per square meter at the trimester level. We can thus think of the identification of a zone as being equivalent to that of a relevant market. By choosing Idealista's zones as our definition of neighbourhood, we can approximate the geographical scope of the individual housing market, as the real estate company has likely chosen the zones to minimise the area-specific heterogeneity and the information costs. This mapping allows us to compare different zones both across and within cities controlling for unobserved zone-level factors and helps us identify the impact of Airbnb.

4.2. Airbnb Data

Data on Airbnb come from AirDNA, a provider of short-term rental data and analytics, which collects information directly from Airbnb's website (AirDNA, 2021). AirDNA provides two datasets: a property one and a daily one. The property dataset provides information on dwelling characteristics, ownership and rental conditions The daily dataset provides, for each dwelling, rental outcomes such as whether the dwelling was available, rented, blocked and, if rented, at what price. This fine-grained detail allows us to measure Airbnb supply reliably: rather than using reviews or the listing's creation date as a proxy of activity, we can look at the actual days in which the property was available or rented. AirDNA data covers the period from October 2014 to December 2019. The datasets report the coordinates of each dwelling, albeit with a margin of error. For privacy reasons, in fact, Airbnb scrambles these coordinates so that the reported location of the dwelling is within a 150m radius from the actual ones. As the anonymised data change over time, AirDNA provides an average of these values, therefore increasing geolocation precision.

We merge the AirDNA and the Idealista datasets by assigning the listings to the zones, and we finally obtain two measures of Airbnb intensity at the zone-trimester level: the number of listings and the listing density. The former is derived as the number of listings being offered for rent in a given trimester and reserved at least once during the year – a constraint needed to expunge listings that are not really active. The latter is defined as the ratio between the number of listings and the number of houses in a given zone.

4.3. Final Dataset

To characterise each zone according to the attributes of its real estate and the sociodemographic and economic dimensions, we rely on two additional sources: the OMI (*Osservatorio del Mercato Immobiliare*, the Italian register of the real estate market) dataset and the Italian 2011 census by the Italian National Institute of Statistics (ISTAT).

OMI provides, for its own geographical partitions, the annual number of housing units, their average number of rooms, the number of commercial activities, their average size in square meters, and the number of garages. Data are available from 2016 to 2019 for every city but Rome, for which they start from 2017. To match the time series of Airbnb data, we extrapolate the OMI data for 2015 (also 2016 for the city of Rome) and the last quarter of 2014, assuming a linear trend. We assigned Idealista zones to their respective OMI partitions – with some minor approximation. OMI partitions are smaller than Idealista's zones and, typically, they are contained within the Idealista's zone. When the two geographical units do not completely overlap, we merge the respective data (under the assumption that the real estate market is uniformly distributed within the OMI partition) and assign a share equal to the percentage of overlap to the Idealista zone. OMI partitions are further characterised as central, semi-central, peripheral, suburban and rural. We make use of this distinction to define a binary variable – which we call area – that identifies an Idealista zone as belonging to either the "city centre" or the "suburbs". We define a zone as belonging to the city centre area if it is either a central or semi-central zone. Conversely, a zone is defined as belonging to the suburbs area if it is either peripheral, suburban or rural. In the empirical analysis, we exploit this dichotomy to investigate whether the impact of Airbnb presence differs contingent on the centrality of the zone and to account for centrality-driven unobserved factors through time-varying area fixed effects.

Through OMI data, now attributed to the Idealista zone, we calculate Airbnb density as the ratio between the number of listings and the number of houses. Similarly, we compute the housing, store and garage densities by dividing the corresponding stock to the area of the Idealista zone, expressed in hectares.

To find additional pre-determined control variables we exploited the 2011 census which provides a wealth of (time-invariant) data on demographics, education, occupation and housing characteristics (ISTAT, 2021) at the census tract level for the cities in the analysis. We collected the number of residents, characterised by age, education level, employment status and citizenship; the number of owner-occupiers; the number of houses, further characterised by occupancy and physical condition. To give an idea of the geographical resolution of census data, note that, while Rome is divided into 117 Idealista zones, it consists of about 13,000 census tracts. Therefore, by appropriately rearranging the data, it is possible to characterise an Idealista zone accurately with census variables.

The resulting dataset consists of a balanced panel of 6,027 observations at the zone-trimester level. It comprises 287 zones and 21 time intervals from the last trimester of 2014 to the last of 2019. Table 4 presents summary statistics at the zone-trimester level except for census variables, which are time invariant.

Table 4. Summary Statistics

	Mean	SD	Min	Max	N
Idealista					
Rent [€/m²]	12.60	3.65	4.41	32.22	6,027
Sale [€/m²]	3,131.35	1,391.56	694.44	10,889.59	6,027
Airbnb					
Airbnb Listings	161.06	370.62	0	5,353.00	6,027
Airbnb Density	1.55%	3.43%	0.00%	31.12%	6,027
OMI					
House Density	43.89	33.46	0.61	161.09	6,027
Store Density	12.75	12.94	0.22	75.95	6,027
Garage Density	15.69	11.25	0.28	43.77	6,027
Avg. House Rooms	5.11	0.67	3.78	9.19	6,027
Avg. Store m²	45.06	13.44	9.84	98.29	6,027
CENSUS					
Num. Residents	20,300.74	13,683.38	1,072.05	71,855.31	6,027
% Owner-occupancy	0.65	0.10	0.29	0.83	6,027
% 20-39 years	0.24	0.04	0.17	0.45	6,027
% >60 years	0.22	0.05	0.06	0.36	6,027
% Graduates	0.19	0.10	0.03	0.44	6,027
% Working	0.41	0.06	0.18	0.57	6,027
% Foreigners	0.10	0.06	0.01	0.37	6,027
% Full houses	0.93	0.06	0.61	1.00	6,027
Num. Houses	9,675.36	6,477.95	248.57	30,495.69	6,027
% House in poor condition	0.15	0.13	0.01	0.78	6,027

This table shows summary statistics for the main variables used in the empirical analysis. Sources: AirDNA Idealista, ISTAT and OMI.

5. Empirical Methods

The empirical analysis starts by estimating the overall impact of Airbnb presence on monthly rents and sale prices for the five cities altogether. Next, we turn to the main research questions, and we exploit the heterogeneity in our dataset: we first disentangle the individual Airbnb effect in each city, then we investigate within-city differences by estimating the impact of listing density on the city centre and on the suburbs. Finally, we address the potential externalities on the real estate market by investigating if (and how) the intensification of Airbnb presence in the city centre affects the housing market in the periphery. Our research strategy accounts for endogeneity concerns in several ways. For a start, we lag the variable of interest to reduce reverse causality concerns and we include a large set of control variables to address the problem of omitted variable bias, including spatial (city, area and zone) and time (year and quarter) fixed effects as well as spatial-time interactions. Controlling for spurious correlations is crucial because there may be city-and area-specific trends and seasonal factors that affect sale and rental prices but are unrelated to Airbnb, for example a divergence between nicer and less areas over time, which is important to account for. Then we estimate 2SLS regressions using shift-share instrumental variables that exploit two alternative sources to construct the cross-sectional part of the instrument (Section 5.1), and we also estimate a dynamic panel data model using with the GMM-System estimator to allow for dynamic effects in the housing market (Section 6.4.1). We cluster standard errors at the zone level to account for correlation across the time-dimension within zones and, because neighbourhood effects may exhibit patterns of mutual dependence across zones, we also allow for spatial correlation by calculating Driscoll and Kraay (1998) standard errors that we report below the robust standard errors clustered by zone. A battery of robustness tests concludes the analysis.

To estimate the overall impact of Airbnb we start with the following baseline OLS equation:

$$\log\left(Y_{n,t}\right) = \beta \mathit{Airbnb\ Intensity}_{n,t-1} + \gamma X_{n,t} + \delta F_n + \tau_y + \varepsilon_{n,t} \tag{1}$$

where $Y_{n,t}$ is either rents or sale prices in zone n at year-quarter t. $\mathit{Airbnb\ Intensity}_{n,t-1}$ is the listing density in zone n at time $t-1$.[2] As the dependent variable is expressed as natural logarithms, our estimates represent the semi-elasticity of rents and prices with respect to Airbnb intensity. $X_{n,t}$ is a matrix of time-varying controls in zone n at time t, and F_n is a matrix of time-invariant demographic characteristics of zone n. τ_y are year fixed effects. $\varepsilon_{n,t}$ is a mean-zero error term.

Our coefficient of interest in Equation (1) is β, which captures the overall effect of the Airbnb intensity measure on the dependent variable. We address the risk of omitted variable bias by controlling for both time-varying and time-invariant demographic and structural characteristics at the zone level as well as year fixed effects.

Starting from the second model, we include zone fixed effects as well as city and area-specific time effects in order to control for differences in time trends and seasonality amongst and within cities which imply different pricing dynamics that, if unaccounted for, may bias the estimate of the overall impact of Airbnb via spurious correlations. As shown in Section 3, the five cities are indeed differently exposed to tourist flows, business-related traffic, occurrence of shows and exhibitions and subject to different seasonality of the flows. Moreover, they are characterised by different economic performances and conditions, income levels and distribution, and degree of marginalisation of the peripheries; as such, housing market dynamics are expected to differ not only across cities but also within cities. We thus refine the specification by introducing an appropriate set of

[2] In Section 6.4.4 we also use the number of listings to measure the intensity of Airbnb.

interacted time and location fixed effects, at different levels, as per the following model:

$$\log\left(Y_{n,t}\right) = \beta Airbnb\ Intensity_{n,t-1} + \gamma X_{n,t} + \pi_{s,i} + \tau_{y,i,a} + \mu_n + \varepsilon_{n,t} \tag{2}$$

where all terms have the same meaning as before, except for μ_n which is a zone-specific fixed effect, $\pi_{s,i}$ is the interaction between city i and quarter s (to account for city-specific seasonality) and $\tau_{y,i,a}$ is the interaction among the year, the city and the area, i.e., city centre vs. periphery.

When we add the zone fixed effects, the time-invariant controls for sociodemographic factors cannot be estimated, even though they may still display, over time, an impact on sale and rental prices unrelated to Airbnb. Because part of this impact can escape the spatial-time fixed-effects interactions we add in Equation (2), we adopt an additional strategy to account for these factors, by interacting the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. Plausibly, these additional variables should be correlated with the remaining unobserved factors that influence the housing market at the zone level, contributing to reduce the problem of spurious correlations (see also Section 5.1.2). Moreover, by comparing these results with those from Equation (2), we can also check to what extent we actually miss the effect of dynamic confounding factors unobserved to us.

The previous models do not allow to estimate a city-specific effect of the impact of Airbnb nor allow for different effects between the city centre and its periphery. To investigate these further issues, we estimate the following models:

$$\log\left(Y_{n,t}\right) = \beta Airbnb\ Intensity_{n,t-1} \times \text{city}_i + \gamma X_{n,t} + \pi_{s,i} + \ \tau_{y,i,a} + \mu_n + \varepsilon_{n,t} \tag{3}$$

and

$$\log\left(Y_{n,t}\right) = \beta Airbnb\ Intensity_{n,t-1} \times \text{city}_i \times \text{centre}_b + \gamma X_{n,t} + \pi_{s,i} + \ \tau_{y,i,a} + \mu_n + \varepsilon_{n,t} \tag{4}$$

where all terms have the same meaning as before, except for $Airbnb\ Intensity_{n,t} \times \text{city}_i$ which is the interaction between listing density in zone n at time t with the city i, and $Airbnb\ Intensity_{n,t} \times \text{city}_i \times \text{centre}_b$, which is the interaction among listing density in zone n at time t with city i and the indicator variable that denotes the city centre.

Finally, we turn to our research question on the distributional impact of Airbnb on the city residents in the centre and in the suburbs by investigating whether the diffusion of Airbnb in the centre impacts also the property values and rents of suburban zones. To perform this analysis, we focus on the sub-sample of peripheral zones, and we calculate the aggregate Airbnb density in the city centre, while keeping zone-level density as a measure of Airbnb presence in the suburbs – as in previous regressions – to control for its effect as well. We then estimate the following model:

$$\begin{aligned}\log\left(Y_{n,t}\right) = \beta_1 \mathit{Airbnb\ Intensity}_{n,t-1} \times \text{city}_i + \beta_2 \mathit{Airbnb\ Intensity\ Centre}_{i,t-1} \\ + \gamma X_{n,t} + \pi_{s,i} + \ \tau_{y,i,a} + \mu_n + \varepsilon_{n,t}\end{aligned} \tag{5}$$

where n represents here the subsample of suburban zones. Through this model, we investigate whether the diffusion of Airbnb in the city centre (i) has consequences also on the property values and rents of suburban zones (n), after controlling for the effect of zone-specific Airbnb density on each suburban zone.

Controlling for Different Evolution Over Time of Centre and Suburbs. In our setting, a key threat to our identification strategy is that neighbourhoods with a higher tourist attractiveness a_n^{TA} – and therefore a higher Airbnb penetration – could also be undergoing a process of sociodemographic transformation which might have an impact on house prices and rents (Garcia-López et al., 2020). Our specifications controls for this phenomenon through the inclusion of zone-level time-varying housing characteristics associated with urban revival processes (such as that of gentrification) and, more to the point, centre- and suburbs-specific year-level fixed effects. In particular, these fixed effects control for any spurious correlation that comes from the fact that a high tourist attraction score is a good proxy of centrality of a neighbourhood. However, we might still miss some trends due to the unavailability of time varying socio-demographic variables.

5.1. Instrumental Variable Estimation

Our previous specifications control for unobserved factors at the zone and city-centre-year and city-quarter level, but do not yet account for unobserved zones-specific and time-varying factors contained in the error term $\varepsilon_{n,t}$ and correlated with the measure of Airbnb intensity $\mathit{Airbnb\ Intensity}_{n,t}$. To address these and other endogeneity concerns, we employ an instrumental variable method that follows the approach introduced by Bartik (1991), which has previously been adopted in the literature studying the impact of Airbnb on the housing market by Garcia-López et al. (2020) and Barron et al. (2020). The approach makes use of a shift-share instrumental variable that combines a cross-sectional variation across zones of a measure of tourist attraction, and an aggregate time variation of a measure of Airbnb intensity. We will provide evidence of how this instrument is plausibly uncorrelated with local shocks to the housing market $\varepsilon_{n,t}$ while affecting the intensity of Airbnb penetration.

5.1.1. Construction of the Instrument

The cross-sectional part (share) of the shift-share instrument is a measure of tourist attractiveness of a given zone, which we draw from Tripadvisor. For each city, we scrape the list of the top 150 tourist attractions, their geographical coordinates and their respective number of reviews until the end of 2013, that is, before the beginning of our analysis' time window, in order to prevent reverse causality concerns. We define a measure of the tourist attractiveness of a zone as follows:

$$a_n^{\text{TA}} = \sum_k^K \frac{\text{reviews}_k}{\text{dist}_{n,k}}$$

where n represents the zone, k the tourist attraction, reviews_k the number of reviews of attraction k, $\text{dist}_{n,k}$ the distance of attraction k from the centroid of zone n expressed in kilometres. This variable predicts where Airbnb listings locate, as the presence of tourist attractions increases tourists' willingness to pay, which in turn raises both listing price and Airbnb activity (Garcia-López et al., 2020).

In addition, we construct an alternative measure of tourist attractiveness that we use as a robustness test, based on Lonely Planet guidebooks. Lonely Planet lists, the top 10 sites of interest for each city in their books and websites, ordering them by popularity. We geolocate these sites using Google Maps' API to get the coordinates, and we define the alternative share component as follows:

$$a_n^{\text{LP}} = \sum_k^{10} \frac{1/\text{position}_k}{\text{dist}_{n,k}}$$

where position_k is the position of the attraction in Lonely Planet's list, and the other terms have the same meaning as before.

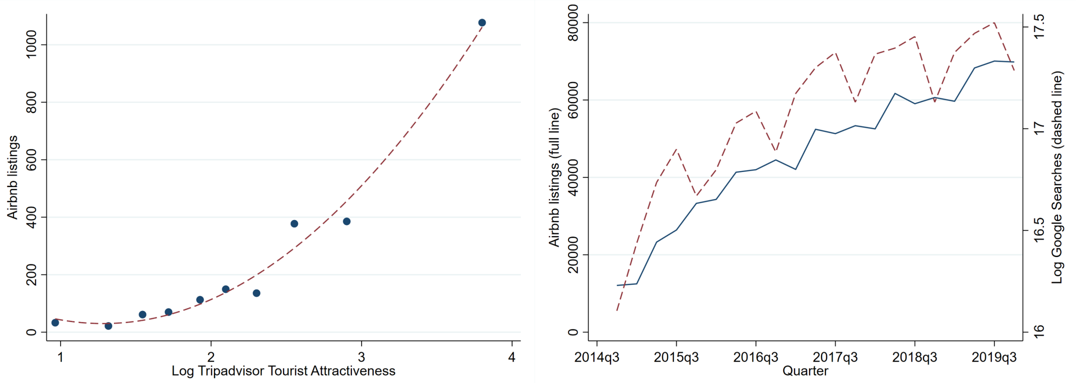
The shift-share instrument's temporal part (shift) is a measure of Airbnb intensity over time, which we refer to as g_t and we derive from Google Trends by retrieving the number of worldwide searches of the word "Airbnb" at the monthly level. Google Trends provides percentages relative to the month with the highest number of searches. We convert these into absolute numbers by matching them with data from WordTracker, a website that provides numbers of searches for the last 12 months. This variable provides a proxy of Airbnb intensity by representing the extent of public awareness of the platform on both the demand and supply sides. As pointed out by Barron et al. (2020), the limited time window of the analysis makes it unlikely that the shift component reflects the growth of overall tourism demand, while it should reflect the growth of the short-term housing supply only where caused by Airbnb. Notably, we use a global measure of searches rather than a city specific one so that correlation with tourist flows at the local level is unlikely.

Our instrument, referred to as *touristiness*, is thus the product of the cross-sectional and temporal components (see Garcia-López et al., 2020 and Barron et al., 2020 for a similar approach), as follows:

$$z_{n,t} = a_n^{\text{TA}} \times g_t$$

Its intuitive rationale is that the attractiveness score a_n^{TA} predicts where Airbnb listings appear, while the number of searches g_t predicts when they are offered. The working of the instrument is presented graphically in Figure 3.

FIGURE 3. Working of the Instrument



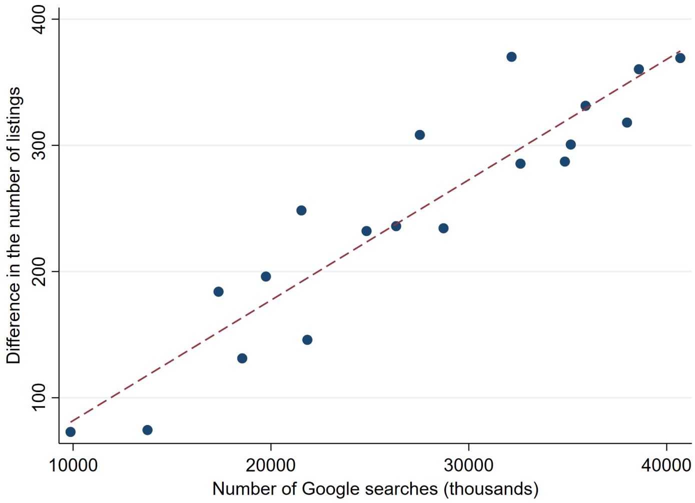
On the left pane, the figure shows the number of Airbnb listings as a function of the natural logarithm of the tourist attractiveness score. The dots are the deciles of the tourist attractiveness distribution, while the dashed line is the quadratic fit. On the right pane, the figure shows the number of Airbnb listings by quarter (blue line) and the natural logarithm of Google worldwide searches of the word "Airbnb" (dashed line).

The left pane shows the number of listings in a zone as a function of the natural logarithm of the tourist attractiveness score, where the dots are the deciles of the tourist attractiveness distribution. We can see how zones with a higher tourist attractiveness score have a larger number of listings. The right pane shows the number of listings over time and the number of Google searches of the word Airbnb. We can see how the number of Google searches approximates well the number of listings in a given quarter. While Figure 3 provides graphical evidence of the relevance of the instrument, in Section 6.1 we provide further proof by reporting the first-stage estimates.

The effectiveness of the instrument hinges on the fact that property owners become increasingly likely to offer their property on Airbnb after becoming aware of the platform.

Following Barron et al. (2020), we test this hypothesis by looking at the relationship between Google searches and the difference in the number of listings between tourist and non-tourist zones.[3] Figure 4 provides a visual representation that the hypothesis holds, as this difference increases with the number of Google searches.

FIGURE 4. Instrument Effectiveness



This figure shows the average difference in the number of listings between high- and low-tourist attractiveness zones – i.e., the zones above or below the median – as a function of the number of Google worldwide searches of the word "Airbnb" (blue dots). The dashed line is the linear fit.

5.1.2. Instrument Validity

The popularity of shift-share instruments has spurred the growth of a recent literature that studies their validity conditions (Christian and Barrett, 2017; Borusyak et al., 2018; Goldsmith-Pinkham et al., 2020). These works highlight how the consistency of the estimator can derive from the exogeneity of either of its terms, even when the other is endogenous.[4] However, the literature also underlines how – usually – the main identification threats come from the share component (Goldsmith-Pinkham et al., 2020). In our case, the exogeneity of the shares requires that the tourist attractiveness a_n^{TA} is uncorrelated with unobservable zone-specific time-varying shocks captured by the error term $\varepsilon_{n,t}$. That is to say, the tourist attractiveness of a zone should be correlated with changes in house prices and rents only through the density of Airbnb – after controlling for our set of covariates and fixed effects. Our empirical strategy already accounts for identification

[3] We split zones according to touristic attractiveness depending on whether they are below or above the median.

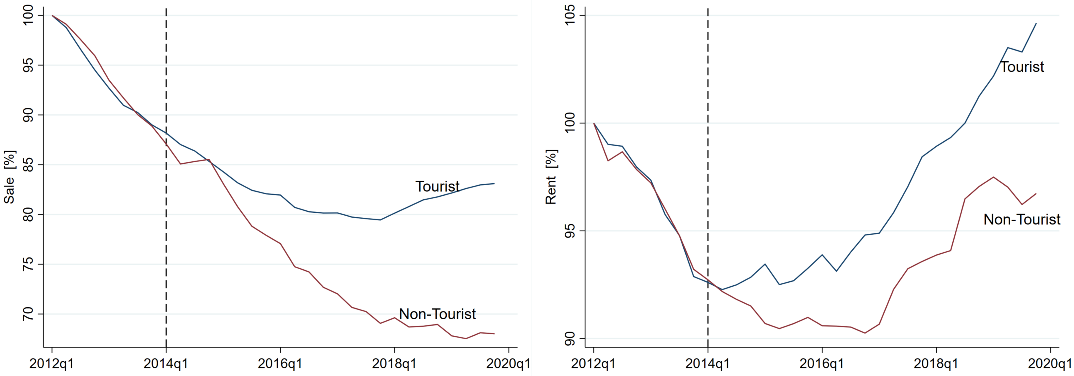
[4] Among others, consistency of the estimator can derive from the sole shift component where a long time series having weak serial dependence is present, even when there is a single shock per period (Borusyak et al., 2018).

threats attributable to the high correlation between tourist attraction and centrality by including centre- and suburbs-specific year-level fixed effects and zone-level time-varying controls associated with urban revival processes. However, in this section we make three further arguments for why the exogeneity condition is likely to hold in our setting.

Parallel Pre-Trends. Goldsmith-Pinkham et al. (2020) note how, as the shift-share instrument makes use of level differences in the share component, the validity of the following assumption should be assessed: that the shock (i.e., awareness of the Airbnb platform, as opposed to pre-existing conditions) is what determines the difference in the changes in house prices and rents. We can do so by looking at the trends in changes prior to the shock, similarly to a parallel pre-trend test in a difference-in-differences analysis.

We split neighbourhoods according to the tourist attraction score a_n^{TA}, distinguishing between non-tourist zones (first quartile of the distribution of a_n^{TA}, our control group) and tourist zones (the other three quartiles, the treatment group). While our data on actual Airbnb supply starts from 2014, we can see from Figure 1 how the number of listings in that year is still low compared to the following years.[5] We therefore take the first quarter of 2014 as the treatment period. Figure 5 shows the average house price (left pane) and rents (right pane), normalised to the first quarter of 2012.

Figure 5. Parallel Pre-Trends



On the left pane, the figure shows the number of Airbnb listings as a function of the natural logarithm of the tourist attractiveness score. The dots are the deciles of the tourist attractiveness distribution, while the dashed line is the quadratic fit. On the right pane, the figure shows the number of Airbnb listings by quarter (blue line) and the natural logarithm of Google worldwide searches of the word "Airbnb" (dashed line).

No differential pre-trends appear between tourist and non-tourist zones for both house prices and rents: the trends start to diverge only after the diffusion of Airbnb. This

[5] When looking at the creation date of the listing, we can approximate Airbnb supply from 2008 onwards. Even when using this measure of supply, we can identify in the beginning of 2014 the moment at which Airbnb's diffusion started to pick up pace.

graphical evidence suggests that the neighbourhoods with a different tourist attraction score did not generally have different long-run house price and rent trends.

IV Impact on Non-Airbnb Zones. We perform a second test by checking whether our touristiness instrument finds a significant effect in neighbourhoods that never registered an impactful Airbnb activity. For the instrument to be valid, it should find an effect only in those zones where listings are present. If an effect is found outside of those cases, the instrument does not predict house prices and rents only through the Airbnb density, and it is therefore capturing a spurious correlation.

We test this by estimating Equation 2, our most complete specification, regressing the natural logarithm of sale prices directly on the instrumental variable (without 2SLS). We do so in three subsamples:

1. Zones that never registered any Airbnb activity.
2. Zones that registered very little Airbnb activity.
3. Zones that registered a significant Airbnb activity.

Subsample 2 consists of zones having at most 100 listings throughout the entire time period, i.e., a maximum average of 5 listings per quarter. Subsample 3 is composed of neighbourhoods belonging to the top three quartiles of the distribution of listings. We do not limit the analysis to the first and third subsamples because the neighbourhoods that do not have any listing throughout the time period are very few. Furthermore, we perform this test only for the sale variable because – as we will see in Section 6 – we find a significant effect for rents only when estimating by city effects. Table 5 reports our findings, where the three columns represent the different subsamples.

Columns 1 and 2 of the table show that, after controlling for our set of covariates and fixed effects, there is no evidence of a statistically significant relationship between our instrument and house prices. The effect is significant only for the third column, where neighbourhoods with Airbnb activity are considered. This finding provides further evidence that the instrument predicts house prices only through the Airbnb intensity.

Table 5. Correlation Between Instrument and House Prices in Zones With No Airbnb Presence

Dep. Var:	Log Sale		
	(1)	(2)	(3)
Touristiness at $t-1$	3.60e-10 (2.98e-10)	4.37e-10 (3.17e-10)	1.85e-11** (8.91e-12)
Controls			
Time varying controls	X	X	X
Fixed Effects			
Zone FE	X	X	X
Quarter#City FE	X	X	X
Year#City#Area FE	X	X	X
Observations	680	1,040	2,860
Adjusted R2	0.969	0.951	0.986

FE estimates of Equation 2. The dependent variable is the natural logarithm of the sale price, while the variable of interest is the lagged touristiness. Columns 1 to 3 show the coefficients for the three different subsamples according to Airbnb activity. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Placebo Test. We further assess the exogeneity of our instrument by means of a placebo test, following Christian and Barrett (2017) and Barron et al. (2020). Through this analysis, we test whether the effects we estimate can be reasonably attributed to a causal relation or depend on spurious time trends. To this end, we randomise our measure of Airbnb intensity by swapping the number of listings among zones that have at least some degree of Airbnb penetration by the end of the time period (i.e., one listing by the last quarter of 2019). The swap is consistent between periods: if zone i is swapped with zone j, this is true for every quarter t. We keep constant every other variable of our analysis. Through this transformation we maintain any pre-existing correlation between a hypothetical omitted variable and the overall trend in Airbnb's diffusion, while losing any relationship between the instrument and the number of listings in each zone.[6] As a consequence, should the significance of our 2SLS coefficients be driven mainly by a spurious correlation with whether a zone has any Airbnb presence, we would expect the resulting 2SLS estimates to be significant also in the placebo. Conversely, if it is the intensity of Airbnb that truly drives our results, then the instrument should become weak, and the coefficients in the placebo should be insignificant.

We perform 5,000 iterations, reassigning Airbnb listings to different zip codes in each one. We then estimate the 2SLS specification of Equations 2 (overall effect) and 3 (effect

[6] Barron et al. (2020) aptly describe it as preserving the impact of touristiness on the *extensive margin* of Airbnb's diffusion (whether there are any listings) while eliminating its impact on the *intensive margin* (how many listings are there).

by city) with the scrambled dataset. Table 6 reports the share of iterations in which the estimated coefficient is significant at the 5% level, i.e., those that challenge the validity of the instrumental variable. The results for the overall impact show that the 2SLS placebo coefficient is insignificant at the 5% level in 100% of the randomised draws in the rents regressions and in 99% of the draws in the regressions for house prices. As such, we do not identify any spurious time trends.

Table 6. Placebo Test

Effect	**Share of significant iterations**	
	Sale	**Rent**
Overall	1.10%	0.00%
Florence	1.64%	0.44%
Milan	0.02%	0.02%
Naples	1.88%	1.44%
Rome	0.50%	0.00%
Turin	5.52%	0.34%

The table shows, out of the 5,000 iterations, the share of them in which the coefficient of the variable of interest of Equations 2 and 3 (i.e., overall and by city) estimated with 2SLS is significant at the 5% level.

The evidence is similar when looking at the impact by city: the share of significant coefficients in the placebo tests is well below 2% in all cities except for Turin's sale prices (5.52%). As shown in Table 2, Turin has, by far, the lowest Airbnb density as compared to the other cities, and we will be particularly cautious when assessing this city's coefficients.

Overall, these preliminary tests suggest that the touristiness variable we employ in the instrumental variable estimations is effective and that spurious time trends should not bias the results in our main analysis, providing evidence in favour of the robustness of our identification strategy.

6. Results

6.1. The Overall Effect of Airbnb Density

Table 7 and Table 8 present the results estimating the overall impact of Airbnb on rents and sale prices (€/m^2) for the five Italian cities over the period 2014-2019. The intensity of Airbnb is measured by listing density at the neighbourhood level.[7] Columns (1) and (2) report the OLS and fixed effect results, and Columns (3) and (4) the 2SLS estimates. First-stage results are at the bottom of the table.

[7] In Section 6.4.4 we report the results when using the number of listings as the variable of interest.

Table 7. Overall Effect on Sale

Dep. Var:		Log Sale		
	(1) OLS	(2) FE	(3) 2SLS	(4) 2SLS
Airbnb Density at $t-1$	0.937	0.561	0.630	0.618
	(0.479)*	(0.149)***	(0.161)***	(0.157)***
	(0.121)***	(0.0701)***	(0.117)***	(0.114)***
House Density	-0.00114	0.00336	0.00321	0.00402
	(0.000677)*	(0.00267)	(0.00266)	(0.00261)
	(0.000104)***	(0.00170)*	(0.00169)*	(0.00144)**
Store Density	0.00764	0.0166	0.0172	0.0152
	(0.00192)***	(0.00959)*	(0.00990)*	(0.00971)
	(0.000194)***	(0.00764)**	(0.00756)**	(0.00695)**
Garage Density	-0.00675	0.00305	0.00297	0.00373
	(0.00149)***	(0.00568)	(0.00570)	(0.00567)
	(0.000156)***	(0.00482)	(0.00483)	(0.00486)
Avg. House Rooms	0.00360	0.0372	0.0369	0.0457
	(0.0310)	(0.0528)	(0.0527)	(0.0516)
	(0.00241)	(0.0316)	(0.0316)	(0.0287)
Avg. Store Mq	-0.000275	0.00360	0.00356	0.00334
	(0.000900)	(0.00280)	(0.00280)	(0.00280)
	(0.000238)	(0.00102)***	(0.00102)***	(0.000913)***
Num. Residents	0.0000140			5.38e-08
	(0.00000552)**			(0.000000307)
	(0.00000145)***			(0.000000143)
% Owner-occupancy	-0.473			-0.00893
	(0.178)***			(0.0185)
	(0.0211)***			(0.0112)
% 20-39 years	0.323			0.210
	(0.809)			(0.101)**
	(0.123)**			(0.0787)**
% >60 years	1.121			0.0774
	(0.487)**			(0.0647)
	(0.156)***			(0.0545)
% Graduates	2.836			0.0298
	(0.256)***			(0.0268)
	(0.0788)***			(0.0147)*
% Working	2.120			-0.0555
	(0.431)***			(0.0573)
	(0.156)***			-0.0327
% Foreigners	-0.984			0.0344
	(0.308)***			(0.0172)**
	(0.0515)***			(0.0235)
% Houses in use	0.480			0.102
	(0.275)*			(0.0411)**
	(0.0168)***			(0.0267)***
Num. Houses	-0.0000296			-0.000000439
	(0.0000119)**			(0.000000718)
	(0.00000353)***			(0.000000319)
% House in poor condition	-0.0505			0.00146
	(0.103)			(0.00442)
	(0.0148)***			(0.00489)
First Stage				
Touristiness at $t-1$			5.82e-11	5.80e-11
			(6.21e-12)***	(6.16e-12)***
			(7.68e-12)***	(7.67e-12)***
F-stat. excluded instrument			87.712	88.689
			57.370	57.167
Controls				
Interacted census controls				X
Fixed Effects				
Year FE	X			
Zone FE		X	X	X
Quarter#City FE		X	X	X
Year#City#Area FE		X	X	X
Observations	5,740	5,740	5,740	5,740
Adjusted R2	0.799	0.981		

OLS estimates of Equation (1) in Column (1); FE estimates of (2) in Column (2); 2SLS estimates of (2) in Column (3) and (4), with the latter including the interaction of the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. The dependent variable is the natural logarithm of the sale price. The instrument in Columns (3) and (4) is $a_n^{\mathrm{TA}} \times g_t$. For each coefficient, the first parenthesis shows robust standard errors clustered by zone, the second shows D-K standard errors. The same order is followed when showing the F-statistic of the excluded instrument, while the adjusted R2 is shown only for clustered standard errors. *** p<0.01, ** p<0.05, * p<0.1

In Table 7 we find that Airbnb density positively and significantly affects house prices. The OLS estimates in Column (1) show the relationships between house prices and our large set of control variables. We find that, on average, prices are higher in zones where houses are smaller and less likely to be empty, shop density is higher and parking lots are fewer, hence most probably in the city centre. Moreover, houses are more expensive where the proportion of elderly and more educated people is higher. In contrast, the average price is lower in zones where the share of foreign residents and unemployed people is higher. Finally, house prices are lower in zones where the share of owner-occupier is higher.

In Column (2), where we account for zone specific fixed effects and for year-city-area and quarter-city time effects, the size of the coefficient reduces to 0.561. Based on this estimate, an increase of one percentage point in Airbnb density leads to a 0.561% increase in price per square meter.

Column (3) reports the 2SLS regression with Tripadvisor's *touristiness* as the instrument. At the bottom of the table, the first-stage results show that the correlation between Airbnb density and the instrument is very strong. The IV coefficient is statistically significant at the 1% level and its size of 0.63 implies a sale price increase of 0.63% for a one percentage point increase of Airbnb density. In Section 6.4.3 we compare these estimates with those using the listing's creation date, as done by most of the previous literature.

Finally, in Column (4) we report the results of the IV regression adding the time invariant zone-level controls interacted with the percentage change of the population for each city. This allows us to control for the spurious correlations which survived to the inclusion of our large set of spatial-time interactions. We find that the coefficient estimated by this augmented model is 0.618, quite similar to Column (3) and comfortingly suggesting that our estimate of the impact of Airbnb does not appear to be (too) biased by unaccounted dynamic factors influencing the housing market beyond that of Airbnb.

Our model estimates an impact in terms of percentage change of sale prices and rents as a consequence of a one percentage point increase in Airbnb density. To express this in meaningful economic terms, we convert it to the change in euros of the sale and monthly rental prices per m^2. To do so, we take the change in Airbnb density registered during the time period at the zone level. We then multiply the average of these density changes with the estimated coefficient of Airbnb intensity, β. Finally, we express this impact in euro terms by multiplying it by the average monthly rent and sale price. Relying on the

Airbnb coefficient in Column (3), we find that Airbnb's growth over our analysed time period accounts for an increase of 44.24 €/m^2 in sale prices.

Table 8. Overall Effect on Rent

Dep. Var:	Log Rent			
	(1) OLS	(2) FE	(3) 2SLS	(4) 2SLS
Airbnb Density at $t-1$	0.438	0.174	0.106	0.116
	(0.362)	(0.138)	(0.128)	(0.128)
	(0.0740)***	(0.0708)**	(0.115)	(0.117)
House Density	-0.000727	-0.000971	-0.000819	-0.00222
	(0.000606)	(0.00234)	(0.00239)	(0.00220)
	(0.000121)***	(0.00100)	(0.000963)	(0.00165)
Store Density	0.00631	0.00965	0.00900	0.0117
	(0.00176)***	(0.00925)	(0.00924)	(0.00896)
	(0.000157)***	(0.00471)*	(0.00434)*	(0.00471)**
Garage Density	-0.00586	0.00157	0.00165	0.00186
	(0.00137)***	(0.00449)	(0.00448)	(0.00442)
	(0.000100)***	(0.00225)	(0.00226)	(0.00242)
Avg. House Rooms	-0.0792	-0.108	-0.108	-0.118
	(0.0212)***	(0.0708)	(0.0709)	(0.0661)*
	(0.00932)***	(0.0182)***	(0.0182)***	(0.0181)***
Avg. Store Mq	-0.00104	0.00174	0.00177	0.00172
	(0.000707)	(0.00173)	(0.00173)	(0.00161)
	(0.000106)***	(0.000812)**	(0.000854)*	(0.000843)*
Num. Residents	0.00000717			-0.000000312
	(0.00000502)			(0.000000280)
	(0.00000121)***			(0.000000170)*
% Owner-occupancy	-0.251			0.0249
	(0.148)*			(0.0187)
	(0.0132)***			(0.0177)
% 20-39 years	-1.329			0.0386
	(0.542)**			(0.0741)
	(0.139)***			(0.0343)
% >60 years	0.369			-0.125
	(0.370)			(0.0569)**
	(0.0602)***			(0.0200)***
% Graduates	1.345			0.0193
	(0.204)***			(0.0291)
	(0.0418)***			(0.0111)*
% Working	1.963			-0.0106
	(0.362)***			(0.0690)
	(0.0607)***			(0.0415)
% Foreigners	0.692			-0.0239
	(0.253)***			(0.0178)
	(0.0828)***			(0.00884)**
% Houses in use	0.399			0.0561
	(0.199)**			(0.0297)*
	(0.0534)***			(0.00854)***
Num. Houses	-0.0000200			0.000000723
	(0.0000113)*			(0.000000580)
	(0.00000308)***			(0.000000358)*
% House in poor condition	0.0121			-0.0167
	(0.0827)			(0.00556)***
	(0.0126)			(0.00643)**
First Stage				
Touristiness at $t-1$			5.82e-11	5.80e-11
			(6.21e-12)***	(6.16e-12)***
			(7.68e-12)***	(7.67e-12)***
F-stat. excluded instrument			87.712	88.689
			57.370	57.167
Controls				
Interacted census controls				X
Fixed Effects				
Year FE	X			
Zone FE		X	X	X
Quarter#City FE		X	X	X
Year#City#Area FE		X	X	X
Observations	5,740	5,740	5,740	5,740
Adjusted R2	0.699	0.962		

OLS estimates of Equation (1) in Column (1); FE estimates of (2) in Column (2); 2SLS estimates of (2) in Column (3) and (4), with the latter including the interaction of the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. The dependent variable is the natural logarithm of the rent price. The instrument in Columns (3) and (4) is $a_n^{\text{TA}} \times g_t$. For each coefficient, the first parenthesis shows robust standard errors clustered by zone, the second shows D-K standard errors. The same order is followed when showing the F-statistic of the excluded instrument, while the adjusted R2 is shown only for clustered standard errors. *** p<0.01, ** p<0.05, * p<0.1

Table 8 estimates the same models with the log of rental monthly rates as the dependent variable. The evidence is weaker, though, as compared to house sale prices. The OLS specification in Column (1) tells us that an increase of one percentage point in Airbnb density leads to an increase of 0.438% of rent per square meter. Most control variables correlate with rents similarly to sale prices. An exception is the positive coefficient on foreign residents, probably because it correlates to zones where the rental turnover is higher, thus enabling the landlord to increase the rent more often. When we move to the specification in Column (2), with zone and spatial-time fixed effects, the coefficient drops to 0.174, while the IV estimates in Columns (3) and (4) are insignificant, despite the good performance of the instrument in the first-stage regression. Using the estimates of Column (3), an increase of one percentage point in Airbnb density leads to an increase in average monthly rent per square meter of 0.106%, corresponding to a 3 €cent/m^2 rent increase over the analysed time period. The weak significance of the overall Airbnb effect in the rental market may be due to the high heterogeneity of the five cities. Moreover, the rental market in Italy is influenced by a housing policy that grants below-market rents in the social housing sector, assigns favourable tax-regimes to assisted tenancies and restricts free-market rents to long-term contracts (4 years) (Baldini and Poggio, 2012). As a result, the rental prices may be less responsive to the pressure of increasing Airbnb density. The next step of our analysis investigates the city-specific effects.

6.2. The Effect of Airbnb Density by City

In this section, we disentangle the impact of Airbnb presence on sale prices and rents by city. Results in Table 9 and Table 10 show that the impact of Airbnb on the housing market is very different across the five cities, thus suggesting that trying to estimate an overall Airbnb effect with a large pool of different cities may veil the evidence.

In Table 9, the IV results show that the effect of Airbnb density on house prices is always positive and significant. In order to appreciate the magnitude of the impact, however, coefficients must be adjusted considering the market prices in each city, which differ widely (see Table 2). For example, 2019 prices in Florence and Milan are significantly higher than in Rome, which in turn are much higher than in Naples and double those in Turin. Following the approach of Section 6.1, we compute Airbnb's effect on the average zone by using the coefficients from Column (3). Airbnb's growth over the period has led to an increase of 162.31 €/m^2 in Milan, 127.69 €/m^2 in Turin, 91.00 €/m^2 in Naples,

Table 9. Effect on Sale by City

Dep. Var:	Log Sale			
	(1) OLS	(2) FE	(3) FE	(4) 2SLS
Airbnb Density at $t-1$ in:				
Florence	-0.134	0.439	0.437	0.438
	(0.303)	(0.106)***	(0.139)***	(0.140)***
	(0.124)	(0.0642)***	(0.0803)***	(0.0808)***
Milan	2.098	2.233	2.509	2.508
	(1.414)	(1.029)**	(1.276)*	(1.275)*
	(0.767)**	(0.340)***	(0.719)***	(0.706)***
Naples	0.380	1.447	1.778	1.981
	(1.293)	(0.760)*	(0.712)**	(0.772)**
	(0.329)	(0.575)**	(0.608)***	(0.534)***
Rome	2.244	0.198	0.410	0.404
	(0.524)***	(0.149)	(0.144)***	(0.144)***
	(0.314)***	(0.101)*	(0.133)***	(0.133)***
Turin	-27.48	11.24	12.05	12.04
	(8.124)***	(2.670)***	(3.384)***	(2.967)***
	(1.999)***	(1.633)***	(2.795)***	(2.928)***
Controls				
Time invariant controls	X			
Time varying controls	X	X	X	X
Interacted census controls				X
Fixed Effects				
Year FE	X			
Zone FE		X	X	X
Quarter#City FE		X	X	X
Year#City#Area FE		X	X	X
Observations	5,740	5,740	5,740	5,740
Adjusted R2	0.83	0.98		

OLS, FE and 2SLS estimates of Equation (3) with the inclusion of progressively more fixed effects (full specification from Column (2)) and, in Column (4), of the interaction of the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. The dependent variable is the natural logarithm of the sale price. The instrument in Columns (3) and (4) is $a_n^{\text{TA}} \times g_t \times \text{city}_i$. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

72.04 €/m^2 in Florence and 19.37 €/m^2 in Rome. The effect is remarkable when compared to the sale prices variation during the time period: Airbnb's growth is accountable for 48% of the increase in sale prices in Milan, and for 22% of the increase in Florence. Interestingly, Airbnb impact on property values is highest not only in Florence, the city with the highest listing density and number of tourists per resident, but also in Milan, or even Turin, where the business-related component of visitors is relatively more important. If we were expecting to find a difference in the nature of the demand by tourist and by the business communities, our results suggest that both these demand segments count on Airbnb to satisfy their needs. This evidence confirms that the role of Airbnb has gone well beyond the accommodation of tourists in search of a "sharing" experience but has grown into a major online rental estate portal.

Table 10. Effect on Rent by City

Dep. Var:	Log Rent			
	(1) OLS	(2) FE	(3) FE	(4) 2SLS
Airbnb Density at $t-1$ in:				
Florence	-0.334	0.240	0.293	0.299
	(0.187)*	(0.172)	(0.142)**	(0.142)**
	(0.0549)***	(0.0865)**	(0.148)*	(0.146)*
Milan	4.100	-0.0131	-1.533	-1.517
	(1.220)***	(0.609)	(1.222)	(1.222)
	(0.404)***	(0.326)	(1.022)	(1.034)
Naples	1.251	1.838	1.909	1.763
	(1.206)	(0.320)***	(0.420)***	(0.436)***
	(0.133)***	(0.300)***	(0.390)***	(0.423)***
Rome	1.306	-0.218	0.121	0.131
	(0.341)***	(0.181)	(0.151)	(0.150)
	(0.153)***	(0.116)*	(0.0978)	(0.102)
Turin	-32.65	-2.466	-2.140	-1.039
	(7.210)***	(1.269)*	(0.941)**	(1.195)
	(1.971)***	(0.784)***	(1.086)*	(1.430)
Controls				
Time invariant controls	X			
Time varying controls	X	X	X	X
Interacted census controls				X
Fixed Effects				
Year FE	X			
Zone FE		X	X	X
Quarter#City FE		X	X	X
Year#City#Area FE		X	X	X
Observations	5,740	5,740	5,740	5,740
Adjusted R2	0.83	0.98		

OLS, FE and 2SLS estimates of Equation (3) with the inclusion of progressively more fixed effects (full specification from Column (2)) and, in Column (4), of the interaction of the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. The dependent variable is the natural logarithm of the rent price. The instrument in Columns (3) and (4) is $a_n^{\mathrm{TA}} \times g_t \times \mathrm{city}_i$. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Turning to the IV estimates for rents in Table 10, we find that the effect is significant and positive in Florence and Naples, insignificant in Milan and Rome, even negative in Turin. Average rents are more similar across cities than house prices, with Milan ranking first and Turin last. Based on coefficients from Column (4), the estimated impacts over the period show increases of 37 and 19 €cent/m^2 in Naples and Florence. While these values may seem low, they are significant when compared to the rental rates variation from 2015 to 2019, as Airbnb's growth is accountable for 65% and 10% of rent increases in Naples and Florence, respectively. We further disaggregate Airbnb's impact by investigating within-cities effects.

6.3. Airbnb's Effect and Neighbourhood Characteristics

The literature typically assumes that the main impact of Airbnb on the real estate market materialises through the substitution of long term with short term rentals and ultimately affects sale and rental prices (see Table 1). However, a few studies address the problem of externalities generated by Airbnb in cities where visitors' inflow and Airbnb density are particularly high (Sheppard and Udell, 2016; Filippas and Horton, 2018; Barron et al., 2020). On the one hand, should Airbnb density and visitor turnover make the neighbourhood noisy, congested and unsafe, residents may decide to leave the area, depressing property values in the long run. On the other hand, if landlords of properties in the city centre switch from long-term to short-term rental they reduce the housing supply, raising the rents and ultimately house prices (i.e., substitution effect), and forcing low-income long-term home-seekers to move from centre to suburbs. More recently, some studies have focused on the consequences of Airbnb's impact on welfare distribution among residents as well as the possibility that it may concur to a spatial dimension of inequality through the reinforcement of residential sorting (Almagro and Domínguez-Iino, 2021; Calder-Wang, 2021; Xu and Xu, 2021). These studies highlight how the substitution effect can be endogenously enhanced if the growth of touristic flows increases the supply of amenities (shops, bar, restaurants, theatres, museums) and, in turn, the attractiveness of highly touristic areas, typically the centre. This would in turn attract city residents who are willing to pay higher rents or house prices in order to live in the centre (Couture et al., 2019). The net effect of these different impacts, however, depends on many factors, such as whether or not properties in the centre are vacant, occupied by owners or by long-term tenants, whether the suburbs are attractive and safe, houses in periphery are easy to renovate or whether there is enough space to build new houses. Since these features vary across and within cities, it is not easy to predict ex-ante how the presence of Airbnb may reshape the housing market and whether it might ultimately contribute to the spatial dimension of inequality, but we can expect that neighbourhoods' characteristics and location can lead to differentiated impacts within a city. Therefore, in this section we explore these issues by dividing each city in a central area ("centre") and a peripheral one ("suburbs"), as described in Section 4.3. It is worth noting that, contrary to many US cities, in Italy and in Continental Europe, the centre is typically the area where middle-class and well-off people live, while residents in the suburbs are often poorer and marginalised. A quick look to Appendix

Table A.2 reporting the average income in the richest and poorest postal code in the five cities provides descriptive evidence on this issue.

6.3.1. Is the Effect of Airbnb Constant Across City Centre and Suburbs?

We start by providing evidence, for each city, about the different effect of Airbnb in the city centre and in the suburbs. Our purpose is to determine whether the overall Airbnb's impact at the city level is driven by its diffusion in the city centre or if it is also significant in the suburbs, and whether the sign of the impact is the same. In addition, we assess the magnitude of the impact in the two sub-city areas, calibrating the quantitative effects based on the area-specific listing density changes and housing market characteristics. We estimate Equation (4) with a FE and a 2SLS estimators that control for unobserved zone-specific and time-varying variables factors and report the results in Table 11. The specification allows us to estimate the effect of Airbnb presence in each zone in each area (centre or periphery) on the prices in that zone.

The estimated impact of Airbnb is more evident on sale prices than on rents, same as in the previous analyses. Looking at the IV estimates, we find that the effect of Airbnb density on house value is positive and significant both in the city centre and in the suburbs in Florence, Milan and Rome,[8] but only in the city centre of Naples and Turin. However, whether the magnitude of the coefficient is larger in the centre or suburbs changes on a by city basis. In the period 2014-2019, Airbnb's growth in Florence accounts for an increase of 159.30 €/m^2 in the suburbs and of 132.48 €/m^2 in the city centre, even though listing density has tripled. The evidence suggests that the growth of Airbnb's presence in the suburbs generates a value-increasing process in the area, possibly leading the property values to become more similar to the centre over time as the central Florence becomes more and more congested. In contrast, Milan shows an opposite trend: Airbnb's growth appears to account for an increase of 505.03 €/m^2 in the city centre, where the density has doubled over time, but of 262.72 €/m^2 in the suburbs, suggesting that Milan's city centre is increasing its attractiveness faster than the suburbs. The evidence is similar in Rome, with a larger price increase in the centre, where Airbnb density has doubled over time, and a smaller effect in the suburbs, where density is low and stable. Notably, in Milan and Rome the average income difference between the richest and the poorest postcode (Appendix Table A.2) is much larger than in any other city (5.3 and 4.2 times higher for Milan and Rome, respectively). Finally, in Naples and Turin, the two cities

[8] The difference between each pair of coefficients is (at least) statistically significant at the 10% level for all cities.

TABLE 11. Effects on Rent and Sale Within City

Dep. Var:	Log Sale		Log Rent	
	(1) FE	(2) 2SLS	(3) FE	(4) 2SLS
Airbnb Density at $t-1$ in:				
Florence suburbs	0.528	4.194**	-1.720	1.356
	(1.406)	(2.013)	(3.400)	(5.866)
Florence central	0.438***	0.469***	0.250	0.301*
	(0.106)	(0.149)	(0.169)	(0.172)
Milan suburbs	4.534***	8.950**	1.260	-0.120
	(1.734)	(4.185)	(1.183)	(2.448)
Milan central	1.106	2.458**	-0.636	-1.557
	(0.760)	(1.167)	(0.549)	(1.213)
Naples suburbs	-2.773**	-1.876	-0.300	-1.676
	(1.343)	(3.433)	(0.992)	(3.044)
Naples central	1.902***	1.849***	2.073***	2.027***
	(0.440)	(0.636)	(0.307)	(0.394)
Rome suburbs	3.463*	5.548*	-0.643	1.309
	(1.798)	(3.065)	(2.390)	(3.384)
Rome central	0.167	0.312**	-0.217	0.0975
	(0.148)	(0.141)	(0.181)	(0.143)
Turin suburbs	26.57***	-10.53	2.796	-6.553
	(6.503)	(26.90)	(6.207)	(10.76)
Turin central	9.163***	12.84***	-3.179***	-1.974**
	(2.782)	(2.950)	(1.020)	(0.909)
Controls				
Time varying controls	X	X	X	X
Fixed Effects				
Year FE				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Observations	5,740	5,740	5,740	5,740
Adjusted R2	0.98		0.96	

FE and 2SLS estimates of Equation 4. The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The instrument in Columns (2) and (4) is $a_n^{\mathrm{TA}} \times g_t \times \mathrm{city}_i \times \mathrm{centre}_b$. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

with the lowest income per capita and the lowest listing densities, the impact on sale prices is significantly positive only in the city centre, as the effect in the suburbs does not survive in IV estimation. To sum up, in Milan and Rome, the presence of Airbnb appears to generate a divergence between the property values in the centre and the suburbs while in Florence it seems to contribute to a convergence of such values. In Turin and Naples Airbnb seems to bring an advantage to the centre – in terms of revamping property values comparatively very low and even decreasing over time, but no benefit to the suburbs.

The analysis on rents is less informative. We find a positive effect of Airbnb density on rents in the city centre of Florence and Naples, while the effect seems negative in Turin. Although this finding could be suggestive of a negative externality of Airbnb, the usual caution for the city of Turin applies. The impact of Airbnb's growth on Florence and Naples city centres accounts for an increase of 31 €cent/m^2 and 87 €cent/m^2 respectively.

6.3.2. What Is the Effect of Airbnb's Diffusion in the City Centre on the Suburbs?

To provide further evidence on the heterogeneous impact of Airbnb, we now investigate if the increase of Airbnb's diffusion in the city centre affects the suburbs by exacerbating their disparity in terms of property values, attractiveness and living conditions or by reducing such gap. We recall that by generating an endogenous response of the centre's amenities, Airbnb can reinforce residential sorting (Almagro and Domínguez-Iino, 2021) increasing – or decreasing – the attractiveness of the area and the house prices (see also Xu and Xu, 2021). In this analysis we restrict the estimation sample to the full set of peripheral zones, and we use the aggregate listing density in the city-centre as the variable of interest to estimate its impact on the suburbs, while controlling for the effect of the zone-specific density on the zone prices. We estimate Equation (5) with FE and 2SLS, including the usual set of spatial and time-spatial interacted fixed effects, and adding the interaction of time invariant socio-demographic characteristics of the zone with the growth rate of the population, to reduce the concern about spurious correlations. With this model, we investigate whether the diffusion of Airbnb in the city centre has consequences also on the property values and rents of suburban zones, other than the effect due to the Airbnb presence in the suburban zone itself. Such spillover could happen for a variety of reasons. For example, residents in the centre could reallocate their properties to short-term rental and move to the suburbs, thus increasing the demand for housing and the house prices. Alternatively, the increase in Airbnb's presence in the centre could attract larger tourist volumes due to the augmented accommodation's capacity and, in turn, a richer provision of local amenities, shifting the attention of developers and investors from the suburbs to the city centre as a consequence of the impact this inflow could have on the business. This process would widen the gap between centre and periphery even further.

The FE and 2SLS results for sale prices and rents are shown in Table 12. In the upper panel, where we report the impact of zone-specific Airbnb densities in the suburbs, the results are, not surprisingly, very similar to the fixed effect coefficients in Table 11. In the lower panel we estimate the relationship between the aggregate density in the city centre and the sale and rental prices in the suburbs.

TABLE 12. Cross-Effect From Centre to Suburbs

Dep. Var:	Log Sale Suburbs		Log Rent Suburbs	
	(1) FE	(2) 2SLS	(3) FE	(4) 2SLS
Airbnb Density at $t{-}1$ in:				
Florence	0.616	16.04	-1.411	8.350
	(1.692)	(14.38)	(4.313)	(23.73)
Milan	5.304***	17.43***	1.883	3.617
	(1.754)	(5.709)	(1.217)	(2.804)
Naples	-3.263***	-2.482	-2.135**	-6.759
	(1.046)	(3.545)	(1.046)	(4.946)
Rome	3.873**	7.541**	-0.605	-0.708
	(1.835)	(3.383)	(2.404)	(3.555)
Turin	28.71***	-15.26	6.201	-20.17
	(6.394)	(32.94)	(6.311)	(14.93)
Airbnb Density in Centre at $t{-}1$ in:				
Florence	-4.086***	-0.214	-3.920	-34.34**
	(1.235)	(9.748)	(3.079)	(15.78)
Milan	-2.680**	-11.35***	-2.113**	-2.279
	(1.057)	(3.123)	(1.038)	(2.291)
Naples	1.894	-32.33	6.681	50.50
	(5.984)	(67.52)	(5.655)	(84.64)
Rome	-3.855***	-5.022***	-0.650	-3.523
	(0.471)	(1.724)	(0.725)	(2.869)
Turin	-17.45***	-34.06	-18.32***	-52.71***
	(6.546)	(26.21)	(4.183)	(17.44)
Controls				
Time varying controls	X	X	X	X
Interacted census controls		X		X
Fixed Effects				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Observations	3,840	3,840	3,840	3,840
Adjusted R2	0.94		0.94	

FE and 2SLS estimates of Equation 5. The sample is limited to those zones belonging to the suburbs. The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The instrument in Columns (2) and (4) are $a_n^{\mathrm{TA}} \times g_t \times \mathrm{city}_i$ and $a_b^{\mathrm{TA}} \times g_t \times \mathrm{city}_i$. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Looking at the IV estimates for sale prices in Column (2), we find in the lower panel that increases in the presence of Airbnb in the centre are significantly related to a loss in property values in the peripheries of Milan and Rome, controlling for the effect of zone-specific listing density (which is positive for Milan and Rome). This evidence confirms the interpretation we gave of the results in Table 8 of a possible negative net effect of the increasing density in the centre. At face value, our estimates imply that, controlling for the impact of peripheral zone density, an increase of one percentage point in Airbnb density in the centre leads to a decrease in the price per square meter in the suburbs of 3.86% in Rome and 2.67% in Milan.

Overall, our results suggest that, at least for the two largest cities in Italy, the housing market in the city's periphery is negatively affected by the revaluation that Airbnb brings to the centres of our five cities. We take this as the consequence of an increasing capacity of the city centre to attract tourists, investments as well as local residents keen on living in the centre following the diffusion of Airbnb and the positive spillovers on the area. This result is in line with the findings by Xu and Xu (2021) of a positive effect of Airbnb on private capital investments, and with the ongoing debate regarding the impact of overtourism in major cities. As highlighted by Calder-Wang (2021) and Almagro and Domínguez-Iino (2021), city centres are increasing their profitability as they evolve to accommodate every tourist's desires. As investments become ever more concentrated in the centre, the suburbs may pay the cost by becoming more marginalised, even though Airbnb presence has, per se, a positive effect on house prices in the suburbs. Considering together the two last sets of results, we have provided some evidence that this process may be ongoing in Milan and Rome. As to Florence, the city with the highest density of tourists and listings, we found that the presence of Airbnb seems to bring about a convergence in house prices between the centre and the periphery over time, which suggests an overall positive impact. Notably, the centre of Florence is a relatively small and closely watched jewel that, thanks to the legal protection of artistic areas and buildings, does not allow real estate developments. Hence, what our results suggest is that positive spillovers can occur in the suburbs, possibly, through renovation investments.

Overall, it appears that the impact of Airbnb's presence on the metropolitan housing markets is not neutral. Airbnb almost certainly benefits the centre, while leaving the peripheries behind by making them relatively less attractive, at least is some cases. Our analysis has shown the importance of estimating the effect by city, and within cities, without forgoing the cross-effects that may exists between centre and suburbs. Moreover, we have also shown that the dynamics of the response of the housing market to the presence of Airbnb deeply depends on the initial conditions in the local economy, such as the present stance – depressed or booming – of the real estate prices and the average income differences between the centre and the suburbs.

6.4. Robustness

In the previous sections, we have shown that Airbnb's diffusion has a positive and significant effect on the real estate market's prices in Italy, although with some differences between the impact on sale prices and rents and a lot of heterogeneity across towns. To further corroborate this evidence, in this section we challenge our main results through a battery of robustness tests.

6.4.1. Allowing for Dynamic Effects in the Housing Market: A Dynamic Panel Model

A dynamic approach is complementary to the scope of this work: housing units (and, in turn, rental prices) are often evaluated by comparison with similar dwellings, and their prices are then adjusted to account for differences. As such, aside from exogenous shocks, it is reasonable to assume that a house valuation strongly depends on its previous values, making sale and rental prices persistent over time. This approach has been recently used in literature, to analyse the effect of Airbnb's expansion on London's house prices (Benítez-Aurioles and Tussyadiah, 2021). In this section we proceed by estimating a dynamic version of the model that studies the relationship between sale or rental prices and Airbnb density.

The model has the following form:

$$log(Y_{n,t}) = \alpha log(Y_{n,t-1}) + \beta \text{Airbnb Intensity}_{n,t-1} + \gamma X_{n,t} + \tau_t + \mu_n + \varepsilon_{n,t} \qquad (6)$$

This specification includes the lagged dependent variable $Y_{n,t-1}$ to account for its persistence over time. To account for the *dynamic panel bias* that arises from the correlation between the lagged dependent variable and the fixed effect in the error term (Nickell, 1981), we adopt the GMM-SYS approach (Arellano and Bond, 1991; Blundell and Bond, 1998). We use the Blundell-Bond estimator with the Windmeijer's finite sample correction (Windmeijer, 2005), dealing with situations where the lagged dependent variable is persistent (i.e., the autoregressive parameter is large). This model estimates a system of first-differenced and level equations and uses lags of variables in levels as instruments for equations in first-differences and lags of first-differenced variables as instruments for equations in levels, in which the instruments must be orthogonal to the firm-specific effects. For the validity of the GMM estimates, it is crucial that the instruments are exogenous, so we report the appropriate tests: the Arellano and Bond (1991) autocorrelation tests to control for first-order and second-order correlation in the residuals, and the two-step Sargan-Hansen statistic to test the joint validity of the instruments. Standard errors are robust to heteroskedasticity and arbitrary patterns of autocorrelations within firms.

Both the lagged dependent variable and the variable of interest (i.e., Airbnb's density) are treated as endogenous and are instrumented with the GMM approach; to keep the number of instruments under control, we constrain the moment conditions regarding time intervals, depending on the individual specifications, reporting the ratio between instruments and groups at the bottom of the table. We include both the temporal effects and the instrumental variables for touristiness (based on Google searches and Tripadvisor's tourist attraction score) as external instruments in the estimation. The temporal effects control for seasonality at the city level as well as for year effects at the area level (i.e., centre and suburbs) in each city.

In Appendix Table A.3, we report the one-step GMM-SYS estimates of the dynamic specification and the (inconsistent) fixed effects results for comparison. The autocorrelation tests for second-order correlation in the residuals and the two-step Sargan-Hansen statistic suggest that our estimates are valid (although in the rent equation we can reject the null that instruments are invalid at the 5%, but not at the 10% level). The ratio between instruments and groups is well below one.

Comfortingly, the GMM-SYS estimates show that also when we account for dynamic effects and apply a different estimator, Airbnb intensity affects positively and significantly both rental rates and sale prices, consistent with our previous evidence. Moreover, the coefficient of the lagged dependent variable confirms that prices in the housing market are quite persistent, particularly sale prices.

6.4.2. Alternative Measure of Tourist Attraction in the Instrumental Variable Analysis

We test the robustness of our instrument by constructing another shift-share IV against which to compare our previous estimates. Specifically, we define a second measure of the share component (i.e., tourist attractiveness) based on the rankings by Lonely Planet guidebooks and websites, which orders by popularity the top 10 attractions of each city (refer to Section 5.1.1). The shift component remains the same, based on worldwide Google searches of the word Airbnb.

The rationale behind this alternative instrument is similar to the one based on Tripadvisor but differs in two respects. On the one hand the Lonely Planet instrument, largely based on time invariant touristic-artistic-archaeological and geographical sites, may be less precise and responsive in the identification of the most popular locations. On the other hand, compared to the Tripadvisor rating system, the tighter but steadier classification of Lonely Planet attractions is less sensitive to tourist trends and fads, hence ultimately less influenced by Airbnb diffusion (a feature that motivated us to scrape the data for the Ttripadvisor share component at a date earlier than our estimation period).

Appendix Table A.4 and A.5 report the 2SLS estimates overall and by city, in Columns (2) and (4) using the new IV, while showing in Columns (1) and (3) the corresponding 2SLS results of the main analysis, for ease of comparison. Looking at Appendix Table A.4, we find that the coefficient is significant for sale prices, but not for rents, similar to previous evidence. The first-stage results show a strong correlation between Airbnb density and the instrument based on Lonely Planet. Turning to the analysis by city, we note that the results for sale prices are very similar to those obtained when we use the Tripadvisor instrument. As for rents, the only difference is that the coefficient of Milan has turned significant while the estimate for Turin is no longer significant.

6.4.3. Alternative Measures of Airbnb Supply: Listings by Creation Date

The literature has highlighted the difficulty of precisely measuring Airbnb activity either through scraping Airbnb's website or using publicly available databases. Since most of the previous studies used datasets that captured listings' activity (i.e., whether they are blocked, available for rent, or reserved) through occasional scrapes of the platform's website, quantifying the number of active listings per time period required some degree of approximation.[9] To estimate supply, one can ultimately apply one of the following strategies. First, one can assume that the listing's entry date is the date of its first review and that, thereafter, the listing never exited the market, which leads to overestimating Airbnb supply. A slightly different strategy defines active any listing that has received at least a review during the quarter, a method that underestimates Airbnb supply as listings can be active even when they do not receive a review for a period. However, Airbnb pushes guests as well as hosts to leave reviews about their stay, hence this approach leads to a smaller bias than in the case where no exit from the market is assumed. Other studies use the host's registration date as a proxy for the listing's entry, which also overestimates supply since a host may open additional listings after the first one, but all subsequent listings would be erroneously backdated to the time of the earliest apartment.

In this paper, to the contrary, we leverage on the fine-grained detail of the AirDNA database, based on daily scrapes, an information that allows us to pinpoint the activity of each listing on a daily basis. In the following robustness test, we estimate the impact of

[9] See Sheppard and Udell (2016), Horn and Merante (2017), Barron et al. (2020) and Garcia-López et al. (2020). Ayouba et al. (2020) is the only work we are aware of that, like us, uses daily scrapes to measure Airbnb supply.

Airbnb intensity by using the approach based on the listing's creation date that assumes no exit and we compare the results with our previous findings in Table 7, 8, 9 and 10. The results are in Columns (2) and (4) of Appendix Table A.6 and A.7. We find that using listings' creation date to measure Airbnb supply leads to lower estimates of the impact, with coefficients being about half as large as in Columns (1) and (3), thus confirming that measuring Airbnb intensity with the creation date approach leads to underestimating its impact on the real estate market.[10]

6.4.4. Alternative Measures of Airbnb Intensity: Number of Listings

The impact of Airbnb's diffusion can also be tested by using the number of listings to measure Airbnb intensity, instead of listing density (e.g., Barron et al., 2020; Garcia-López et al., 2020, among the others). The rationale of our preference for listing density is that it allows us to account for size differences of the various zones, which are highly heterogeneous in our dataset. However, for completeness, we re-estimate our models using the number of listings as our variable of interest. The results in Appendix Table A.8 and A.9 show the effect of an increase of 100 listings in a given zone on sale prices and rent. Overall, we find that the direction of the effect is similar to previous evidence using listing density, but the coefficients are less precisely estimated, both overall and by-city.

Looking at the 2SLS results in Column (4) in Appendix Table A.8, our estimates imply that an increase in 100 Airbnb listings in a neighbourhood translates to an increase in sale price of 0.6%. The impact of Airbnb on rent and sale prices across cities computed using the number of listings instead of their density gives comparable results to the one presented in Section 6.1: we find an increase in sale prices of 40.67 €/m^2 and in rent prices of 2.7 €cent/m^2 during the analysed time period. Appendix Table A.9 shows the effect on sale prices and rent by city.[11] In cities where Airbnb's impact is significant, we again find consistent results with the analysis shown in Section 6.2. Regarding sale prices, Airbnb's growth has led to an increase of 83.58 €/m^2 and 49.60 €/m^2 in Turin and Florence respectively. Regarding rent prices, Airbnb increased them in Naples and

[10] The difference is statistically significant at the 1% level for the overall coefficients, at least at the 5% level for significant coefficients of the analysis by city except for Milan and Naples (10% level).

[11] As a further robustness test, we have also estimated the impact of Airbnb intensity by focusing on the supply of "professional listings", i.e., entire apartments that are either reserved for at least 90 days per year or belonging to a multi-host. The results are similar, confirming the previous evidence, with slightly larger coefficients indicating that in zones where the density of professional listings is higher the impact on the real estate market is also more pronounced.

Florence by 62.7 €cent/m^2 and 18.6 €cent/m^2 in Naples and Florence respectively, while it decreased them of 8.1 €cent/m^2 in Turin.

7. Conclusions

The diffusion of home-sharing platforms has recently sparked interest on their potential distributional impact on the participants in the housing market. In this paper, we have studied how Airbnb' growth has affected house prices and rents in five important cities which aptly represent the heterogeneity of the Italian housing market: Florence, Milan, Naples, Rome, and Turin. After quantifying the overall and city-specific effects of Airbnb diffusion, we have estimated whether the impact is constant across city core and periphery and, finally, what is the effect of Airbnb listing density growth in the centre on the periphery. To address endogeneity concerns we applied an instrumental variable approach which interacts an out-of-sample measure of tourist attraction that varies within cities (derived from Tripadvisor), and a measure of public awareness of Airbnb that varies over time (derived from Google searches). We accounted for possible identification threats deriving from the high correlation between tourist attraction and centrality through the inclusion of centre- and suburbs-specific year-level fixed effects and zone-level time-varying controls associated with urban revival processes.

Our findings suggest that Airbnb diffusion has caused an increase of rents and, especially, of house prices. Our overall results indicate that an increase of 1 percentage point in Airbnb density leads to an average 0.63% rise in sale prices. Over the period of the analysis, this translates to an increase of 44.24 €/m^2. We find that this impact strongly differs across cities. Looking at city specific effects, the increase ranges from the 162.31 €/m^2 of Milan to the 19.37 €/m^2 of Rome. The impact on rents is significant for the cities of Florence and Naples, with an estimated increase of 19 and 37 €cent/m^2 over the time period.

Interestingly, when looking within cities at the different impact on central and suburban areas, we find cases where not only the centre, but the suburbs too are driving these results. Centre and suburbs differences are again key when assessing the magnitude of the coefficients: the increases in Milan and Rome's city centres are much higher than those in the suburbs, whereas in Florence the price increase is higher in the periphery.

Finally, we find evidence of spillover effects: the increase in listing density in central areas has a negative impact on property values in the suburbs. The evidence is stronger for Milan and Rome, which report the largest gap between neighbourhoods with the highest and lowest average income. Overall, this suggests that the presence of Airbnb in the centre and the related boost to the supply of localised amenities may reinforce its attractiveness at the expense of the peripheries. On the opposite side, in Florence Airbnb presence has a stronger positive effect on the suburbs than on the city centre's prices and is seemingly leading prices to converge.

Our results speak of an overarching effect, but also of differentiated impacts which require context-specific policies. Indeed, the consequences of these impacts on prices need to be evaluated on a case-by-case basis, as they have the potential of being either positive for residents – for example, if they help invert a trend toward strong property devaluation – or negative, as the debate on overtourism and gentrification suggests. Our analysis can inform the debate on the regulation of the platform on both a national and municipal level and help understand whether Airbnb's diffusion is benefiting some parts of the city while leaving other neighbourhoods behind.

Acknowledgement: This work has been partially supported by Ministero dell'Istruzione, dell'Università e della Ricerca Award TESUN-83486178370409 finanziamento dipartimenti di eccellenza CAP. 1694 TIT. 232 ART. 6.

Bibliography

AirDNA. (2021). AirDNA — Short-Term Rental Analytics — Vrbo & Airbnb Data. Retrieved March 9, 2021, from https://www.airdna.co/

Almagro, M., and Domínguez-Iino, T. (2021). Location Sorting and Endogenous Amenities: Evidence from Amsterdam. *Working Paper*. Retrieved February 8, 2022, from https://m-almagro.github.io/Location_Sorting.pdf

Arellano, M., and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, *58*(2), 277–297. https://doi.org/10.2307/2297968

Ayouba, K., Breuillé, M.-L., Grivault, C., and Le Gallo, J. (2020). Does Airbnb Disrupt the Private Rental Market? An Empirical Analysis for French Cities. *International Regional Science Review*, *43*(1-2), 76–104. https://doi.org/10.1177/0160017618821428

Baldini, M., and Poggio, T. (2012). Housing Policy Towards the Rental Sector in Italy: A Distributive Assessment. *Housing Studies*, *27*(5), 563–581. https://doi.org/10.1080/02673037.2012.697549

Barron, K., Kung, E., and Proserpio, D. (2020). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science*, *40*(1), 23–47. https://doi.org/10.1287/mksc.2020.1227

Bartik, T. (1991). *Who Benefits from State and Local Economic Development Policies?* (Books from Upjohn Press). W.E. Upjohn Institute for Employment Research. Kalamazoo, Michigan (USA).

Benítez-Aurioles, B., and Tussyadiah, I. (2021). What Airbnb does to the housing market. *Annals of Tourism Research*, *90*(100), 103108. https://doi.org/10.1016/j.annals.2020.103108

Blundell, R., and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115–143.

Borusyak, K., Hull, P., and Jaravel, X. (2018, September). *Quasi-Experimental Shift-Share Research Designs* (Working Paper No. 24997). National Bureau of Economic Research. https://doi.org/10.3386/w24997

Calder-Wang, S. (2021). *The Distributional Impact of the Sharing Economy on the Housing Market* (SSRN Scholarly Paper No. ID 3908062). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3908062

Christian, P., and Barrett, C. B. (2017, August 23). *Revisiting the Effect of Food Aid on Conflict: A Methodological Caution* (Policy Research Working Papers WPS8171). The World Bank. Washington, DC. https://doi.org/10.1596/1813-9450-8171

Coles, P., Egesdal, M., Ellen, I. G., Li, X., and Sundararajan, A. (2018). Airbnb Usage across New York City Neighborhoods: Geographic Patterns and Regulatory Implications. In *The Cambridge Handbook of the Law of the Sharing Economy* (pp. 108–128). Cambridge University Press. https://doi.org/10.1017/9781108255882.009

Couture, V., Gaubert, C., Handbury, J., and Hurst, E. (2019). *Income Growth and the Distributional Effects of Urban Spatial Sorting* (Working Paper No. 26142). National Bureau of Economic Research. https://doi.org/10.3386/w26142

Driscoll, J. C., and Kraay, A. C. (1998). Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *The Review of Economics and Statistics*, *80*(4), 549–560. https://doi.org/10.1162/003465398557825

Duso, T., Michelsen, C., Schäfer, M., and Tran, K. D. (2020, August 1). *Airbnb and Rents: Evidence from Berlin* (DIW Berlin Discussion Paper No. 1890). DIW. Berlin, Germany. https://doi.org/10.2139/ssrn.3676909

Einav, L., Farronato, C., and Levin, J. (2016). Peer-to-Peer Markets. *Annual Review of Economics*, *8*(1), 615–635. https://doi.org/10.1146/annurev-economics-080315-015334

Farronato, C., and Fradkin, A. (2018, February). *The Welfare Effects of Peer Entry in the Accommodation Market: The Case of Airbnb* (w24361). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w24361

Filippas, A., and Horton, J. J. (2018). The Tragedy of Your Upstairs Neighbors: Externalities of Home-Sharing. *Working Paper*.

Franco, S. F., and Santos, C. D. (2021). The impact of Airbnb on residential property values and rents: Evidence from Portugal. *Regional Science and Urban Economics*, *88*, 103667. https://doi.org/10.1016/j.regsciurbeco.2021.103667

Garcia-López, M.-À., Jofre-Monseny, J., Martínez-Mazza, R., and Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, *119*, 103278. https://doi.org/10.1016/j.jue.2020.103278

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik Instruments: What, When, Why, and How. *American Economic Review*, *110*(8), 2586–2624. https://doi.org/10.1257/aer.20181047

Gyourko, J., and Molloy, R. (2015, January 1). Chapter 19 - Regulation and Housing Supply. In G. Duranton, J. V. Henderson and W. C. Strange (Eds.), *Handbook of Regional and Urban Economics* (pp. 1289–1337). Elsevier. https://doi.org/10.1016/B978-0-444-59531-7.00019-3

Horn, K., and Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. *Journal of Housing Economics*, *38*, 14–24. https://doi.org/10.1016/j.jhe.2017.08.002

Idealista. (2021). Idealista — Case e appartamenti, affitto e vendita, annunci gratuiti. Retrieved March 9, 2021, from https://www.idealista.it/

ISTAT. (2016). Movimento turistico in Italia - Anno 2015. Retrieved December 22, 2021, from https://www.istat.it/it/files/2016/11/Movimento-turistico_Anno-2015.pdf

ISTAT. (2020). Prospetti Turismo 2019-2020. Retrieved December 22, 2021, from https://www.istat.it/it/files/2020/12/Prospetti-Turismo_2019_2020.xlsx

ISTAT. (2021). Istat.it - 15° censimento della popolazione e delle abitazioni 2011. Retrieved March 9, 2021, from https://www.istat.it/it/censimenti-permanenti/censimenti-precedenti/popolazione-e-abitazioni/popolazione-2011

Koster, H. R. A., van Ommeren, J., and Volkhausen, N. (2021). Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. *Journal of Urban Economics*, *124*, 103356. https://doi.org/10.1016/j.jue.2021.103356

MEF. (2021). Analisi statistiche - Open Data Dichiarazioni. Retrieved February 18, 2022, from https://www.finanze.gov.it/opencms/header.html

Nickell, S. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica*, *49*(6), 1417–1426. https://doi.org/10.2307/1911408

Poterba, J. M. (1984). Tax Subsidies to Owner-Occupied Housing: An Asset-Market Approach*. *The Quarterly Journal of Economics*, *99*(4), 729–752. https://doi.org/10.2307/1883123

Sheppard, S., and Udell, A. (2016). Do Airbnb properties affect house prices? *Williams College Department of Economics Working Papers*, *3*.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, *126*(1), 25–51. https://doi.org/10.1016/j.jeconom.2004.02.005

Xu, M., and Xu, Y. (2021). What happens when Airbnb comes to the neighborhood: The impact of home-sharing on neighborhood investment. *Regional Science and Urban Economics*, *88*, 103670. https://doi.org/10.1016/j.regsciurbeco.2021.103670

Zervas, G., Proserpio, D., and Byers, J. W. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of Marketing Research*, *54*(5), 687–705. https://doi.org/10.1509/jmr.15.0204

Appendix

I. Maps of the Five Cities

Each map shows a city divided in the Idealista zones. The zones' colours reflect the level of Airbnb density as of 2019, divided in quintiles. The size of the white circles refers to the level of touristiness of each zone, computed as in Section 5.1.1. Note that the circles' sizes are relative to each city, and as such are not comparable across cities.

FIGURE A.1. Florence

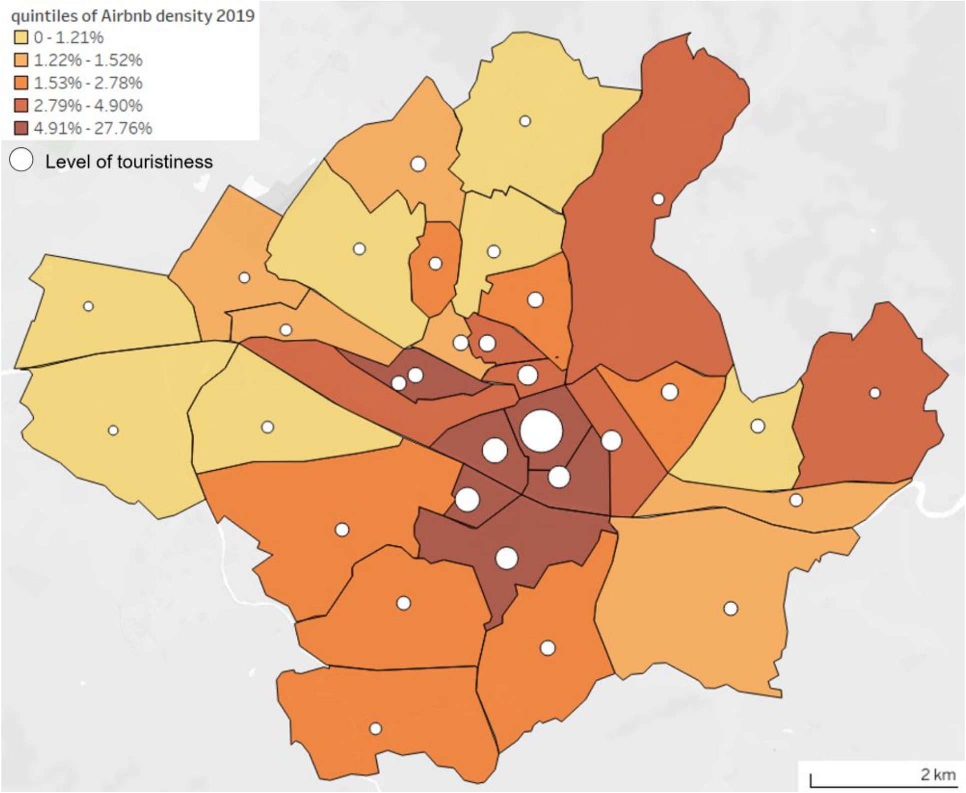


Figure A.2. Milan

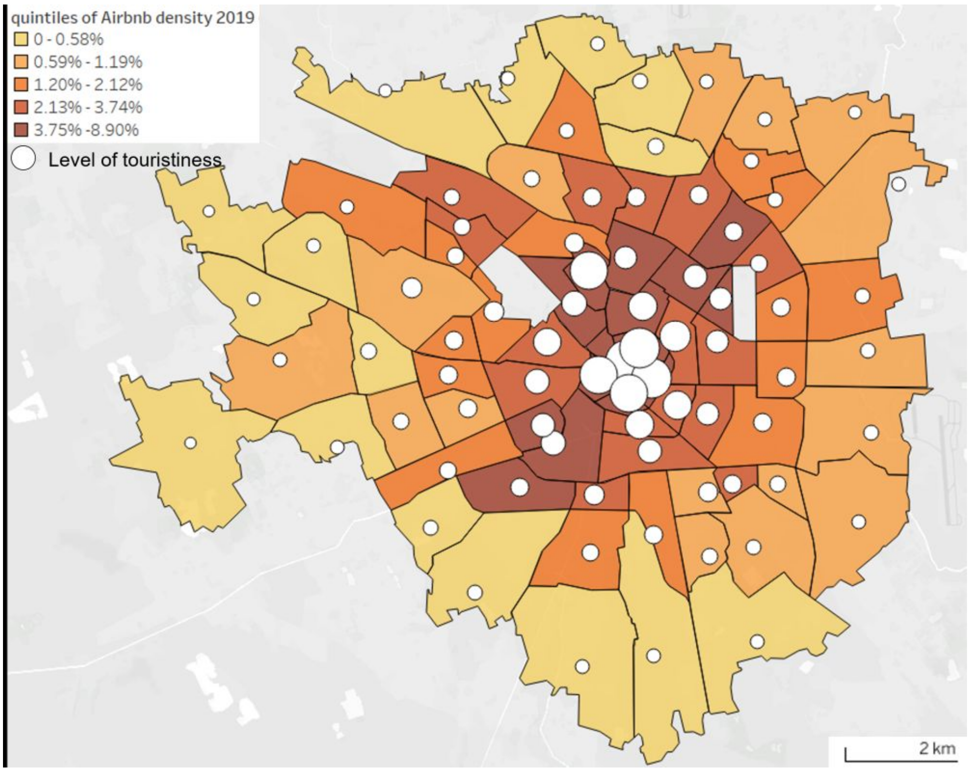


Figure A.3. Naples

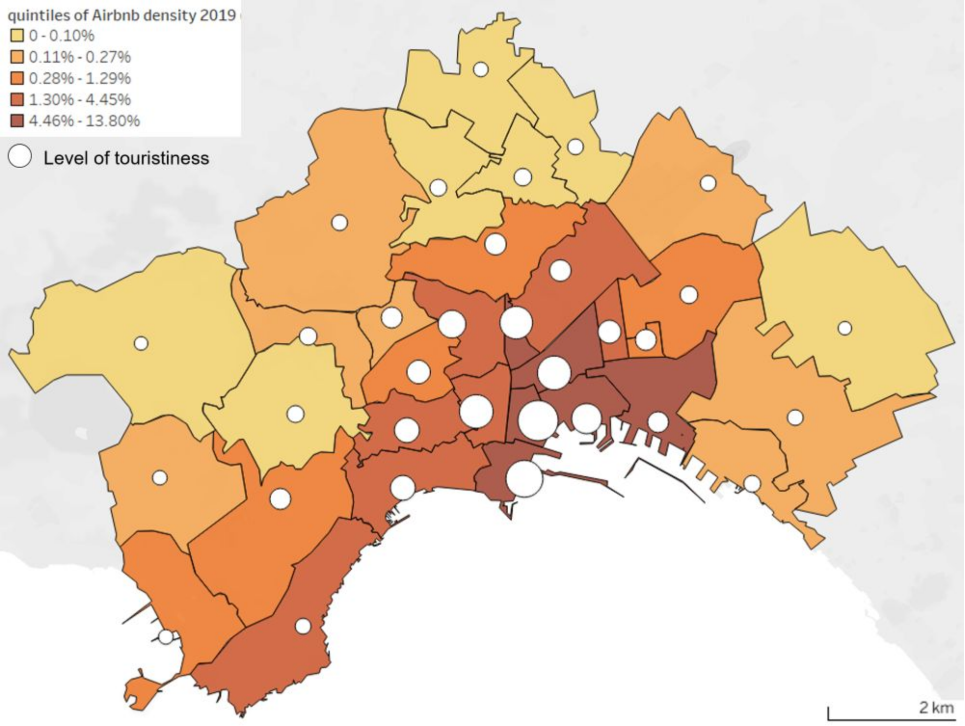


Figure A.4. Rome

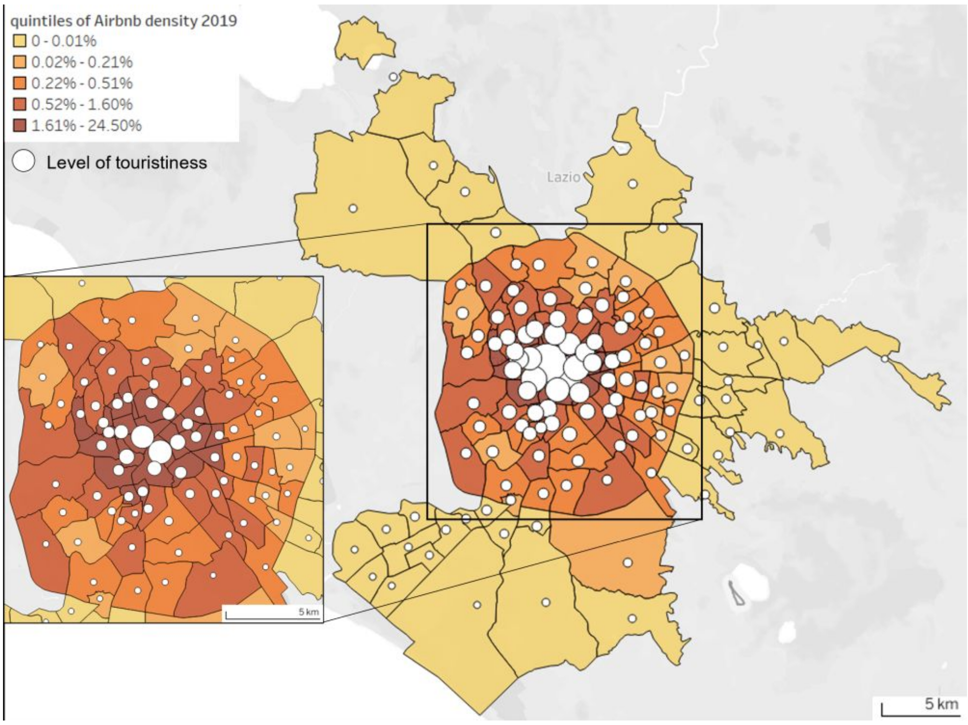
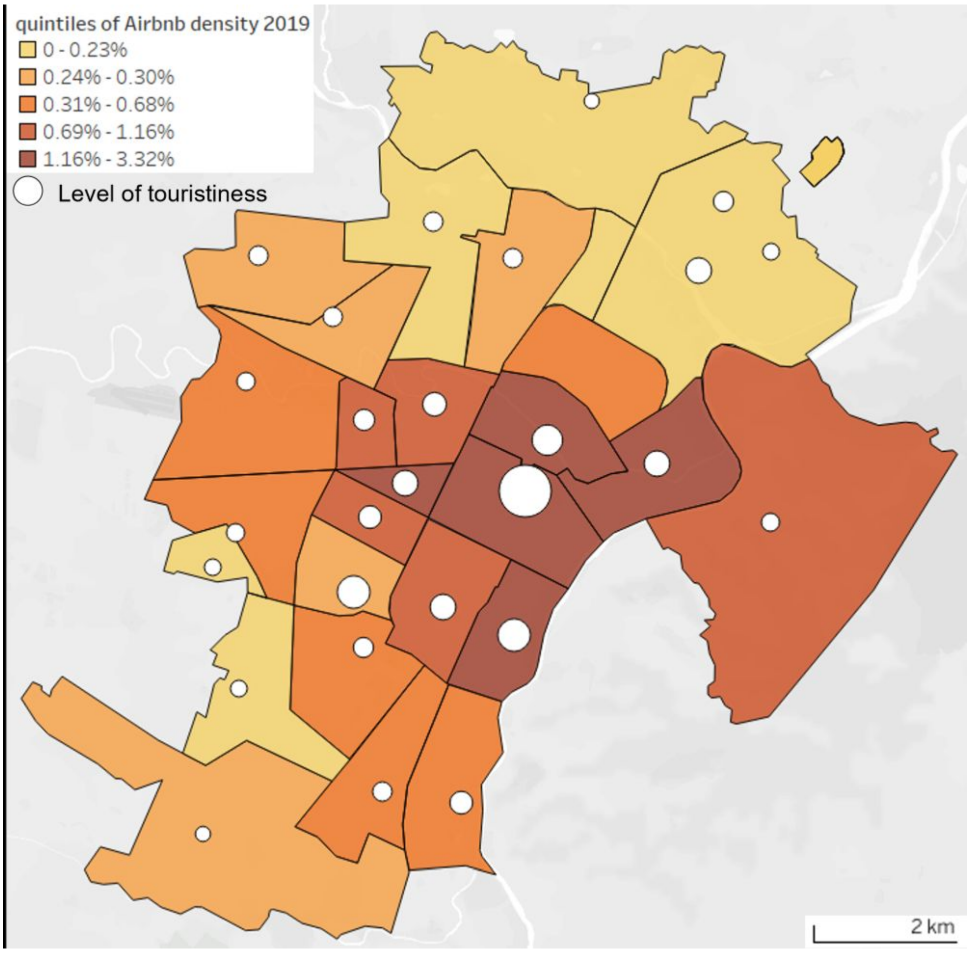


Figure A.5. Turin



II. Additional Descriptive Tables

Table A.1. Sociodemographic and Economic Time-Invariant Characteristics Within City

	Suburbs				
	% Owner Occupancy	% >60 years	% Graduates	% Working	Household Size
Average	64.83%	22.43%	14.56%	40.03%	2.24
Florence	69.14%	26.26%	14.64%	42.04%	2.16
Milan	63.31%	25.44%	16.43%	43.13%	2.03
Naples	46.00%	15.45%	6.14%	23.02%	3.00
Rome	68.56%	20.57%	15.88%	41.45%	2.29
Turin	67.23%	25.95%	10.84%	39.00%	2.13
	Centre				
	% Owner Occupancy	% >60 years	% Graduates	% Working	Household Size
Average	64.43%	24.16%	26.69%	41.87%	2.06
Florence	68.21%	26.30%	23.72%	43.70%	2.02
Milan	62.91%	23.59%	34.32%	47.54%	1.95
Naples	58.52%	20.33%	17.72%	30.40%	2.52
Rome	67.32%	25.60%	28.66%	41.81%	1.99
Turin	61.03%	22.90%	21.49%	43.01%	1.98

This table shows – at the city centre and suburbs level for each city – average values for owner occupancy, share of residents older than sixty, share of graduates, share of employed, average household size. All data come from the 2011 Census. Source: ISTAT.

Table A.2. Revenue Disparity by City

	Low Revenue Postal Code	High Revenue Postal Code	Ratio High to Low
Florence	20,523	40,527	1.97
Milan	18,926	100,489	5.31
Naples	13,462	47,316	3.51
Rome	16,298	68,264	4.19
Turin	18,158	64,094	3.53

This table shows average revenues by city in low and high revenue postal codes, while reporting they ratio. Source: MEF.

III. Robustness Regression Tables

Table A.3. GMM-SYS - Lagged Density in GMM (Both in Levels and in Difference)

Dep. Var:	Log Sale		Log Rent	
	(1) FE	(2) GMM-SYS	(3) FE	(4) GMM-SYS
Dep. Var. at $t{-}1$	0.651***	0.890***	0.439***	0.496***
	(0.030)	(0.013)	(0.022)	(0.051)
Airbnb Density at $t{-}1$	0.098	0.100**	0.173*	0.444***
	(0.072)	(0.044)	(0.089)	(0.131)
House Density	0.0004	-0.0002***	-0.0003	-0.0008***
	(0.0009)	(0.00008)	(0.0014)	(0.0002)
Store Density	0.0049	0.0004**	0.0021	0.0018***
	(0.0042)	(0.0001)	(0.0049)	(0.0005)
Garage Density	-0.0016	0.0003*	-0.0001	0.0003
	(0.0029)	(0.0002)	(0.0041)	(0.0004)
Avg. House Rooms	-0.0084	-0.0052	-0.0001	-0.0339***
	(0.0219)	(0.0036)	(0.0501)	(0.0094)
Avg. Store Mq	0.0022**	0.0002**	0.0015	0.0002
	(0.0010)	(0.0001)	(0.0011)	(0.0003)
Fixed Effects				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Num Instruments		261		135
Num Instruments/p		0.91		0.47
AR(1)		-9.38***		-9.99***
AR(2)		0.95		-0.21
Hansen test (P-value)		0.124		0.085
Observations	4879	4879	4879	4879

Estimates of Equation (6). One-step GMM-SYS estimates (Columns (2) and (4)) and inconsistent fixed effects results (Columns (1) and (3)) for comparison. The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The variable of interest is the lagged Airbnb density, both in levels and in differences. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table A.4. IV Lonely Planet - Overall Effect

Dep. Var:	Log Sale		Log Rent	
	(1) TA	(2) LP	(3) TA	(4) LP
Airbnb Density at $t-1$	0.630***	1.154***	0.106	0.290
	(0.161)	(0.310)	(0.128)	(0.247)
First Stage				
Touristiness at $t-1$	5.82e-11***	0.000000769***	5.82e-11***	0.000000769***
	(6.21e-12)	(0.000000138)	(6.21e-12)	(0.000000138)
F-stat. excluded instrument	87.712	30.960	87.712	30.960
Controls				
Time varying controls	X	X	X	X
Fixed Effects				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Observations	5,740	5,740	5,740	5,740

2SLS estimates of Equation (2). The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The instrument in Columns (2) and (4) is $a_n^{\text{LP}} \times g_t$, while Columns (1) and (3) report the results from Column (4) of Table 7 and 8 – where the instrument is $a_n^{\text{TA}} \times g_t$ – for ease of comparison. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table A.5. IV Lonely Planet - by City Effect

Dep. Var:	Log Sale		Log Rent	
	(1) TA	(2) LP	(3) TA	(4) LP
Airbnb Density at $t-1$ in:				
Florence	0.437***	0.475***	0.293**	0.313**
	(0.139)	(0.134)	(0.142)	(0.157)
Milan	2.509*	3.477***	-1.533	-2.268**
	(1.276)	(1.194)	(1.222)	(1.103)
Naples	1.778**	1.772***	1.909***	2.472***
	(0.712)	(0.564)	(0.420)	(0.431)
Rome	0.410***	0.463***	0.121	0.198
	(0.144)	(0.149)	(0.151)	(0.164)
Turin	12.05***	9.036*	-2.140**	-1.684
	(3.384)	(5.436)	(0.941)	(1.133)
Controls				
Time varying controls	X	X	X	X
Fixed Effects				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Observations	5,740	5,740	5,740	5,740

2SLS estimates of Equation (3). The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The instrument in Columns (2) and (4) is $a_n^{\text{LP}} \times g_t \times \text{city}_i$, while Columns (1) and (3) report the results from Column (4) of Table 9 and 10 – where the instrument is $a_n^{\text{TA}} \times g_t \times \text{city}_i$ – for ease of comparison. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table A.6. Creation Date - Overall Effect

Dep. Var:	Log Sale		Log Rent	
	(1) Baseline	(2) Creation Date	(3) Baseline	(4) Creation Date
Airbnb Density at $t-1$	0.630***	0.294***	0.106	0.0496
	(0.161)	(0.0745)	(0.128)	(0.0599)
Controls				
Time varying controls	X	X	X	X
Fixed Effects				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Observations	5,740	5,740	5,740	5,740

2SLS estimates of Equation (2). The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The instrument is $a_n^{\mathrm{LP}} \times g_t$. Columns (1) and (3) report the results from Column (4) of Table 7 and 8 for ease of comparison. In Columns (2) and (4), the measure of Airbnb density in the first stage is obtained from listings' creation date. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table A.7. Creation Date - by City Effect

Dep. Var:	Log Sale		Log Rent	
	(1) Baseline	(2) Creation Date	(3) Baseline	(4) Creation Date
Airbnb Density at $t-1$ in:				
Florence	0.437***	0.204***	0.293**	0.139**
	(0.139)	(0.0652)	(0.142)	(0.0666)
Milan	2.509*	1.293**	-1.533	-0.785
	(1.276)	(0.650)	(1.222)	(0.633)
Naples	1.778**	1.005**	1.909***	1.083***
	(0.712)	(0.396)	(0.420)	(0.244)
Rome	0.410***	0.186***	0.121	0.0575
	(0.144)	(0.0673)	(0.151)	(0.0697)
Turin	12.05***	4.635***	-2.140**	-0.839**
	(3.384)	(1.326)	(0.941)	(0.367)
Controls				
Time varying controls	X	X	X	X
Fixed Effects				
Zone FE	X	X	X	X
Quarter#City FE	X	X	X	X
Year#City#Area FE	X	X	X	X
Observations	5,740	5,740	5,740	5,740

2SLS estimates of Equation (3). The dependent variable is the natural logarithm of sale (Columns (1) and (2)) and rent prices (Columns (3) and (4)). The instrument is $a_n^{\mathrm{LP}} \times g_t \times \mathrm{city}_i$. Columns (1) and (3) report the results from Column (4) of Table 9 and 10 for ease of comparison. In Columns (2) and (4), the measure of Airbnb density in the first stage is obtained from listings' creation date. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

TABLE A.8. Listings - Overall Effect

Dep. Var:	Log Sale				Log Rent			
	(1) OLS	(2) FE	(3) 2SLS	(4) 2SLS	(5) OLS	(6) FE	(7) 2SLS	(8) 2SLS
Listings/100 at $t{-}1$	0.0151***	0.00417**	0.00603***	0.00592***	0.0111***	0.00117	0.00102	0.00111
	(0.00325)	(0.00164)	(0.00215)	(0.00213)	(0.00278)	(0.00130)	(0.00130)	(0.00130)
Controls								
Time invariant controls	X				X			
Time varying controls	X	X	X	X	X	X	X	X
Interacted census controls				X				X
Fixed Effects								
Year FE	X				X			
Zone FE		X	X	X		X	X	X
Quarter#City FE		X	X	X		X	X	X
Year#City#Area FE		X	X	X		X	X	X
Observations	5,740	5,740	5,740	5,740	5,740	5,740	5,740	5,740
Adjusted R2	0.80	0.98			0.71	0.96		

OLS estimates of Equation (1) in Columns (1) and (5); FE estimates of (2) in Columns 2 and 6; 2SLS estimates of 2 in Columns (3), (4), (7) and (8). Columns (4) and (8) include the interaction of the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. The dependent variable is the natural logarithm of sale (Columns (1) to (4)) and rent prices (Columns (5) to (8)). The variable of interest is the lagged number of listings/100. The instrument in Columns (3), (4), (7) and (8) is $a_n^{\mathrm{TA}} \times g_t$. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table A.9. Listings - by City Effect

Dep. Var:	Log Sale				Log Rent			
	(1) OLS	(2) FE	(3) 2SLS	(4) 2SLS	(5) OLS	(6) FE	(7) 2SLS	(8) 2SLS
Listings/100 at $t-1$ in:								
Florence	-0.00289	0.00640***	0.00470*	0.00470*	-0.00387	0.00304	0.00459**	0.00464**
	(0.00431)	(0.00201)	(0.00264)	(0.00265)	(0.00266)	(0.00247)	(0.00219)	(0.00218)
Milan	0.0337***	0.0246**	0.206	0.206	0.0493***	0.0144**	-0.129	-0.127
	(0.0105)	(0.00967)	(0.217)	(0.216)	(0.0119)	(0.00606)	(0.187)	(0.185)
Naples	0.00475	0.00882	0.0138	0.0163	0.0178**	0.0156***	0.0268**	0.0255**
	(0.00975)	(0.00660)	(0.0131)	(0.0142)	(0.00730)	(0.00279)	(0.0108)	(0.0110)
Rome	0.0179***	0.000331	0.000949	0.000898	0.0126***	-0.00200	0.00214	0.00217
	(0.00328)	(0.000856)	(0.00299)	(0.00300)	(0.00246)	(0.00128)	(0.00233)	(0.00231)
Turin	-0.0700**	0.0314***	0.0379***	0.0381***	-0.0894***	-0.00624*	-0.00798*	-0.00349
	(0.0312)	(0.00991)	(0.0105)	(0.00934)	(0.0286)	(0.00318)	(0.00481)	(0.00487)
Controls								
Time invariant controls	X				X			
Time varying controls	X	X	X	X	X	X	X	X
Interacted census controls				X				X
Fixed Effects								
Year FE	X				X			
Zone FE		X	X	X		X	X	X
Quarter#City FE		X	X	X		X	X	X
Year#City#Area FE		X	X	X		X	X	X
Observations	5,740	5,740	5,740	5,740	5,740	5,740	5,740	5,740
Adjusted R2	0.82	0.98			0.77	0.96		

OLS, FE and 2SLS estimates of Equation (3) with the inclusion of progressively more fixed effects (full specification for Columns (2)-(4) and (6)-(8)) and, in Columns (4) and (8), of the interaction of the time invariant zone-level controls for demographic, education, occupation and housing characteristics with the growth rate of each city population in the base year 2011. The dependent variable is the natural logarithm of sale (Columns (1) to (4)) and rent prices (Columns (5) to (8)). The variable of interest is the lagged number of listings/100. The instrument in Columns (4) and (8) is $a_n^{\mathrm{TA}} \times g_t \times \mathrm{city}_i$. Robust standard errors clustered by zone in parenthesis. *** p<0.01, ** p<0.05, * p<0.1