EUI DEPARTMENT OF ECONOMICS

Essays on the Effects of Information on Economics

Anatole Jacques Idrissa Cheysson

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Florence, 02 April 2024

European University Institute
Department of Economics

Essays on the Effects of Information on Economics

Anatole Jacques Idrissa Cheysson

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Examining Board

Prof. Giacomo Calzolari, EUI, Supervisor
Prof. Alessandro Tarozzi, EUI, Co-Supervisor
Prof. Emmanuelle Auriol, Toulouse School of Economics
Prof. Mattia Nardotto, Université Libre de Bruxelles

Researcher declaration to accompany the submission of written work
Department Economics - Doctoral Programme

I Anatole Jacques Idrissa Cheysson certify that I am the author of the work Essays on the Effects of Information on Economics I have presented for examination for the Ph.D. at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

The copyright of this work rests with its author. Quotation from it is permitted, provided that full acknowledgement is made. This work may not be reproduced without my prior written consent. This authorisation does not, to the best of my knowledge, infringe the rights of any third party.

I declare that this work consists of 27760 words.

Signature and date:
21/03/2024

A Ch

Abstract

Chapter 1 investigates the impact of foreign language proficiency on brain drain in Albania. It leverages the accidental exposure to Italian television signals in the country, which led to increased Italian proficiency. The study finds that this exposure significantly increased the migration of highly skilled individuals, while other skill groups were not affected.

Chapter 2 explores the market for machine data (MD) generated by ICT and AI, highlighting issues such as fragmented datasets, externalities, and fuzzy property rights. It examines how data aggregators can contract with various data producers to share data and analytics, considering factors like producer heterogeneity, preference for anonymity, and market competition.

Chapter 3 examines the relationship between price discrimination and tacit collusion in the U.S. airline industry using data from 2000 to 2019. The study reveals that the impact of price discrimination on collusion varies over time, influenced by the quality of information available. In periods of poor information quality, price discrimination inhibits collusion, but the effect reverses as information quality improves.

Contents

Abstract i

Acknowledgements iv

1 Plurilingualism and Brain Drain: Unexpected Consequences of Access to Foreign TV 1

1.1 Abstract . . . 1

1.2 Introduction . . . 1

1.3 Literature Review . . . 4

1.4 Historical Background . . . 5

1.5 Data Description . . . 6

1.5.1 RAI and Geographic Datasets for Albania . . . 6

1.5.2 2005 Living Standards Measurement Survey Albania . . . 7

1.5.3 City-Level Dataset for Albania . . . 10

1.6 Identification Strategy . . . 10

1.7 Results . . . 12

1.7.1 Italian Television and Language Proficiency . . . 12

1.7.2 Italian Television and Brain Drain . . . 14

1.8 Exclusion Restrictions . . . 16

1.8.1 Italian Television, Competing Channels to Language Proficiency . . . 16

1.8.2 Italian Television and Returns to Education . . . 18

1.9 Robustness . . . 18

1.10 Conclusion . . . 21

2 Machine Data: market and analytics 24

2.1 Introduction . . . 24

2.1.1 Literature . . . 27

2.2 The Baseline Model . . . 29

2.3 Benchmarks . . . 32
2.4 Market-based analytics . . . 36
2.4.1 Delegation and in-house analytics . . . 38
2.5 Analytics with Anonymity . . . 41
2.5.1 Analytics with uniform contracts . . . 41
2.5.2 Analytics with secret contracts . . . 44
2.6 Analytics with competing producers . . . 46
2.7 Conclusions . . . 52

3 Price Discrimination and Tacit Collusion: An Empirical Analysis of the U.S. Airline Industry 56
3.1 Data . . . 58
3.2 Identification Strategy . . . 61
3.3 Results . . . 63
3.4 Robustness . . . 67
3.5 Conclusion . . . 68

A First Appendix A.1
A.1 Italian TV Shows Watched by Albanian Migrants . . . A.1
A.2 Emigration Patterns of Albanians 1990-2005 . . . A.2
A.3 GIS Data for Municipality . . . A.3
A.3.1 LSMS Questions . . . A.3
A.4 Balance Test . . . A.6
A.5 Greek Community in Albania . . . A.7
A.6 Instrumental Variable Regressions . . . A.8
A.6.1 One-Sample . . . A.8
A.6.2 Two-Samples . . . A.9
A.6.3 Additional Results . . . A.12

B Second Appendix B.1
A1 Proofs . . . B.1
A2 The value of data . . . B.9
A2.1 Functions and parameters for the running Example . . . B.9
A3 Equilibrium existence . . . B.10
A4 Secret contracts . . . B.12
B1 On the properties of the analytics' value . . . B.14

B1.1 Scale Property B.14
B1.2 Scope Property B.17

Acknowledgements

I would like to express my gratitude to my thesis directors, Giacomo and Alessandro. Thank you for your incredible supervision Giacomo, you were always available and I am very lucky to have benefited from your guidance. Both from a professional and a personal standpoint, I cannot overemphasize the lasting impact you have had and will continue to have on my career and life.This thesis would also not have come to fruition without the early guidance of Michèle, to whom I am deeply thankful. I would also like to extend my gratitude to Russell, who briefly served as my supervisor and engaged me in stimulating conversations.

I am indebted to all the professors who have left a lasting mark on me and served as a source of inspiration along this academic path: Andrea Mattozzi, Andrea Ichino, and David Levine.

My gratitude also extends to the persons who provided support within the department during the course of this thesis: Antonella, your kind words and infectious enthusiasm always brightened my day. I wish to express my appreciation to Sarah, Lucia, Cécile, Antonio, and Maurizio.

I would like to acknowledge my fellow thesis companions: Federica, Leonardo, Anna, Marcin, Adrien, Nihan, Karina, Alice, and Chloe. Our shared experience, unfortunately interrupted by COVID, was an enriching, and I am grateful to have had your company throughout.

Last but not least, I want to mention Jan, Damiano, Hugo and Riccardo. You are more than friends; you have significantly enriched my years in Florence and my life. I am genuinely thankful for having met all of you.

Finally, my eternal gratitude goes to my brother, my sister, and my closest friend, Felix, Salomé, and Paul, for their unwavering support. With the three of you by my side, I have never felt alone. You have consistently provided assistance and support, for which I am profoundly happy. I also extend my gratitude to my mother for her unconditional love and pride in my achievements; I would not have reached this point without her.

Chapter 1

Plurilingualism and Brain Drain: Unexpected Consequences of Access to Foreign TV

Co-written with **Damiano Argan**

1.1 Abstract

We study how foreign language proficiency affects brain drain by exploiting the heterogenous exposure of Albania to Italian television in the second half of the twentieth century. We document that, due to geographical proximity, the Italian TV signal accidentally reached Albania and, conditional on geographic conditions, Albanians' exposure to the signal was as good as random. We find that exposure to Italian TV led to a considerable increase in Italian proficiency rates and strongly increased the probability of migrating of highly skilled individuals while not affecting other skill groups.

1.2 Introduction

Linguistic distance between countries' languages is a key determinant of migratory flows (Belot and Ederveen, 2012; Adserà and Pytliková, 2015). As this distance increases, migrants tend to experience poorer labor market results (Adsera and Ferrer, 2015), which in turn makes migrating to countries with significantly different languages less appealing. The penalty imposed by linguistic differences is especially hard on high-skill individuals for whom communication skills are more valuable (Chiswick, 1995;

Berman et al., 2003). Consequently, linguistic distance is an important driver of migrants' self-selection into emigration (Borjas, 1987; Belot and Hatton, 2012): the higher the proximity between two countries' languages, the more migratory flows are composed of high skilled individuals.

Although the relationship between language and migration has received much attention, empirical research has remained observational in nature, unable to quantify and inform the causal effect of foreign language proficiency on migration decisions. To this day, we have little evidence on the effects of policies that promote plurilingualism on emigration patterns, and in particular on the emigration decisions of the educated. However, the considerable impact of the migration of high-skill individuals, i.e. brain drain, on the economy of origin countries is the subject of ongoing interest in the literature (Docquier and Rapoport, 2012; Shrestha, 2017; Anelli et al., 2023). In a related context, EU policy makers are keen on evaluating the impact of language barriers on labor mobility, which is crucial for the success of monetary unions.[1] We fill this gap in the literature by providing novel causal evidence on the relationship between foreign language proficiency and emigration, with a specific focus on the emigration of high-skill individuals.

In 1957, the Italian public broadcasting company (RAI) built a TV transmitter in Puglia, a region in southeast Italy, and its signal inadvertently reached parts of neighboring Albania. During that period, and until 1990, Albania was a communist dictatorship isolated from the rest of the world, both physically and culturally. Conditional on geographical characteristics, we show that individual's exposure to Italian television was quasi-random; the specificity of this historical episode dispels common endogeneity concerns as signal access was unintentional and internal movement in Albania was restricted, preventing Albanians from relocating to areas with signal availability. These factors address the problem of endogenous location choice for both the transmitter and individuals, which typically results in biased estimates of media exposure on observed outcomes. Following the collapse of the regime in 1990, massive emigration waves and a brain drain occurred (Gërmenji and Milo, 2011; Gëdeshi and King, 2019). We leverage this distinctive context to examine the impact of Italian television access on Italian language proficiency and the Albanian brain drain.

For this study, we use three datasets : (i) a geo-referenced dataset of signal availability and power provided by RAI (*RAI* dataset, henceforth); (ii) a geo-referenced dataset of terrain characteristics aggregated at the municipality level (*Geographic* dataset); (iii) the 2005 Living Standard Measurement Survey conducted by the World Bank and Albanian statistical agency (*LSMS* dataset). We measure the average exposure to Italian television for each municipality in Albania using the *RAI* dataset. With the

[1] Since Mundell (1961), mobility has been considered key to the success of monetary unions. In the EU, labor is deemed not mobile enough, especially when compared to the US labor market (House et al. (2018)). Language barriers are seen by EU policy makers as one of the reasons for the lack of labor mobility (`https://education.ec.europa.eu/focus-topics/improving-quality/multilingualism/about-multilingualism-policy`, among others).

Geographic dataset, we generate a comprehensive set of geographical and topographic controls at the municipality level. We exploit three sections of the LSMS: the internal migration folder to relocate individuals to their municipality of residence in 1990 to infer their access to Italian television prior to the dictatorship's fall; the 1990 foreign language proficiency questionnaire; and since the LSMS, by design, only includes individuals residing in Albania in 2005, we use the questionnaire on respondents' siblings' residences to create a dataset that encompasses migrants.

Our study offers two novel contributions. Firstly, we examine the impact of Italian television access on foreign language proficiency in 1990. Since the LSMS base sample consists only of non-migrants, we cannot estimate the average treatment effect of television on language proficiency. However, in line with the literature, we assume that non-migrants have a lower propensity to learn a foreign language than migrants (Bütikofer and Peri, 2021), implying that estimating the effect of Italian TV access on language skills for non-migrants provides a lower bound. We estimate a lower-bound positive increase of 7 percentage points in Italian proficiency rates between municipalities fully exposed to Italian television and those with no exposure, which is more than double the average Italian language proficiency rate of 5.3% in 1990. We also successfully conduct placebo tests for other foreign languages.

Our second contribution involves estimating the causal impact of Italian TV exposure on the likelihood of emigration. Using the sample of LSMS respondents' siblings, we observe no effect on the probability of migration when estimating for the entire sample. However, for high-skilled individuals, we find a substantial positive effect of approximately 20 percentage points on the likelihood of emigrating abroad, accompanied by a similar effect on the probability of emigrating to Italy. While we cannot estimate an instrumental variable regression to extend beyond this reduced form estimate due to data limitations, the already sizable 20 percentage point increase allows us to confidently assert that foreign language proficiency significantly boosts the migration probability for high-skilled individuals.

We then discuss the exclusion restrictions, specifically that exposure to Italian TV only influences migration behavior through Italian language knowledge. Competing channels include television's role as an information provider and its impact on expected returns from migration (Farré and Fasani, 2013; Pesando et al., 2021; Adema et al., 2022). We exploit interviews conducted in 1991 with Albanian migrants, which reveal that their primary viewing preferences were entertainment programs that lacked pertinent migration information, such as job opportunities, regional economic conditions, mobility, and housing-related details. Furthermore, we use the LSMS to show that Albanians who migrated abroad and returned did not use TV as an information source to organize their emigration. Lastly, we discuss whether Italian TV led Albanians to overestimate the benefits of migrating to Italy (Mai, 2004). However, such a channel would imply a uniform impact across skill categories, which is not what we observed - we only found an effect among high-skilled individuals. This finding corresponds with the notion that language proficiency is crucial for high-skilled migrants, as effective communication skills are particularly important in high-skill jobs, as identified in the literature and predicted by the Borjas model (Borjas, 1987; Chiswick, 1995; Berman et al., 2003).[2]

The paper is organized as follows: Section 1.3 presents the literature review, in Section 1.4 we summarise the historical background, Section 1.5 describes the data, Section 1.6 then discusses our identification strategy, in section 1.7 we show the results. Section 1.8 discusses the exclusion restriction, Section 1.9 presents robustness tests. Finally, Section 1.10 concludes.

1.3 Literature Review

Our paper makes contributions to five areas of literature. Firstly, this study adds to the literature on linguistic and cultural determinants of migration by providing the first causal evidence of language proficiency's effect on emigration patterns. Specifically, we contribute to the literature on the causes of brain drain, which has identified cultural distance as a key predictor of migrants' educational selectivity (Belot and Hatton, 2012). While existing literature has been observational (Belot and Ederveen, 2012; Adsera and Ferrer, 2015), our research presents causal findings that can inform policymakers about the consequences of foreign media exposure and policies promoting plurilingualism.

Second, our study is linked to the literature on the influence of mass media on societal outcomes, from which we derive our identification approach (Olken, 2009; La Ferrara, 2016; Durante et al., 2019).[3] In particular, Farré and Fasani (2013) shows how TV exposure in rural Indonesia reduced internal migration by helping to correct overestimated returns to internal mobility; Adema et al. (2022) shows how internet access increases desire to migrate and actual migration by reducing the cost of information, trust in government and perceived well-being.[4] Our findings complement this research in providing evidence of the effect of media exposure on migration through language skill acquisition, a specific form of human capital, as opposed to other types of information.

Additionally, we connect to research that concentrates on the media's influence on educational outcomes: Gentzkow and Shapiro (2008) shows how television exposure in the US had a positive effect on the test scores of children raised in non-English speaking households. Kearney and Levine (2019) shows

[2] In the Online Appendix, we show that the Borjas model predicts an increase in migration probabilities for above-average productive individuals as a consequence of a positive exogenous shock to the correlation coefficients for a wide range of parameters.

[3] This literature includes a wide range of possible outcomes: political outcomes (Gentzkow and Shapiro, 2008; Olken, 2009; Enikolopov et al., 2011), gender norms (Jensen and Oster, 2009; Chong and La Ferrara, 2009; Ferrara et al., 2012; Kearney and Levine, 2015)), and consumption choices (Bursztyn and Cantoni, 2016).

[4] In a 2007 working paper, Braga (2007) explores the influence of Italian television on promoting seasonal migration from Albania. However, the study has notable limitations: it fails to suggest a mechanism through which TV impacts migration, neglects to investigate the role of Italian television in international migration, and does not address the varying effects of TV on different skill groups.

that the edutainment program *Sesame Street* was beneficial for children's educational attainment. Durante et al. (2019) demonstrates how children exposed to Berlusconi's television became less cognitively sophisticated and civically minded. Our research complements these findings by showing how exposure to foreign media increased foreign language proficiency.

Finally, this article relates to the research on language proficiency and migrants integration. Causal studies have documented how proficiency in the host country's language increases migrants earnings (Sarvimäki and Hämäläinen, 2016), labour force participation (Lochmann et al., 2019) and employment (Lang, 2022; Schmid, forthcoming). Given our findings that foreign language proficiency increases emigration, our work suggests that potential migrants anticipate these improved labor market outcomes.

1.4 Historical Background

Enver Hoxha came to power in Albania in 1944 in the immediate aftermath of the war.[5] He rapidly seized absolute power and organized the complete isolation of the country from the outside world: internal migration was controlled and limited, and emigration to foreign countries was forbidden.[6] This isolation also extended to culture: no foreign books, movies, nor newspapers were allowed to circulate. Hoxha's communist regime lasted until 1990.

Despite Enver Hoxha's best efforts, there was *a tear in the wall*. In 1957, the RAI (Radiotelevisione italiana - Italian State Television) built a television transmitter in Martina Franca (Italy, Puglia- the Italian region closest to Albania, on the other side of the sea). Thanks to its power and the short distance between Italy and Albania, the transmitter unintentionally reached parts of Albania, it still broadcasts to this day, and did so without interruption since 1957. Since the 70s, when TV sets began to be widespread in Albanian homes, Albanians have regularly watched Italian television.[7] Italian programs provided entertainment shows that Albanian television did not feature at the time: it only had one channel broadcasting four hours each day, alternating between propaganda and few Albanians films repeated continuously. It is the entertainment content of Italian programming that proved attractive to Albanians.[8]

In 1990, following pressure for reform from the population, the communist structures began to be dismantled, and in 1992 the first democratically elected government took power. From June 1990 on-

[5] This section owes much to Dorfles and Gatteschi (1991); Abrahams (2016); Fevziu et al. (2018)

[6] Only around 6000 Albanians managed to escape to foreign countries between 1944 and 1990. While foreign emigration boomed right at the fall of the regime.

[7] Historical evidence on Italian television watching in Albania are manifold: Dorfles and Gatteschi (1991); Mai (2004); Abrahams (2016); Fevziu et al. (2018) among others. Although in 1973 Italian television watching was forbidden in Albania, people continued to do so regularly. Using World Bank data we compute that around 61% of Albanian household had a TV set in 1990. Data on distribution of TV sets by district in Albania in 1990 can be found in the Online Apppendix.

[8] Interviews of Albanians arriving in Italy in 1990 were conducted, they revealed the extent to which Albanians were familiar with Italian television. More details is available in Appendix A.1.

ward, Albanians recovered their ability to emigrate. During the 1990s decade, around 800 thousands Albanians migrated abroad, about one fourth of the entire Albanian population at the time. It is estimated that about 600,000 Albanians emigrated to Greece and 200,000 to Italy.[9] This emigration wave has been coined repeatedly as a brain drain in the literature: by 2000, an estimated 20% of high-skilled Albanians had left the country (Docquier and Marfouk, 2006; Gërmenji and Milo, 2011; Gëdeshi and King, 2019). Although migration began immediately after the fall of the regime in 1990, its intensity varied with economic and political events: it picked up pace following an economic crisis in 1997, and reached a peak with the war in neighboring Kosovo.

1.5 Data Description

Our analysis builds upon the creation of a novel dataset. We collected information on the Italian TV signal coverage in Albania obtained from RAI along with information about terrain elevation from NASA's Shuttle Radar Topography Mission. For each Albanian municipality, we computed distance measurements and a terrain ruggedness indicator. We then aggregated these datasets at the municipality level, and merged them with the 2005 World Bank Living Standard Measurement Survey for Albania that contains individuals information. Finally, we construct an urban area dataset for Albania for 1986 by classifying NASA satellite images using machine learning techniques.

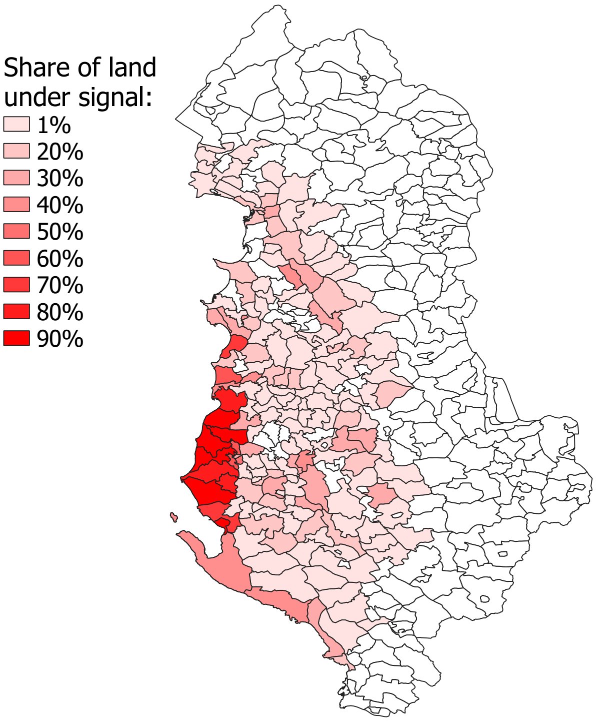
1.5.1 RAI and Geographic Datasets for Albania

We obtained from RAI geographically referenced data on Italian TV signal strength in Albania. The Italian town of Martina Franca is home to the oldest and most powerful Italian TV transmitter able to broadcast all the way into Albania, all other transmitters powerful enough to reach Albania have their signals contained in it. Therefore, we only collected and processed the signal emitted from this antenna. Operational since 1957, the transmitter has not experienced any modifications that altered its power or reception. To compute its signal propagation across the terrain, the RAI uses a standardized forecasting model.[10] We re-classified the dataset of signal quality provided by RAI in two steps. First, to align with RAI's guidelines, we initially transformed signal propagation into a binary dataset, which determines whether Italian TV is accessible for each 100x100 meter grid on the Albanian map. More specifically, we designated Italian television as accessible when the signal quality meets or exceeds a threshold of 55 dBµV/m. Second, we computed for each municipality the share of its area where radio signal is available. Figure 1.1 displays the re-coded TV signal availability across Albanian municipalities.

[9] See Galanxhi et al. (2004). See also Figure A.1 in appendix A.2 which plots yearly emigration flows by destination.

[10] Prescribed by the International Telecommunication Union, See in particular Recommendation P.526. The model takes into account the diffraction due to the orography of the terrain which reinforces or blocks propagation.

Figure 1.1: RAI Signal Coverage



Notes: Representation of Albania at the municipality division unit. Signal radio is aggregated at the municipality level to compute the share of area with Italian television access.

We collect topographic characteristics of the terrain from the Shuttle Radar Topography Mission of the NASA which contains information on elevation at a 30x30 meters resolution. From this data we compute the terrain ruggedness index following Riley et al. (1999). We then aggregated both elevation and ruggedness at the municipality level by taking the average over municipality area. We complement this topographic data with distance data, by computing for each municipality the average distance of each of its 30x30 meters cells to Italy, to Greece, to the closest port,[11] and to the antenna in Martina Franca.[12]

1.5.2 2005 Living Standards Measurement Survey Albania

Administered to each household member of 3840 households in 480 primary sample units (geographical census area), the 2005 Living Standards Measurement Survey (LSMS) contains information on 17302

[11] We consider the four most important ports in Albania : Saranda and Vlorë in the south, Durrës in the center, and Shëngjin in the north.

[12] More details is available in the Appendix A.3

individuals.[13] Restricting the sample to those who were at least 18 years old in 2005, we retain 11040 individuals living across 322 of the 383 Albanian municipalities.[14] As the survey was conducted in 2005, it provides direct information only on non-migrants, however, household heads and spouses are asked to list all their siblings (henceforth, we refer to household heads and spouses that list their siblings as *listing sibling*) and report for each one their demographics, country of living, and year of departure if they migrated. A maximum of seven siblings can be listed, but it's noteworthy that individuals with more than seven siblings only comprise 2% of the total sample. Household members also list their children and spouse living out of the household. We derive two datasets from these sets of questions: (i) one in which each sibling is an observation (27666 obs.); (ii) another in which each child or spouse out of the household is an observation (4714 obs.). Unlike the respondents of the *LSMS*, individuals in these two additional datasets can either reside in Albania or abroad. We thus derive three datasets from the LSMS, the first about the respondents themselves (hereafter, *base dataset*), the second about the household heads and spouses' siblings (hereafter, *siblings dataset*), and the third about household members' children and spouses out of the household (hereafter, *children/spouses dataset*).

Regarding the base dataset, we concentrate our analysis on three types of information: (i) internal migration history since birth; (ii) foreign language proficiency in 1990; (iii) individual education. The exact phrasing of all questions relevant to the analysis of this paper can be found in Appendix A.3.1. Using the internal migration history of respondents, we relocate individuals to their municipality of residence in 1990 and thus to their exposure to Italian television signal before the fall of the regime. Respondents reported their foreign language proficiency in 1990 in Italian, Greek, English or if they had knowledge of "another foreign language". They can answer either 1) Yes, fluently, 2) Yes, some or 3) No. We generate a dummy variable for foreign language proficiency that we code 1 if individuals answer Yes, fluently or Yes, some and 0 if they answer No. Finally, the LSMS records individual's highest education levels, we code a dummy variable equal to 1 if an individual attended university for at least one year. For each individual in the siblings dataset and in the children/spouses dataset, the LSMS includes the country of residence and the date of emigration. The children/spouses dataset also contains information on foreign language skills in 1990. Where necessary, we attribute the characteristics of their relatives to the individuals in these datasets, in particular their location in 1990 and the highest level of education of the listing sibling to the individuals in the siblings dataset. Note that only children are missing from the siblings dataset and that the children/spouses dataset contains information on a specific sub-population of individuals, namely those who have left the household. Nevertheless, only children account for a minor portion of the population, comprising 6% of the sampled household heads and 3% of the spouses.

[13] Data collection ran between May and early July in 2005. Data and all the material are available at `https://microdata.worldbank.org/index.php/catalog/64`. Household membership is defined as having been away from the household for less than 6 months during the year preceding the survey.

[14] For underage individuals information is missing.

Table 1.1: 2005 LSMS, Selected Statistics

	Base			Siblings			Children/Spouses		
Variable	All	Men	Women	All	Men	Women	All	Men	Women
Observations	11040	5226	5814	27666	14421	13245	4714	2236	2478
Age Distribution									
25 percentile	24	24	24	37	37	37	27	27	27
50 percentile	40	41	39	45	45	45	33	34	33
75 percentile	53	54	53	55	55	55	40	41	40
Mean	41	41	40	46	46	46	34	34	34
Education									
Primary	54%	49%	58%	.	.	.	52%	53%	51%
Secondary	21%	22%	20%	.	.	.	52%	53%	51%
Vocational	16%	19%	13%	.	.	.	13%	14%	12%
University	9%	10%	9%	.	.	.	9%	8%	10%
Proficiency in 1990									
Italian	5.3%	5.2%	5.3%	.	.	.	7.9%	8.1%	7.7%
Greek	1.9%	2.5%	1.5%	.	.	.	3.1%	4.2%	2.1%
English	4.4%	3.9%	5.0%	.	.	.	3.1%	4.2%	2.1%
Internal Migration									
Before 1990	7.9%	4.7%	11.2%	.	.	.	.	.	.
After 1990	20.6%	16.3%	24.7%	.	.	.	.	.	.
International									
Share migrated	.	.	.	17%	21%	12%	44%	61.3%	28.9%
Before 1990	.	.	.	0.8%	0.6%	1.1%	0.2%	0.2%	0.3 %
Destination									
Italy	.	.	.	32%	32%	33%	39%	41%	36%
Greece	.	.	.	50%	51%	46%	40%	40%	42%
UK	.	.	.	5%	5%	3%	7%	9%	3%
USA	.	.	.	7%	6%	9%	5%	4%	8%
Television									
Ownership rate	62%	.	.	.	.	.	.	.	.

Source: 2005 Living Standard Measurement Survey, World Bank and INSTAT.

Furthermore, neither dataset contains individuals from families that have completely emigrated.

Table 1.1 presents descriptive statistics for each dataset. All three datasets are balanced in terms of their sex-ratio, they contain between 49.5% and 53.4% men. With 5.3% of respondents self-declaring their proficiency in Italian, Italian stands as the most widely spoken foreign language in Albania in 1990. English is a close second with 4.4%, and Greek stands third with 1.9% of individuals. In the children/spouses dataset, 28.1% of the sample could speak Italian in 1990, 22.6% could speak Greek, and 15.1% English. Finally, international migration represents 44.2% of the sample of children/spouses out of the household and 16.9% of the sample of siblings. Within the samples, Italy and Greece are the most common destination countries, with 32% of siblings that migrated living in Italy, 50% in Greece; around 39% of children/spouses living abroad are in Italy with an equal share in Greece. In 1990, The LSMS reports that 62% of households were endowed with a TV set.

1.5.3 City-Level Dataset for Albania

We build an urban land cover dataset for Albania for 1986. At a resolution of 30x30 meters, we record for each year and each cell whether it contains urban land or not, and we calculate the proportion of the city that is covered by the Italian TV signal. With this dataset we can calculate the proportion of the urban area of a municipality exposed to the signal in 1986, which we call *Signal II*. This alternative measure has the advantage of estimating exposure only in urban areas, thus avoiding the definition of an exposed municipality when it is mainly exposed in the inhabited area.

1.6 Identification Strategy

A common difficulty in the estimation of a causal effect of signal availability on societal outcomes is the placement of the transmitter. Transmitters are typically placed in strategic locations in order to target specific populations such as densely populated urban areas. In parallel viewers might self-select by relocating to areas where the signal is accessible. This simultaneous selection can substantially bias estimations of causal effects, making the treated population different from the untreated population on unobservables characteristics. The treatment effect on the treated thus differs from the average treatment effect.

The Albanian setting suffers none of these two issues. First the transmitter was placed to satisfy the needs of the Italian population, and no attention was paid to the possibility that the signal might reach Albania, it accidentally did so. Second, emigration was forbidden and internal migration was restricted and centrally managed under the Communist regime, preventing any selection on the Albanian side.[15] Table 1.1 reports that only 0.2% of migrants in the siblings dataset and 0.7% of the migrants in the children/spouses dataset emigrated abroad before 1990. Internal migration tripled from 7.9% of the sample that internally migrated between 1975 and 1990 to 20.6% between 1990 and 2005.

Once controlling for geographic and topographic variables that correlate both with the radio signal exposure and the outcome variables, the exposure to radio signal can be considered as good as random. The controls we consider are: (i) distances to Italy, the transmitter and the nearest port; (ii) topographic data on elevation and ruggedness; (iii) district fixed effects. These controls are potentially correlated with both signal decay and other variables related to our outcomes: migration cost and cultural proximity. Once included, we thus can estimate the effects of residual variations in signal reception due to the topography of the terrain within districts' areas on each outcome variable. We estimate the following specification:

$$y_{i,m,d} = \alpha_0 + \beta \times Sig_m + \gamma \times Dist_m + \theta \times Geo_m + \sum_{d=1}^{36} \alpha_d \times \text{Distr}_d + \varepsilon_{i,m,d} \tag{1.1}$$

Where Sig_m is the share of a municipality's area reached by the TV signal. $Dist_m$ is a vector containing the distances of municipalities to Italy, the nearest port, and the transmitter in Martina Franca. Some specifications also include distance to Greece in $Dist_m$. Geo_m controls for the elevation and ruggedness of municipality *m*. $Distr_d$ are district fixed-effects, such that we measure within a district the differences created between municipalities by the radio signal.[16] $\varepsilon_{i,m,d}$ is the error term. In this specification, β identifies the causal effect of exposure to Italian television on outcome $y_{i,m,d}$. Importantly, it identifies an intent-to-treat effect as we only estimate the effect of exposure to the television signal rather than the one of actually watching Italian television.

We study 2 sets of outcomes: (i) Language proficiency as measured by the self-declared language proficiency in 1990 of individual i living in municipality m of district d in the *base* dataset; (ii) Migration outcome, as measured by whether an individual in the *siblings* dataset lived abroad in 2005 or not. To compute the heterogeneity of the effects we restrict the samples to specific subsets of the population of interest, rather than including a dummy, this approach ensures that fixed-effects and controls are population-specific. Finally, we cluster standard errors at the municipality of residency in 1990 level (i.e. the treatment level) in all regression exercises.

One concern is that municipalities close to the Albanian coastline both concentrate the most TV exposure in the sample and have the lowest distance to either Italy or the transmitter, making it hard to disentangle the effects of distance and TV exposure. If results happen to be sensitive to the exclusion of the municipalities that are the closest to Italy, this might cast doubts on the identification strategy. We

[15] See Galanxhi et al. (2004) page 9.

[16] Albania is divided in 36 districts, each district contains 8.6 municipalities on average.

address this concern in a number of ways. First, the inclusion of district fixed effects ensures we compare the effect of Italian TV signal between municipalities of the same district, where the distances to Italy are relatively similar. Second, Figure 1.2 in Section 1.9 plots the mean TV signal coverage at different deciles of the distribution of distance to Italy, distance to the closest port and elevation. Although TV signal is concentrated in the first deciles of each distribution, there is considerable variation within and beyond those deciles that allows for meaningful comparisons. Third, in Section 1.9 we go further by showing that results are robust to the exclusion from the sample of the municipalities that are the closest to the ports and the closest to the Greek border. Appendix A.4 proposes balance tests on age and sex ratios using the siblings dataset, confirming that the treated and untreated samples are comparable on observables.

1.7 Results

In the following section we present the results of regressions of the effects of Italian television exposure on Italian language proficiency in the 1990 and on the migration probability of individuals between 1990 and 2005. We show that Italian TV exposure had a sizeable and significant effect on the probability to know Italian in 1990, no effect on the average likelihood to migrate on the Albanian population, but a significant and sizeable effect on the probability to migrate for high skilled individuals, and in particular on the probability to migrate to Italy. Results are paired with placebo tests.

1.7.1 Italian Television and Language Proficiency

This section considers the effect of Italian television on the probability of knowing Italian in 1990. Empirical work highlights the effects of television watching on cognitive outcomes: whether through educational (Kearney and Levine, 2019) or entertainment content (Durante et al., 2019), television has been found to have an influence on human capital accumulation. In our context, we test whether exposure to the Italian language through television pushed Albanians into developing language skills in Italian.

The LSMS is a survey of non-migrants as only individuals who did not migrate until 2005 are eligible to take the survey, hence on this sample we estimate the effect of television access on the acquisition of language skills of individuals that did not emigrate. In line with the literature, if we assume that non-migrants have a lower propensity to learn a foreign language than migrants (Bütikofer and Peri, 2021), then the estimate of the effect of Italian television access on the language proficiency of the non-migrants is a lower bound of the average treatment effect on the full population. We later test this assumption using the children/spouses dataset which contains information on language proficiency and includes migrants and non-migrants, it allows to estimate the effect of Italian television access given migration decisions.

Table 1.2: Italian television effect on foreign language proficiency in 1990

	Base				*Children/Spouse*	
					Abroad	Albania
	Italian (1)	English (2)	Other (3)	Greek (4)	Italian (5)	Italian (6)
Signal	0.070^{**} (0.033)	0.020 (0.018)	0.006 (0.023)	0.013 (0.010)	0.133^{**} (0.064)	0.041 (0.046)
Observations	11040	11040	11040	11040	2088	2626
Clusters	322	322	322	322	233	229
Controls:						
Common: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness						
Greek Community	N	N	N	Y	N	N

Notes: The table reports OLS estimates of the effect of exposure to Italian TV on foreign language proficiency in 1990. (1)-(4) use the base dataset, (5)-(6) the children/spouse dataset, in specification (5) the ones living abroad and in (6) the ones living in Albania. The dependent variable is the reported capability of speaking Italian, English, Other (category any other language), and Greek in 1990 coded as a dummy. The main explanatory variable, Signal, is the share of a municipality's area with access to Italian TV. Controls for Greek community in specification (4) include distance to Greece and dummies for: (i) Greek ethnicity, (ii) orthodox religion; (iii) Greek as maternal language; (iv) speaks Greek daily at home; (v) speaks Greek in the community. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

We estimate Equation 1.1 with foreign language proficiency in 1990 as a dummy outcome variable. Table 1.2 columns (1)-(4) present the results of the regression of Italian proficiency on TV signal exposure and three placebo test, using individuals from the base dataset of the LSMS as the sample. In municipalities fully exposed to Italian television, estimated effect in column (1) indicates that the rate of Italian proficiency increased by 7 percentage points, more than double the proportion of Italian speakers in Albania. The estimate is significant at the 5% level and economically sizeable. As explained above, since the sample only include non-migrants, this estimate is a lower bound of the average treatment effect. Additionally, as we can only measure television exposure, not television watching, we can only estimate an intention-to-treat effect, lower than the average treatment effect.

Columns (2) to (4) of Table 1.2 report the results of placebo tests, we check that Italian television exposure did not cause an increase in proficiency of other languages. As expected, the coefficients are small and insignificant. In the case of Greek language proficiency, we added controls related to whether individuals in the sample belong to the Greek diaspora present in Albania, additional controls include dummies for Greek ethnicity, Orthodox religion, and Greek spoken daily at home. We also included a distance variable that measures for each municipality its distance to the Greek border.[17]

To overcome the limitation of the base LSMS sample, which only contains non-migrants, we exploit the dataset of children/spouses which contains household members' spouses and children that no longer live in the households, and can either be international migrants or still in Albania. Regressions (5) and (6) present the estimates of regressions on the sub-sample of children/spouses that respectively live abroad and in Albania. We find a sizeable effect of exposure on Italian proficiency on the sub-sample that lives abroad of 13 additional percentage points in fully exposed municipalities, triple the estimate on the sample of non-migrants. This result approximates the intent-to-treat effect of Italian television exposure on migrants, it confirms that migrants have a higher propensity to learn a foreign language than non-migrants (Bütikofer and Peri, 2021).

The lower bound of 7 percentage points increase we estimate indicates that the impact of television access on rates of Italian proficiency in Albania was considerable, more than doubling the rate of Italian proficiency. We thus find that exposure to foreign media can be used as an effective tool to foster foreign language proficiency. Given empirical results that link linguistic proximity and emigration (Belot and Ederveen, 2012; Adsera and Ferrer, 2015), we expect television access to have impacted patterns of emigration to Italy through its impact on language proficiency.

1.7.2 Italian Television and Brain Drain

In this section, we investigate the effect of Italian television on the emigration decisions of Albanians between 1990 to 2005. The literature underlined the penalty that linguistic differences represent for highly-skilled migrants (Adsera and Ferrer, 2015) owing to the complementarity between language and skill (Chiswick, 1995; Berman et al., 2003). The seminal paper of Borjas (1987) also underlined the importance of such mechanisms in driving the self-selection of migrants. As a consequence, we expect the effect of Italian television to have differed across skill groups.

To conduct this investigation, we resort to the siblings dataset. As discussed in Section 1.5, the LSMS respondents are all non-migrants, we thus exploited a sample composed of their siblings, that can either be migrants or non-migrants. We assume that siblings of respondents were living in the same municipality as respondents in 1990, consistent with the low internal migration rates characterizing Albania before 1990. As international migration was forbidden prior to 1990 (see Section 1.4) individuals residing in a foreign country in 2005 migrated between 1990 and 2005. We attribute to siblings the human capital of their listing siblings: we assume that education levels of siblings were highly correlated.

[17] The case of Greek, owing to the particular history between the two countries, is further discussed in Appendix A.5.

Specifically, we define a sibling as high skilled if her listing sibling attended university for at least one year. We exploit the identification strategy described in Section 1.6. Standard errors are clustered at the municipality of residency in 1990 level (treatment level).

Table 1.3: Effect of Italian Television Exposure on Probability to Migrate

	Siblings dataset			
	Abroad (1)	Abroad (2)	Italy (3)	Greece (4)
Signal	-0.002 (0.031)	0.244*** (0.074)	0.131** (0.059)	0.0518 (0.074)
Sample	Full	High Skill	High Skill	High Skill
Observations	27666	2153	2153	2153
Clusters	310	128	128	128

Common Controls: District F.E., Distance to Italy, to transmitter, to Port, Elevation, Ruggedness

Notes: The table reports OLS estimates of the effect of exposure to Italian television on the probability to reside abroad. The outcome variable is a dummy taking value 1 if abroad for columns (1) and (2), 1 if in Italy for column (3) and 1 if in Greece for column (4), and 0 otherwise. All specifications use the siblings dataset (see Data Section 1.5). Specifications (2)-(4) restrict the sample to siblings of individuals that attended university for at least one year. Signal is the share of a municipality's area exposed to Italian television signal. Standard errors are clustered at the municipality of residency in 1990 level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.3 column (1) reports the estimate for the effect of TV signal exposure on individuals' probability to migrate abroad, we do not find evidence of an effect. In column (2) we subset for the population of individuals whose listing sibling attended university for at least one year and repeat the estimation of Equation 1.1. We find an economically and statistically significant effect: Italian television signal increased the migration probabilities of fully exposed individuals by 24 percentage points. We stress that the sample of high-skill individuals represents only 10% of the total sample in specification (1), implying that the effect is attenuated in the full sample. The positive effect we estimate on high-skill individuals is not paralleled by a significant negative effect on individuals with other education levels.[18] Finally, we test whether the destination of the emigration of the high-skilled is Italy, columns (3) and (4) test for

[18] Results on the rest of the sample are available in Appendix A.6.3.

emigration by destination. Although reduced in magnitude, Italian television signal access significantly increased emigration towards Italy, and left emigration rates towards Greece unchanged. In Section 1.9 we successfully test our results with alternative specifications of TV signal and human capital measures.

To move from this reduced-form estimate to the effect of language proficiency on emigration probability, we would need to perform an instrumental variable regression. However, we do not have information on the Italian language proficiency for siblings, and would thus need to rely for the first-stage on the results of the regressions of foreign language proficiency on the sample of non-migrants. It would underestimate the effect of TV on language proficiency, inducing an overestimation of effects in the second stage. Nonetheless, since the effect of TV exposure on Italian proficiency is necessarily bounded between 0 and 1, we do know that the reduced form estimate is necessarily a lower bound of the effect of language proficiency on the migration probabilities of the highly skilled.[19] Given that the reduced-form estimate already shows a substantial 20 percentage point increase in the likelihood of emigration, we can confidently assert that foreign language proficiency strongly enhances the migration probability of highly skilled individuals.

Taken together, our results imply that Italian television accentuated the brain drain towards Italy. Television access pushed many educated people into emigration towards Italy, thus increasing the positive selection of emigrants and contributing to the brain drain. Previous research investigating the impact of the media on migration behaviour emphasized the role of the media as a source of information (Farré and Fasani, 2013; Pesando et al., 2021; Adema et al., 2022). In this setting, given results on language proficiency, we expect the language-skill complementarity characterizing highly-skilled individuals to have played an important role in raising their returns to emigration (Chiswick, 1995; Berman et al., 2003). We posit that language proficiency is the main mechanism through which Italian television exposure increased the emigration of the educated. In the next section, we discuss the exclusion restriction to our identification strategy.

1.8 Exclusion Restrictions

1.8.1 Italian Television, Competing Channels to Language Proficiency

The most widely discussed channel in the literature is the one of information: migrants' expectations about income abroad can be biased (McKenzie et al., 2013), television and media might correct these expectations by providing valuable information about life abroad (Farré and Fasani, 2013; Adema et al., 2022). Applied to the Albanian context, Italian television would have provided high-skill individuals

[19] In appendix A.6, we nonetheless test for the two sample instrumental variable regression. Results are too imprecise to provide meaningful information.

with information about economic opportunities in Italy. In this section, we provide evidence to rule out the role of this competing channel.

Historical sources emphasize that Albanians were watching entertainment programs on Italian television, and data confirms that picture. Dorfles and Gatteschi (1991) reports results of interviews conducted in March 1991 on 311 Italian speaking Albanian migrants just arrived in Italy. Of the people interviewed, 301 declared they were watching Italian television in Albania, they were further asked which Italian television programs they would usually watch. The overwhelming majority of programs listed, 93%, are entertainment programs, only 7% of listed programs were news shows. In Farré and Fasani (2013), it is precisely news content which induced potential migrants to revise their beliefs. As interviews reveal, Albanians mainly watched entertainment programs: they were not being provided with useful information thanks to Italian television.[20]

We dispel further concerns about the contribution of the informational channel by exploiting the migration questionnaire of the LSMS. Members of surveyed households are all asked whether they migrated for at least one month since the age of 16 (since respondents are all in Albania, they would by definition be temporary migration episodes). Those who responded positively were subsequently asked ”*who provided information on where to go and/or how to find work during this first migration episode*”. Respondents can choose their answer from a list including the item *TV, radio, newspaper or book*. Table 1.4 presents the distribution of answers: only 1% of individuals chose this item. Even though the sample interviewed is one of return migrants, it is informative of what migrants themselves would have answered, and indicates further that television was not used as a source of information.

Beside information, watching entertainment television could have led Albanians to form an idealized view of life in Italy as suggested in Mai (2004). We would expect such an effect to have been homogeneous across skill groups, it could nonetheless turn heterogeneous in our sample if low-skilled individuals faced liquidity constraint preventing them from financing migration project. Alternatively, it could also be that TV ownership was correlated across skill-groups. Table 1.4 addresses these concerns. First, it shows that individuals migrated across all education groups with only small differences in emigration rates, indicating that individuals with lower education levels were not necessarily liquidity constrained. The same is true for television ownership: although its rate increases with education, TV sets were widespread across education groups.[21] This evidence suggests that how Italian television shaped beliefs did not impact migration patterns.

[20] A detailed presentation of the results of these interviews in available in Appendix A.1.

[21] In addition, historical sources report that group viewing of Italian television were regular and frequent, people did not need to own a TV to watch Italian television regularly.

1.8.2 Italian Television and Returns to Education

Another competing mechanism is that television watching raised the return to education. Shrestha (2017) shows that the possibility to migrate can in some context raise returns to education, thus increasing the average education of the population. If this is the case in the Albanian context, university educated individuals in municipalities with Italian TV access might differ on unobservables from university educated individuals who lived in municipalities without such access because they were pushed into accumulating more human capital by TV access, and these unobservables may drive our results. To remove these unobservables from the sample, we restrict it to individuals that completed their education prior to the fall of the regime in 1990. The only way in which Italian TV might have increased education returns is by raising the returns on emigration, as migration was forbidden before 1990, this effect must have been absent. We therefore estimate the siblings' hypothetical age of graduation, defined as the age each sibling would have graduated if they had completed their education at the same age as the listed sibling. We then filter out from the sample individuals with hypothetical year of graduation posterior to 1990. Specification (5) of Table 1.6 confirms our baseline results on this subsample, hence, even among individuals that accumulated human capital before 1990, when Italian television could not have raised the returns to education, we find the same effect on the emigration of the educated.

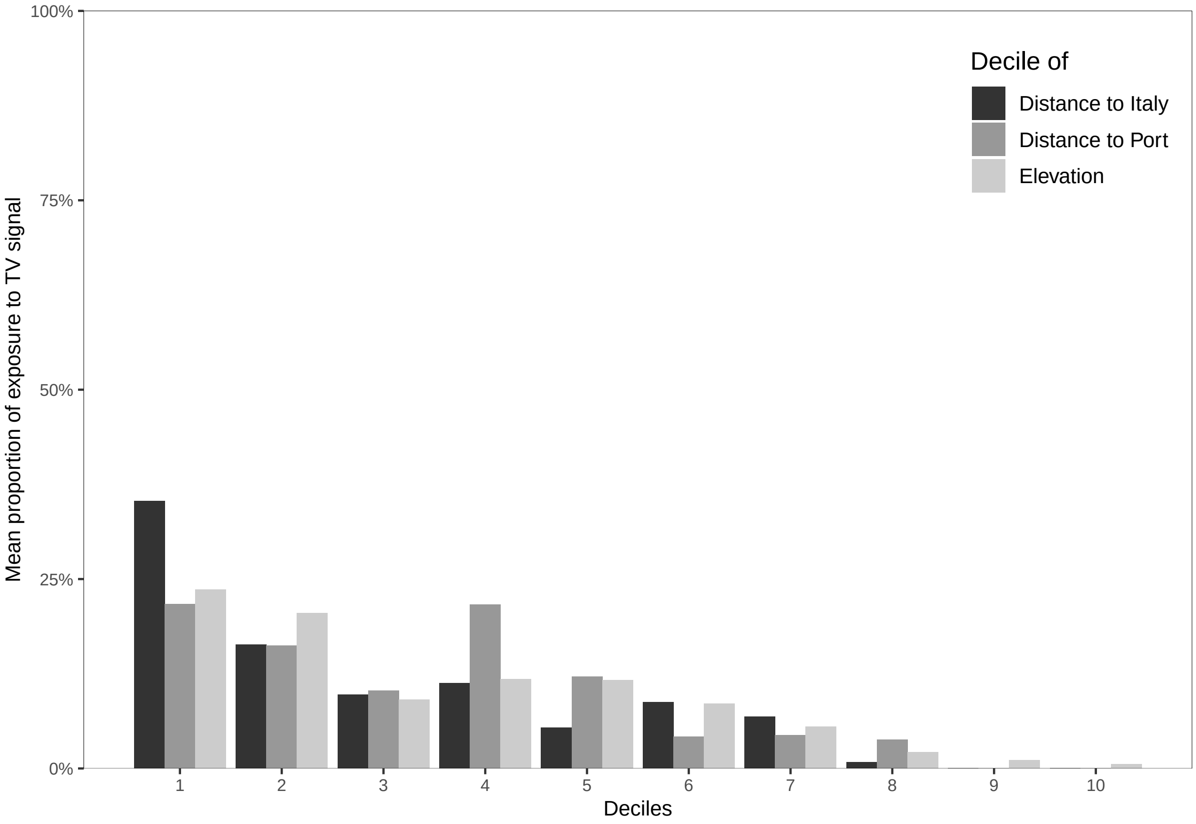
The results presented thus far suggest that neither the informational channel nor the belief channel played any role in fostering emigration towards Italy of the highly-skilled Albanians. We conclude this section by underlining the role played by language proficiency which, given the language-skill complementarity (Chiswick, 1995; Berman et al., 2003), raised the returns to migration of high skilled individuals.

1.9 Robustness

This section presents robustness tests of our results. First, we show that our results do not depend on the higher exposure of the Albanian coastal areas to the Italian television signal. Second, we show that our results are not sensitive to the exclusion from the sample municipalities closest to ports and closest to Greece, where migration costs were low. Finally we test that results are robust to alternative identifications of high-skilled individuals, and to different definitions of signal exposures.

Since television access is concentrated in the coastal areas, a concern surrounding the identification strategy is the high correlation between signal power and distance measurements. As the latter are directly related to migration costs (the further away from Italy, the more complicated to migrate), high levels of correlations might result in spurious estimations. Figure 1.2 comes to alleviate this concern: although most of TV exposure is concentrated in municipalities within the first deciles of the distances

Figure 1.2: Radio Signal and topographic data



Notes: Each bar represents the mean share of municipalities' areas under the signal for municipalities that can be found in the decile of the relevant topographic variable considered.

to the closest port and to Italy, there is significant variation in signal exposure between municipalities in all deciles up to the 7$^{\text{th}}$.

Another concern is that the inclusion of municipalities of coastal areas and bordering Greece puts in the sample individuals unlikely to be impacted by Italian television: their migration costs might be so low that there is little role left for television. In Table 1.5, we limit the sample to highly-skilled individuals that lived more than 30km from a port (1st quartile of distance to ports) and more than 48km for the Greek border (1st decile of distance to Greece). Results dispel all doubts related to spurious estimation: the coefficients of interests are still precisely estimated, significant at the 5% level for migration abroad, and at the 1% level for migration to Italy. It is worth noting that in this regression exercise, the effect of the signal on migration and the effect on migration to Italy collapse to the same point estimate. It suggests that the difference between the two coefficients we observed in Table 1.3 is due to the presence of *always-takers* who would have migrated even if they had not been exposed to the signal.

Table 1.6: The effect of Italian Television exposure on migration decision: alternative variables definition

	Siblings Dataset						
	Abroad	Italy	Abroad	Italy	Abroad	Italy	Abroad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Signal II	0.148**	0.139**					
	(0.0742)	(0.0638)					
Signal			0.154**	0.0819*	0.177***	0.104**	0.221**
			(0.0690)	(0.0472)	(0.0592)	(0.0450)	(0.0958)
Sample	High Skill	High Skill	Small Fam.	Small Fam.	Wealthy	Wealthy	H. Skill < 1990
Observations	2153	2153	2449	2449	4043	4043	1510

Common Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness

Notes: The table reports OLS estimates of the effect of exposure to Italian TV on the probability to be abroad and in Italy. It replicates specifications and results of Table 1.3 with alternative definitions of signal exposure and human capital using the siblings dataset. In particular, specifications (1) and (2) repeat specifications (2) and (3) of Table 1.3, but use 1986 municipalities' share of urban area exposed to the signal as explanatory variable instead of the usual signal definition. Specifications (3)-(6) exploit the usual signal variable but change the definitions of high skilled individuals. (3)-(4) subset the sample by for family with less than 4 children, while (5) and (6) subsets for individuals that lived in an apartment in thev 4th quartile of the distribution of the number of rooms per person in 1990. Controls are as defined in notes Table 1.3. Specification (7) identifies an individuals as high skilled individuals if she attended university and completed her education prior to 1990. Clustered standard errors in parentheses. Standard errors are clustered at the municipality of residency in 1990 level (# of clusters: (1) and (2) 128; (3) and (4) 243; (5) and (6) 240; (7) 107. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

As an additional robustness check, we test whether our results depends on the specifications of either TV signal exposure or the measures of human capital. We first vary the measure of TV signal exposure, while we previously used the share of a municipality's land exposed to the signal, we here use the share of the municipality's urban area (in 1986) exposed to the signal. This avoids accounting for signal reaching inhabited rural areas. Results of this exercise are presented in table 1.6 columns (1) and (2). With this refined measure, using as dependent variable migration abroad or specifically to Italy yield the same results of 14 percentage points, thereby confirming earlier results.

In the main specification, we identify an individual as highly skilled if her listing sibling attended university at least one year. In this robustness check, we use alternatively family and dwelling dimension, characteristics we can measure for the sibling herself, as a proxy for education. Large families are more likely to come from agrarian backgrounds, where it is less likely that children can be sent to university.

We thus split the sample according to family size, identifying a family as small if it is composed of less than four children. Columns (3) and (4) implement this sample split, repeating main specification (1.1). Results are qualitatively similar, although less precise. In Appendix A.6.3 we show that the effect of TV exposure decreases in the family dimension. We define the dwelling dimension as the ratio of the number of rooms to the family dimension, yielding the number of rooms per person. Much like for family dimension, we assume that housing size is correlated with education. In (5) and (6), we subset the siblings sample for the fourth quartile of the distribution of housing dimension and confirm our baseline results. Appendix A.6.3 shows the alternative regressions for smaller dwellings. These robustness exercises confirm that our results do not hinge on the definition of either TV signal access or on human capital.

1.10 Conclusion

How much does foreign language proficiency affect individuals' migration probabilities? Answering this question is relevant for policy makers deciding whether to promote plurilingualism in society. It is also relevant for understanding the causes of brain drain. So far, this question has only been addressed by observational studies, unable to address the inherent self-selection issues that characterize these settings.

In this paper, we exploit a natural experiment that occurred in Albania in the second half of the twentieth century to assess the causal effect of foreign language proficiency on high-skilled migration. We show that as good as random exposure to Italian television increased by at least 7 percentage points Italian language proficiency and by 24 percentage points the likelihood of migration of high-skilled individuals in fully exposed municipalities. We interpret the effect of signal exposure on foreign migration as the effect of higher language proficiency and rule out competing channels.

While our study contributes novel insights by establishing the causal impact of foreign language proficiency on the migration of highly-educated individuals and documenting the effects of foreign media on language proficiency, there are limitations to our analysis. It is restricted to estimating a reduced-form equation of the impact of language proficiency on emigration, as we cannot estimate an instrumental variable regression due to data limitations. Therefore, we provide a lower bound estimate of the impact of TV exposure on language proficiency, and as a result, we can only offer a lower bound estimate of the effect of language proficiency on the likelihood of emigration.

The economic literature still presents diverging results as to the effects of brain drain on the economy Shrestha (2017); Anelli et al. (2023), we leave for further research the evaluation of the impact of the Albanian brain drain on Albania's economic development.

Table 1.4: 2005 LSMS, Selected Statistics

	Base Dataset
Variable	Share
Information provider:	
Family/relatives in Albania	0.03
Family/relatives abroad	0.30
Friends in Albania	0.14
Friends abroad	0.41
Previous personal experience	0.08
Neighbours	0.02
TV, radio, newspapers	0.01
Internet	0
Others	0.01
Owns a TV in 1990 by education:	
Primary or less	0.57
Secondary	0.68
Vocational	0.68
University	0.78
Emigrated by education:	
Primary or less	0.14
Secondary	0.20
Vocational	0.20
University	0.21

Table 1.5: Effect of Italian Television Exposure on Migration Decision. Sensitivity to Coastal Areas and the Greek Border.

	Siblings dataset		
	Abroad	Italy	Greece
	(1)	(2)	(3)
Signal	0.168^{**}	0.168^{***}	-0.0334
	(0.0675)	(0.0519)	(0.0792)
Sample	High Skill	High Skill	High Skill
Observations	1476	1476	1476
Clusters	72	72	72

Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness

Notes: The table reports OLS estimates of the effect of exposure to Italian Television on the probability to be abroad. Outcome variable is a dummy taking value 1 if abroad (1), in Italy (2), in Greece (3). All specifications exploit the siblings dataset (Section 1.5), restrict the sample to individuals whose *listing sibling* attended university for at least one year. Signal is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Municipalities that are less than 30 Km of distance (1st quartile of distance to ports) and less than 48 km from the Greek border are removed from the sample. Clustered standard errors at the municipality level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Chapter 2

Machine Data: market and analytics

Co-written with **Giacomo Calzolari** and **Riccardo Rovatti**

2.1 Introduction

Machines and industrial equipment, e.g. in manufacturing and agriculture, continuously produce a vast amount of *machine data* (MD hereafter), also known as non-personal or industrial data. These data have an enormous potential for more efficient production and management, new and high-performance machines design, and, ultimately, cheaper and better products for final consumers. Recent technological developments, namely the Internet of Things (IoT), 5G transmissions, and Artificial Intelligence (AI), are now putting MD at the forefront, surpassing in terms of size and economic value the very much debated personal data. In this paper, we provide a first detailed and formal analysis of MD and discuss how the development of a market for MD and MD analytics should not be taken for granted. Our analysis can help explaining why only a limited amount of MD is currently used with an untapped potential.[1] We also clarify the implications of what AI experts (e.g. Andrew Ng in Hao (2021)) now consider the main difficulty in the next frontier of AI, AI-for-industry, that is the difficulty of dealing with relatively small and dispersed industrial datasets.

Industrial activities generate MD as a byproduct, such as the data generated by an electric motor or a welding machine. Collecting these data requires sensors that monitor several parameters and the transmission and storage of these data. The raw data can then be transformed into useful information for prediction and decision-making with *data analytics*, i.e., mathematical methods leveraging statistics and, more recently, Machine Learning tools. For example, with data from the electric motors, one

[1] For example, in 2022 the European Commission estimated that only 20% of industrial data was currently used, and helping the development of a market for MD may create value for €270 billion of additional GDP by 2028.

could identify and avoid the working conditions of a similar motor that generate the highest stress and increase failure probability, provide precise predictions to optimize maintenance, and even design more robust motors. However, transforming raw data into valuable information requires a (i) sufficiently large amount of data and (ii) data possibly collected under diverse conditions covering multiple working configurations. The first element refers to a *Scale property* of Machine Learning tools that typically show increasing returns at low levels of data, followed by decreasing returns and saturation with a large amount of data. The second element instead refers to a *Scope property*, or synergy of data sources, according to which the value of data increases when relying on different data sources. Although there is commonly accepted consensus about these two properties in the computer science literature, we provide novel evidence with Machine Learning classifiers, which could be of value *per-se* (section 2.1.1). We investigate and combine these technical properties of data analytics, Scale and Scope, with three critical economic and organizational characteristics.

First, as with any other type of information, MD and analytics are a *semi-public good*: they are non-rival (they can be re-utilized with no deterioration of information content) and excludable (one can grant access to some users and exclude others). For example, the analytics that a firm uses to manage its electric motors could benefit from the data of the electric motors of other firms, which in turn could benefit from the same analytics as it would be with “enabling technologies” (Gambardella et al., 2021). Since individual firms may fail to account for the positive externality of their MD for other producers, data production and sharing can be suboptimal. In the case of MD, this is even more problematic because, for the Scope property, the value of MD is further enhanced when they originate from different sources.

Second, large amounts of raw data are currently fragmented into a myriad of machines located in equally many firms, some of which are small and medium enterprises. *Data fragmentation* is a significant problem when combined with externalities and Scale and Scope. In addition, data analytics implies non-negligible fixed costs, which can make in-house analytics too expensive, especially for small producers. With increasing returns to scale at a low scale, a market for MD may end up with large amounts of stranded data and, at the same time, very high levels of concentration, possibly replicating the lock-in and market tipping currently observed with personal data.

Third, there is presently *no clear assignment of property rights* of MD. Different subjects could claim ownership of MD: the firms using the machines, the machine manufacturers, the “retrofitters” placing sensors on machines, or the data aggregators that collect data from different sources and run the analytics (such as the company Machinemetrics).[2] In this situation, firms rely on bilateral agreements that reflect

[2] So far the European Parliament (2018) refrained from assigning (*in rem*) property rights, and instead focused on the possibility that firms would share MD. The European Commission recently proposed a new regulation (the Data Act, February 2022) on industrial data. It prescribes that primary data holders (e.g. machine manufacturers) have no exclusive right to MD and must grant adequate access to MD if requested (e.g., by producers using the manufacturers’ machines), with compensation that may cover the cost incurred for making the data.

relative bargaining power and significant transaction costs. Contracts for sharing MD and analytics may be incomplete, exposing parties to excessive risks and unforeseen contingencies, limiting their potential, as we discuss in this paper.

Combining Scale and Scope with MD externalities, fragmentation, and lack of ownership makes for a rich and novel environment.[3] Our analysis provides what is, to our knowledge, the first detailed and formalized investigation of the organization of a market for MD. At the same time, we show how one can effectively study this market and its complexity by combining traditional tools of the managerial and economics literature. In particular, we consider a scenario with *data producers*, i.e. companies that generate MD with their production, and a *data aggregator* that collects the data and provides data analytics valuable to producers. Although data producers are *de-facto* owners of MD (they can exclude others from access to their MD), they may be too small to extract information profitably due to varying returns to scale and fixed costs. Instead, a *data aggregator* can pool MD from several data producers and profit from the combination of Scale and Scope. To do so, the aggregator must convince producers to accept a contractual offer contemplating the sharing of MD, a data analytics service that increases producers' profitability, and a monetary transfer. Relying on shared data and analytics, we dub this endeavor "cooperative analytics for MD".

We address several important and new questions. In particular, we identify market features that may facilitate or hinder the development of MD analytics. Even if the analytics were offered free of any access charges (being thus subsidized), data producers would fail to internalize the external value of their data, resulting in underprovision of data. A data aggregator can redress this classic public-good issue by incentivizing data provision, running the analytics for profit or breaking even. We show that data producers paying a fee to join the cooperative analytics are those with high value for it, while those with relatively small value are subsidized because their data have a sizeable collective value for other producers. Moreover, when producers can run their analytics in-house, the aggregator may profitably operate by collecting data from a selected group of small and more homogeneous producers, disregarding larger companies.

We investigate the consequences of imprecise allocation of ownership rights. When joining the analytics and sharing their data, producers may risk that critical information concerning production leaks out, possibly to rivals (or suppliers who could exploit it). This risk and the associated costs are especially relevant when property rights are not well-allocated, as with MD.[4] To limit this risk, producers may

[3] The combination of these elements makes MD different from personal data (as recognized, for example, in European Parliament (2018)), although in some cases this distinction is blurred (Graef et al. (2018)), such as with data from human-machines interaction such as with data from batteries of electric cars.

[4] As a part of the European Strategy for Data, a recent regulation (the Data Governance Act, approved in May 2022) sets the duties of data intermediary services, such as our data aggregators, including the obligations for ex-ante compliance and transparency. This regulation was intended to limit the risks of misuse and loss of competitive advantage, thus reducing firms' worries when sharing their data.

require *anonymity* the contractual offer for the cooperative analytics and MD sharing. We show that anonymity seriously constrains the value of the analytics, and in some cases, it may even lead to a complete market breakdown. Even if this does not occur, we show that, accounting for anonimity, the data aggregator prefers avoiding pooling data from dissimilar data producers.

We then consider producers that compete in related markets. We allow the degree of competition to vary in terms of final-product substitutability, and we study the implications for the market of MD. We show that the aggregator can end up playing the role of coordination device among competing data-producers with an inefficient analytics and product markets. Moreover, too-intense competition leads to the breakdown of the analytics or the exclusion of some producers. The aggregator thus prefers running a cooperative analytics with firms that are not too close competitors but also not in entirely unrelated markets, with a preferred intensity of competition that systematically diverges from the (socially) efficient one.

Overall, these results provide a rich picture of the possibility of obtaining a market solution to analytics for MD. Although with careful contracting, an aggregator could address some of the issues with MD and offer valuable analytics services to producers, some significant inefficiencies emerge. We think these results could also help inform a policy agenda on how to support a market for MD and analytics.

A specificity and novelty of our paper is that we combine market analysis with a modelling of Machine Learning and its properties that closely reflects actual AI algorithms. Methodologically, we think this approach is valuable *per-se* as we can rely on detailed properties of AI algorithms that are realistic and yet tractable with a combination of theory and simulations.[5]

2.1.1 Literature

Although the policy debate around MD analytics is quite active (e.g. European Commission (2017), and Duch-Brown et al. (2017)), the academic literature is scant. In economics, Farboodi et al. (2019) studies an environment where production generates data, but producers can only rely on internal analytics. We differ from this approach focusing on the market for MD and analytics, where an aggregator can offer analytics services but must convince data producers to join. Considering data about consumers' preferences, other papers have investigated how these data affect production. Prüfer and Schottmüller (2020) studies a dynamic model where the quality-cost of products reduces with the amount of data about consumers' preferences, as with learning-by-doing. Jones and Tonetti (2020) studies the sharing of personal-data in a macroeconomic growth model with innovation and studies the role of privacy

[5] The benefits of this novel approach of studying the impact of actual AI algorithms in markets with simulations have been popularized by Calvano et al. (2020) and, more recently, Johnson et al. (2023).

regulations. Our analysis differs because we consider MD and we focus on market organization.[6]

The semi-public good characteristic of MD relates our analysis to the economics and managerial literature on excludable public goods (or "club goods"), e.g. Anderson et al. (2004) and Cornes and Hartley (2007). When considering data producers that are competitors in related markets, our model shares similarities with competition models with R&D spillovers and research joint ventures (see Amir et al. (2019) for a recent account). Beside the loose connection, and unlike these two strands of literature, we focus on the possibility of providing MD analytics with a market-based solution and identify market characteristics that facilitate or hinder the provision of MD analytics. Related issues for innovators of "enabling technologies", such as MD analytics, have been recently discussed in Gambardella et al. (2021). We complement their approach with a formal analytical framework, considering a technology (i.e. the analytics) that is available but must be fed with data.

Some scholars, mainly in the legal literature, have recently discussed (often against) an assignment of property rights of MD that would lead to the rights to exclude others in using MD (e.g. Zech (2016), Kerber (2016), and Drexl (2016)). It has been argued that adapting existing approaches, such as intellectual property rights, to MD would be either inappropriate or ineffective.[7] Thus, MD are currently managed with contractual agreements and technical measures against misappropriation. In this paper, we rely on this status-quo, with *de-facto* property rights of MD to data producers. We contribute to this debate by studying the market for MD analytics and explicitly accounting for the costs that data producers and aggregators face in litigations.

Finally, the implications (advantages and difficulties) of combining large amounts of data and from multiple sources have been investigated in the Computer Science literature, e.g. Mitchell (1999), and the recent surveys Alam et al. (2017) and Meng et al. (2020). Although there is consensus in this literature about Scale and Scope (on this see also Duch-Brown et al. (2017) and Schaefer and Sapi (2022) in economics), precise and neat accounts of these properties are scant. In the paper, we show how to combine the value of information from different sources coherently in a model for the analytics and we illustrate common classification algorithms in Machine Learning that exhibit these properties (Appendix B1).

The paper is organized as follows. Section 2.2 lays out the baseline model. In Section 2.3 we discuss

[6] Bergemann et al. (2019) and Ichihashi (2020) have studied competition between brokers reselling consumer data to downstream competitors.

[7] For example, this would be the case for the difficulty of proving novelty and originality using IPR. Other approaches seem ineffective too. The copyright protection of databases (e.g. European Parliament and Council of the European Union (1995) in the EU) cannot protect the data but only the investments to aggregate pre-existing data. Leakage of MD would also not be protected under trade secret law, which does not grant exclusive property rights and would require proving effort in keeping MD secret).

some benchmarks. The market-based analytics with the organization of an aggregator is in Section 2.4. In Section 2.5 we show the difficulties in dealing with anonymity. Section 2.6 discusses cooperative analytic with producers that compete for consumers and Section 2.7 concludes with a discussion of developments and future research. In the text we summarize simple but relevant observations with Remarks. The proofs of Lemmas and Propositions are in the Appendix, where we also discuss and illustrate the properties of Machine Learning algorithms for data analytics that we use in the analysis.

2.2 The Baseline Model

Each of P producers obtains a value from a data-analytics, paying access fees and contributing with own data. The payoff of producer i when providing n_i units of data is,

$$B_i = \alpha_i \eta\left(n\right) - \gamma_i n_i - Q_i, \tag{2.1}$$

where $\eta\left(n\right)$ is the *value of the analytics* that relies on data $n = \left(n_1, \ldots, n_P\right)$ from (up to P) distinct sources that we discuss below. Parameter $\alpha_i \geq 0$ measures the ability of the i-th producer in transforming the analytics into profits, $\gamma_i n_i$ is a cost for handling and sharing own data n_i, and Q_i is a transfer payment for the analytics and the data, which can be negative in case producer i receives a net contribution for the data. Producer i is willing to participate in the data-aggregation agreement if B_i is higher than an alternative payoff obtained with no participation, $B_i^{\min} \geq 0$. In the first part of the paper, we consider producers that operate in different final product markets such that B_i and $B_i^{\min}$ do not depend on other producers' analytics and data. In section 2.6, we will instead consider competitors in the related markets.

A single data aggregator collects data and fees from producers, performs an analytics incurring operating costs and obtains a payoff,

$$B_{\text{agg}} = \sum_{j=1}^{P} Q_j - \bar{\delta} - \delta \sum_{j=1}^{P} n_i - \varepsilon \eta\left(n\right) \tag{2.2}$$

where $\bar{\delta} + \delta \sum_{j=1}^{P} n_i$ is the fixed and variable cost to produce the analytics and $\varepsilon \eta\left(n\right)$ is a cost related to managing the analytics of a given value $\eta\left(n\right)$. The aggregator is willing to operate on producers' data and provide the analytics only if his payoff is larger than some minimum value, i.e. $B_{\text{agg}}^{\min} \geq 0$.[8]

We consider the following sequence of events:

1. The aggregator proposes the contract for the analytics and the payment Q_i with each producer i.

[8] We will also discuss the case in which the analytics is designed to maximize welfare, the sum of the aggregator's and the producers' payoffs.

2. Each producer i refuses or accepts the contract, in which case he provides n_i units of data to the aggregator.

3. The aggregator receives the data, generates the analytics with value $\eta(n)$ and provides it to the accepting producers.

4. Contracts and payments are executed, costs and payoffs realized.

We will illustrate some of the results with a running *Example* where there are two producers ($P = 2$), producer 1 having a higher value for the analytics, in the Example (the code to reproduce all figures and simulations is available).

Assumptions and Interpretations.
The value of the analytics. We rely on features of common classification and prediction algorithms in Machine Learning and obtain the value of the analytics $\eta(n)$ and its properties in two steps. First, the gain obtained processing data from a single producer is represented with a non-decreasing function $\upsilon : \mathbb{R}^+ \mapsto [0, \eta^{\max}]$, where $\eta^{\max}$ is the maximum value attainable with a single data source. Then we combine data from different producers.

We assume that for large values of its argument $\upsilon(.)$ is concave to model that when most of the data space has been sampled, the marginal gain of further data lots decreases, as with standard classifiers. For small values of its argument, $\upsilon(0) = 0$ and $\upsilon(.)$ is convex modelling that the marginal gain of the very first data lots is smaller than the marginal gain of the subsequent ones. In fact in Machine Learning the very first data slots are used to tune the parameters of the algorithm, with very limited value. Subsequent slots allow then to produce predictions, thus significantly increasing the gain (up to the concave region discussed above). Furthermore, data augmentation techniques (Mumuni and Mumuni, 2022) with small data allow to obtain a value form the analytics also at very low levels of data, thus implying $\upsilon'(0) > 0$. When $P = 1$, simply have $\eta(n_1) = \upsilon(n_1)$. When instead $P > 1$, the interaction between different datasets becomes relevant. To account for this new element we consider a monotonically increasing convex function $\Upsilon : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\Upsilon(0) = 0$ and define the commutative and associative aggregating operation $\oplus : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that,

$$\upsilon' \oplus \upsilon'' = \Upsilon\left(\Upsilon^{-1}(\upsilon') + \Upsilon^{-1}(\upsilon'')\right).$$

The total gain associated to a set of data contributions is thus,

$$\eta\left(n_1, \ldots, n_P\right) = \bigoplus_{j=1}^{P} \upsilon\left(n_i\right) = \bigoplus_{j=1}^{P} \eta\left(n_i\right) \tag{2.3}$$

This two-steps procedure delivers a number of important properties of the analytics' value that we discuss next. In Appendix B1 we instead rely on a theoretical analysis and numerical simulations showing that standard classification algorithms do feature these key properties.[9]

Clearly, there is no value with no data, $\eta(n) = 0$ for $n = (0, \dots, 0)$. All else equal, more data from a given dataset increases the analytics' value: $\eta(n)$ is increasing in n_i, and, (assuming differentiability for ease of notation) $\frac{\partial\eta(0,n_{-i})}{\partial n_i} > 0$ for any n_{-i}. The function $\eta(n)$ is also convex with small amounts of data and asymptotically concave for large n, which we dub the *Scale property* of the analytics. It also features a *Scope property* which combines increasing difference and superadditivity (formally stated in Appendix A2). The former means that the higher value obtained from more data of a given dataset is enhanced when it is combined with more data from another dataset. The latter means that joining two (or more) datasets into single analytics provides a higher value than the sum of the values of separate analytics (i.e. relying on different datasets generates economies of scope). In other terms, the Scope property states that more data diversity maps the data space more effectively.

Data and analytics costs. The producer's cost for handling own data $\gamma_i n_i$ contemplates both industrial costs for acquiring and transmitting data and indirect costs associated with the risks of sharing data outside the firm. Since MD are only *de-facto* protected, producers face the risk that information about their production process leaks from shared data and must put some effort into limiting litigation issues over data. Overall, the weaker is *de-facto* protection of MD, the higher the *anonymity cost* γ.

The aggregator's (marginal) cost of handling data from different datasets is δ, which refers to industrial costs and legal litigation costs for weak protection (proportional to the size of the dataset). Setting up an analytics involves a fixed cost $\bar{\delta}$, which may be large enough to make in-house analytics unprofitable when relying only on internal data. For most of the analysis, we assume $\bar{\delta}$ is large enough so no producer can independently operate its own analytics.

The aggregator also faces costs for managing the analytics, proportional to its underlying value $\varepsilon\eta(n)$. For example, with valuable analytics, the aggregator may face high (expected) costs for legal disputes concerning its ownership, especially in a weak property-rights environment. Similarly, the higher the analytics' value, the stronger the risk that the aggregator faces cyber-attacks, and the data and the analytics may become public.[10]

Each producer is endowed with a certain (large) amount of data. We will discuss later on the possibility to associated data with actual production.

Analytics appropriability, contracts and market structure. The analytics is a *semi-public good* (or club good), being excludable but non-rival. The aggregator can exclude producers from the benefits and value

[9] Although there is consensus about these properties for Machine Learning classifiers, this appendix provides practical and direct illustrations that we think are of interest *per-se*.

[10] With an alternative interpretation, the aggregator may directly benefit from the analytics, in which case $\varepsilon < 0$, for example, when the aggregator manufactures and sells machines that produce the data.

of the analytics (e.g. if they do not provide data), and each joining producer benefits from the analytics without degrading its value for other joining producers. Since we focus on relatively small producers, we consider a monopolistic aggregator with bargaining power to design the contractual offers for data-sharing and analytics. On the other hand, our producers are *de-facto* owners of the data, they are free to refuse the aggregator's offer and make data n_i unavailable for the analytics.

The model is flexible enough to account for different market structures. For example, we could consider the manufacturers of the machines. These players would directly access the producers' data to whom they sell machines. A manufacturer that sells machines to all producers would act as the data aggregator described above, with the difference that he would not need the consent of producers to access the data. We will discuss these possibilities and their implications.

We assume payoffs are common knowledge.[11] Since third parties typically observe the realized value of the analytics to a given producer i with some error, we assume that producers and the aggregator cannot rely on contracts and payments that explicitly depend upon the realized value of the analytics.[12] We begin assuming that the contracts offered by the aggregator are publicly observable. We then investigate the case in which producers require contractual contractual anonymity to join the analytics.

The running "Example". To illustrate some of the results we will rely on an example of the model that accommodates all its key elements. The Example contemplates two producers ($P = 2$), producer 1 having a higher value for the analytics, i.e. $\alpha_1 > \alpha_2$, but also a higher cost for handling data, i.e. $\gamma_1 > \gamma_2$. We dub producer 1 (2) as the strong (weak) producer. All other details and parameters of the Example are in Appendix A2.1. All figures refer to the Example and, in particular, Figure 2.1 illustrates the associated value of the analytics $\eta(n_1, n_2)$.[13]

2.3 Benchmarks

For future reference, we discuss some useful benchmarks.

Feasible and efficient analytics. For given data $n = (n_1, \ldots, n_P)$, the surplus of the analytics is:

$$W(n) = B_{\text{agg}} + \sum_{j=1}^{P} B_j = \eta(n)(\sum_{j=1}^{P} \alpha_j - \varepsilon) - \bar{\delta} - \sum_{j=1}^{P} (\delta + \gamma_j) n_j \tag{2.4}$$

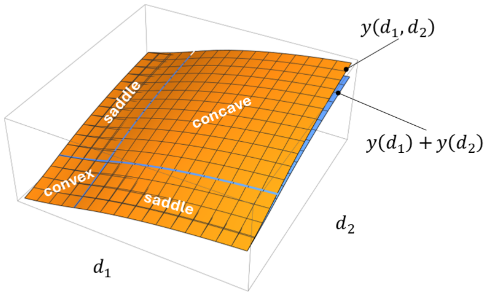
We define an *feasible analytics* as the combination of n such that $W(n) \geq B_{\text{agg}}^{\min} + \sum_{j=1}^{P} B_j^{\min}$. In the Example with identical marginal costs, Figure 2.2 shows the combinations of n such that the surplus is

[11] We leave for future research the interesting case in which α_i is producer i's private information.

[12] See Dosis and Sand-Zantman (2019) for a theory of personal data sharing with incomplete contracts and the hold-up problem.

[13] The code to replicate all the figures and numerical results is available upon request.

Figure 2.1: Value of analytics $\eta(n_1, n_2)$ in the Example (details in Appendix A2.1) showing Scale and Scope.



equal to the outside options of producers and the aggregator for several values of the analytics' marginal costs (lower marginal costs in thicker lines).

The properties of the analytics imply the following.

Remark 1. *A feasible analytics requires (i) a minimal size of data, (ii) not too many data, (iii) balanced datasets when many date are available, (iv) unbalanced datasets with few data.*

Conditions (i) and (ii) are direct consequence of Scale: the marginal benefits are smaller than marginal costs with little or too many data. Condition (iii) follows from Scope. It can be seen with the increasing boundaries in the northwest/southeast parts of the Figure where further increases of the "over-represented" dataset makes the analyitcs not feasible, as well as in the northeast declining boundary when slightly unbalancing large datasets. Condition (iv) shows up in the southwest portion of the figure, where increasing one of the small datasets grants higher marginal benefits than costs at least with one source of data.

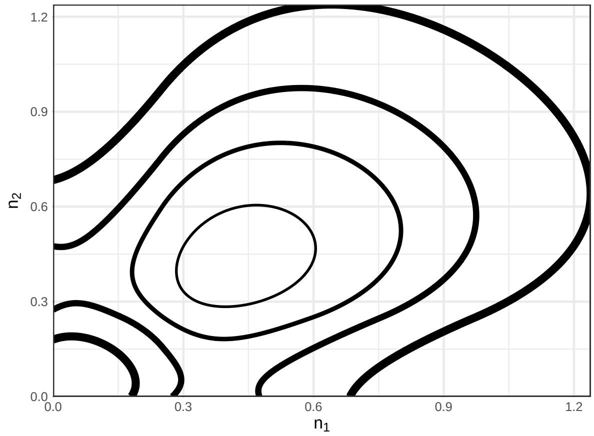
We indicate the amount of data that maximizes $W(n)$ with $n^{\mathrm{w}} = (n_1^{\mathrm{w}}, \ldots, n_P^{\mathrm{w}})$ and the associated maximal surplus with $W^{\max}$. At an interior solution, n^{w} is implicitly identified by,[14]

$$\frac{\partial \eta(n)}{\partial n_i}\left(\sum_{i=1}^{P} \alpha_i - \varepsilon\right) = \delta + \gamma_i, \quad i = 1, \ldots, P. \tag{2.5}$$

The cost of procuring and handling the marginal unit of data of producer i (the r.h.s. in condition (2.5))

[14] Given the properties of $\eta(n)$, condition (2.5) can realize when $\eta(n)$ is convex in n_i, or when it is concave in n_i. However, the former case would correspond to a minimum of W.

Figure 2.2: Feasible analytics (i.e. combinations of n_1, n_2 s.t. $W(n) \geq B^{\min}_{\text{agg}} + \sum_{j=1}^{P} B^{\min}_{j}$) with thicker lines associated with lower (symmetric) marginal costs.



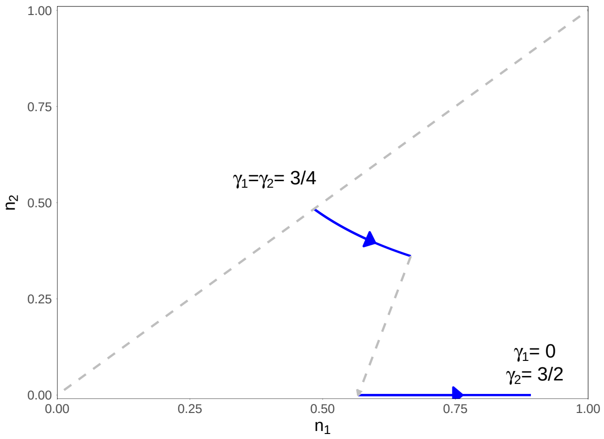
is equal to its value (the l.h.s.). The latter accounts for its positive impact for *all* producers due to the public-good nature of data. A relevant implication of this property is that the efficient amount of data n_i^{w} that satisfies the internal solution of (2.5) is the same for any producer i whenever $\gamma_i = \gamma$ for any i. Hence, despite extracting less value from analytics, a producer with a low α_i but also a low cost for data should contribute with relative large amount of data. Also, the optimal amount of data from two producers can only differ if their costs do so. The Example shows the implications of asymmetric costs configurations with the Scale and Scope properties. Perturbing the (initially) symmetric marginal costs of the producers with $\gamma_1 = 3/4 - \Delta$, $\gamma_2 = 3/4 + \Delta$ and $0 \leq \Delta \leq 3/4$ we obtain the locus of optimal n^{w} as in Figure 2.3.

The following summarizes these simple but remarkable properties of the efficient amount of data.

Remark 2. *For interior solutions, i.e. $n_i > 0$ for any i, the efficient amount of data does not depend on the distribution of the producers' value of the analytics, $\alpha_i, i = 1, ..., P$. Producers with lower costs γ_i, contribute more data independently of their value of the analytics.*

Figure 2.3 shows another interesting property, unbalancing marginal costs may cause the optimal solution to be non-internal, and the producer with the higher marginal cost does not to provide data (while still enjoying the analytics). The discrete drop in data of producer 2 in the figure is a remarkable and general outcome of the Scale and Scope properties, that we will further discuss. When γ_2 increases, the reduction of n_2 is smooth, up to the region in which the value of analytics $\eta(.)$ becomes convex for low n_2. At that point, with further increases of γ_2 condition (2.5) would identify a minimum due to the convexity of $\eta(.)$. When the data of producer 2 drop to zero, also n_1^{w} reduces discretely (with $n_2 = 0$ the marginal value of n_1 reduces since the synergy between datasets is lost).

Figure 2.3: Efficient data when costs γ_1 and γ_2 respectively decrease and increase by a same amount.



The convexity of the value of the analytics at a small amount of data also implies that relatively inefficient producers do not provide data, even if they benefit from the analytics. Relatedly, active producers must provide relatively large batches of data. This is stated in the following Lemma which is general and does not only apply to the efficient analytics.

Lemma 1. *When a producer is active with the analytics, i.e. it provides data to the analytics ($n_i > 0$), the efficient amount of data is bounded away from zero.*

Free analytics. Assuming the analytics is freely available, each producer i would decide the amount of data to share solving the following problem,

$$\max_{n_i} \quad B_i \tag{2.6a}$$

$$\text{s.t.} \quad B_i - B_i^{\min} \geq 0 \tag{2.6b}$$

(where we set the payment $Q_i = 0$). Expecting $\hat{n}_{-i}$ data from other producers, producer i would choose n_i satisfying the following optimality (interior) condition,

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} \alpha_i - \gamma_i = 0. \tag{2.7}$$

This condition implicitly defines the best response for producer i, the optimal amount of data n_i in response to the expectation of some $\hat{n}_{-i}$. The Nash equilibrium of the game between producers that independently decide how much data to share with the free-of-charge analytics is an n^0 such that for any producer i, (2.7) is satisfied at $n_i = n_i^0$ and $\hat{n}_{-i} = n_{-i}^0$.

Comparing the optimality conditions (2.7) with those that guarantee the efficient analytics (2.5), independent producers fail to consider the positive effect that their own data have on the value of the analytics to all other producers, i.e., $\frac{\partial\eta(n)}{\partial n_i}\sum_{j\neq i}^P \alpha_j$. Since the analytics is free, they also fail to internalize the cost of processing data and handling the analytics $\frac{\partial\eta(n)}{\partial n_i}\varepsilon + \delta$. Whether independent producers provide more or less data than what would be efficient depends on the composition of these two elements and the sign of

$$\Delta_i^0 = \frac{\partial\eta(n^0)}{\partial n_i}\sum_{j\neq i}^P \alpha_j - \left(\delta + \frac{\partial\eta(n^0)}{\partial n_i}\varepsilon\right) \tag{2.8}$$

As η is concave and saturating, its derivative eventually vanishes so that, if n_i^0 is large enough, $\Delta_i^0 \to -\delta < 0$, causing over-provision of data, $n_i^0 \geq n_i^{\text{w}}$. This latter case is likely to happen when the distribution of the producers' value of the analytics $\alpha_i, i = 1,...,P$ is particularly unbalanced. Figure 2.4 shows these possibilities depicting the best response curves, the Nash equilibrium and the efficient data. In the Example, the ability of extracting value from analytics of producer 1 ($\alpha_1 = 1$) is larger than that of producer 2 ($\alpha_2 = 1/4$). For producer 1, the balance between the non-internalized value of the analytics and costs depends on the feature α_2 of the other producer 2 and is negative $\Delta_1^0 \simeq -0.542 < 0$, causing n_1^0 to be larger than n_1^{w}. Vice-versa, for producer 2 we have $\Delta_2^0 \simeq 0.167$, causing n_2^0 to be smaller than n_2^{w}.

Free analytics suffers from another critical problem in addition to inefficient data provision: multiple equilibria. Expecting no other producers to provide data, a producer may prefer not to confer any data as well. Hence, expectations about other producers' data provision matter dramatically with free analytics or with analytics where an aggregator does not control data producers' incentives: pessimistic beliefs can drive a free analytics to complete breakdown.

The following summarizes these observations.[15]

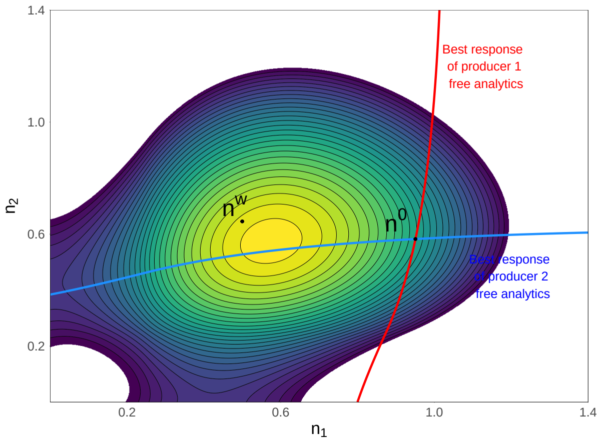
Proposition 1. *With free-analytics, (i) at least one equilibrium exists; (ii) producers that value the analytics less (more) under-provide (over-provide) data with respect to the efficient analytics; (iii) with sufficiently high costs for data producers, multiple equilibria exist, including one where the analytics does not operate.*

2.4 Market-based analytics

A data aggregator may step in, motivated by own or producers' profits. This aggregator could be an independent actor or it could be the manufacturer of the machines used by producers. We first consider a

[15] In the appendix we provide the proof of equilibrium existence of the free analyitcs. The proof can be easily adapted to prove existence for all other cases in the paper.

Figure 2.4: Combinations of data such that $W^{\max} \geq B_{\text{agg}}^{\min} + B_1^{\min} + B_2^{\min}$ (and heat-map) with the two best response curves for free analytics (asymmetric case).



profit-maximizing aggregator that can instruct each producer i the amount of data n_i to provide. Although this possibility may seem unrealistic, we next show how its outcome can be replicated with producers free to decide n_i and the aggregator incentivizing their choice. The aggregator offers a personalized contract to each producer i, (Q_i, n_i), with a monetary transfer and an amount of data, that the producer can accept or refuse.

The aggregator solves the following program,

$$\max_{(Q_1,n_1),\ldots,(Q_P,n_P)} \quad B_{\text{agg}} = \sum_{j=1}^{P} Q_j - \bar{\delta} - \delta \sum_{j=1}^{P} n_j - \varepsilon \eta \left(n\right) \tag{2.9a}$$

$$\text{s.t.} \quad B_{\text{agg}} - B_{\text{agg}}^{\min} \geq 0, \tag{2.9b}$$

$$B_i - B_i^{\min} \geq 0 \qquad i = 1, \ldots, P \tag{2.9c}$$

Since B_{agg} increases and B_i decreases in Q_i, at the optimum the transfer Q_i is set so that constraint (2.9c) binds, i.e.,

$$Q_i = \alpha_i \eta \left(n\right) - \gamma_i n_i - B_i^{\min}. \tag{2.10}$$

Substituting this payment into B_{agg}, the aggregator's profit rewrites as the right-hand side of (2.4), that is the social surplus. Hence, the data n^{a} that maximizes the aggregator's profit is precisely n^{w}. In fact, by appropriating the value of the analytics to each producer (up to the payoff that induces them to participate, i.e., $B_i^{\min}$), the aggregator maximizes the analytics' total surplus.

Clearly, the same result would occur when the aggregator's mandate was to run the analytics to maximize producers' payoffs $\sum_{j=1}^{P} B_j$, subject to a break-even constraint, as it could be the case with a not-for-profit incorporation organized by the producers. Since the objective would decrease in $\sum_{j=1}^{P} Q_j$, the aggregator would set the total transfer so that its participation condition $B_{\text{agg}} \geq B_{\text{agg}}^{\text{min}}$ just binds. Substituting for $\sum_{j=1}^{P} Q_j$ from the binding constraint, the program becomes one of maximizing welfare W subject to the participation constraints of producers, $B_i \geq B_i^{\text{min}}$.[16]

Remark 3. *The aggregator induces the efficient analytics with personalized take-it-or-leave-it offers* (Q_i, n_i) *to data producers.*

2.4.1 Delegation and in-house analytics

The aggregator does not need to impose the amount of data to producers. We briefly show that the same outcome of the previous section can be obtained with personalized affine monetary transfers for data and delegating to producers the amount of data to contribute. Consider a data-payment schedule for producer i, $Q_i = \bar{q}_i + q_i n_i$, where $\bar{q}_i$ is a fixed payment and q_i is a monetary transfer per-unit of data. Given the aggregator's contractual offers, producer, producer i chooses n_i to maximize its payoff, with a program similar to the free-analytics program (2.6) except that here the transfer to the aggregator is not nil. Assuming that each producer i wants to participate, at an interior solution the optimal n_i will satisfy the following (necessary) condition:

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} \alpha_i - \gamma_i - q_i = 0 \tag{2.11}$$

where $\hat{n}_{-i}$ are the data contributions that the i-th producer expects from the other producers.

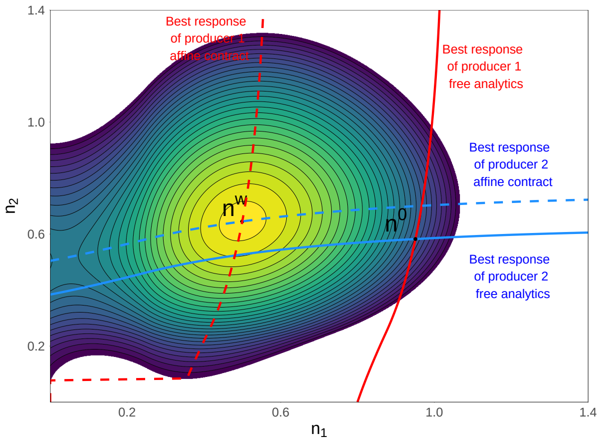
With an appropriate choice of q_i the aggregator can control producers' incentives to provide data and with the fixed component $\bar{q}_i$ of the tariff, it can appropriate the producers' value of the analytics. In particular, the optimality condition of each producer (2.11) becomes equivalent to the condition for efficient data (2.5) whenever,

$$q_i = \frac{\alpha_i(\delta + \gamma_i)}{\sum_{j=1}^{P} \alpha_j - \varepsilon} - \gamma_i. \tag{2.12}$$

The fixed fee $\bar{q}_i$ can then be set so that $B_i = \alpha_i \eta(n^*) - (\gamma_i + q_i) n_i^* - \bar{q}_i = B_i^{\text{min}}$. By doings so, the aggregator can delegate the choice of data and induce efficient analytics, n^{w}, by offering and disclosing to all producers these personalized price schedules. Observing these transfers, each producer correctly

[16] Given the optimal analytics $n^{\text{a}} = n^{\text{w}}$, since $W^{\text{max}} \geq B_{\text{agg}}^{\text{min}} + \sum_{j=1}^{P} B_j^{\text{min}}$, there exist payments $(Q_i, \ldots, Q_P)$ that guarantee all producers' participation constraints. The solution for these payments is typically not unique, and some of the solutions may favor some producers, leaving them a payoff higher than B_i^{min} and disfavor others granting exactly B_i^{min}.

Figure 2.5: A for-profit aggregator modifies producers' best responses to maximize their own profit and replicates the efficient outcome.



anticipates the data provided in equilibrium by other producers and optimally chooses the amount of data $n_i = n_i^w$.[17]

With an appropriate choice of the data-price schedule, the aggregator makes each producer internalize the positive externality that the data of each of them has on the payoff of the others. This is optimal for the aggregator because it can then extract this individual surplus (net of the outside option) with the fixed fees $\bar{q}_i$.[18]

Figure 2.5 shows how with variable parts as in (2.12), the aggregator modifies the best responses of producers and induces the efficient analytics. In particular, to obtain this outcome the aggregator here sets $q_1 = 0.59$ and $q_2 = -0.05$. The analytics involves a variable payment with the strong producer 1 benefiting the most from the analytics, that is increasing in own data, and a payment decreasing in own data for the weak producer 2.[19] Notably, these properties of the data-payment schedules follow from a general property, as it can be seen in (2.12). In fact, q_i is high (small and possibly negative) when the contribution α_i to the total value extracted by producers from the analytics $\sum_j \alpha_j$ is high (small).

When a producer values the analytics sufficiently low relatively to other producers, it can also be the case that it is subsidized entirely with a negative total price Q_i, while the aggregator profits with

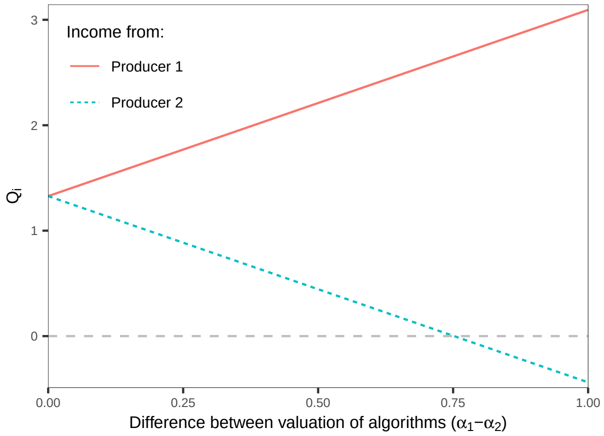
[17] The same approach shows that the efficient analytics also realizes when the aggregator maximizes producers' payoff subject to a break-even constraint. In section 2.5.2 we discuss the role of private contracting where the aggregator cannot disclose the contracts offered to each producer.

[18] If producers expect the others to procure little or no data, then the left-hand side in (2.11) may be negative at $n_i = 0$, implying that $n_i = 0$ would be optimal, a break-down as with the free analytics. Since slightly more flexible payments $Q_i(n_i)$ would allow the aggregator to eliminate this outcome, we disregard this possibility.

[19] In the Example, both producers end up paying an overall transfer to the aggregator with $\bar{q}_1 \simeq 0.75$ and $\bar{q}_2 \simeq 0.19$.

other producers. This case is reported in Figure **??**, where we consider symmetric costs but we increase continuously the difference between α_1 and α_2 keeping constant $\sum_j \alpha_j$. Although this change leaves the optimal amount of data n^* unaffected, as seen in (2.5), producer 2 is remunerated for its data. Instead, producer 1 provides data but pays for the analytics.

Figure 2.6: **??**Transfer as a function of the difference in the abilities to gain from the analytics (with identical costs).



Remark 4. *(i) A data aggregator can replicate the maximal profit and efficient outcome with personalized data-payments. (ii) When some transfer for data are negative (implying subsidy instead of payments to join the analytics), subsidies are offered to producers with a relatively small value from the analytics and high costs for sharing data.*

Producers' in-house analytics.

Some producers may have the possibility to run their analytics, relying on own data exclusively. By doing so, the payoff of producer i would be

$$B_i^{\text{self}} = \max_{n_i} \alpha_i \eta \left(n_i, 0, ..., 0\right) - \bar{\delta} - \delta n_i \tag{2.13}$$

where the producer faces the same processing cost of the aggregator. Since the analytics is run independently and in-house, the producer does not face the legal risks of handling its data externally (i.e. to ease notation we consider the case where this cost reduction is such that $\gamma_i = 0$), nor the costs associated with the risk of litigation about the analytics (i.e. we set $\varepsilon = 0$). With this option available to producer i, the aggregator must now adjust the fee Q_i to leave a payoff $B_i^{\min} = B_i^{\text{self}} > 0$. Whenever $B_i^{\text{self}} > 0$ is large enough, the aggregator may fail to find the analytics profitable when transfers are such that $B_{\text{agg}} \geq B_{\text{agg}}^{\min}$

or may prefer not to run the analytics with all producers.

Proposition 2. *(i) When data producers can independently run their analytics, the aggregator may not operate profitably with all producers unless the synergy between different datasets is sufficiently strong. (ii) When exclusion of some producers is optimal, the aggregator tends to exclude producer(s) with the most significant in-house value for the analytics* B_i^{self}.

2.5 Analytics with Anonymity

We have seen that the possibility of implementing profitable and efficient analytics requires personalized and publicly observable contracts so that producers can decide joining the analytics with a precise expectation of the identity and the data provided by other producers. However, producers may prefer to keep their decision to join the analytics and the contractual details private. They may fear that details of their production strategies become publicly known *via* the analytics, a case that we dub as *loss of anonymity*.[20] This loss of anonymity could be even more relevant when producers are competitors in related markets (as in Section 2.6). Preserving anonymity can thus be an important element of an effective analytics.

Since the profitability of a shared analytics may clash with anonymity, in this section we consider contractual offers that are constrained to preserve anonymity. We first evaluate the possibility that the aggregator offers the same *public and uniform contract* to all producers, thus forfeiting the personalization of the arrangements. We then allow for contracts that are personalized but also *secret and unobservable to third parties*, so that anonymity is preserved by secrecy.[21]

2.5.1 Analytics with uniform contracts

Assume the aggregator is bound to offer an unique and undifferentiated contract to all producers, $Q(n_i)$.[22] Without loss of generality, instead of dealing with $Q(n_i)$ we allow the aggregator to offer a (finite) set of alternatives (n_k^*, Q_k) with $k = 1, \ldots, K$, the same set for all producers, where n_k^* is an amount of data and Q_k is the associated monetary transfer. These pairs are designed so that each producer i prefers to join the analytics and autonomously selects its optimal choice. For this to be the case, we need that for any producer i there is an alternative i in the aggregator's offer so that for any other alternative $j \neq i$ in the

[20] Anonymity would also be violated when the aggregator discloses the structure and content of the dataset.

[21] As an interesting alternative, the aggregator may merge and mix the data into the same dataset, *de-facto* anonymizing them. However, in this case, the value of the analytics would be $\eta(\sum_i n_i, 0, ..., 0)$, and the Scope property would be lost, significantly reducing the value of the analytics.

[22] Here we assume that parties do not renegotiate the public contract secretly. We address this possibility in the next subsection.

offer (with $j = 1, \ldots, K$),

$$\alpha_i \eta \left(n_i^*, n_{-i}^*\right) - \gamma_i n_i^* - Q_i \geq \alpha_i \eta \left(n_j^*, n_{-i}^*\right) - \gamma_i n_j^* - Q_j. \tag{2.14}$$

This constraint guarantees that for each producer there is an entry in the set of alternatives (n_k^*, Q_k) that the producer prefers to the other. In addition, that alternative is designed to also guarantee participation of producer i,

$$\alpha_i \eta \left(n_i^*, n_{-i}^*\right) - \gamma_i n_i^* - Q_i \geq B_i^{\min} \qquad i = 1, \ldots, P.$$

The following proposition illustrates the relevant implications of an anonymous analytics.

Proposition 3. *(i) When the costs of data-sharing and the producer's value of the analytics are positively related, efficient analytics is unattainable with anonymity. (ii) In this case, combining more diverse producers into the same analytics reduces the analytics' value.*

To grasp the intuition of the proposition, consider any two specific producers that, without loss of generality, may be indicated as producer 1 and 2, and define $\eta'(n_1, n_2) = \eta\left(n_1, n_2, n_{-\{1,2\}}^*\right)$. Writing (2.14) for $i = 1$, $j = 2$ as well as for $i = 2$ and $j = 1$ to yield the conditions

$$\alpha_1 \eta'(n_1^*, n_2^*) - \gamma_1 n_1^* - Q_1 \quad \geq \quad \alpha_1 \eta'(n_2^*, n_2^*) - \gamma_1 n_2^* - Q_2 \tag{2.15}$$

$$\alpha_2 \eta'(n_1^*, n_2^*) - \gamma_2 n_2^* - Q_2 \quad \geq \quad \alpha_2 \eta'(n_1^*, n_1^*) - \gamma_2 n_1^* - Q_1 \tag{2.16}$$

that can be satisfied if and only if,

$$\begin{aligned} &\alpha_1 \left[\eta'(n_2^*, n_2^*) - \eta'(n_1^*, n_2^*)\right] - \gamma_1 \left(n_2^* - n_1^*\right) \leq Q_2 - Q_1 \\ &\qquad\qquad \leq \alpha_2 \left[\eta'(n_1^*, n_2^*) - \eta'(n_1^*, n_1^*)\right] - \gamma_2 \left(n_2^* - n_1^*\right) \end{aligned} \tag{2.17}$$

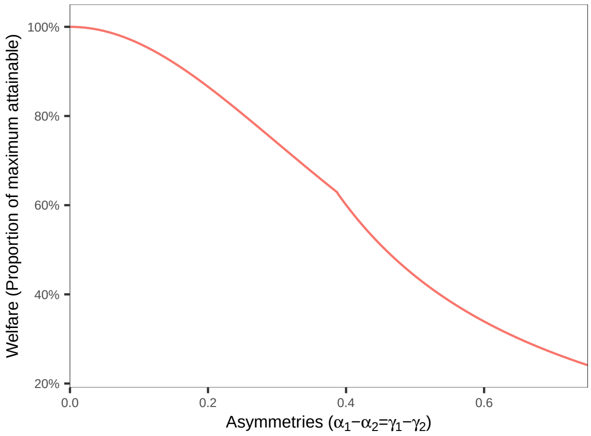
A necessary condition for this is,

$$C(\gamma_1, \gamma_2) := \alpha_2 \left[\eta'(n_1^*, n_2^*) - \eta'(n_1^*, n_1^*)\right] - \alpha_1 \left[\eta'(n_2^*, n_2^*) - \eta'(n_1^*, n_2^*)\right] - (\gamma_2 - \gamma_1)\left(n_2^* - n_1^*\right) \geq 0 \tag{2.18}$$

where we have highlighted that the optimal amount of data depends on the costs γ_1, γ_2. In the proof of Proposition 3) we show that starting from a symmetric cost environment where $\gamma_1 = \bar{\gamma} + d\gamma, \gamma_2 = \bar{\gamma}$ with $d\gamma = 0$, and introducing a (small) asymmetry in the costs with $d\gamma > 0$, we obtain,

$$C(\bar{\gamma} + d\gamma, \bar{\gamma}) \simeq -\Psi(\alpha_1 - \alpha_2) d\gamma \tag{2.19}$$

Figure 2.7: The cost of anonymity: welfare reduction with different producers (parameterized with $\Delta = \alpha_1 - \alpha_2 = \gamma_1 - \gamma_2 \geq 0$), with anonymity preserved with a unique contract for all producers.



where $\Psi > 0$. Equation (2.19) implies that if producers also differ in their values of the analytics in the same direction as with costs, i.e. $\alpha_1 > \alpha_2$ with $\gamma_1 > \gamma_2$, then it is impossible to induce different producers to select different alternatives from an anonymous contract of alternatives.

For point (ii) in the proposition note that it is reasonable that a producer extracting a significant value from the analytics also faces higher costs for data-sharing. When this occurs, the only possibility for the aggregator is to offer a unique data and payment pair (n, Q) to all producers. Since producers differs, it thus is impossible to replicate the necessary condition for an efficient analytics (2.5), and the data provided are necessarily suboptimal and the more so the more diverse are the data producers.

The Example allows to assess how the loss of value of the analytics increases with the differences in the producers. We set $\alpha_1 = \gamma_1 = 3/4 - \Delta$ and $\alpha_2 = \gamma_2 = 3/4 + \Delta$. For each value of the perturbation $0 \leq \Delta \leq \frac{3}{4}$ we compute the efficient data n_1^{w} and n_2^{w} and the associated surplus maximum surplus $W^{\max}$. It can be checked that Equation (2.18) is never satisfied with these parameters, so that to preserve anonymity, the aggregator must offer a unique (optimally chosen) option (n, Q) to all producers. Figure 2.7 shows the percentage loss of efficiency of this anonymous analytics as a function of the perturbation parameter Δ.

An interesting implication of Proposition 3 is that, instead of insisting on anonymous but distorted analytics that involves all producers, the aggregator may prefer to exclude some producers and form analytics between more homogeneous ones.

2.5.2 Analytics with secret contracts

The aggregator could preserve anonymity using secret contracts, so that the details of a contract with a producer are only known by that producer and the aggregator. With secrecy, the aggregator can still rely on personalized contracts, and preserve anonymity.[23] However, as we discuss next, this comes with possibly strong limitations on the amount of data that producers are willing to share.

We first consider the more straightforward case where the aggregator faces zero cost for managing the analytics, i.e. $\varepsilon = 0$, as it is the case when there are no legal risks nor costs with sharing the analytics. For simplicity, we discuss the case of two producers ($P = 2$) and affine payments ($Q_i = \bar{q}_i + q_i + n_i$), but the results generalize.

Since other producers' contracts are not observable, each producer must form beliefs about the actual data provided by other producers, $\hat{n}_{-i}$, for whatever (not observed) contract they were offered. Given this expectation, the first-order condition for data of producer i is,

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} = \frac{\gamma_i + q_i}{\alpha_i}, \tag{2.20}$$

which implicitly defines the optimal amount of data $n_i(q_i, \hat{n}_{-i})$ that producer i is willing to share. Note that, differently from section 2.4.1, here producer i holds some fixed beliefs about others' data $\hat{n}_{-i}$ which cannot change with the contractual conditions offered to other producers which are not observed.[24]

The aggregator in this case solves the following problem,

$$\begin{aligned} \max_{\{\bar{q}_i, q_i\}_i} \quad & B_{\text{agg}} = \sum_{i=1} [(q_i - \delta) n_i(q_i, \hat{n}_{-i}) + \bar{q}_i] - \bar{\delta} && \text{(2.21a)} \\ \text{s.t.} \quad & B_{\text{agg}} \geq 0, && \text{(2.21b)} \\ & B_i(n_i(q_i, \hat{n}_{-i}), \hat{n}_{-i}) \geq B_i^{\min} \quad \text{for any } i. && \text{(2.21c)} \end{aligned}$$

As in section 2.4.1, the aggregator controls the level of data $n_i \geq 0$ with the per-unit fee q_i and appropriates profits (or absorb losses) with $\bar{q}_i$. Also in this case, the aggregator problem is best understood as directly choosing the optimal level of data rather than transfer. Substituting the fix component $\bar{q}_i$ from

[23] Clearly, producers can guess other producers' participation decisions.

[24] We consider passive beliefs, that is, when observing the aggregator offering unexpected (off-equilibrium) contracts, each producer i thinks that the aggregator is not changing other producers' offers.

the (optimally) binding participation constraints (2.21c), the program becomes,

$$\max_{\{n_i\}_i} \quad \sum_{i=1} [\alpha_i \eta(n_i, \hat{n}_{-i}) - (\gamma_i + \delta) n_i] - \bar{\bar{\delta}} \tag{2.22a}$$

$$\text{s.t.} \quad B_{\text{agg}} \geq 0 \tag{2.22b}$$

This program shows an interesting property. Since each producer optimizes its level of data provision given its expectation about the data provided by other producers, the problem of identifying the optimal amount of data from each producer is separable from the analogous program for the others. Producers do not care about the actual contracts offered to one another but only about expectations on the data, expectations which the aggregator cannot not influence. The optimal data provision that solves the above program is given by:[25]

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} \alpha_i = \gamma_i + \delta. \tag{2.23}$$

The difference between (2.23) and condition (2.5) for efficient data is consequential: the amount of data under secrecy is lower than when anonymity is not a concern. In fact, the optimality condition with secret contracts (2.23) does not account for the effect of data n_i on the benefit of the analytics for other producers. In other terms, the aggregator cannot make producers internalize the positive externality of their data on other producers. With secret contracts, the optimality conditions (2.23) are, in fact, very similar to the case of free analytics (2.7).[26]

Imagine the aggregator tried to convince producer i that the efficient data n^w would be shared instead, so that producer i expects $\hat{n}_{-i} = n_i^w$. The aggregator would then prefer to approach any another producer j and propose to share data n_j that maximize the bilateral surplus $\alpha_j \eta(n_j, \hat{n}_{-j}) - (\gamma_i + \delta) n_j$ (the rest of the surplus does not depend on n_j but on producers' expectations about it). This clearly undermines the possibility that producer i can reasonably expect that the data of producers $j \neq i$ are efficient, i.e. $\hat{n}_j = n_j^w$.[27]

The outcome with the optimality conditions (2.23) is thus strongly inefficient and, as with a free-analytics, contemplates a substantial loss of value of the analytics. The next Proposition shows that the inefficiency with secrecy can be even deeper when the aggregator faces a cost $\varepsilon > 0$ to manage the analytics.

Proposition 4. *(i) When the aggregator preserves producers' anonymity with secret contracts, data*

[25] One can then recover the price per-unit of data for each producer by combining equations (2.20) and (2.23).

[26] A relatively small difference is that here the aggregator accounts for the cost δ of managing the data for the analytics.

[27] The reasoning developed here is similar to that in the economic literature on vertical contracting. Other types of beliefs may limit the inefficiency of secret contracting, although not eliminating it (Rey and Verge, 2004).

sharing is inefficiently low. (ii) If the cost for managing the analytics is positive and the synergy among datasets sufficiently strong, the aggregator must exclude some of the data producers from the analytics.

The reason why with a cost for managing the data the aggregator may prefer to exclude some producer is related to its objective function, which in this case writes as,

$$\sum_{j=1}^{2}[\alpha_j\eta(n_j,\hat{n}_{-j})-(\gamma_i+\delta)n_j]-\bar{\delta}-\varepsilon\eta(n_1,n_2).$$

This shows that when $\varepsilon > 0$, the decisions concerning any two data n_i and n_j are no longer separable. When the aggregator considers maximizing the bilateral surplus with any producer j, it realizes that the bilateral decision to reduce n_j now directly affects and compounds with any other data *via* the new term $\varepsilon\eta(n)$. The proof of the Proposition shows that when this is the case, and if the Scope between datasets is sufficiently large (so that the compounding effect is strong), the second order conditions of the aggregator's problem are violated when $n_i > 0$ for all producers. The optimum must thus involve $n_i = 0$ for some of them. Interestingly, when cooperative analytics shows its highest potential, i.e. when the Scope between datasets is strong, the aggregator may have to do without it.

2.6 Analytics with competing producers

When producers share their data but also compete for buyers, the producer' specific value of the analytics α_i may well be endogenous and depend on other producers' decisions. In this section we study this important case with producers that compete for final consumers and choose prices of differentiated products.[28]

The value $\eta(n_i,n_{-i})$ of the analytics here reduces the per-unit cost of production of each joining producer. The analytics benefits producers that decide to join by reducing their unitary cost:

$$c_i(n_i,n_{-i}) = \bar{c}-\eta(n_i,n_{-i}), \tag{2.24}$$

where $\bar{c}$ is the baseline per-unit cost of a producer that does not join the analytics.[29] Each producer i then sets the final-consumers' price p_i of its product and consumers decide how much to buy of each product.[30]

[28] To focus on competition, we abstract from issues with anonymity. Contracts are public and non (secretly) renegotiable. We initially consider data that are available independently from actual production, and then discuss the possibility that they are a by-product of the production process.

[29] The per-unit cost with analytics is similar to cost-reducing R&D with spillovers, as in López and Vives (2019).

[30] The alternative timing (prices decided first and then data) would deliver results qualitatively similar to the analysis in the previous sections with exogenous α_i. For more on this see also footnote 31.

Let $x_i(p_i, p_{-i})$ be the demand of producer i that is decreasing in the producer's own price p_i and (weakly) increasing in any price in the vector p_{-i} of the other producers' prices. The payoff of producer i is,

$$B_i = (p_i - c_i(n_i, n_{-i}))x_i(p_i, p_{-i}) - \gamma_i n_i - Q_i,$$

In terms of our previous notation, here we have $\alpha_i = x_i(p_i, p_{-i})$, so that a producer's ability to extract value from the analytics is now *endogenously* determined by product-market competition.[31] Although we will provide a general formulation, we further specify this otherwise complex environment considering two producers (i.e. $P = 2$).

To identify the intensity of competition with a single exogenous parameter, we further specify the model assuming quasi-linear utility so that the representative consumer' problem is

$$max_{(x_1,...,x_P)} U(x_1, ..., x_P) - \sum_i p_i x_i. \tag{2.25}$$

where the consumer's preferences $U(.)$ are,

$$U(x_1, x_2) = \theta(x_1 + x_2) - \frac{1-\rho}{2}(x_1^2 + x_2^2) - \rho x_1 x_2 \tag{2.26}$$

Parameter $\theta > 0$ is a demand shifter and $\rho \in [0, 1/2]$ is our key parameter that measures product differentiation, and thus the *intensity of competition*. With $\rho = 0$ products are independent as with separate monopolies. With $\rho = 1/2$ competition is instead maximal for perfectly substitutable products.[32] With an interior solution (consumers demand a positive quantity of both goods), the demand function for each producer $i = 1, 2$ is,

$$x_i(p_i, p_{-i}) = \frac{\theta(1-2\rho) + \rho p_{-i} - (1-\rho)p_i}{1-2\rho}. \tag{2.27}$$

Free analytics.

For given data $(n_1, ..., n_P)$ and analytics $\eta(n_1, ..., n_P)$, when active in the final consumer markets

[31] In a different timing where producers first set p_i and then decide n_i, we would have $\alpha_i = x_i$ where the quantity x_i would be exogenous when producers decide about data, exactly as in the previous sections.

[32] The formulation of 2.26 guarantees that when increasing ρ the market size is kept constant. This implies that, in general, the consumer' surplus increases in ρ.

competing producers independently set prices according to the following optimality conditions,[33]

$$x_i(p_i,p_{-i}) + (p_i - \bar{c} + \eta(n_i,n_{-i}))\frac{\partial x_i(p_i,p_{-i})}{\partial p_i} = 0,\ i = 1,...,P \tag{2.28}$$

Solving this system gives the Bertrand-Nash equilibrium prices $p_i(\eta(n_i,n_{-i}))$ and producers' profits,

$$B_i(n_i,n_{-i}) = (p_i(\eta(n_i,n_{-i})) - \bar{c})x_i(\eta(n_i,n_{-i})) + x_i(\eta(n_i,n_{-i}))\eta(n_i,n_{-i}) - \gamma_i n_i$$

where with a slight abuse of notation we indicate $x_i(\eta(n_i,n_{-i})) = x_i[p_i(\eta(n_i,n_{-i})), p_{-i}(\eta(n_i,n_{-i}))]$.

Anticipating these prices (and assuming an interior solution), the necessary optimality condition for the data of producer i is,[34]

$$\begin{aligned} x_i(n_i,n_{-i})\frac{\partial \eta}{\partial n_i} - \gamma_i + \\ + [p_i(\eta(n_i,n_{-i})) - \bar{c} + \eta(n_i,n_{-i})] \times \frac{\partial \eta}{\partial n_i}\sum_{j\neq i}\frac{\partial p_j(\eta(n_i,n_{-i}))}{\partial \eta}\frac{\partial x_i(\eta(n_i,n_{-i}))}{\partial p_j} = 0 \end{aligned} \tag{2.29}$$

The first line is equivalent to the optimality condition for the free analytics and non-competing producers, i.e. (2.7), where the producer-specific value of the analytics is $\alpha_i = x_i$, i.e. the units of outputs (on which the analytics guarantees a unitary cost reduction). The second line instead shows a novel impact of market competition and accounts for a series of reactions induced by the data of producer i. In particular, more data n_i increase the value of the analytics, $\frac{\partial \eta}{\partial n_i}$, which affects rivals' equilibrium prices, $\frac{\partial p_j}{\partial \eta}$, which in turn affect the firm's demand, $\frac{\partial x_i}{\partial p_j}$. Eventually, this demand change is valued according to the price-cost margin (the square parenthesis).

Clearly when $\rho = 0$, products are independent and the entire expression in the second line of (2.29) is nil because $\frac{\partial x_i(\eta(n_i,n_{-i}))}{\partial p_j} = 0$. This case corresponds to the case of producers operating in separate markets of the previous sections. When instead $\rho = 0$, since $\eta(.)$ is a common cost shifter reducing costs to all firms, a more valuable analytics reduces the equilibrium price of any firm, i.e. $\frac{\partial p_j}{\partial \eta} \leq 0$. A relevant implication is that the entire second line in (2.29) is non positive, and competition necessarily implies a reduction of shared data n_i. The intuition is simple: with the second line each competing producer i accounts for the fact that fewer data n_i increase rivals' prices, relaxing the intensity of competition. In general, the more intensively producers compete, the lower is the amount of data they are willing to

[33] In the proofs we also consider the possibility that prices significantly differ, so that a producer i is not active, that is it does not sell any unit, when $p_i \geq \frac{1-\rho}{\rho}p_{-i} - \frac{1-2\rho}{\rho}\theta$.

[34] For the Envelope Theorem, the impact of own data on profits *via* the producer's price change $\frac{\partial p_i(\eta(n_i,n_{-i}))}{\partial \eta}\frac{\partial \eta}{\partial n_i}$, is nil in view of (2.28). We discuss the case of non-interior solutions later on.

share. In the limit, when competition is maximal, i.e. $\rho \to 1/2$, the expression $\frac{\partial x_i(\eta(n_i,n_{-i}))}{\partial p_j}$ becomes exceedingly large so that each producer i sets $n_i = 0$. In fact, this occurs already for high but lower intensity of competition, because the marginal benefits of the analytics decline with ρ and the cost of providing the data γ_i is instead positive and constant. The dashed line in panel (a) of Figure 2.8 illustrates that this competitive-effect of data similarly operate when comparing with the (socially) efficient amount of data.

Proposition 5. *With competing producers and free analytics:*

(i) More intense competition reduces the amount of data that producers share.

(ii) The analytics breaks down if competition is sufficiently intense: competing producers share no data.

(iii) With respect to the social optimum, realized consumers' surplus and welfare does not necessarily increase with the intensity of competition.

Result (iii) is remarkable and shown in panels (b) and (c) of Figure 2.8. In a standard environment with no analytics, more intense competition, i.e. higher ρ, would normally *reduce* the distortion on the consumers' surplus and welfare induced by producers' market power relative to the social optimum. This is because more intense competition reduces prices which are inefficiently high when firms have market power. With the analytics, instead, more intense competition induces firms to limit the amount of shared data (as discussed above), increasing the firms' costs and with a net negative effect on consumers and efficiency. Although the effect is not very pronounced initially in the Figure (consumer welfare and aggregate welfare reduce slightly for low ρ as a percentage of the social optimum), one should note that, absent the analytics, these variables would increase with the intensity of competition. Also, when competition becomes very intense, the free-analytics simply becomes non-viable (result (ii) in the proposition) because producers prefer not to share data. In this case, the consumers' surplus and welfare drop.

Data aggregator.

Consider now an aggregator that offers to joining producers simple contracts of the type (Q_i, n_i), with the amount of data n_i that producer i must provide for a transfer Q_i. The program of the aggregator is similar to (2.9)-(2.9c), with two notable differences. First, as discussed above, $B_i(n_i, n_{-i})$ is now a more complex object that accounts for market profits and competition. Second, the payoff when refusing the aggregator's offer $B_i^{\min}$ is no more an exogenous element as it depends on the analytics available to the rivals. Even if producer i rejects the aggregator's offer, other producers may still accept it. In this event, competition occurs with cost $\bar{c}$ for producer i, while the other firms have a lower cost because of the analytics. The profit of firm i when refusing the analytics contract is now,

$$B_i^{\min} = max\{0, B_i(0, n_{-i})\} = max\{0, [p_i(\eta(0, n_{-i})) - \bar{c}]x_i(\eta(0, n_{-i}))\}$$

This shows that when other producers provide more data, the outside option of producer i, $B_i^{\min}$, is accordingly reduced. Since, as seen in previous cases with the payment Q_i the aggregator extracts producers' surplus up to $B_i^{\min}$, more rivals' data allow the aggregator to reduce the transfer that it must grant to convince producer i to join the analytics.[35]

At the optimum, the aggregator makes the producers' participation constraints bind and, substituting, it chooses data n to maximize:

$$\sum_i \{[p_i(\eta(n)) - \bar{c} + \eta(n)]x_i(\eta(n)) - \varepsilon\eta(n) - (\gamma_i + \delta)n_i\} - [p_i(\eta(0, n_{-i})) - \bar{c}]x_i(\eta(0, n_{-i})),$$

where, for clarity, we have identified the data provided by a generic producer i and the others, i.e. $n = (n_i, n_{-i})$.

At an interior solution, the optimality condition for n_i can be written as,

$$\begin{aligned} \left[\frac{\partial\eta(n)}{\partial n_i}\left(\sum_{j=1}^{P} x_j - \varepsilon\right) - (\delta + \gamma_i)\right] + \frac{\partial\eta(n)}{\partial n_i}\sum_{j=1}\sum_{k\neq j}[p_k - \bar{c} + \eta(n)]\frac{\partial x_k}{\partial p_j}\frac{\partial p_j}{\partial\eta} = \\ = \sum_{j\neq i}\sum_{k\neq j}\left(p_j^r - \bar{c}\right)\frac{\partial x_j^r}{\partial p_k^r}\frac{\partial p_k}{\partial\eta}\frac{\partial\eta_{-j}}{\partial n_i} \end{aligned} \qquad (2.30)$$

where p_j^r and x_j^r are short-hands respectively for the equilibrium price and quantity of producer j when it rejected the aggregator's offer, and η_{-j} is the associated value of the analytics.

The first term on the left-hand side corresponds to the the optimality condition (2.5) with non-competing producers and accounts for the internalization of the analytics' positive externality across all producers. All other terms in (2.30) account for product market competition. In particular, the second term on the left-hand side is the price-effect we have seen for the free analytics in equation (2.29). It is negative as with the free analytics, but it is now much higher (in absolute terms) because it accounts for the effect of n_i on prices and profits of *all* producers. This strong *price-effect* tends to reduce the optimal amount of data significantly: the aggregator reduces the analytics' data and quality to dampen market competition (reducing prices) and thus extract higher profits, to the detriment of final consumers. Inter-

[35] A subtle difference emerges here when the aggregator decides the data or when it delegates this choice to producers. In the former case, the aggregator may adjust the data of other producers when producer i does not join, e.g. it may further increase others' data to punish that decision. In the latter and with bilateral contracts, producer i not joining leaves others' data unaffected (it would be an unexpected, or off-equilibrium decision). Since the decisions on data are most likely delegated to producers, we follow this latter approach here.

estingly, for this effect the aggregator plays a coordination role allowing producers to *partially collude* (only partially because it does not directly control prices), a possibility that was informally mentioned in Lundqvist (2018) and that our analysis substantiates.[36] The last term, on the right-hand side of (2.30), accounts for the fact that other producers' data n_i affect the profits of a producer deciding not to join the analytics. Since more data n_i reduce the gain that the aggregator must leave to each producer, this *participation-effect* pushes towards more data n_i.

A simple but useful observation is also that the aggregator cannot reproduce the efficient amount of data that maximizes welfare, even if contracting is unrestricted (e.g. there are no anonymity issues). As seen, the aggregator cannot control producers' prices and, although it appropriates producers' profits, this gain now differs from total welfare, which also accounts for consumers' payoff. Hence, for a more apt comparison, we consider here an analytics that would maximize welfare where producers would be still free to optimally set their prices (2.28). Comparing with this social optimum, the solid lines of panel (a) in Figure 2.8 shows a case where the price-effect prevails over the participation-effect and the aggregator induces an inefficient analytics that relies on too little data.

Point (i) of the following Proposition shows that inefficient analytics is a general result under mild conditions. One may also expect that more intense competition (i.e. higher ρ) reduces this inefficiency. However, panels (b) and (c) of Figure 2.8 show that this need not to be the case. When the price and participation effects play a role, more intense competition may well adversely affect the size of the analytics with respect to the socially optimal one.

Proposition 6. *With competing producers and a data aggregator, if the value of the analytics is sufficiently concave in data,*

(i) the analytics is (generically) inefficient with respect to the socially optimal analytics, with too little shared data, and the inefficiency may not reduce with the intensity of competition;

(ii) past a certain level of competition (ρ sufficiently high), some producer is inefficiently excluded from the analytics, and not necessarily the least productive one;

(iii) the aggregator optimally combines data of producers and products associated with an intensity of competition that (generically) differs from that yielding maximal welfare or consumers' surplus.

Point (ii) shows a strong version of inefficiency. When competition is very intense, the aggregator prefers to deal with data of a subset of producers, excluding others. Although this may look benign compared to free analytics, where the analytics may simply break down, the overall effect is less so.

[36] The environment shares similarities with competing firms that cooperate at the R&D phase, as in the Research Joint-venture case in Amir (2000), with two significant differences. The analytics reduces all producers' costs by the same token and independently of the–possibly different– amount of provided data. Second, the aggregator must convince producers to join the analytics, with rivals' data affecting their outside option.

In fact, the ensuing asymmetric costs may induce exclusion of some producers from the final-product market itself.

A general comparison with the case of a free analytics is also instructive. On one hand the aggregator internalizes the benefits of all producers, with an higher (marginal) value of any data. On the other hand the "collusive" price-effect commands a reduction in the data. Panel (a) of Figure 2.8 shows that, although the aggregator relies on more data than with free analytics, the difference can be quite small.

Point (iii) of Proposition 6 illustrates another interesting fact. Imagine the aggregator could choose the firms joining the analytics, i.e. choosing two firms whose products are characterized by a given intensity of competition parameter $\rho \in [0, 1/2]$. What would be the optimal combination of these firms, i.e. the optimal ρ? Starting from unrelated products ($\rho = 0$), it can be shown that the aggregator's optimal amount of data is first increasing in ρ, up to a threshold and then it decreases. This in turn reflects into an hump-shaped aggregator's profit as a function of ρ, as in panel (d) of Figure 2.8. In other terms, the aggregator would prefer to select and admit to the analytics firms that are competing although moderately. Since welfare (and consumers' surplus) is instead increasing in the intensity of competition, point (iii) of the proposition shows that the aggregator inefficiently combines producers.[37]

We conclude this section considering the possibility that data are a by-product of actual production. Suppose, for simplicity, that there is a one-to-one mapping between data and production so that the data that producer i can provide cannot be larger than the amount produced, i.e. $n_i \leq x_i$.[38] If this constraint is not binding, then the analysis would be as in the previous paragraphs. If it binds, we have an additional effect on the analytics. As usual with competing firms, each producer faces an incentive to reduce its price to steal demand from rivals. However, when $n_i = x_i$ by doing so, the producer also increases its data and reduces that of the rivals. Since, as discussed in the previous sections, the value of analytics is degraded with unbalanced data sources, a producer may refrain from lowering its price. We thus have that the presence of the analytics further limits the intensity of competition.

2.7 Conclusions

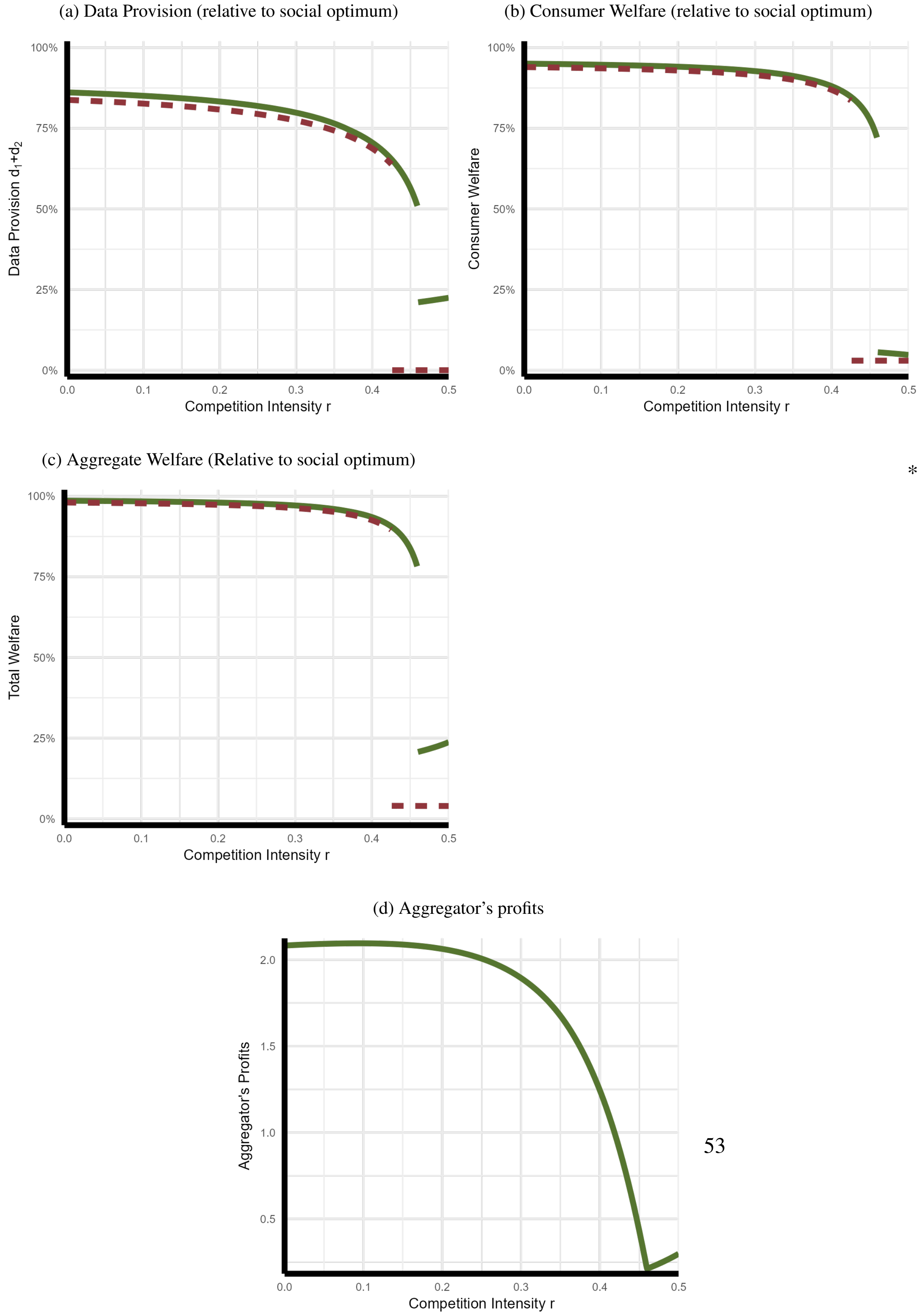
Machine Data (MD), i.e. data that machines generate with production, have received much less attention than personal data. However, with recent technological developments (such as IoT, G5, and AI tools such as Machine Learning), these data have the potential to provide enormous value for production and, ultimately, for consumers.

This paper shows that a well-functioning market for MD cannot be taken for granted. MD are

[37] This result is consistent with previous works discussing a tension between product-substitutability and personal-consumer data sharing, see Zhu et al. (2008) and Jones and Tonetti (2020).

[38] We consider here a simultaneous production of data and commodities.

Figure 2.8: Competition between data producers: Market-based Aggregator (solid line) and Free-analytics (dashed line) on the intensity of competition r (horizontal axis).



(a) Data Provision (relative to social optimum)

(b) Consumer Welfare (relative to social optimum)

(c) Aggregate Welfare (Relative to social optimum)

*

(d) Aggregator's profits

parcelized into a myriad of machines of many, possibly small, firms/data producers. Collecting and analyzing these data contemplate costs and require knowledge that may make these activities non-profitable for some firms, especially when facing risks with ill-defined ownership of MD and analytics. With the public-good nature of MD, data producers may also fail to realize and monetize the effective value of their data.

We have developed a first formal study of the market for MD and the associated analytics when pooling different data sources, i.e. a *cooperative analytics for MD*. We introduced two critical properties of MD analytics, Scale and Scope. We have investigated the implications of these properties accounting for relevant characteristics of MD producers such as the heterogeneity of data producers, their value for anonymity, and product market competition.

As a first step into the organization of this novel market for MD, our analysis can be extended in several directions, possibly attracting considerable attention for future research. For example, we have only considered the possibility of a unique data aggregator. Although socially suboptimal, we have identified several cases in which the data aggregator prefers to exclude some MD producers from the analytics. This outcome may spur entry and competition in the data-analytics market that we are investigating in ongoing research.

We have assumed that all subjects are fully informed about the details of this market. However, producers may have private information about how much they value the analytics. Introducing this element of incomplete information may have some relevant implications that can be identified using a mechanism design approach.[39]

Although our model is static, some dimensions in the market for MD may require a dynamic perspective. For example, effective analytics may help to fine-tune the production process over time and, to the extent this knowledge is shared with cooperative analytics, induce homogenization of production and products. The implications of this homogenization are unclear. The lack of diversity may reduce the value of cooperative analytics, and innovators may face limited incentives to join it.[40] We have taken the analytics technology as given, emphasizing dispersed MD's bottleneck. However, Machine Learning tools are the subject of intense R&D, with long-run implications for market structure (Gambardella et al., 2021). Ultimately, the analytics of MD may be a source of cycles between innovation and standardization.

Since the seminal work of Ronald Coase, we know that with transaction costs, the allocation of property rights plays a vital role in market efficiency. In this paper, we have investigated a status quo where producers *de-facto* own MD. Other players may be relevant, though, and claim MD ownership.

[39] If producers can run their in-house analytics, one obtains a challenging multi-agent environment with type-dependent outside options.

[40] Relatedly, it is unclear to what extent the synergy embedded in the analytics varies with product differentiation. Addressing this interesting point should rely on an investigation at the intersection of industrial organization and computer science.

This could be the case with machine manufacturers or companies specialized in monitoring machines and transmitting data (e.g. retrofitting machines with sensors). In agriculture, for example, some machine manufacturers have started to impose technical design and contractual clauses allowing them to appropriate MD, notwithstanding common codes of conduct attribute to farmers inalienable ownership of MD (Atik and Martens, 2020). Building on the environment developed in this paper, one can analyze different MD ownership arrangements and how they affect market outcomes.

Chapter 3

Price Discrimination and Tacit Collusion: An Empirical Analysis of the U.S. Airline Industry

Much attention has been devoted to the role of multimarket contacts in facilitating collusion between firms. Extensive overlap in the markets served by two firms generally increases both the benefits of collusion and the costs associated with retaliation for deviating from a collusive agreement. In contrast, the subject of price discrimination has garnered less scholarly focus. Price discrimination itself leads to market segmentation, effectively creating a form of multimarket contact. However, the impact of this type of multimarket contact on collusive agreements remains underexplored, in particular in the empirical literature.

This paper empirically examines the U.S. airline market to determine whether price discrimination practices make collusion easier or more challenging to sustain. Theoretical predictions about the role of price discrimination in facilitating or hindering collusive agreements are ambiguous. On the one hand, existing literature on multimarket contact suggests that it generally facilitates collusion. The specific theoretical literature on price discrimination and collusion tends to go in the same direction. On the other hand, the quality of information available for implementing price discrimination appears to be a critical factor. Specifically, the facilitating effect of price discrimination on collusion may not hold when the information used to estimate consumer willingness to pay is imprecise.

The U.S. airline industry serves as an exemplary subject for this study, not only due to its oligopolistic structure and the notable price variations arising from discriminatory pricing strategies but also because of the wealth of data available for analysis. Using yield management techniques that became widespread in the 1980s, airlines have implemented varying degrees of price discrimination that can be exploited for this analysis. To conduct this research, I integrate two distinct datasets. My primary source is the comprehensive Airline Origin and Destination Survey from the Bureau of Transportation Statistics, it includes a 10% sample of domestic U.S. ticket sales from 2000 to the pre-COVID year of 2019. This core dataset is further enriched with demographic and economic data on U.S. metropolitan areas, sourced from the Bureau of Economic Analysis, allowing to identify key regions based on population size and tourism dependency.

To assess the impact of price discrimination on collusive behavior, I adopt the analytical framework outlined in Ciliberto et al. (2019). I examine fluctuations in the difference in average ticket prices across competing airlines for each route, aiming to understand how changes in the intensity of price discrimination efforts influence this price differential. Guided by the theoretical model proposed by Werden and Froeb (1994) and following Ciliberto et al. (2019), I interpret a widening in this price gap as indicative of reduced collusion. My results demonstrate that a one percentage point increase in a firm's price dispersion correlates with a 1.6 percentage point expansion in the average price differential between competing airlines. As a supplementary test for collusion, I also examine price rigidity, as recommended by prior studies (Athey et al., 2004; Ciliberto et al., 2019). Estimates show that a 1 percentage point increase in price dispersion is associated with a decrease in price rigidity.

An issue in this framework is the measurement of price discrimination. Given my reliance on price dispersion as an indicator, I conduct supplementary tests to ensure that the observed effects are genuinely attributable to price discrimination. To bolster the robustness of my findings, I adopt methodologies from Borenstein and Rose (1994); Gerardi and Shapiro (2009) to identify routes with greater consumer heterogeneity. In such cases, a larger portion of observed price dispersion is likely due to price discrimination. When I limit our regression analysis to these routes, the results exhibit a stronger magnitude, thereby reinforcing the validity of the empirical study.

A second issue involves the potential for endogeneity. This adjustment is necessary because the adoption of price discrimination techniques at the route level could be correlated with market structure in ways that are difficult to control for—especially since market structure itself may an outcome of price discrimination behavior. To mitigate this, I shift the focus from price dispersion within individual routes to average price dispersion across airlines as a whole. By elevating the focus to the airline level, I isolate the variables of interest and eliminate potential endogeneity. This refined approach substantiates my initial findings, confirming that increased price discrimination is inversely related to collusive behavior.

Finally, upon examining the data across different years, my analysis indicates that the relationship between price discrimination and collusion has varied over the past 20 years. The influence of price discrimination appears to be dual-natured: discouraging collusion during the first half of the study period, while promoting it in the final years. I posit that this fluctuating influence can be attributed to the evolving quality of information available to airlines for implementing price discrimination, particularly as the digital era advances. This dynamic highlights the critical role that information quality plays in mediating the relationship between price discrimination and collusion, a finding that aligns closely with existing literature (Liu and Serfes, 2007; Peiseler et al., 2022).

This paper seeks to make contributions to two distinct areas of academic literature. The first explores the relationship between price discrimination and collusion, as discussed in works by Liu and Serfes (2007); Helfrich and Herweg (2016); Peiseler et al. (2022). These studies examine how price discrimination affects the viability of tacit collusion. Liu and Serfes (2007) argues that collusion is more easily sustained under uniform pricing when the quality of information for price discrimination is low; however, the opposite holds true as information quality improves. Helfrich and Herweg (2016) posits that collusion is more challenging to maintain in a regime of price discrimination compared to uniform pricing. Peiseler et al. (2022) suggests that collusion becomes difficult to sustain when the quality of information is either exceptionally high or low, but is more feasible at intermediate levels of information quality.

The second area to which this paper contributes is the study of multimarket contact. By segmenting the market through price discrimination, firms inadvertently create conditions for multimarket contact, as noted by Liu and Serfes (2007). Therefore, this paper also aims to shed light on the implications of multimarket contact, particularly in markets that allow for greater substitutability among products.

The structure of this paper is as follows: Section 2 introduces the data, Section 3 outlines the empirical strategy, Section 4 discusses the main findings, and Section 5 provides the conclusion.

3.1 Data

I use data from the Airline Origin and Destination Survey (DB1B) for the years 2000-2019, a dataset that comprises a 10% sample of domestic airline tickets and is widely cited in airline industry research. In alignment with existing literature, I exclude tickets with extreme fares—specifically, those below $20 and above $2500. Multi-ticketed routes are also omitted from the sample. I consider routes to be directional, and accordingly, and split round-trip tickets into two one-way tickets, halving the fare for each. The data is then aggregated at the level of airline, route, and year/quarter. Airlines accounting for less than 5% of sales on a given route are excluded for that period and route. The final dataset encompasses 6,654 markets, with a quarterly sample frequency. It includes a total of 34 carriers, yielding 637,579 observations of carrier-route-year-quarter pairs.

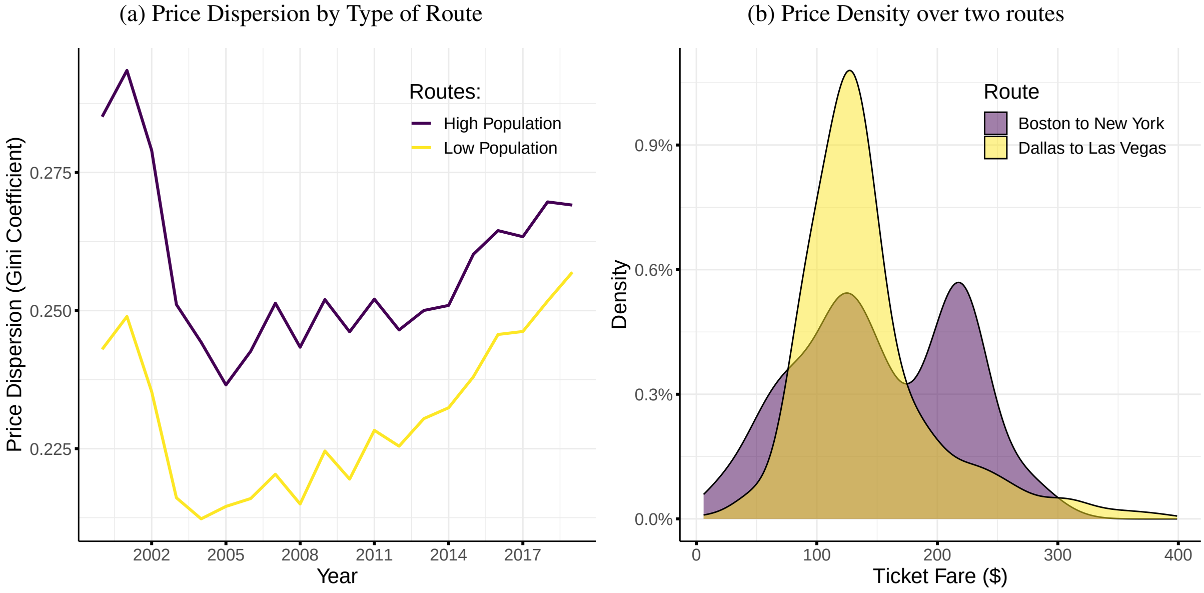
To assess signs of tacit collusion in the airline industry, I calculate two key metrics using the DB1B database, as guided by Ciliberto et al. (2019). The first metric is the average absolute price difference between each pair of airlines operating on the same route for each time period. The second is the coefficient of variation for the average price between each pair of airlines across all routes in the entire sample. These metrics serve as my primary dependent variables. Additionally, I use three measures of price dispersion, inspired by Borenstein and Rose (1994), to act as proxies for price discrimination. These include: the Gini coefficient for each airline's prices on each route per time period; the coefficient of variation of these prices; and the interquartile range. To facilitate regression analysis at the airline-pair level, I aggregate these price dispersion measures by summing them for the two airlines in question. I also incorporate control variables like the Herfindahl index for each route and categorical indicators for low-cost and legacy airlines, using data from both the DB1B and the International Civil Aviation Organization (ICAO). Table 1 summarizes these key variables and their statistics.

The factors influencing price discrimination are varied and can be categorized primarily into two aspects. First, the airline's strategic approach plays a significant role. For example, low-cost carriers generally engage in less yield management, resulting in more consistent pricing that doesn't fluctuate based on round-trip versus one-way travel or time of day. Second, the nature of the route also impacts the degree of price discrimination. Routes serving large, diverse cities often attract a mix of business travelers and tourists. Business travelers typically exhibit lower price elasticity of demand compared to tourists, which encourages airlines to implement different pricing strategies for these two groups.

To assess consumer diversity across different markets, I develop a proxy measure, drawing on the research of Borenstein and Rose (1994); Gerardi and Shapiro (2009). Using data from the Bureau of Economic Analysis, which includes metrics on personal income and industry earnings at the metropolitan level, I isolate areas where tourism constitutes a significant share of revenue. Additionally, I identify the 30 most populous cities within the sample. Based on this data, I establish two binary indicators for each airport: one signifies whether the airport is located in a metropolitan area where the ratio of accommodation earnings to total earnings exceeds the 85th percentile, and the other indicates whether the area is among the 30 most populous cities. Routes are labeled as "touristic" if at least one endpoint is in a touristic area, and as "big-city" if both endpoints are among the top 30 most populous areas. According to existing literature, "big-city" routes are likely to attract a more diverse set of consumers, including both leisure and business travelers, whereas "touristic" routes tend to cater to a more homogeneous, leisure-oriented customer base. Figure 3.1a illustrates the variation in pricing across different types of routes. Specifically, routes connecting two of the 30 most populous cities exhibit greater price dispersion, although this difference diminishes over time within the sample. Figure 3.1b provides an example of how price dispersion varies between routes primarily serving tourists (e.g., Dallas to Las Vegas) and those with a more diverse consumer base. In routes characterized by higher consumer diversity, the price density graph is double-peaked, likely reflecting the distinct pricing of leisure and business travelers.

Utilizing this dataset, I first examine the variations in price dispersion among airlines, arguing that these differences likely stem from varying degrees of yield management and price discrimination em-

Figure 3.1: Price Dispersion and Price Density


(a) Price Dispersion by Type of Route

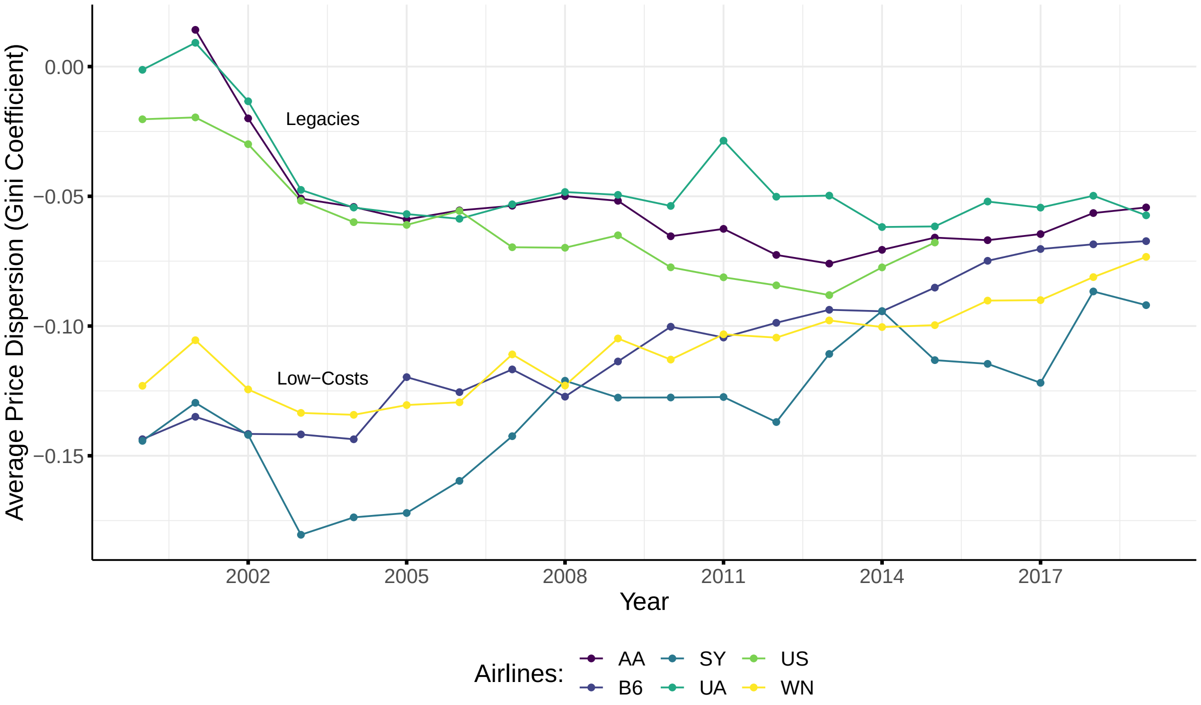
(b) Price Density over two routes

Notes: Panel 3.1a: Lines present the average of the airlines-route level Gini coefficient weighted by the number of passengers. High population refers to routes linking two of the 30 most populous metropolitan areas, while low population refers to all other routes. Panel 3.1b: presents the price distribution over two example routes, one linking two high population metropolitan areas and one route mostly used by leisure travelers (for period 2002Q4 on both routes).

ployed by the firms. Figure 3.2 displays the average Gini coefficient across routes for a subset of airlines in the dataset, specifically comparing three legacy airlines to three low-cost carriers. The data reveals two key insights: First, legacy airlines exhibit greater price dispersion compared to low-cost airlines, corroborating existing literature on the subject. Second, this disparity has been gradually diminishing over the time span covered by the dataset; by the end of the sample period, both types of airlines demonstrate comparable levels of average price dispersion.

To estimate both the quality of information available to airlines for price discrimination and their capability to implement it, I utilize data from the Current Population Survey, sourced from the Bureau of Labor Statistics and the Census Bureau. This data measures the proportion of internet users across metropolitan areas. The rationale behind using this proxy is based on the assumption that areas with higher internet penetration are likely to have a greater number of consumers purchasing tickets online. This, in turn, enables airlines to employ more responsive online yield management techniques. Furthermore, I posit that metropolitan areas with better internet access are likely to provide airlines with more accurate data on consumer willingness to pay. Lastly, I anticipate that the quality of available information will naturally improve over time as online consumption becomes increasingly prevalent. Therefore,

Figure 3.2: Price Dispersion by Type of Airlines



Notes: Average price dispersion measured by the average Gini coefficient over routes operated by each airlines weighted by the number of passengers operated in each route. The group of legacy airlines include: American Airlines (AA), United Airlines (UA) and US airways (US). The group of Low-Costs: Sun Country airlines (SY), JetBlue (B6) and Southwest airlines (WN).

I consider the passage of time as an indicator of improving information quality.

3.2 Identification Strategy

My objective is to investigate whether the practice of price discrimination, which I can only approximate through measures of price dispersion, acts as a barrier to behaviors resembling tacit collusion. To examine this, I adopt the methodology outlined in **?**. Building on insights from Werden and Froeb (1994), the integration of a competitor's profits into one's own objective function should result in converging average prices. Therefore, two firms engaged in tacit collusion would exhibit prices that grow increasingly similar. My first set of regressions aims to correlate differences in average route-level prices between airlines with their respective practices of price discrimination. The second set of regressions is inspired by Athey et al. (2004), positing that firms engaged in tacit collusion should display some degree of price rigidity. Following **?**, this second set of regressions examines the relationship between the coefficient of variation in pricing among firms and the extent to which they engage in price discrimination.

To empirically assess the influence of price discrimination strategies on a firm's propensity for collusive behavior, I estimate the following regression equation:

$$\log(y_{p,m,t}) = \beta \log(\text{PD}_{p,m,t}) + \gamma X_{p,m,t} + \varepsilon_{p,m,t} \tag{3.1}$$

In this equation, $\log(y_{p,m,t})$ represents either the absolute difference in average prices between the two airlines in pair pp on market mm at time tt, or the coefficient of variation of the average prices for these airlines on market mm over the entire sample period. $\log(\text{PD}_{p,m,t})$ denotes the sum of the price dispersion for the two airlines in pair pp on market m at time t. This price dispersion can be measured using one of three metrics: the Gini coefficient, the coefficient of variation, or the interquartile range. $X_{p,m,t}$ represents the fixed-effects I include in the regressions. Finally $\varepsilon_{p,m,t}$ represents the error term.

A primary concern is the adequacy of the proxy used for measuring price discrimination. In airline markets, factors other than price discrimination, such as peak-load pricing or stochastic demand pricing, could also contribute to price dispersion. Existing literature has identified various mechanisms through which price dispersion can occur without the presence of price discrimination. While the limitations of the DB1B dataset restrict our ability to refine our controls, one approach I can take is to consistently test regression results on the sample of large cities. These urban areas typically attract both leisure and business travelers, ensuring a significant level of consumer heterogeneity that would justify the use of price discrimination techniques.

A second concern relates to endogeneity. As highlighted in existing literature, market structure significantly impacts price discrimination and dispersion. Generally, an increase in the number of competitors is expected to reduce price dispersion. This tendency could potentially bias the results downward. Higher levels of price discrimination would likely correlate with lower levels of competition, either in terms of the Herfindahl index or the number of competitors. This would, in turn, reduce price differences if the insights from Ciliberto et al. (2019) hold true. To mitigate this bias, we calculate an airline's average price dispersion as the weighted average of its price dispersion across all markets where it operates, with weights determined by the number of passengers it transports in each market. This metric is arguably exogenous to each specific market mm and reflects an airline's overarching strategy in terms of price discrimination. Through this regression analysis, we aim to isolate the impact of price discrimination on the viability of collusive arrangements.

The price dispersion of an airline at the route level—whether measured by the Gini coefficient, the coefficient of variation, or the interquartile range—is more strongly correlated with that airline's average level of price dispersion than with the price dispersion exhibited by competitors on the same route. I interpret this as evidence that an airline's price dispersion is not merely a function of the specific routes it operates, but rather a strategic choice made at the corporate level. Consequently, utilizing the airline-level measure of price dispersion serves to address the endogeneity issue highlighted earlier.

Finally, to assess the interplay between signal quality and price discrimination in affecting the sustainability of collusive behavior, we employ the following regression model:

$$\log(y_{p,m,t}) = \sum_{t=1}^{T} \beta_t \log(\mathrm{PD}_{p,m,t}) + \sum_{t=1}^{T} \beta_t^i \log(\mathrm{PD}_{p,m,t}) * \mathrm{IE}_{m,t} + \gamma X_{p,m,t} + \varepsilon_{p,m,t} \tag{3.2}$$

Through this regression equation, we aim to discern how advancements in signal quality have influenced the impact of price discrimination on collusive arrangements. As time progresses, we anticipate that the quality of information available to airlines will improve. Additionally, in a cross-sectional analysis, we expect metropolitan areas with higher internet usage to have superior signal quality. The coefficient $\beta_1$evaluates the effect of price discrimination at the lowest level of data quality in the sample (starting in 2000), while β_T assesses its effect at the highest level of data quality. If signal quality is indeed better in areas with greater internet access, then β_i should be equal to $\frac{\partial \beta_t}{\partial t}$.

3.3 Results

In this section, I present the findings derived from the regression model outlined in Equation 3.1. Based on existing literature (Helfrich and Herweg, 2016), I anticipate that price discrimination will negatively impact collusive behavior. However, I also expect that the relationship between price discrimination and collusion will be moderated by the quality of information available to airlines.

My initial empirical analysis examines the relationship between the absolute difference in average prices for pairs of airlines operating on the same routes and the degree of price dispersion exhibited by those airlines on those routes. This approach is theoretically grounded in Werden and Froeb (1994) and has been previously implemented in Ciliberto et al. (2019). Table 3.1 displays the results of the first set of regressions, which focus on these price differences. I employ each of the three measures of price dispersion proposed in Borenstein and Rose (1994). The fixed effects incorporated into the model include period-specific effects (both year and quarter), route-specific effects, and airline pair-specific effects. Regardless of the measure of price dispersion used, the estimated relationship is both statistically significant and of substantial magnitude. Specifically, within a given route, an increase in price dispersion correlates with a widening gap in average prices between two competing airlines. The magnitude of this effect varies depending on the measure of price dispersion used but remains robust across all measures. These initial findings suggest that price dispersion undermines the potential for tacit collusion.

In Table 3.2, I replicate the regressions from Table 3.1, this time using the coefficient of variation of average prices as the dependent variable. This second test is grounded in the theoretical framework

Table 3.1: Difference in Average Prices and Price Dispersion

	Dependent variable:					
	$\log(\lvert\Delta p_{p,m,t}\rvert)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Gini	0.912^{***}	0.838^{***}				
	(0.010)	(0.011)				
Coef. Var.			0.534^{***}	0.456^{***}		
			(0.009)	(0.009)		
Interquartile Range					0.567^{***}	0.546^{***}
					(0.005)	(0.005)
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Pair Fixed Effect	Yes	No	Yes	No	Yes	No
Route Fixed Effect	Yes	No	Yes	No	Yes	No
Route-Pair Fixed Effect	No	Yes	No	Yes	Yes	No
Nbr. Markets	6,654	6,654	6,654	6,654	6,654	6,654
Nbr. Carrier Pairs	278	278	278	278	278	278
Nbr. Markets/Pair	51,574	51,574	51,574	51,574	51,574	51,574
Observations	637,580	637,580	637,580	637,580	637,580	637,580
Adjusted R^2	0.281	0.375	0.276	0.371	0.288	0.380

*p<0.1; **p<0.05; ***p<0.01

provided by Athey et al. (2004) and has previously been employed in empirical studies such as Ciliberto et al. (2019). The test operates on the assumption that firms engaged in collusion are more likely to exhibit price rigidity compared to those in competition. I re-estimate Equation 3.1, substituting the dependent variable with the coefficient of variation of average prices between two airlines of a pair on a given route over the sample period. The results, presented in Table 3.2, corroborate earlier findings and are consistent in terms of magnitude.

Table 3.2: Coefficient of Variation and Price Discrimination

	Dependent variable:					
	$\log(CV_{p,m})$					
	(1)	(2)	(3)	(4)	(5)	(6)
Gini	1.162^{***}	1.492^{***}				
	(0.035)	(0.045)				
Coef. Var.			0.756^{***}	1.298^{***}		
			(0.029)	(0.042)		
Interquartile Range					0.610^{***}	0.530^{***}
					(0.019)	(0.023)
Pair Fixed Effect	No	Yes	No	Yes	No	Yes
Route Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Nbr. Markets	5,992	5,992	5,992	5,992	5,992	5,992
Nbr. Carrier Pairs	251	251	251	251	251	251
Observations	41,537	41,537	41,537	41,537	41,537	41,537
Adjusted R^2	0.129	0.196	0.119	0.193	0.129	0.183

*p<0.1; **p<0.05; ***p<0.01

However, the current results may be subject to endogeneity concerns. As indicated by the literature we consult, including works by Gerardi and Shapiro (2009); Borenstein and Rose (1994), factors such as the number of competitors on a route and their market concentration could also influence price discrimination and dispersion. Specifically, Gerardi and Shapiro (2009) identifies a negative relationship between competition and price dispersion. This dynamic could introduce a downward bias in my results; higher price dispersion could be indicative of reduced competition, which in turn could enhance the sustainability of collusive arrangements.

To address this potential bias, I shift the focus from individual routes to each company's average price dispersion across routes. Rather than regressing collusive behavior on route-level price dispersion, we base our regression on the companies' overall strategies for price discrimination. This approach mitigates the endogeneity concerns of the type highlighted by Gerardi and Shapiro (2009). Table 3.3 presents these revised results. As anticipated, the magnitude of the effects is considerably greater, suggesting that a bias did indeed exist and was exerting a negative influence on the initial findings.

Table 3.3: Collusion Proxies and Company Level Measure of Price Discrimination

	Dependent variable:					
	ΔP	ΔCV	ΔP	ΔCV	ΔP	ΔCV
	(1)	(2)	(3)	(4)	(5)	(6)
Gini	1.629^{***}	0.512^{***}				
	(0.030)	(0.043)				
Coef. Var.			1.224^{***}	0.276^{***}		
			(0.025)	(0.031)		
Interquartile Range					0.059^{***}	0.159^{***}
					(0.013)	(0.014)
Time Fixed Effect	Yes	No	Yes	No	Yes	No
Pair Fixed Effect	Yes	No	Yes	No	Yes	No
Route Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Nbr. Markets	6,654	5,992	6,654	5,992	6,654	5,992
Nbr. Carrier Pairs	278	251	278	251	278	251
Nbr. Markets/Pairs	51,574	41,537	51,574	41,537	51,574	41,537
Observations	637,580	41,537	637,580	41,537	637,580	41,537
Adjusted R^2	0.275	0.106	0.274	0.104	0.271	0.106

Note: $\Delta P = \log(|\Delta p_{p,m,t}|)$, $\Delta CV = \log(CV_{p,m})$. *p<0.1; **p<0.05; ***p<0.01

The findings suggest that price discrimination appears to inhibit the occurrence of collusion. However, the empirical framework doesn't provide much insight into the role of signal quality in this relationship.

Building on these initial results, I now aim to explore how information quality modulates the interaction between price discrimination and collusion, as observed in Tables 3.1 and 3.2. To do this, I proceed with the estimation of Equation 3.2. The rationale for this regression is that the quality and availability of information for airlines to engage in price discrimination should have consistently improved with the digitalization of the economy. A growing proportion of airline sales are now conducted online, enabling companies to systematically record sales data and implement more nuanced pricing strategies. I anticipate that information quality will be higher in metropolitan areas with greater internet penetration, reflecting the increased volume of online sales in these regions. In this context, the effect of internet penetration can be interpreted as the marginal impact of time on the relationship between price discrimination and collusion.

Due to data limitations, this regression analysis is confined to the period from 2000 to 2017. Figure 3.3 displays the outcomes of this analysis. Intriguingly, the graph reveals a reversal in the impact of price discrimination over the sample period. While price dispersion initially has a negative effect on collusive practices (making tacit collusion more difficult), the opposite holds true towards the end of the sample.

Contrary to expectations, the interaction between time and internet penetration does not trend in the anticipated direction. Towards the end of the sample, routes in metropolitan areas with high internet access show a more positive correlation between price discrimination and collusion. This suggests that in areas where signal quality is higher, the relationship between price discrimination and collusion is weaker. Meanwhile, as time progresses (and presumably information quality improves), the estimated relationship becomes increasingly negative.

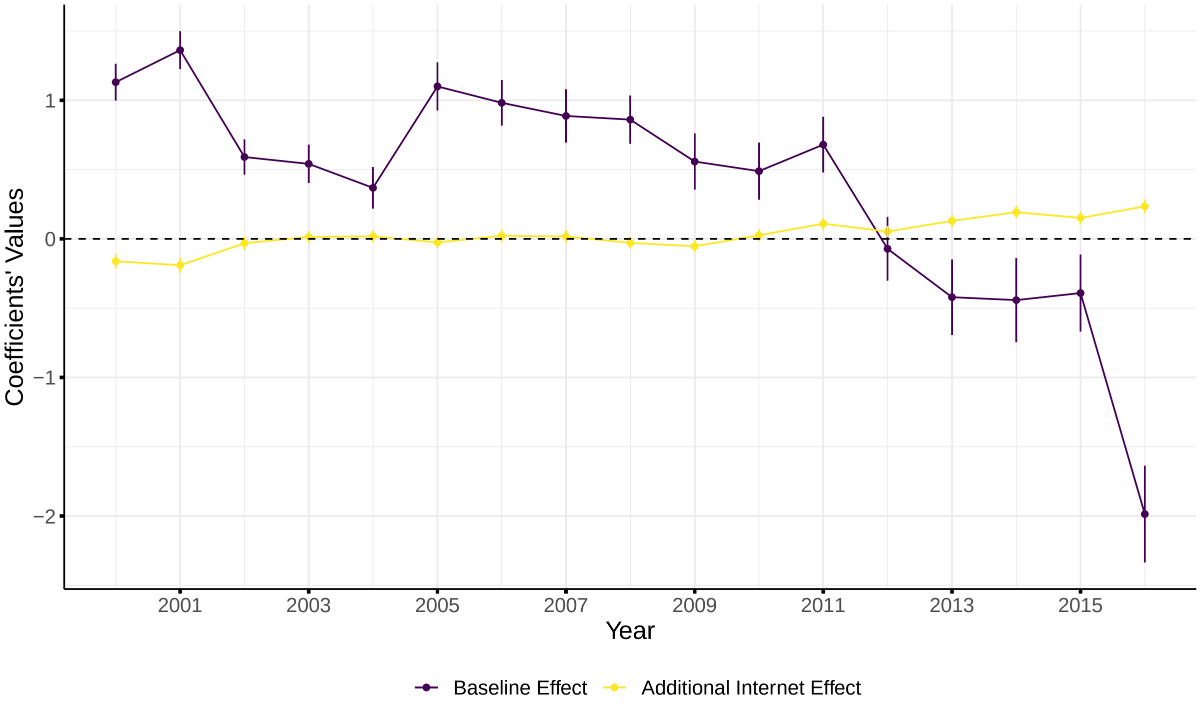
3.4 Robustness

One potential limitation of this analysis stems from the proxies I employ for price discrimination, specifically price dispersion. While an increase in price discrimination would likely lead to greater price dispersion, the reverse is not necessarily true. There are compelling reasons to believe that a significant portion of the price dispersion observed in the DB1B dataset is not attributable to price discrimination. For instance, airline practices like peak-load pricing or stochastic demand models can generate price dispersion without being directly related to price discrimination.

This issue may not pose a significant problem for the analysis if any form of price dispersion inherently makes tacit collusion more difficult to sustain. In that case, price discrimination, to the extent that it generates price dispersion, would also hinder tacit collusion. However, I aim to isolate the specific impact of price discrimination, as it creates multi-market contacts in a way that general price dispersion does not, suggesting different underlying mechanisms.

While the dataset I use doesn't allow for precise control over these factors, it does enable me to differentiate markets based on expected levels of consumer heterogeneity. Following the literature (Borenstein and Rose, 1994; Gerardi and Shapiro, 2009) and as detailed in Section 3.1, I distinguish between routes expected to have high consumer heterogeneity (big city markets) and those expected to have low heterogeneity (touristic routes). In big-city routes, a larger proportion of the observed price dispersion is likely

Figure 3.3: Modulation by Information Quality of the Effect of Price Discrimination on Collusion



Notes: This graph presents the results of the estimation of Equation 3.2. The Baseline effect corresponds to the estimate of β_t whereas the additional internet effect corresponds to the time series of β_t^i

due to price discrimination. Therefore, one would expect the impact of price dispersion on tacit collusion to be more pronounced in these markets. This expectation is confirmed when splitting the sample between big city and touristic routes, as shown in Table 3.4.

3.5 Conclusion

In this paper, I have demonstrated that airlines employing strategies leading to greater price dispersion are less likely to engage in tacit price collusion with competitors during the period from 2000 to 2012. However, this trend reverses towards the end of the sample period. Given that much of the price dispersion in the airline industry can be attributed to price discrimination, I argue that the quality of available information plays a pivotal role. Specifically, poor information quality in the early part of the sample led to a negative impact on collusion, an effect that reverses as information quality improves over time.

These findings carry significant policy implications. In environments with low information quality, price discrimination could serve as a tool to counteract collusion. Conversely, when information quality

Table 3.4: Differentiated Effect of Price Discrimination Given Route Characteristics

	$\log(\lvert\Delta p_{p,m,t}\rvert)$			
	Samples By:			
	Population		Tourism	
	High	Low	High	Low
	(1)	(2)	(3)	(4)
Gini	1.481***	1.226***	1.066***	1.439***
	(0.049)	(0.057)	(0.056)	(0.050)
Time Fixed Effect	Yes	Yes	Yes	Yes
Pair Fixed Effect	Yes	Yes	Yes	Yes
Route Fixed Effect	Yes	Yes	Yes	Yes
Nbr. Markets	1,513	1,404	795	2,122
Nbr. Carrier Pairs	199	206	206	207
Nbr. Markets/Pairs	17,614	14,837	10,977	21,474
Observations	238,777	192,744	162,147	269,374
Adjusted R^2	0.287	0.262	0.298	0.263

Note: *p<0.1; **p<0.05; ***p<0.01

is high, policymakers may need to restrict the extent of price discrimination to prevent the emergence of tacit collusion. While these conclusions have been previously suggested through theoretical models, this paper is the first to empirically substantiate them.

Nevertheless, this study is not without limitations. First, the effects of price discrimination are only imperfectly observed, raising the possibility that the observed impacts may actually stem from price dispersion rather than discrimination, despite robustness tests. If so, the underlying mechanisms and conclusions would differ substantially. Second, the measure used for signal quality has inherent limitations. The passage of time not only affects signal quality but may also introduce other uncontrolled changes in airline markets. Third, the conclusions may be sensitive to sample selection. Although the sample aligns with existing literature, further robustness tests could be conducted by segmenting the sample based on different criteria, such as round-trip journeys, direct flights, or journeys between major metropolitan areas. Lastly, the measures used to assess tacit collusion may not fully satisfy all readers, suggesting the need for alternative tests to evaluate collusion.

Bibliography

Abrahams, Fred (2016). *Modern Albania: from dictatorship to democracy in Europe*. NYU Press, URL `https://doi.org/10.18574/nyu/9780814705117.001.0001`. OCLC: 967257139.

Adema, Joop, Cevat Giray Aksoy, and Panu Poutvaara (2022). "Mobile Internet Access and the Desire to Emigrate." *CESifo Working Papers*, (9758).

Adsera, Alicia and Ana M. Ferrer (2015). "The Effect of Linguistic Proximity on the Occupational Assimilation of Immigrant Men in Canada." *IZA Discussion Paper*.

Adserà, Alícia and Mariola Pytliková (2015). "The Role of Language in Shaping International Migration." *The Economic Journal*, 125(586), F49–F81.

Alam, Furqan, Rashid Mehmood, Iyad Katib, Nasser N. Albogami, and Aiiad Albeshri (2017). "Data Fusion and IoT for Smart Ubiquitous Environments: A Survey." *IEEE Access*, 5, 9533–9554.

Amir, Rabah (2000). "Modelling imperfectly appropriable R&D via spillovers." *International Journal of Industrial Organization*, 18(7), 1013–1032.

Amir, Rabah, Huizhong Liu, Dominika Machowska, and Joana Resende (2019). "Spillovers, subsidies, and second-best socially optimal R&D." *Journal of Public Economic Theory*, 21(6), 1200–1220.

Anderson, Gary M, William F Shughart, and Robert D Tollison (2004). "The Economic Theory of Clubs." In *The Encyclopedia of Public Choice*, pp. 499–504. Springer US, URL `https://doi.org/10.1007/978-0-306-47828-4_81`.

Anelli, Massimo, Gaetano Basso, Giuseppe Ippedico, and Giovanni Peri (2023). "Emigration and Entrepreneurial Drain." *American Economic Journal: Applied Economics*, 15(2), 218–252.

Athey, Susan, Chris Sanchirico, and Kyle Bagwell (2004). "Collusion and Price Rigidity." *The Review of Economic Studies*, 71(2).

Atik, Can and Bertin Martens (2020). “Competition Problems and Governance of Non-personal Agricultural Machine Data: Comparing Voluntary Initiatives in the US and EU.” *EUR - Scientific and Technical Research Reports*, p. 40.

Barber, C. Bradford, David P. Dobkin, and Hannu Huhdanpaa (1996). “The Quickhull Algorithm for Convex Hulls.” *ACM Trans. Math. Softw.*, 22(4), 469–483.

Belot, Michèle and Sjef Ederveen (2012). “Cultural barriers in migration between OECD countries.” *Journal of Population Economics*, 25(3), 1077–1105.

Belot, Michèle and Timothy Hatton (2012). “Immigrant Selection in the OECD.” *The Scandinavian Journal of Economics*, 114(4), 1105–1128.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2019). “The economics of social data.” *Discussion Paper No. 2203R, Cowles Foundation, New Haven, CT*.

Berman, Eli, Kevin Lang, and Erez Siniver (2003). “Language-skill complementarity: returns to immigrant language acquisition.” *Labour Economics*, 10(3), 265–290.

Borenstein, Severin and Nancy L Rose (1994). “Competition and Price Dispersion in the U.S. Airline Industry.” *Journal of Political Economy*, 102(4), 653–683.

Borjas, George J (1987). “Self-Selection and the Earnings of Immigrants.” *The American Economic Review*, 77, 24.

Braga, Michela (2007). “Dreaming Another Life. The Role of Foreign Media in Migration Decision. Evidence from Albania.” *World Bank*, p. 41.

Bruckner, Andrew and E. Ostrow (1962). “Some function classes related to the class of convex functions.” *Pacific Journal of Mathematics*, 12(4), 1203–1215.

Bursztyn, Leonardo and Davide Cantoni (2016). “A Tear in the Iron Curtain: The Impact of Western Television on Consumption Behavior.” *Review of Economics and Statistics*, 98(1), 25–41.

Bütikofer, Aline and Giovanni Peri (2021). “How Cognitive Ability and Personality Traits Affect Geographic Mobility.” *Journal of Labor Economics*, 39(2), 559–595.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo’, and Sergio Pastorello (2020). “Artificial intelligence, algorithmic pricing, and collusion.” *American Economic Review*, 110(10), 3267—-3297.

Chiswick, Barry R (1995). “The Endogeneity between Language and Earnings: International Analyses.” *Journal of Labor Economics*.

Chong, Alberto and Eliana La Ferrara (2009). "Television and Divorce: Evidence from Brazilian *Novelas*." *Journal of the European Economic Association*, 7(2-3), 458–468.

Ciliberto, Federico, Eddie Watkins, and Jonathan W. Williams (2019). "Collusive pricing patterns in the US airline industry." *International Journal of Industrial Organization*, 62, 136–157.

Cornes, Richard and Roger Hartley (2007). "Aggregative Public Good Games." *Journal of Public Economic Theory*, 9(2), 201–219.

Docquier, Frédéric and Abdeslam Marfouk (2006). "International Migration by Education Attainment, 1990-2000." In *International migration, remittances, and the brain drain*, edited by Çaglar Özden and Maurice W. Schiff, Trade and development series, chap. 5, pp. 151–200. World Bank : Palgrave Macmillan, Washington, DC.

Docquier, Frédéric and Hillel Rapoport (2012). "Globalization, Brain Drain, and Development." *Journal of Economic Literature*, 50(3), 681–730.

Dorfles, Piero and Giovanna Gatteschi (1991). *Guardando all'Italia : influenza delle TV e delle radio italiane sull'esodo degli albanesi*. No. 1 in Instant research / VQPT-SO, 1 ed., Roma : RAI radiotelevisione italiana.

Dosis, Anastasios and Wilfried Sand-Zantman (2019). "The Ownership of Data." *SSRN Electronic Journal*.

Drexl, Josef (2016). "Designing Competitive Markets for Industrial Data - Between Propertisation and Access." *SSRN Electronic Journal*.

Duch-Brown, NNstor, Bertin Martens, and Frank Mueller-Langer (2017). "The Economics of Ownership, Access and Trade in Digital Data." *SSRN Electronic Journal*.

Durante, Ruben, Paolo Pinotti, and Andrea Tesei (2019). "The Political Legacy of Entertainment TV." *American Economic Review*, 109(7), 2497–2530.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya (2011). "Media and Political Persuasion: Evidence from Russia." *American Economic Review*, 101(7), 3253–3285.

Etgeton, Stefan (2018). "Cluster-SE." *Github*.

European Commission (2017). "Building a European Data Economy." URL `https://eur-lex.europa.eu/content/news/building_EU_data_economy.html`.

European Parliament (2018). “REGULATION (EU) 2018/1807 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 November 2018 on a framework for the free flow of non-personal data in the European Union.” Regulation REGULATION (EU) 2018/1807, URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1807&from=EN.

European Parliament and Council of the European Union (1995). “Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.” URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp (2019). “Big Data and Firm Dynamics.” *AEA Papers and Proceedings*, 109, 38–42.

Farré, Lídia and Francesco Fasani (2013). “Media exposure and internal migration — Evidence from Indonesia.” *Journal of Development Economics*, 102, 48–61.

Ferrara, Eliana La, Alberto Chong, and Suzanne Duryea (2012). “Soap Operas and Fertility: Evidence from Brazil.” *American Economic Journal: Applied Economics*, 4(4), 1–31.

Fevziu, Blendi, Robert Elsie, Majlinda Nishku, and Blendi Fevziu (2018). *Enver Hoxha: the iron fist of Albania*. Bloomsbury Publishing, URL https://doi.org/10.5040/9781350986268?locatt=label:secondary_bloomsburyCollections. OCLC: 1128170332.

Galanxhi, Emira, Elena Misja, Desareta Lameborshi, Mathias Lerch, Philippe Wanner, and Janine Dahinden (2004). *Migration in Albania: population and housing census 2001*. No. 18 in 2001 population and housing census, INSTAT, Tirane.

Gambardella, Alfonso, Sohvi Heaton, Elena Novelli, and David J. Teece (2021). “Profiting from Enabling Technologies?” *Strategy Science*, 6(1), 75–90.

Gentzkow, Matthew and Jesse M. Shapiro (2008). “Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study *.” *Quarterly Journal of Economics*, 123(1), 279–323.

Gerardi, Kristopher S. and Adam Hale Shapiro (2009). “Does Competition Reduce Price Dispersion? New Evidence from the Airline Industry.” *Journal of Political Economy*, 117(1), 1–37.

Graef, Inge, Raphael Gellert, Nadezhda Purtova, and Martin Husovec (2018). “Feedback to the Commission’s Proposal on a Framework for the Free Flow of Non-Personal Data.” *SSRN Electronic Journal*.

Gëdeshi, Ilir and Russell King (2019). "The Albanian scientific diaspora: can the brain drain be reversed?" *Migration and Development*, 10(1), 19–41.

Gërmenji, Etleva and Lindita Milo (2011). "Migration of the skilled from Albania: brain drain or brain gain?" *Journal of Balkan and Near Eastern Studies*, 13(3), 339–356.

Hao, Karen (2021). "Andrew Ng: Forget about building an AI-first business. Start with a mission." *MIT Technology Review*.

Helfrich, Magdalena and Fabian Herweg (2016). "Fighting collusion by permitting price discrimination." *Economics Letters*, 145, 148–151.

Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou (2017). "Deep Learning Scaling is Predictable, Empirically." *arXiv:1712.00409 [cs, stat]*. ArXiv: 1712.00409.

House, Christopher, Christian Proebsting, and Linda Tesar (2018). "Quantifying the Benefits of Labor Mobility in a Currency Union." Tech. Rep. w25347, National Bureau of Economic Research, Cambridge, MA, URL `http://www.nber.org/papers/w25347.pdf`.

Ichihashi, Shota (2020). "Online privacy and information disclosure by consumers." *American Economic Review*, 1(110), 569–95.

Inoue, Atsushi and Gary Solon (2010). "Two-Sample Instrumental Variables Estimators." *The Review of Economics and Statistics*, 92(3), 557–561.

Jensen, Robert and Emily Oster (2009). "The Power of TV: Cable Television and Women's Status in India *." *Quarterly Journal of Economics*, 124(3), 1057–1094.

Johnson, Justin, Andrew Rhodes, and Matthijs R. Wildenbeest (2023). "Platform design when sellers use pricing algorithms." *forthcoming Econometrica*, pp. 1–55.

Jones, Charles I. and Christopher Tonetti (2020). "Nonrivalry and the Economics of Data." *American Economic Review*, 110(9), 2819–2858.

Kearney, Melissa S. and Phillip B. Levine (2015). "Media Influences on Social Outcomes: The Impact of MTV's *16 and Pregnant* on Teen Childbearing." *American Economic Review*, 105(12), 3597–3632.

Kearney, Melissa S. and Phillip B. Levine (2019). "Early Childhood Education by Television: Lessons from *Sesame Street*." *American Economic Journal: Applied Economics*, 11(1), 318–350.

Kerber, Wolfgang (2016). "A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis." p. 23.

La Ferrara, Eliana (2016). "Mass Media and Social Change: Can we Use Television to Fight Poverty?" *Journal of the European Economic Association*, 14(4), 791–827.

Lang, Julia (2022). "Employment effects of language training for unemployed immigrants." *Journal of Population Economics*, 35(2), 719–754.

Liu, Qihong and Konstantinos Serfes (2007). "Market segmentation and collusive behavior." *International Journal of Industrial Organization*, 25(2), 355–378.

Lochmann, Alexia, Hillel Rapoport, and Biagio Speciale (2019). "The effect of language training on immigrants' economic integration: Empirical evidence from France." *European Economic Review*, 113, 265–296.

Lundqvist, Björn (2018). "Competition and Data Pools." *Journal of European Consumer and Market Law*, 7(4), 146–154.

Mai, Nick (2004). "'Looking for a More Modern Life...': the Role of Italian Television in the Albanian Migration to Italy." *Westminster Papers in Communication and Culture*, 1(1), 3.

McKenzie, David, John Gibson, and Steven Stillman (2013). "A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad?" *Journal of Development Economics*, 102, 116–127.

Meng, Tong, Xuyang Jing, Zheng Yan, and Witold Pedrycz (2020). "A survey on machine learning for data fusion." *Information Fusion*, 57, 115–129.

Mitchell, Tom M. (1999). "Machine learning and data mining." *Communications of the ACM*, 42(11), 30–36.

Mumuni, Alhassan and Fuseini Mumuni (2022). "Data augmentation: A comprehensive survey of modern approaches." *Array*, p. 100258.

Mundell, Robert (1961). "A Theory of Optimum Currency Areas." *The American Economic Review*, 51(4), 657–665.

Olken, Benjamin A. (2009). "Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages." *American Economic Journal: Applied Economics*, 1(4), 1–33.

Pacini, David and Frank Windmeijer (2016). “Robust inference for the Two-Sample 2SLS estimator.” *Economics Letters*, 146, 50–54.

Peiseler, Florian, Alexander Rasch, and Shiva Shekhar (2022). “Imperfect information, algorithmic price discrimination, and collusion*.” *The Scandinavian Journal of Economics*, 124(2), 516–549.

Pesando, Luca Maria, Valentina Rotondi, Manuela Stranges, Ridhi Kashyap, and Francesco C. Billari (2021). “The Internetization of International Migration.” *Population and Development Review*, 47(1), 79–111.

Prono, Luciano, Philippe Bich, Mauro Mangia, Fabio Pareschi, Riccardo Rovatti, and Gianluca Setti (2022a). “A Naturally Pruned Non-conventional Neural Network trained with Softmax Shrinking.” In *Proceedings of the IEEE International Conference on Artificial Intelligence Circuits and Systems - AICAS 2022*.

Prono, Luciano, Mauro Mangia, Fabio Pareschi, Riccardo Rovatti, and Gianluca Setti (2022b). “A Non-Conventional Sum-and-Max Based Neural Network Layer for Low Power Classification.” In *Proceedings of the IEEE International Symposium on Circuits and Systems - ISCAS 2022*.

Prüfer, Jens and Christoph Schottmüller (2020). “Competing with Big Data.” *Journal of Industrial Economics*.

Rey, Patrick and Thibaud Verge (2004). “Bilateral Control with Vertical Contracts.” *The RAND Journal of Economics*, 35(4), 728.

Riley, Shawn, Stephen DeGloria, and Robert Elliot (1999). “A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity.” *Intermountain Journal of Sciences*, 5(1-4).

Sarvimäki, Matti and Kari Hämäläinen (2016). “Integrating Immigrants: The Impact of Restructuring Active Labor Market Programs.” *Journal of Labor Economics*, 34(2), 479–508.

Schaefer, Maximilian and Geza Sapi (2022). “Complementarities in Learning from Data: Insights from General Search.” p. 43.

Schmid, Lukas (forthcoming). “The Impact of Host Language Proficiency on Migrants’ Employment Outcomes.” *American Economic Review: Insights*.

Shrestha, Slesh A. (2017). “No Man Left Behind: Effects of Emigration Prospects on Educational and Labour Outcomes of Non-migrants.” *The Economic Journal*, 127(600), 495–521.

Werden, Gregory and Luke Froeb (1994). "The Effects of Mergers in Differentiated Products Industries: Logit Demand and Merger Policy." *The Journal of Law, Economics, and Organization*, 10(2), 407–426.

Zech, Herbert (2016). "Data as a Tradeable Commodity." In *European Contract Law and the Digital Single Market*, edited by Alberto De Franceschi, 1 ed., pp. 51–80. Intersentia, URL `https://www.cambridge.org/core/product/identifier/CBO9781780685212A026/type/book_part`.

Zhu, Hongwei, Stuart E. Madnick, and Michael D. Siegel (2008). "An Economic Analysis of Policies for the Protection and Reuse of Noncopyrightable Database Contents." *Journal of Management Information Systems*, 25(1), 199–232.

Appendix A

First Appendix

A.1 Italian TV Shows Watched by Albanian Migrants

In this section we report the distribution of TV shows watched by Albanians migrants. In 1991, Dorfles and Gatteschi (1991) interviewed 311 Italian speaking Albanian migrants just arrived in Italy, 301 declared watching Italian television back home in Albania. They were asked to list all Italian shows they usually watched in Albania. Table A.1 (extracted from Dorfles and Gatteschi (1991)) reports the results. We report a brief description and a Wikipedia link for the TV shows that count at least 4% of answer: they are all entertainment shows. *TG1*, the main Italian news show, appears in less than 3% of answers.

Domenica In: "Domenica in is an entertainment Italian TV show on air on Rai 1 since 1976.." (Wikipedia:https://it.wikipedia.org/wiki/Domenica_in)

Fantastico: "Fantastico was an Italian TV variety show broadcast saturday prime time on Rai 1 from 1979 to 1980 and from 1981 to 1992.."(Wikipedia: https://it.wikipedia.org/wiki/Fantastico_(programma_televisivo)

Piacere RaiUno:"During the show there were prank calls, dance, music and interview to popular TV characters. There were also some time dedicated to information about issues of different Italian city" (Wikipedia: https://it.wikipedia.org/wiki/Piacere_Raiuno)

La Domenica Sportiva: "La Domenica Sportiva is the oldest sport show of Italian television." (Wikipedia:https://it.wikipedia.org/wiki/La_Domenica_Sportiva)

Crème Caramel: variety and Vaudeville Tv show (Wikipedia: https://it.wikipedia.org/wiki/Cr%C3%A8me_Caramel_(programma_televisivo)

Quark: TV show to popularize science (Wikipedia: https://it.wikipedia.org/wiki/Quark_(programma_televisivo))

Sanremo: Broadcasted music festival (Wikipedia: https://it.wikipedia.org/wiki/Festival_

Table A.1: Italian TV Shows Preferences of Albanians Migrants

	Italian TV Shows Preferences of Albanians Migrants		
	Obs. (1)	Share (2)	Type (3)
Domenica In	183	25%	Entertainment
Fantastico	92	13%	Entertainment
Piacere Raiuno	86	12%	Entertainment
Domenica sportiva	84	11%	Entertainment
Creme Caramel	35	5%	Entertainment
Quark	34	5%	Entertainment
Sanremo	30	4%	Entertainment
La Piovra	25	3%	Entertainment
Lunedi film	23	3%	Entertainment
Tg1	21	3%	Information
Mercoledì sport	19	3%	Entertainment
Big	17	2%	Entertainment
Tg1 7	13	2%	Information
Discoring	13	2%	Entertainment
Speciale Tg1	12	2%	Information
Linea Verde	12	2%	Entertainment
Viaggio intorno all uomo	10	1%	Entertainment
Colpo Grosso	10	1%	Entertainment
Telemike	9	1%	Entertainment
Notte Rock	7	1%	Entertainment

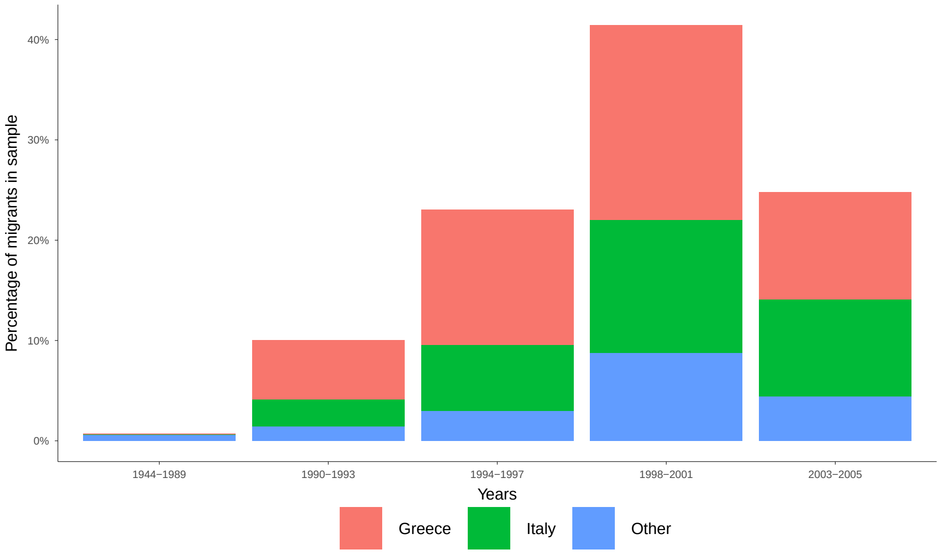
Source: Data derived from Dorfles and Gatteschi (1991).

`di_Sanremo`)

A.2 Emigration Patterns of Albanians 1990-2005

Figure A.1 reports the distribution of migrants over time and over destination country in the sample. For each 4 years bracket, we compute the share of the migrants in the sample and their destination country choice. We can see that the share of migrants prior to 1990 (end of the regime) is almost null, and that it increasing until 1998-2001, when it peaks before decreasing afterwards.

Figure A.1: Emigration Patterns of Albanians 1990-2005. Source: *Siblings sample* from 2005 Living Standard Measurement Survey Albania



Notes: Numerator is the numer of migrants in a given period, denominator the number of migrants in the sample.

A.3 GIS Data for Municipality

All distance indicators are computed in kilometers with the same method: rather than selecting an arbitrary center for each municipality from where to compute distance, we transformed each municipality into rasters with a 30x30 meters grid size.[1] Then, for each cell of each municipality, we computed the straight line distance from the center of the cell to each of the considered geographical points. For each municipality, we then computed the mean distance of all the raster cells it encompasses.

Figure A.2 presents the administrative map of Albania, overlapped with the sampling of the Albania 2005 Living Standards Measurement Survey and signal coverage. The Online Appendix reports the number of observations in the LSMS by district.

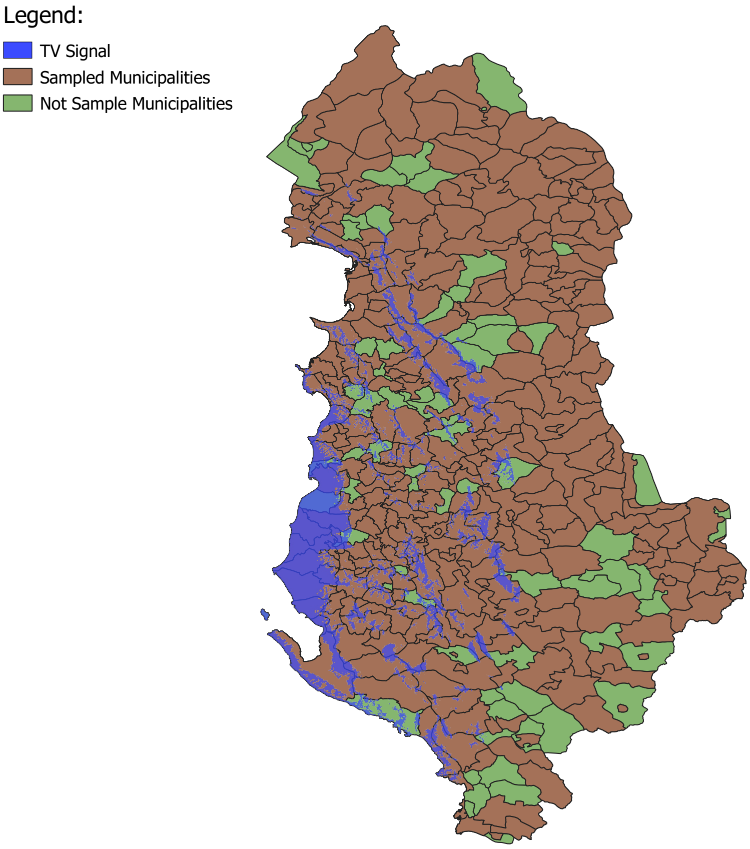
A.3.1 LSMS Questions

1. **Foreign language proficiency**

    - Did [NAME] speak English in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO

[1] A raster is a geographical object that subdivides a geographical area into cells of equal size.

Figure A.2: Albanian municipalities and Italian Television signal



- Did [NAME] speak Italian in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
- Did [NAME] speak Greek in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
- Did [NAME] speak another foreign language in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO

2. **Internal migration**

- Prior to the current residence, has [NAME] ever lived in a different municipality in Albania? 1 YES, 2 NO
- Which district and municipality/comuna did [NAME] move from?
- In what year did [NAME] move to the current residence?
- Prior to this residence in [MUNICIPALITY/ COMUNA], did [NAME] live in a different municipality/ comuna in Albania? (the loop start over until they track all internal migration

history)

3. **Spouse/children away from home**

    - Please list your spouse, if he or she is no longer living in the household, and all the children 15 years old and over who are no longer living in this household. (Include all children of head and/or spouse.)
    - Did [NAME] speak English in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
    - Did [NAME] speak Italian in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
    - Did [NAME] speak Greek in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
    - Where does [NAME] currently live? If in Albania, then ask for district and municipality/comuna. If abroad, country and place.
    - In what year did [NAME] move abroad to [COUNTRY]?

4. **Siblings**

    - Ask all the questions to the household head, and then to the spouse of the household head. If no spouse, leave the second section blank.
    - Please list the first name of up to SEVEN brothers and sisters for both the head of the household and the spouse. Begin with those brothers and sisters living abroad.
    - In which country does [NAME] currently live? Indicate the country in which [NAME] spent the most time during the past year
    - How many years has [NAME] lived in [COUNTRY]?

5. **TV ownership in 1990**

    - Did your household own any of the following items in January 1990?Colour TV, Black & White TV, Tape player/CD player, Refrigerator, Washing machine, Sewing/knitting machine , Satellite dish, Bicycle.

6. **Education**

    - What is the highest grade you have completed in school? None 0; ”8 or 9 years” school 1; Secondary general 2; Vocational 2-3 years 3; vocational 4/5 years 4; University- Albania 5; University- abroad 6; Post-graduate- Albania 7; Post-graduate- abroad 8.

7. **Past migration**

- Who provided information on where to go and/or how to find work during this most recent migration episode? (MAIN SOURCE) Family/Relatives in Albania; Family/Relatives Abroad; Friends in Albania; Friends Abroad; Previous Personal Experience; Neighbours; TV, Radio, Newspaper or Book; Internet; Other

A.4 Balance Test

Table A.3 reports the results of regressions of age and sex ratio on Italian TV signal using the identification strategy specified in Equation (1.1) and the siblings dataset. We expect to see no effects on age and sex as in our specification being exposed is as good as random. Columns (1) to (4) report results using two measures of signal: share of the municipality exposed to the signal (Signal) and share of urban territory of the municipality exposed in 1986. Concerning the sex ratio we can see there is no effect of signal on the probability of being a man (Sex ratio). There is, however, an effect on age of two point half year, significant at the 10% level, when using Signal as treatment, while there is no significant effect when using Signal II as treatment. Although an older sample would bias down results (negative correlation between age and migration), we want to rule out the possibility that it is a symptom of issues in our identification strategy. In particular, we show that this result comes from the fact that sets of siblings who have all migrated are absent from the dataset.

Table A.2 shows that in close to a port and to the Greek border areas individuals have significantly higher migration rate and are on average older. As age and migration probability are negatively correlated (-0.2 in our sample) and as the correlation of age between the *listing sibling* and the siblings is extremely high (0.72 in our sample), very low migration migration cost that characterized these areas caused young individuals to be excluded from the sample, via migration of entire set of siblings. Indeed, as we discuss in Data Section 1.5, the siblings dataset contains a sample of Albanians that can be either in Albania or abroad, thus containing migrants, but sets of siblings that all migrated, and only children cannot be a part of the sample. Thus, as close to the port/Greek border areas are differently than average exposed to the signal (Figure 1.1, Table A.2), we observe that Signal affects age although there can not be an effect of our treatment on age. Table A.3 specifications (5)-(6), we show that when we subset for individuals living in 1990 in area farther away from the first quartile of distance to the port (31 km) and first decile of the Greek border (49 km) there is no effect of signal exposure on age.

Table A.2: Distance to Ports and Greek Border: Signal, Migration, and Age

	Distance to Ports		Distance to Greece	
Variable	$\leq$ 31km	$>$ 31km	$\leq$ 49km	$>$ 49km
Migration	0.210	0.150	0.210	0.160
Age	49.4	46.9	49.2	47.6
Signal	0.180	0.060	0.020	0.110

Notes: We display the average share of individuals being abroad, the average age, the average signal exposure for individuals in/outside the first quartile of distance to the nearest port (30.461 km) and in/outside the first decile of distance to the Greek border (48.834 Km).

A.5 Greek Community in Albania

Albania in 1990 was populated by Greek minorities, for many individuals in the survey, Greek is not exactly a foreign language.[2] According to the 1989 Albanian census there were 60 000 Greeks in Albania in 1990, while according to the Greek government they were 300 000. The Communist government recognized 99 villages as *minority zones* in the southern districts of Gjirokastër, Sarandë and Delvina and authorized schooling in both Greek and Albanian for the whole dictatorship period. However, aside from the official recognized minority zones, Greek communities were scattered in many other areas of the country. This is why in Table 2 we control for the Greek community indicators: i) Greek ethnicity ii) Greek maternal language iii) Use Greek language daily at home iv) Use Greek language with extended family members v) Orthodox religion. In Table A.4, we show that Greek language proficiency in 1990 is correlated with all the Greek community indicators. While language proficiency in Greek is measured in 1990 and the Greek community indicators are available solely for 2005, these indicators are all stables condition that can be assumed to be equal in 1990 and in 2005.

There would be no issue if there were no correlation between signal exposure and Greek settlements in Albania but unfortunately it is not the case. Indeed all *minority zones* are located in districts where signal is available: Delvine (share of signal= .08, share of Greek speaker in 1990=.04), Gjirokastër (share of signal= .01, share of Greek speaker in 1990=.47), Sarande (share of signal= .17, share of Greek speaker in 1990=.22). District F.E. are unable to account for this accidental correlation as the 99 villages could be located precisely in the municipalities exposed to the signal. We consider that finding these 99 villages where both Albanian and Greek was taught would not add much to the research, as Greek community

[2] This section is based on the Wikipedia page *Greeks in Albania*: https://en.wikipedia.org/wiki/Greeks_in_Albania

Table A.3: Balance test: Age and Sex Ratio

	Full Sample				Restricted Sample	
	(1)	(2)	(3)	(4)	(5)	(6)
	Age	Sex ratio	Age	Sex ratio	Age	Sex ratio
Signal	2.434*	-0.0253			0.441	-0.0573
	(1.254)	(0.0232)			(2.531)	(0.0450)
Signal II			1.497	-0.0208		
			(1.175)	(0.0197)		
Observations	27666	27666	27666	27666	18985	18985
Clusters	310	310	310	310	180	180

Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness

Notes: The table reports OLS estimates of the effect of exposure to Italian Television on the probability of being a man (Sex ratio) and an individual's age. All specifications exploit the *siblings* dataset. Specifications (1) to (4) exploit the full sample, while specifications (5) and (6) restrict the sample to individuals that lived in 1990 farther away from 30.461 km from the closest port (first quartile of distance to the port) and farther away from 48.834 km from the Greek border (first decile of distance to Greece). Signal is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Signal II is the share of the urban area in the municipality exposed in 1986. Clustered standard errors in parentheses. Standard errors are clustered at the Municipality level.

were also scattered across Albania, and it is beyond the scope of the paper.

In Table A.5 we present the regressions (1)-(2)-(3) of Table 1.2 adding as controls Greek community indicators: results are not affected.

A.6 Instrumental Variable Regressions

A.6.1 One-Sample

We can perform an instrumental variable regression on the dataset of children. For this dataset, we have data on both Italian language proficiency in 1990 and on migratory outcomes. We can thus perform a classical IV regression, using Italian television access as an instrument for Italian language proficiency.

The results are presented in Table A.6. When we analyze the entire sample, the first stage of the regression lacks robustness, leading to insignificant results. Upon examining the subset of educated

Table A.4: Correlation between Greek Proficiency in 1990 and Greek Community Indicator

	Proficient in Greek in 1990
Variable	Correlation coefficient
Greek Ethnic Group	0.51
Greek Maternal Language	0.52
Greek spoken daily at home	0.47
Greek spoken with extended family	0.36
Orthodox religion	0.22

Source: 2005 Living Standard Measurement Survey, World Bank and INSTAT. Base dataset.

individuals, the first stage proves significant, but the second stage does not. Ultimately, the sample size is insufficient and lacks the precision needed to draw any significant conclusions from this data.

A.6.2 Two-Samples

As outlined in Section 1.7.2, we opted not to conduct two-sample instrumental variable regressions within the primary text body, due to the risk of overestimating the treatment effect. However, we've included this regression analysis in the appendix for the reader's reference. The two stages are represented as follows:

$$IT_{i,m,d} = \alpha_0 + \beta_0 \times Sig_m + \gamma_0 \times Dist_m + \theta_0 \times Geo_m + \sum_{d=1}^{36} \alpha_{0,d} \times Distr_d + \varepsilon_{i,m,d} \tag{A.1}$$

$$MIG_{i,m,d} = \alpha_1 + \beta_1 \times \hat{IT}i,m,d + \gamma_1 \times Dist_m + \theta_1 \times Geo_m + \sum d = 1^{36}\alpha_{1,d} \times Distr_d + u_{i,m,d} \tag{A.2}$$

In equation (A.1), we utilize individuals from the base dataset, where the variable $IT_{i,m,d}$ denotes Italian language proficiency, taking a value of either 0 or 1. In equation (A.2), we draw on individuals from the siblings dataset, with the variable $MIG_{i,m,d}$ representing migration — 1 if an individual migrated and 0 if they did not.

Accurate estimation of the variance-covariance matrix is far from trivial. Inoue and Solon (2010) demonstrates that the variance-covariance matrix needs to be amplified in the case of two-sample instrumental variable regressions. Pacini and Windmeijer (2016) delineates how to perform this adjustment in the presence of heteroskedastic errors. To date, no formal proof has been derived for clustered stan-

Table A.5: Italian television effect on foreign language proficiency: controlling for Greek community indicators

	Italian	English	Other
	(1)	(2)	(3)
Signal	0.0635*	0.0118	0.00191
	(0.0334)	(0.0181)	(0.0231)
Obs.	11040	11040	11040
Clusters	322	322	322

Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness, Distance to Greece, Greek community indicators

Notes: The table reports OLS estimates of the effect of exposure to Italian television on foreign language proficiency in 1990. (1)-(4) exploit the sample of the LSMS surveyed individuals. The dependent variable is the reported capability of speaking Italian, English, Other (category any other language). The main explanatory variable, Signal, is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Clustered standard errors in parentheses. Standard errors are clustered at the municipality level.

dard errors. We have applied the methodology as described by Etgeton (2018), inspired by Pacini and Windmeijer (2016), but not yet analytically proven.

The results are presented in the corresponding table. Due to the notably high estimated standard errors in relation to the small sample size, the results are inconclusive. The standard errors in the second stage are simply too excessive to yield any significant conclusions.

Table A.6: IV Regression Results

	(1)	(2)
	First Stage	
Signal	0.110	0.481^{**}
	(0.071)	(0.213)
	Second Stage	
Italian Proficiency	0.053	0.717
	(.620)	(0.464)
Sample	All	H. Skill
Obs.	4,714	425
Clusters	256	104
Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness, Distance to Greece, Greek community indicators		

Notes: The dependent variable is migration, the endogenous variable Italian proficiency, and the instrument signal exposure (see section 1.5 for details). Both stages are linear probability model (see section 1.6). Standard errors are clustered at the municipality of living in 1990 level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.7: Two samples 2SLS

	Dependent variable:		
	Italian Proficiency	Abroad	
	(1)	(2)	(3)
Signal	0.070^{**}		
	(0.033)		
Italian Proficiency		−0.063	2.455
		(1.282)	(29.188)
Sample	All	All	H. Skills
Obs.	11,040	27,666	2,153
Clusters	322	310	128

Notes: We use two-sample 2SLS IV regression. The dependent variable, migration, is only in the siblings dataset, the endogenous variable Italian proficiency only in the base dataset.

A.6.3 Additional Results

In this section we present additional results absent from the main body of the paper. In particular, we show the null effect of signal exposure on presumably low skill individuals and the heterogeneity of the effect over the family dimension and the housing dimension. (1)-(2) show the absence of an effect on migration of low skilled individuals. (3) to (5) confirm that the effect on migration is decreasing in family size.[3] (6)-(7) show that there are no effects of signal exposure for individuals who were living in smaller housing in 1990 (2nd and 3rd quartile for specification (6) and 4th quartile for specification (7)). We confirm that there is no effect for *low skilled* individuals as proxy by sibling education, family and housing dimension.

Table A.8: The effect of Italian Television exposure on migration decision: Other Results

	Low Skilled		Family Dimension: # Children			Housing Dimension	
	Abroad	Italy	Abroad	Abroad	Abroad	Abroad	Abroad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Signal	-0.0103	0.0157	0.248**	0.0502	0.0435	-0.0140	0.0343
	(0.0307)	(0.0172)	(0.111)	(0.0459)	(0.0354)	(0.0414)	(0.0787)
Obs.	25513	25513	517	5926	10986	5922	2442
Clusters	302	302	172	268	287	259	143
Sample	L. Skill	L. Skill	≤ 2	≤ 5	≤ 6	Quart. 2-3	Quart. 1

Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness, Distance to Greece, Greek community indicators

Notes: The table reports OLS estimates of the effect of exposure to Italian television on: (1) probability to be abroad and (2) probability to be in Italy on low skilled individuals (*listing brother* education less equal than secondary education), (3)-(4)-(5) probability to be abroad given the number of children of the family an individual was raised in, (6)-(7) probability to be abroad given housing dimension. All specifications use subset of the *siblings* dataset. The main explanatory variable, Signal, is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Clustered standard errors at the municipality level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

[3] For sake of brevity we do not show family dimension higher than 5. The coefficient steadily decline with family dimension.

Appendix B

Second Appendix

A1 Proofs

This sections contains proofs that are omitted from the text. To make the reading of these profs more agile, we confine some self-contained results in specific additional appendices.

Proof of Lemma 1. The result follows from the fact that at low n_i the function $\eta(.)$ is convex. Here we prove in the case of the efficient analytics, but the proof can be clearly adapted to other environments discussed in the paper. Two cases are possible. (i) First,

$$\frac{\partial\eta(0,n_{-i})}{\partial n_i}\left(\sum_{i=1}^{P}\alpha_i-\varepsilon\right)>\delta+\gamma_i,$$

in which case condition (2.5) is uniquely satisfied for relatively large $n_i > 0$ so that $\eta(n)$ is concave and it corresponds to the maximum of the total surplus $W(n)$.

(ii) Second, $\frac{\partial\eta(0,n_{-i})}{\partial n_i}\left(\sum_{i=1}^{P}\alpha_i-\varepsilon\right)<\delta+\gamma_i$, in which case condition (2.5) can be satisfied at two values of n_i, a small one which however corresponds to the region where $\eta(n)$ is convex and it is thus a minimum of $W(n)$, or at a large $n_i > 0$ so that $\eta(n)$ is concave. In either case, the efficient n_i is bounded away from zero.

Proof of Proposition 1. Part (i) on equilibrium existence is lengthier and proven in a separate Section A3. Part (ii) follows from the discussion in the text. Part (iii) relies on the same reasoning followed in case (ii) analyzed in the proof of Lemma 1. When costs are sufficiently high, we have $\frac{\partial\eta(0,\hat{n}_{-i})}{\partial n_i}\alpha_i<\gamma_i$ when $\hat{n}_{-i}=0$, so that, anticipating that the other producers will not provide data, the optimal n_i is nil. □

Proof of Proposition 2. Let n^* be the data provision profile that maximizes the aggregator's payoff and let

$$B^*_{agg}=\sum_{j=1}^{P}\left[\alpha_j\eta(n^*)-(\gamma_i+\delta)n^*_i-B_i^{self}\right]-\varepsilon\eta(n^*), \tag{B.1}$$

be the maximum payoff the aggregator can obtain. If one ore more B_i^{self} are sufficiently high then the optimal data profile contemplates that one or more of the producers are excluded. Since the synergy between datasets positively affects the value of the analytics relying on data aggregated from different datasets, it clearly does not affect any of the B_i^{self}.[1] Result (i) then follows.

To see why the aggregator may prefer to exclude the producer with the highest B_i^{self} consider three producers indexed 1,2,3, with equal costs γ_i but valuing the analytics differently: $\alpha_1 >> \alpha_2 = \alpha_3$. Consider the case in which having all producers joining the analytics is not optimal for the producer that would obtain a profit $W(n^*) - (B_{\text{agg}}^{\min} + \sum_{j=1}^{P} B_j^{\min}) < 0$ (recall the aggregator extracts all the surplus up to the producers outside options). Ceteris paribus the value of the analytics $\eta(.)$ is higher with more equally sized dataset, i.e. including producers 2 and 3 and excluding the large-value producer 1, rather than sharing data from producer 1 and one of the other producers 2 or 3, especially so if the synergy is large enough. In addition, since $B_1^{\text{self}} > B_2^{\text{self}} = B_3^{\text{self}}$, the transfer to convince-in producer 1 is higher than that to either of producers 2 and 3. □

Proof of Proposition 3. Point (i). We study to what extent the aggregator can obtain the data that maximize surplus, i.e. $n^* = n^{\text{w}}$. Condition (2.5) implies that these data depend (also) on the costs, $n_i^{\text{w}} = n_i^{\text{w}}(\gamma_i)$ for $i = 1, \dots, P$, so that $C(.)$ in (2.18) is a function of γ_1 and γ_2. For producers with identical costs $\gamma_1 = \gamma_2 = \bar{\gamma}$, we would have $n_1^{\text{w}}(\gamma_1) = n_2^{\text{w}}(\gamma_2) = \bar{n}$. Clearly $C(\bar{\gamma}, \bar{\gamma}) = 0$ so that the sign of $C(.)$ in that neighborhood depends on its derivative with respect to the parameter that changes. In the following we will assume that γ_1 varies and thus we have to compute

$$\begin{aligned}
\frac{\partial C(\bar{\gamma},\bar{\gamma})}{\partial \gamma_1} &= \alpha_2 \left[\frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} + \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \\
&\quad -\alpha_1 \left[\frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} + \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \\
&= \alpha_2 \frac{\partial \eta'(\bar{n})}{\partial n_2} \left[\frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} \right] - \alpha_1 \frac{\partial \eta'(\bar{n})}{\partial n_1} \left[\frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \\
&= (\alpha_2 - \alpha_1) \frac{\partial \eta'(\bar{n})}{\partial n_1} \left[\frac{\partial n_2^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\text{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \qquad \text{(B.2)}
\end{aligned}$$

where we have exploited the fact that, due to the symmetry of η, $\frac{\partial \eta'(\bar{n})}{\partial n_1} = \frac{\partial \eta'(\bar{n})}{\partial n_2}$.

Now recall the necessary conditions (2.5) and derive them with respect to γ_1 to obtain,

$$\sum_{j=1}^{P} \frac{\partial^2 \eta(\bar{n})}{\partial n_1 \partial n_j} \frac{\partial n_j^{\text{w}}(\bar{n})}{\partial \gamma_1} = 1 \qquad \text{(B.3)}$$

$$\sum_{j=1}^{P} \frac{\partial^2 \eta(\bar{n})}{\partial n_i \partial n_j} \frac{\partial n_j^{\text{w}}(\bar{n})}{\partial \gamma_1} = 0, \quad i = 2, \dots, P \qquad \text{(B.4)}$$

The variations of n_i^{w} with respect to γ_1 can be derived by considering (B.3) and (B.4) as a linear system and solve it for $\frac{\partial n_j^{\text{w}}(\bar{n})}{\partial \gamma_1}$ for $j = 1, \dots, P$. Due to the symmetry of η we have $\frac{\partial^2 \eta(\bar{n})}{\partial n_i^2} = a \leq 0$

[1] In appendix A2.1 we show how the level of the synergy can be parameterized in the Example.

(assuming that the equilibrium localized in the convex part of η) and $\frac{\partial^2\eta(\bar{n})}{\partial n_i \partial n_j} = b \geq 0$, both independently of i and j. Now note that if the P-dimensional vector x is such that $x_j = \frac{\partial n_j^{\mathrm{w}}(\bar{n})}{\partial \gamma_1}$, y is the P-dimensional vector with $y_1 = 1$ and $y_i = 0$ for $i > 1$, I is the $P \times P$ identity, and U is the $P \times P$ constant unit matrix, for (B.3) and (B.4) we have to solve a system $Ax = y$ for x, with a coefficient matrix

$$A = (a-b)I + bU$$

The matrices A, I and U are symmetric and, since I and U commute, they all have the same eigenvectors that we collect as columns in the following $P \times P$ matrix

$$E = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{pmatrix}$$

whose inverse is

$$E^{-1} = \frac{1}{P} \begin{pmatrix} 1 & -1 & -1 & \dots & -1 & -1 \\ 1 & -1 & -1 & \dots & -1 & P-1 \\ 1 & -1 & -1 & \dots & P-1 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & -1 & P-1 & \dots & -1 & -1 \\ 1 & P-1 & -1 & \dots & -1 & -1 \end{pmatrix}$$

The only non-null eigenvalues of U is P, and thus one of the eigenvalues of A is $a + b(P-1)$ while the other $P-1$ eigenvalues of A are equal to $a-b$. Hence, defining

$$D = \begin{pmatrix} a+b(P-1) & 0 & 0 & \dots & 0 \\ 0 & a-b & 0 & \dots & 0 \\ 0 & 0 & a-b & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a-b \end{pmatrix}$$

we can solve $Ax = EDE^{-1}x = y$ to obtain

$$x = ED^{-1}E^{-1}y$$

Note now that $E^{-1}y = \frac{1}{P}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ and thus $D^{-1}E^{-1}y = \frac{1}{P}\begin{pmatrix} 1/[a+b(P-1)] \\ 1/a-b \\ \vdots \\ 1/a-b \end{pmatrix}$ and, finally,

$$x = \frac{1}{P}\begin{pmatrix} \frac{1}{a+b(P-1)} + \frac{P-1}{a-b} \\ \frac{1}{a+b(P-1)} - \frac{1}{a-b} \\ \vdots \\ \frac{1}{a+b(P-1)} - \frac{1}{a-b} \end{pmatrix}$$

from which, recalling the components of x, we get

$$\frac{\partial n_1^{\mathrm{w}}(\bar{n})}{\partial \gamma_1} = \frac{(P-2)b+a}{(a-b)[(P-1)b+a]} \tag{B.5}$$

$$\frac{\partial n_i^{\mathrm{w}}(\bar{n})}{\partial \gamma_1} = \frac{b}{(b-a)[(P-1)b+a]} \quad i = 2,\ldots,P \tag{B.6}$$

and thus

$$\frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} = \frac{1}{b-a} \tag{B.7}$$

Finally, this can be plugged into (B.2) to obtain (B.8) with (B.9), as in the main text:

$$\frac{\partial C(\bar{\gamma},\bar{\gamma})}{\partial \gamma_1} = K\,(\alpha_2 - \alpha_1) \tag{B.8}$$

with

$$K = \frac{\frac{\partial \eta'(\bar{n})}{\partial n_1}}{\frac{\partial^2 \eta'(\bar{n})}{\partial n_1 \partial n_2} - \frac{\partial^2 \eta'(\bar{n})}{\partial n_1^2}} \geq 0. \tag{B.9}$$

With this and the discussion in the main text, point (i) follows.

For point (ii) note first that with a unique contract (n,Q) condition (2.5) is unattainable. In particular, assume first that the aggregator includes all producers in the analytics. It must then choose Q so that, for given n, $Q = min_i\{\alpha_i \eta(n,...,n) - \gamma_i n - B_i^{\min}\}$. Substituting, the aggregator's objective function becomes

$$-\eta\,(n,...,n)\,\varepsilon - \bar{\delta} - P\delta n + P \times min_i\{\alpha_i \eta\,(n,...,n) - \gamma_i n - B_i^{\min}\}$$

Clearly if producers are identical, this expression is equivalent to $W(n,...,n)$ and there is no loss in value. When this is not the case, the more producers differ in terms of γ_i or α_i, the larger are the distortions in the equilibrium dataset $(n^*,...,n^*)$ with respect to efficient analytics n^w. Clearly, if $min_i\{\alpha_i \eta(n,...,n) - \gamma_i n - B_i^{\min}\}$ is very low, the aggregator may prefer to exclude some of the producers with very low value thus increasing the transfer Q. □

Proof of Proposition 4. The first order condition for n_i is,

$$\alpha_i \frac{\partial \eta(n_i,\hat{n}_{-i})}{\partial n_i} - \varepsilon \frac{\partial \eta(n_i,n_{-i})}{\partial n_i} = \gamma_i + \delta. \tag{B.10}$$

The second order conditions can be rewritten as

$$(\alpha_i - \varepsilon)\eta_{ii} \leq 0 \tag{B.11}$$

where $\eta_{ii} = \frac{\partial^2 \eta(n_i, \hat{n}_j)}{\partial n_i^2}$ and

$$\frac{(\alpha_1 - \varepsilon)(\alpha_2 - \varepsilon)}{\varepsilon^2} > \bar{\sigma} \tag{B.12}$$

where

$$\bar{\sigma} = \frac{\eta_{12}^2}{\eta_{11}\eta_{22}} \tag{B.13}$$

measures the (relative) strength of the positive externality of data. The higher the $\bar{\sigma}$, the more the analytics benefits from data variety. However, if $\bar{\sigma}$ is high, then (B.12) fails. In fact, it must be $\varepsilon < \alpha_i$ and $\eta_{ii} < 0$. In this case, the optimal analytics cannot have all producers providing data and exclusion must occur. □

Proof of Proposition 5 The program of each individual producer writes:

$$\max_{n_i} \frac{(1-2\rho)(1-\rho)}{(2-3\rho)^2}(\theta - \bar{c} + \eta(n))^2 - \gamma_i n_i$$

(ii) The marginal gain of providing an additional unit of data $\frac{(1-2\rho)(1-\rho)}{(2-3\rho)^2}$ is a decreasing function of ρ. As the marginal cost is constant, it follows than an increase in ρ implies a reduction of the level of data provision that maximizes producers' profits.

(ii) The marginal gain of providing one additional unit of data tends to zero for as ρ tends to $1/2$. Since the marginal cost of data γ_i is constant, there exists a $\hat{\rho}$ such that $\forall \rho > \hat{\rho}$, the optimal level of data provision is 0. In addition, $\forall \rho > \hat{\rho}$ a single producer that decided contributing strictly positive amount of data would, a fortiori (being the only contributor), would be better off contributing no data.

(iii) This possibility is directly shown with the Example. □

Proof of Proposition 6 We separately prove the different points in the Proposition.

(i) Consider two identical producers that are both contributing with their data in equilibrium, then we consider the case where only one producer contributes. The problem of the aggregator writes,

$$\begin{aligned} \max_{n_1,n_2} 2\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta - \bar{c} + \eta(n_1,n_2))^2 - (\gamma+\delta)(n_1+n_2) - \varepsilon\eta(n_1,n_2) \\ + \sum_{i=1}^{2}\left[\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\eta(n_i,0)(\theta-\bar{c}) - \left(\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\right)^2(\eta(n_i,0))^2\right], \end{aligned} \tag{B.14}$$

where the first term represents producers' profits (net of the aggregator's costs), and the term in the second line describes the producers' outside options. The optimal amount of data n_{agg}^* satisfy, for any i,

$$\left[4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta-\gamma+\eta(n))-\varepsilon\right]\frac{\partial\eta(n^*_{\text{agg}})}{\partial n_i} \\ -2\left(\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\right)^2\frac{\partial\eta(n^*_{\text{i,agg}},0)}{\partial n_i}\eta(n^*_{\text{i,agg}},0) \\ +\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\frac{\partial\eta(n^*_{\text{i,agg}},0)}{\partial n_i}(\theta-\gamma)=\gamma+\delta,\quad\forall i, \tag{B.15}$$

We now assume that $\eta(n)$ is sufficiently concave at the optimal amount of data, in particular that its concavity is such that $(\eta(n))^2$ is concave too.

Consider the solution n^{**} of the following equation,

$$\left[4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta-\gamma+\eta(n))-\varepsilon\right]\frac{\partial\eta(n^{**})}{\partial n_i} \\ +\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\frac{\partial\eta(n_i^{**},0)}{\partial n_i}(\theta-\gamma)=\gamma+\delta, \tag{B.16}$$

then $n^{**}\geq n^*_{agg}$. By the same token, consider the solution n^{***} of,

$$\left[4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta-\gamma+\eta(n))-\varepsilon\right]\frac{\partial\eta(n^{***})}{\partial n_i} \\ +\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\frac{\partial\eta(n^{***})}{\partial n_i}(\theta-\gamma+\eta(n^{***}))=\gamma+\delta, \tag{B.17}$$

then $n^{***}\geq n^{**}$. This follows from the following observations: from the Scope property of $\eta(n)$, $\frac{\partial\eta(n)}{\partial n_i}\geq\frac{\partial\eta(n_i,0)}{\partial n_i}$; second, including $\eta(n)$ in the parenthesis increases the solution n^{***} of the equation with respect to n^{**}.

Finally, note that the efficient amount of data that solve the social planner's problem,

$$\max_{n_1,n_2}\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}(\theta-\bar{c}+\eta(n))^2-(\gamma+\delta)(n_1+n_2)-\varepsilon\eta(n) \tag{B.18}$$

are determined by the following condition,

$$\left[\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}(\theta-\gamma+\eta(n^*))-\varepsilon\right]\frac{\partial\eta(n^*)}{\partial n_i}=\gamma+\delta. \tag{B.19}$$

Since it is always the case ($\forall\rho\in[0,1/2]$) that:

$$\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}\geq4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}+\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}, \tag{B.20}$$

it follows that $n^* > n^{***}$ and, finally, $n^* > n^*_{agg}$.

In the case where one producer only contributes, the maximum amount of data the aggregator might asks corresponds to the scenario where the one producer chosen finds itself in a monopoly. The aggregator solves:

$$\max_{n_1} \frac{1}{4(1-\rho)}(\theta - c + \eta(n_1,0))^2 - (\gamma+\delta)n_1 - \varepsilon\eta(n_1,0) \tag{B.21}$$

Which always yields a lower amount of data than the resolution of problem (B.18).

□

(ii) Let $n^*(\rho)$ be the solution of problem (2.30). Maximized profits of the aggregator when contracting with both producers write:

$$\begin{aligned}\pi^{agg}(n^*(\rho),\rho) &= 2\frac{(1-\rho)(1-2\rho)}{(2-3\rho)^2}(\theta - c + \eta(n^*(\rho)))^2 - \gamma_1 n_1 - \gamma_2 n_2 - \varepsilon\eta(n^*(\rho)) \\ &- \sum_{i=1}^{2}\frac{(1-\rho)(1-2\rho)}{(2-3\rho)^2}\left[\theta - c - \frac{\rho(1-\rho)}{(2-\rho)(1-2\rho)}\eta(n_i^*(\rho),0)\right]^2\end{aligned}$$

Using the enveloppe theorem we differentiate this function with respect to ρ (Denote $\eta(n_1,n_2) = \eta(n_1,0) + \eta(n_2,0) + \eta_{12}$):

$$\begin{aligned}\frac{\partial\pi^{agg}(n^*(\rho),\rho)}{\partial\rho} &= \sum_{i=1}^{2}(\theta - c)\eta(n_i,0)\frac{4(2-9\rho+9\rho^2-3\rho^3)}{(2-\rho)^2(2-3\rho)^3} \\ &+ \sum_{i=1}^{2}\eta(n_i,0)^2\frac{\rho(-24+132\rho-270\rho^2+259\rho^3-124\rho^4+25\rho^5)}{(1-2\rho)^2(2-3\rho)^3(2-\rho)^3} \\ &- \frac{2\rho}{(2-3\rho)^3}\left(2(\theta-c)\eta_{12} + 2\sum\eta(n_i,0)\eta_{12} + 2\eta(n_1,0)\eta(n_2,0) + \eta_{12}^2\right)\end{aligned} \tag{B.22}$$

The above can only be positive if the first line is positive. For $\rho \approx 0.307$ the first line equals 0. Hence, we know that $\forall\rho > 0.307$ profits are decreasing.

Consider the profits of the aggregator when contracting with only one producer:

$$\begin{aligned}\pi^{agg}(n^*(\rho),\rho) &= \frac{(\rho^2-4\rho+2)(1-\rho)}{(2-3\rho)^2(2-\rho)}\eta(n_i^*(\rho),0)\left(2(\theta-c) + \frac{\rho^2-4\rho+2}{(1-2\rho)(2-\rho)}\eta(n_i^*(\rho),0)\right) \\ &- \gamma_i n_i - \varepsilon\eta(n_i^*(\rho),0)\end{aligned}$$

If the analytics costs reduction is not too small relative to the size of the market ($\eta(n_i^*(\rho),0) \geq \frac{(\theta-c)}{150}$), profits are growing with ρ.

Profits when contracting with both producers are decreasing for $\rho \geq .307$ and tend to 0 as ρ tends to $1/2$. Profits when excluding one producer are always positive and growing with ρ. There must exist $\hat{\rho}$

such that $\forall\rho \geq \hat{\rho}$ the profits made by excluding one producer are higher than the profits of contracting with all producers.

(iii) Ex-post overall welfare, in the case where both producers contract with the aggregator, is given by:

$$\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}(\theta - c + \eta(n_1,n_2))^2$$

The level of competition that maximizes overall welfare is implicitly defined by the following equation:

$$\frac{\partial\eta(n_1,n_2)}{\partial\rho} = -\frac{(1-2\rho)}{(1-\rho)(3-5\rho)}(\theta - c + \eta(n_1,n_2))$$

Assuming symmetry, and using the implicit function theorem, we can analytically define $\frac{\partial\eta(n_1,n_2)}{\partial\rho}$, the above equation is equivalent to:

$$-2\frac{\partial\eta(n_1,n_2)}{\partial n_i}\frac{\frac{\partial\Pi_i}{\partial\rho}}{\frac{\partial\Pi_i}{\partial n_i}+\frac{\partial\Pi_i}{\partial n_{-i}}} = -\frac{(1-2\rho)}{(1-\rho)(3-5\rho)}(\theta - c + \eta(n_1,n_2)) \tag{B.23}$$

where:

$$\begin{aligned}
\frac{\partial\Pi_i}{\partial n_i} =& \frac{\partial^2\eta(n_1,n_2)}{\partial n_i^2}\left[4\frac{(1-\rho)(1-2\rho)}{(2-3\rho)^2}(\theta - c + \eta(n_1,n_2) - \varepsilon\right] + \left(\frac{\partial\eta(n_1,n_2)}{\partial n_i}\right)^2 4\frac{(1-\rho)(1-2\rho)}{(2-3\rho)^2}\\
&+2\frac{\rho(1-\rho)^2}{(2-\rho)(2-3\rho)^2}\frac{\partial^2\eta(n_i,0)}{\partial n_i^2}\left[\theta - c - \frac{\rho(1-\rho)}{(2-\rho)(1-2\rho)}\eta(n_i,0)\right]\\
&-2\left(\frac{\partial\eta(n_i,0)}{\partial n_i}\right)^2\frac{\rho^2(1-\rho)^3}{(1-2\rho)(2-\rho)^2(2-3\rho)^2}\\
\frac{\partial\Pi_i}{\partial n_{-i}} =& \frac{\partial^2\eta(n_1,n_2)}{\partial n_1\partial n_2}\left[4\frac{(1-\rho)(1-2\rho)}{(2-3\rho)^2}(\theta - c + \eta(n_1,n_2) - \varepsilon\right] + \frac{\partial\eta(n_1,n_2)}{\partial n_1}\frac{\partial\eta(n_1,n_2)}{\partial n_2}4\frac{(1-\rho)(1-2\rho)}{(2-3\rho)^2}\\
\frac{\partial\Pi_i}{\partial\rho} =& -\frac{\partial\eta(n_1,n_2)}{\partial n_i}\frac{4\rho(\theta - c + \eta(n_1,n_2))}{(2-3\rho)^3} - \frac{\partial\eta(n_i,0)}{\partial n_i}\frac{4(-2+5\rho-5\rho^2+2\rho^3)}{(2-\rho)^2(2-3\rho)^3}(\theta - c)\\
&+2\frac{\partial\eta(n_i,0)}{\partial n_i}\eta(n_i,0)\frac{(1-\rho)^2\rho(-8+28\rho-34\rho^2+17\rho^3)}{(1-2\rho)^2(2-\rho)^3(2-3\rho)^3}
\end{aligned}$$

Substituting for these expressions, the solution to equation (B.23) must not coincide with the aggregator's profit maximizing ρ, i.e.s such that $\frac{\partial\pi(n^*(\rho),\rho)}{\partial\rho} = 0$. Hence, the level of competition that would maximize overall welfare will not, in general, be chosen by the aggregator. The running Example presented in this paper is such a case where we can show that the preferred level of competition for the aggregator is too low with respect to level of competition that would maximize welfare.

□

A2 The value of data

As illustrated in the text, the value of data is modelled in two steps. The value of data from a single dataset is $\upsilon : \mathbb{R}^+ \mapsto [0, \eta^{\max}]$, where $\eta^{\max}$ is the maximum value attainable. Data from different producers are combined with the monotonically increasing convex function $\Upsilon : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\Upsilon(0) = 0$, endowed with the commutative and associative aggregating operation $\oplus : \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that,

$$\upsilon' \oplus \upsilon'' = \Upsilon\left(\Upsilon^{-1}(\upsilon') + \Upsilon^{-1}(\upsilon'')\right)$$

The value of a set of data contributions is thus,

$$\eta\left(n_1, \ldots, n_P\right) = \bigoplus_{j=1}^{P} \upsilon\left(n_i\right) = \bigoplus_{j=1}^{P} \eta\left(n_i\right) \tag{B.24}$$

Since Υ is convex, we have that $\oplus$ is superadditive, Bruckner and Ostrow (1962). This implies that, for any $1 \leq p \leq P$, we have

$$\eta\left(n_1, \ldots, n_P\right) \geq \eta\left(n_1, \ldots, n_{p-1}\right) + \eta\left(n_p, n_{p+1}, \ldots, n_P\right) \tag{B.25}$$

With the invariance of η with respect to permutations of its arguments, this definition implies that the best way of extracting value from multiple data sets is not to keep them partitioned into different algorithms but to accumulate them within the one of the aggregator. Note that with P producers the maximum gain from the aggregate of all data is $\Upsilon\left(P\Upsilon^{-1}\left(\eta^{\max}\right)\right) \geq P\eta^{\max}$.

From the convexity of the function Υ we obtain that the function of the value of data has the *increasing difference property*. In fact, assume $n_1' \geq n_1$ and $n_2' \geq n_2$ and set $\Delta_1 = \Upsilon^{-1}(\upsilon(n_1')) - \Upsilon^{-1}(\upsilon(n_1))$, and $\Delta_2 = \Upsilon^{-1}(\upsilon(n_2')) - \Upsilon^{-1}(\upsilon(n_2))$. Since Υ and υ are non-decreasing we have $\Delta_1, \Delta_2 \geq 0$ and

$$\begin{aligned}
\Delta\eta' &= \eta(n_1', n_2') - \eta(n_1, n_2') \\
&= \Upsilon\left(\Delta_1 + \Delta_2 + \Upsilon^{-1}(\upsilon(n_1)) + \Upsilon^{-1}(\upsilon(n_2))\right) - \Upsilon\left(\Delta_2 + \Upsilon^{-1}(\upsilon(n_1)) + \Upsilon^{-1}(\upsilon(n_2))\right)
\end{aligned}$$

as well as

$$\begin{aligned}
\Delta\eta &= \eta(n_1', n_2) - \eta(n_1, n_2) \\
&= \Upsilon\left(\Delta_1 + \Upsilon^{-1}(\upsilon(n_1)) + \Upsilon^{-1}(\upsilon(n_2))\right) - \Upsilon\left(\Upsilon^{-1}(\upsilon(n_1)) + \Upsilon^{-1}(\upsilon(n_2))\right)
\end{aligned}$$

Finally, the convexity of Υ implies $\Delta\eta' \geq \Delta\eta$ and thus the increasing difference property for η when $P = 2$. The same property for $P > 2$ descends from the associativity of η.

A2.1 Functions and parameters for the running Example

Following the approach to obtain $\eta(.)$ in the previous section, here we specify the functions and parameters we use for the running "Example" we use throughout the paper.

The individual data-value function $\upsilon(.)$ is defined from its derivative in 0 (υ_0'), and the amount of data ($\bar{n}$) beyond which the convex region becomes concave with the asymptotic maximum ($\upsilon^{\max}$). The data-value function at such a flex values $\upsilon_{\bar{n}} = \upsilon(\bar{n})$. We use the following expression:

$$\upsilon(n) = \upsilon^{\max} \begin{cases} an + bn^2 & \text{for } n \leq \bar{n} \\ 1 - ce^{-d(n-\bar{n})} & \text{for } n > \bar{n} \end{cases}$$

with $a = \frac{\upsilon_0'}{\upsilon^{\max}}$, $b = \frac{\upsilon_{\bar{n}} - a\bar{n}}{\bar{n}^2}$, $c = 1 - \upsilon_{\bar{n}}$, $d = \frac{a\bar{n} - 2\upsilon_{\bar{n}}}{\bar{n}(\upsilon_{\bar{n}} - 1)}$.

The function Υ aggregating the different contributions and accounting for their positive heterogeneity, is modelled with

$$\Upsilon(\upsilon) = (\upsilon + 1)^{1+\sigma} - 1$$

in which $\sigma = 0$ corresponds to no synergy while increasing $\sigma > 0$ generates the super-additive effect in (B.25). In fact, such an Υ implies

$$\sum_{i=1}^{P} \upsilon_i \leq \bigoplus_{i=1}^{P} \upsilon_i \leq -1 + \prod_{i=1}^{P} (\upsilon_i + 1)$$

where the lower bound is attained for $\sigma = 0$ while the upper bound is asymptotically achieved for $\sigma \rightarrow \infty$.

With this, it is convenient to measure the amount of positive interaction between databases in the $[0, 1]$ range with a Scope index,

$$s = \frac{\upsilon^{\max} \oplus \upsilon^{\max} - 2\upsilon^{\max}}{(\upsilon^{\max})^2}$$

In the Example, we use the following parameters: $\upsilon_0' = 3/4$, $\bar{n} = 1/4$, $\upsilon^{\max} = 1$, $\upsilon_{\bar{n}} = 1/4$ and $\sigma = 2$. For the aggregator, we set $\bar{\delta} = 0$, $\delta = 1/2$, $\varepsilon = 1/3$, and $B^{\min}_{\text{agg}} = 0$. For the producers we assume $P = 2$ with $B^{\min}_1 = B^{\min}_2 = 0$, and we consider two configurations:

- The symmetric configuration, with $\alpha_1 = \alpha_2 = 3/4$, and $\gamma_1 = \gamma_2 = 3/4$.
- The asymmetric configuration, with $\alpha_1 = 1$, $\alpha_2 = 1/4$, $\gamma_1 = 1/2$, and $\gamma_2 = 1/4$.

A3 Equilibrium existence

In this appendix we prove the existence of a Nash equilibrium in the case of the free-analytics. The existence in the other cases discussed in the paper follows similar arguments. The proof requires three preliminary Lemmas. In the case of free analytics, producers maximize the following problem:

$$\max_{n_i} \alpha_i \eta(n) - \gamma_i n_i \tag{B.26}$$

Lemma 2. *If a data profile $n^* = (n_1^*, ..., n_j^*)$ is solution to the maximization then $\forall i$, n_i is either such that*

1. $\frac{\partial^2 \eta(n_i, n_{-i})}{\partial n_i^2} \leq 0$
2. $n_i = 0$

Proof. Let n_i^* be such that $n_i^* > 0$ and $\frac{\partial^2 \eta(n_i, n_{-i})}{\partial n_i^2} > 0$. Then, either:

1. $\alpha_i \frac{\partial \eta(n_i,n_{-i})}{\partial n_i} \geq \gamma_i$. Then n_i^* is not a solution because profits increases when increasing data provision. It is beneficial to increase data provision until $\alpha_i \frac{\partial \eta(n_i,n_{-i})}{\partial n_i} = \gamma_i$ which implies $\frac{\partial^2 \eta(n_i,n_{-i})}{\partial n_i^2} \leq 0$ by the properties of η.

2. $\alpha_i \frac{\partial \eta(n_i,n_{-i})}{\partial n_i} < \gamma_i$. Then profits can be increased by decreasing data provision until it reaches $n_i = 0$.

Hence it cannot be that n_i^* be such that $n_i^* > 0$ and $\frac{\partial^2 \eta(n_i,n_{-i})}{\partial n_i^2} > 0$. □

Lemma 3. *For each producer i, there are only two candidate best-responses to each data profile n_{-i}, either $n_i = 0$ or n_i s.t $\alpha_i \frac{\partial \eta(n_i,n_{-i})}{\partial n_i} = \gamma_i$*

Proof. Either the equilibrium lies in the concave space of η or at $n_i = 0$. If the candidate equilibrium lies in the concave space, it should be such that the first derivative of the objective function is equal to 0. □

Denote as $Br_i(n_{-i})$ the positive response of i to n_{-i}. Additionally, denote $n_{-i}^{alt} < n_{-i}$ when each data in n_{-i}^{alt} is at least weakly inferior to those in n_{-i} and one is strictly inferior.

Lemma 4. *If*

$$B_i(0,n_{-i}) \geq B_i(Br_i(n_{-i}),n_{-i}) \tag{B.27}$$

Then, $\forall n_{-i}^{alt} < n_{-i}$ and $\forall n_i$:

$$B_i(0,n_{-i}^{alt}) > B_i(n_i,n_{-i}^{alt}) \tag{B.28}$$

Proof. Let $B(0,n_{-i}) \geq B(Br_i(n_{-i}),n_{-i})$, then:

$$\alpha_i \eta(0,n_{-i}) \geq \alpha_i \eta(Br_i(n_{-i}),n_{-i}) - \gamma_i n_i^* > \alpha_i \eta(Br_i(n_{-i}^{alt}),n_{-i}) - \gamma_i Br_i(n_{-i}^{alt}) \implies \tag{B.29}$$

$$\alpha_i \eta(0,n_{-i}^{alt}) > \alpha_i \eta(Br_i(n_{-i}^{alt}),n_{-i}) - \gamma_i Br_i(n_{-i}^{alt}). \tag{B.30}$$

In other terms moving from n_{-i} to n_{-i}^{alt} implies a stronger reduction of $\eta(0,n_{-i})$ than of $\eta(Br_i(n_{-i}^{alt}),n_{-i})$ because of the positive cross-derivative of the function η. □

Lemma 4 states the continuity of the best response of producers when not providing any data: a reduction in other players' data provisions does not change the best response of a producer not providing any data.

We now combine these results and show that there always exists an equilibrium with a free-analytics. Given Lemma 2 and 3, we know that a producer is either providing a level of data such that $\frac{\partial \eta(n_i,n_{-i})}{\partial n_i} = \frac{\gamma_i}{\alpha_i}$ or is not providing any data. With $P \in \mathbb{N}$ producers, in any equilibrium, each producer belongs to one of two possible sets. Set $\mathscr{I}$ is the set of producers not providing any data. Set $\mathscr{J}$ is the set of producers providing strictly positive amount of data.

Consider the candidate equilibrium n^* such that $\forall i$, n_i^* is such that $\frac{\partial\eta(n_i,n_{-i})}{\partial n_i} = \frac{\gamma_i}{\alpha_i}$. Then either (i) each producer i is playing its best response, which implies n^* is an equilibrium, (ii) or some producers would be better off providing no data.

If n^* is not an equilibrium, some producers of set $\mathscr{J}$ move to set $\mathscr{I}$. Data provision in set $\mathscr{J}$ are re-adjusted and n^{**} is the new candidate equilibrium. Then as before, either (i) each producer i is playing its best response, so that n^{**} is an equilibrium, (ii) or some producers would be better off providing no data. If n^{**} is not an equilibrium, some producers of set $\mathscr{J}$ move to set $\mathscr{I}$. Importantly, by continuity of the best response function when not providing any data (Lemma 4), the movement of producers from set $\mathscr{J}$ to set $\mathscr{I}$ does not change the best response of producers in set $\mathscr{I}$.

This method can be iterated until either (ii) all producers are in set $\mathscr{I}$, (ii) no producer in set $\mathscr{J}$ would be better off not providing any data. After a finite number of iterations, one of these two cases will be reached. In both cases, Lemma 3 ensures that all producers in set $\mathscr{I}$ are playing their best responses. In the former case it ensures a Nash equilibrium is reached since all players are in set $\mathscr{I}$. In the latter case, as no producer in set $\mathscr{J}$ would be better off not providing any data, all producers are playing their best responses. □

A4 Secret contracts

A necessary (but not sufficient) condition such that the results of the FOCs corresponds to an equilibrium is that the Hessian is negative semi-definite in the solution. Hence the first minor must be negative and the second positive. The first is given by:

$$\frac{\partial^2 B_{\text{agg}}(n_1,n_2)}{\partial n_1^2} = \alpha_1 \frac{\partial^2 \eta(n_1,n_2^e)}{\partial n_1^2} - \varepsilon \frac{\partial^2 \eta(n_1,n_2)}{\partial n_1^2} \tag{B.31}$$

It is negative in solution if:

$$(\alpha_1 - \varepsilon)\frac{\partial^2 \eta(n_1,n_2)}{\partial n_1^2} \leq 0 \tag{B.32}$$

To compute the second minor we start by computing the cross-derivative of the objective function:

$$\frac{\partial^2 B_{\text{agg}}(n_1,n_2)}{\partial n_1 \partial n_2} = -\varepsilon \frac{\partial^2 \eta(n_1,n_2)}{\partial n_1 \partial n_2} \tag{B.33}$$

The Hessian is negative semi-definite if the following condition holds:

$$B_{\text{agg}_{11}} B_{\text{agg}_{22}} > B_{\text{agg}_{12}}^2 \tag{B.34}$$

Which corresponds to:

$$\frac{(\alpha_1 - \varepsilon)(\alpha_2 - \varepsilon)}{\varepsilon^2} > \frac{\eta_{12}^2}{\eta_{11}\eta_{22}} \tag{B.35}$$

The right hand side of the equation is bounded between 0 and 1 provided $\eta(.)$ has a negative semi-definite Hessian itself. So we know that any combination such that the left hand side is superior to 1 verifies the necessary condition. More interestingly, if the condition is not verified (meaning ε is high) the solution fails.

B1 On the properties of the analytics' value

When discussing properties of Machine Learning tools, the features of Scale and Scope (embedded in our function Υ discusses in section A2) are usually assumed with generic reference to common practitioners' experience (e.g. Duch-Brown et al. (2017) and the recent surveys in Computer Science Meng et al. (2020)). However, it is difficult to find neat accounts of these intuitive properties. The problem in developing a complete theory of these feature is that nowadays machine learning models are highly non-linear and are the result of complicated and expensive training procedure that are often designed by trial and error. Nevertheless, in this section we provide a direct account of these important properties. We think that, although specific, the analysis contained in this appendix provides some useful insights in its own.

In a first subsection we show the Scale property, and the fact that the function υ in section A2 is, first convex, then concave and bounded. We also show that aggregating data and making the resulting analytics available to producers results in an higher utility with respect to a situation in which each producers uses local data to compute a local analytics.

In a second subsection, the Scope property so that when multiple producers contribute data with a sufficient *diversity*, the value of the analytics is larger than what can be obtained from the same aggregated amount of data coming from a single producer. This gives ground to the features of the aggregating operator $\oplus$ in appendix A2.

We also obtain a byproduct from this analysis, which is of value even if we do not directly exploit is in the paper (at least so far). In particular, we define the notion of *complexity* that is the number of scalar quantities (e.g., sensor readings, configuration settings, etc.) used to characterize each single piece of data, in other terms the dimensionality of each data point in the data sets. We show that, once the number of features in data is enough to allow classification, increasing the complexity of the data negatively affects the value that one can squeeze out of a given amount of data.

The approach that we use in the following two subsections is to define suitably simplified classification models on which the effect of training can be theoretically anticipated, either exactly or for the worst-case scenarios.

Then, when we want to assess the effect of statistically diverse contributions to the data set, we run Monte Carlo simulations varying the actual data points to see how performance varies with the characteristics of the data set.

B1.1 Scale Property

We assume that producers' goods come in units and that due to the fabrication process there is a certain probability that a unit is defective. For simplicity's sake we assume that defective units cannot be sold nor repaired, and must be identified and discarded as the cost of selling a potentially defective unit is too high for the producers. To be on the safe side, each producer will inevitably discard some good units, thus waiving to part of its revenues.

The purpose of the analytics is to maximize revenues and thus to minimize the number of good units that are not sold. It does so by exploiting that fact that the producers have a common technological basis (e.g, manufacturing machines employ the same kind of electrical motor, units are assembled by means of the same welding process, etc.) and thus, even though the final products may be different, each unit is characterized by the same D numerical features (e.g., sensor readings acquired during production, measurements from final quality inspection, etc.) that we will indicate with $x_1,\dots,x_D$ and compound into the D-dimensional vector $x \in X \subset \mathbb{R}^D$, where X indicates the whole range that we assume to be uniformly spanned by the production.

Difference between producers is modelled by assuming that the i-th one produces units corresponding to a proper subset $X_i \subset X$. We do not require $X_i \cap X_j = \emptyset$ for $i \neq j$, though it may be the case. To each data point x there is a label $y \in \{-1,+1\}$ whose negative value indicates a defective unit.

Assume also that the features are sufficient to tell defectives units from good ones by simple monotonic discrimination, i.e., there is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that units for which $f(x) \geq \tau$ are defective, while those for which $f(x) < \tau$ are non-defective, where τ is some unknown threshold.

Each producer collects a finite number n_i of data $x \in \hat{X}_i \subset X_i$ and associated label y assessing their defectiveness. Information is extracted from these data in the form of an estimation $\hat{\tau}_i$ of τ that is then used in a straightforward binary classifier to assess new units in the same X_i.

To be on the safe side (no false negative), the i-th producer estimates

$$\hat{\tau}_i = \max_x \left\{ f(x) \middle| x \in \hat{X}_i \wedge y > 0 \right\} \leq \tau \tag{B.36}$$

With this, the i-th producer sells the units whose features satisfy $f(x) \leq \hat{\tau}_i$ that, in our uniform setting, generate a revenue proportional to the n-dimensional volume $V(X_i \cap \hat{H}_i)$ of the intersection between Z_i and the set $\hat{H}_i$ defined by the above discriminating inequality.

Alternatively, the producers may give their data to the aggregator and let it perform the estimation

$$\hat{\tau} = \max_x \left\{ f(x) \middle| x \in \bigcup_{j=1}^{P} \hat{X}_j \wedge y > 0 \right\} = \max_{j=1,\dots,P} \{\hat{\tau}_j\} \leq \tau \tag{B.37}$$

defining the set $\hat{H}$ such that $f(x) \leq \hat{\tau}$ that is nothing but $\hat{H} = \bigcup_{j=1}^{P} \hat{H}_j$ and may be used by the i-th producer to sell its units falling into $X_i \cap \hat{H}$ and generate a revenue proportional to $V(X_i \cap \hat{H})$.

Clearly, since $\hat{\tau}_i \leq \hat{\tau}$ for every $i = 1,\dots,P$ we have $\hat{H}_i \subseteq \hat{H}$ and thus $V(X_i \cap \hat{H}_i) \leq V(X_i \cap \hat{H})$. Hence, the total revenue of all producers satisfy

$$\sum_{i=1}^{P} V(X_i \cap \hat{H}) \geq \sum_{i=1}^{P} V(X_i \cap \hat{H}_i) \tag{B.38}$$

that is equivalent to say that the value of the analytics based on the aggregated data (left-hand side of (B.38)) is larger than the sum of the values of the analytics based on separate datasets (right-hand side of (B.38)).

In this toy case, concavity is also very easy to see if we assume that all the X_i are compact subsets of $\mathbb{R}^n$ within which sampled units are independent and uniformly distributed.

Consider a sequence of datasets $\hat{X}^{(t)}$ of increasing size, such that $\hat{X}^{(1)} \subset \hat{X}^{(2)} \subset \cdots \subset X$. From (B.36) and (B.37) we get that the corresponding estimates $\hat{\tau}^{(t)}$ are such that $\hat{\tau}^{(1)} \leq \hat{\tau}^{(2)} \leq \cdots \leq \tau$ and that $\lim_{t\to\infty} \hat{\tau}^{(t)} = \tau$ and thus that $V(X \cap \hat{H}^{(1)}) \leq V(X \cap \hat{H}^{(2)}) \leq \cdots \leq V(X \cap H)$ but $\lim_{t\to\infty} V(X \cap \hat{H}^{(t)}) = V(X \cap H)$.

Hence, data-dependent revenues are increasing with the number of samples and have an upper bound that is also their limit. This implies that their trend mus be asymptotically convex.

For the sake of clarity we consider a particular simplified setting in which $f(x) = \sum_{j=1}^{D} x_j^2$, $X = \{x | f(x) \leq 1\}$, with $\tau < 1$.

Given n samples allowing an estimation $\hat{\tau}(n)$, the probability that an additional sample produces an estimation that is larger than a certain ξ is

$$Z(n, \xi) = \begin{cases} 1 & \text{if } \xi < \hat{\tau}(n) \\ \frac{W(\tau) - W(\xi)}{W(\tau) - W(\hat{\tau}(n))} = \frac{\tau^D - \xi^D}{\tau^D - \hat{\tau}(n)^D} & \text{if } \hat{\tau}(n) \leq \xi \leq \tau \\ 0 & \text{if } \xi > \tau \end{cases}$$

where $W(r) = \pi^{D/2}\Gamma^{-1}(D/2 + 1)r^D$ is the volume of the D-dimensional sphere with radius r. Hence, the average of the estimation with $n+1$ samples is

$$\mathbf{E}[\hat{\tau}(n+1)] = \int_0^\infty Z(n, \xi)\mathrm{d}\xi = \frac{D}{D+1} \frac{\tau^{D+1} - \hat{\tau}(n)^{D+1}}{\tau^D - \hat{\tau}(n)^D} \tag{B.39}$$

while the variance is $\mathbf{E}\left[\hat{\tau}(n+1)^2\right] - \mathbf{E}[\hat{\tau}(n+1)]^2$ with

$$\mathbf{E}\left[\hat{\tau}(n+1)^2\right] = \int_0^\infty \xi^2 \frac{\partial(1-Z)}{\partial \xi}\mathrm{d}\xi = \int_0^\infty 2\xi Z(n, \xi)\mathrm{d}\xi = \frac{D}{D+2} \frac{\tau^{D+2} - \hat{\tau}(n)^{D+2}}{\tau^D - \hat{\tau}(n)^D}$$

Since $\lim_{n\to\infty} \mathbf{E}[\hat{\tau}(n+1)] = \tau$ and $\lim_{n\to\infty} \mathbf{E}\left[\hat{\tau}(n+1)^2\right] = \tau^2$ the variance of $\hat{\tau}(n+1)$ vanishes and we may assume that it coincides with its average. With this, (B.39) reads

$$\frac{\hat{\tau}(n+1)}{\tau} = \frac{D}{D+1} \frac{1 - \left(\frac{\hat{\tau}(n)}{\tau}\right)^{D+1}}{1 - \left(\frac{\hat{\tau}(n)}{\tau}\right)^{D}}$$

To compare such a trend with what is commonly observed in machine learning applications, we may concentrate on the relative loss $\varepsilon(n) = 1 - \frac{\hat{\tau}(n+1)}{\tau}$ that is always non-negative and pushed to zero by the training. The above relationship yields

$$
\begin{aligned}
\varepsilon(n+1) &= 1-\frac{D}{D+1}\frac{1-(1-\varepsilon(n))^{D+1}}{1-(1-\varepsilon(n))^{D}} \\
&= 1-\frac{D}{D+1}\frac{1-e^{(D+1)\ln(1-\varepsilon(n))}}{1-e^{D\ln(1-\varepsilon(n))}} \\
&\simeq 1-\frac{D}{D+1}\frac{1-e^{-(D+1)\varepsilon(n)}}{1-e^{-D\varepsilon(n)}} \\
&= 1-\frac{D}{D+1}e^{-\varepsilon(n)/2}\frac{\sinh\left(\frac{D+1}{2}\varepsilon(n)\right)}{\sinh\left(\frac{D}{2}\varepsilon(n)\right)} \\
&\simeq 1-e^{-\varepsilon(n)/2}\simeq\frac{1}{2}\varepsilon(n)
\end{aligned}
$$

whose exponentially vanishing trend is coherent with common observations on state-of-the-art deep learning algorithms Hestness et al. (2017).

All the above shows that global utility increases when producers cooperate but it is ultimately concave and bounded.

A further piece of information can be obtained by noting that the above calculations are based on the a priori availability of a model and of its training strategy. Hence, they fail to take into account what happens at the very beginning of the design of a data-driven application. In real-world applications, the first available data lots are commonly used to set up and tune the ingestion stage (i.e., the data processing pipeline that acquires and transforms raw, incomplete, possibly incoherent data into normalized quantities that can be fed into machine learning blocks), the architecture of the trainable blocks (layers, connections, substructure, etc.) and the training strategy (algorithm, losses, etc.). True, valuable information is obtained from data only after this set up phase is over, and thus the first data lots have an (apparent) marginal utility that is much lower than the marginal utility of data lots that enter a smoothed processing pipeline. This causes the function υ to be convex for small arguments, i.e., when the first data are acquired and used to set up the analytics.

B1.2 Scope Property

To study the Scope Property we may set $X=[0,1]^D$ and assume that the truth is $y=\operatorname{sgn}(x_D-{}^1\!/_2)$. This is clearly an abstract setting and assumes that some change of coordinates has been performed to transform the original data into this domain X in which discriminating between the two classes is trivial.

Trivial as it may be, discrimination must be learnt from samples and thus we have to define a model with some adjustable parameters and identify how these parameters may be set by training.

We use a simple 1-neuron piece-wise linear model

$$
y=\operatorname{sgn}\left(x_D-\max_{k=1,\dots,D-1}\{\alpha_k x_k\}-\max_{k=1,\dots,D-1}\{\beta_k x_k\}-\frac{1}{2}\right)
$$

that mimics the behaviour of a *neuron* with excitation and inhibition weights aggregated with a max instead of a sum, as it has been recently proposed to allow efficient complexity reduction of complex neural networks Prono et al. (2022a)Prono et al. (2022b).

Once set, the parameters identify a piecewise-affine manifold

$$x_D = g_{\alpha,\beta}(x_2,\dots,x_{D-1}) = \max_{k=1,\dots,D-1}\{\alpha_k x_k\} + \max_{k=1,\dots,D-1}\{\beta_k x_k\} - \frac{1}{2}$$

that separates the points that the model marks as positive (above the manifold) from the points that the model marks as negative (below the manifold). Clearly, the optimum value for the parameters is $\alpha_k = \beta_k = 0$.

If this is not the case, all the data points such that $1/2 \leq x_D \leq g_{\alpha,\beta}(x_1,\dots,x_{D-1})$ are false negatives, while all the data points such that $g_{\alpha,\beta}(x_1,\dots,x_{D-1}) \leq x_D \leq 1/2$ are false negatives.

Hence, given data points in $\hat{X} \subset X$, the worst possible model from the point of view of the false negatives is characterized by the parameters

$$\begin{aligned}\hat{\alpha}'_k &= \min_{x\in\hat{X}\wedge y>0}\left\{\frac{x_D - 1/2}{x_k}\right\}\\ \hat{\beta}'_k &= 0\end{aligned}$$

while the worst possible model from the point of view of the false positives is characterized by the parameters

$$\begin{aligned}\hat{\alpha}''_k &= 0\\ \hat{\beta}''_k &= \min_{x\in\hat{X}\wedge y<0}\left\{\frac{1/2 - x_D}{x_k}\right\}\end{aligned}$$

Since data are uniformly distributed in $X = [0,1]^D$ the worst-case false negative rate is the volume of the set $P' \subset X$ of points satisfying

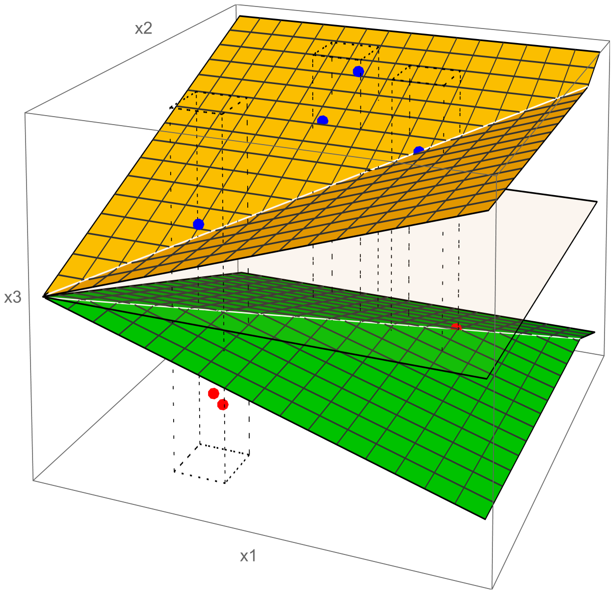
$$\begin{aligned}0 \leq \quad & x_k \quad \leq 1 \quad k = 1,\dots,D-1\\ \frac{1}{2} \leq \quad & x_D \quad \leq g_{\hat{\alpha}',\hat{\beta}'}(x_1,\dots,x_{D-1})\end{aligned}$$

that is $V(P') = \frac{1}{2} - V(Q')$ where $Q' \subset X$ is the convex polytope defined by

$$\begin{aligned}0 \leq \quad & x_k \quad \leq 1 \quad k = 1,\dots,D-1\\ \hat{\alpha}'_k x_k \leq \quad & x_D \quad \leq 1 \quad k = 1,\dots,D-1\end{aligned}$$

In an analogous way, the false positive rate is $\frac{1}{2} - V(Q'')$ where $Q'' \subset [0,1]^D$ is the convex polytope defined by

Figure B.1: An example of the simplified setting with $D = 3$, $v = {}^{1}\!/_{64}$ and $P = 3$ producers of data. Positive (blue) data points and negative (red) data points determine the separation manifold (in yellow) of the worst-false-negative classifier and the separation manifold (in green) of the worst-false-positive classifier. The ideal separation plane is also shown along with the subregions X_i (dashed parallelepipeds) within which each of the three producers generates its data points.



$$\begin{array}{rcll} 0 \leq & x_k & \leq 1 & k = 1, \ldots, D-1 \\ 0 \leq & x_D & \leq \hat{\beta}_k' x_k & k = 1, \ldots, D-1 \end{array}$$

Since Q' and Q'' are convex, we may rely on standard algorithms for the computation of their volume starting from their definition by means of inequalities Barber et al. (1996)

Producers and diversity

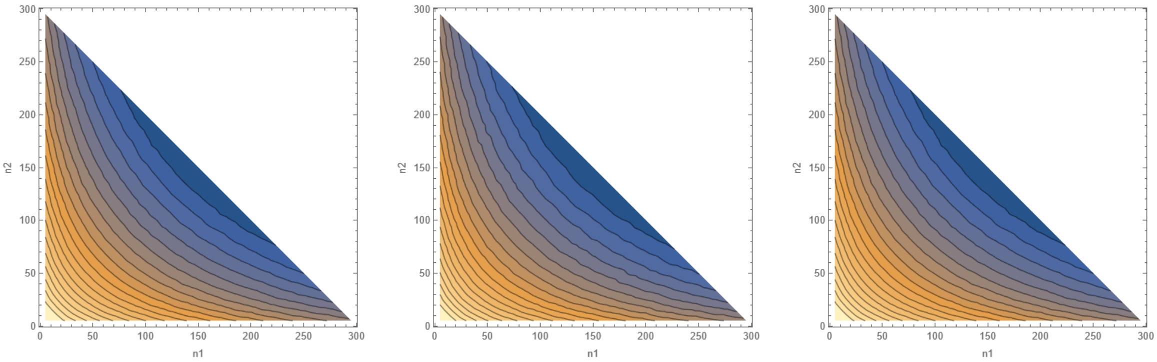
We assume that the subsets X_i in which the producers generate data are such that $V(X_i) = v$ for some v such that $\Delta = \sqrt[D-1]{v}$ has an integer inverse ${}^{1}\!/_{\Delta}$

$$X_i = \left[\bigtimes_{k=1}^{D-1} [(\xi_{i,k} - 1)\Delta, \xi_{i,k}\Delta] \right] \times [0,1]$$

for some choice of the $D-1$ integers $1 \leq \xi_{i,1}, \xi_{i,2}, \ldots, \xi_{i,D-1} \leq {}^{1}\!/_{\Delta}$.

Figure B.1 shows an example of the setting for $D = 3$ and $P = 3$ producers.

Figure B.2: Contour plots of the relationship between the contributions n_1 and n_2 of two producers and the (logarithm of the) worst average performance of the toy classifier for $D = 2$ (a), $D = 3$ (b), $D = 4$ (c). The convexity of the contours quantifies the scope effect.



a) b) c)

Empirical evidence and emerging properties

Consider $D = 2,3,4$, $n = 10,15,\dots,300$, $v = 1/64$, $P = 4$ producers, and different values for the data contributions n_1,n_2,n_3,n_4.

Different data contributions are obtained by dividing the dataset into $\ell = n/5$ lots of 5 data points each. These lots are then assigned to the P producers considering all possible distinguished partition of ℓ, i.e., all the set of integers $\ell_1 \geq \cdots \geq \ell_P \geq 0$ such that $\ell_1 + \cdots + \ell_P = \ell$, and then setting $n_i = 5\ell_i$ for $i = 1,\dots,P$.

For each of the resulting P-tuple, $n_1,\dots,n_P$, the training and performance evaluation of our worst-case classifier is repeated for 10^5 trials. In each trial, P random sets of indexes $0 \leq \xi_{i,1},\dots\xi_{i,D-1} \leq M$ identifying X_i are generated. In each X_i, n_i labelled samples are drawn at random. Based on all the generated samples, the worst-false-negative and worst-false-positive classifiers are computed along wit their error rate. The largest between the maximum false negative rate and the maximum false positive rate is used to quantify the absolute worst-case performance.

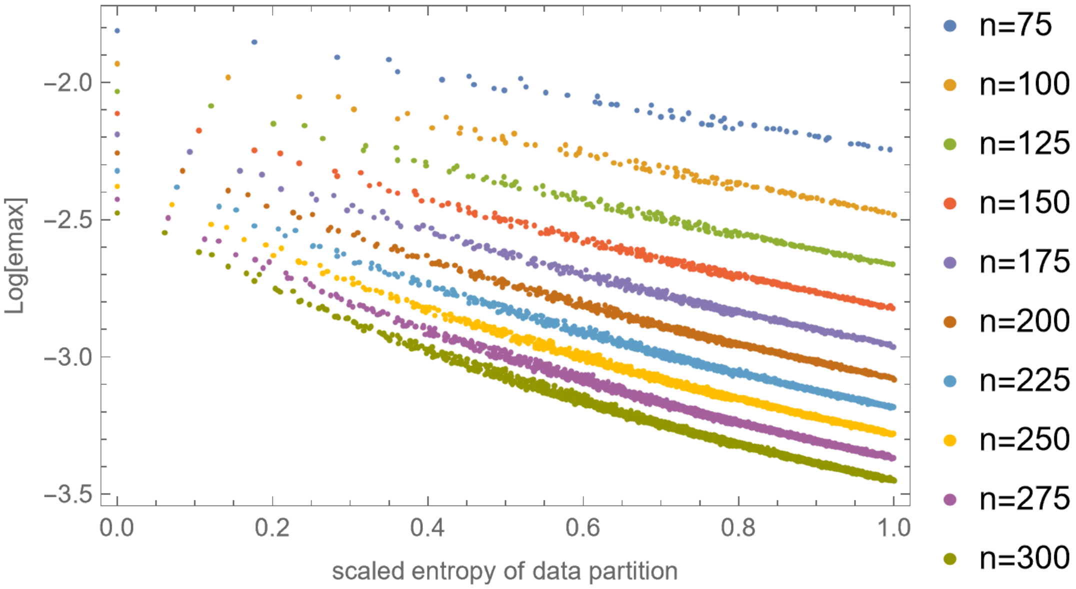
The logarithm of the empirical average over the 10^5 trials of such absolute worst case performance is used in the following plots. This is not an utility function. Yet, it can be safely assumed that error and utility are linked by a monotonic non-increasing function and thus that error reductions correspond to utility increases.

Figure B.2 is obtained selecting the P-tuples in which only n_1 and n_2 are positive. This allows to plot the logarithmic worst-case error against n_1,n_2 in the $P = 2$ case as a sub-case of the $P = 4$ case.

The scale effect manifests as the fact that any straight line passing through the origin (along which one sees contributions with a constant ratio n_1/n_2 with increasing size of the overall dataset $N = n_1 + n_2$) intersect iso-performance lines with progressively lower worst-case error.

Yet, the convexity of the same iso-performance lines reveals the effect of scope. In fact, moving along an iso-scale line $n_1 + n_2 = n =$ constant, the worst-case performance consistently improves as one approaches the even distribution of the data set between the two producers $n_1 = n_2 = n/2$.

Figure B.3: Logarithmic worst-case performance plotted against the scaled entropy of the distribution of n data points among $P = 4$ producers.



To assess whether this scope effect holds with $P > 2$ we should agree on how to measure the *evenness* of a partition of n among more than 2 producer. Among the many ways of measuring *evenness* we choose scaled Shannon entropy, i.e., in the case of P producers

$$E(n_1,\dots,n_P) = -\frac{1}{\log P}\sum_{i=1}^{P}\frac{n_i}{n}\log\frac{n_i}{n}$$

whatever the basis of the logarithm.

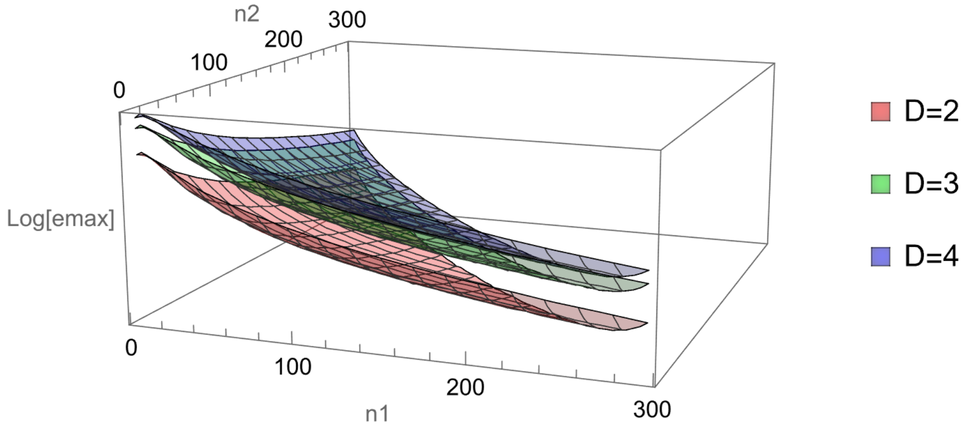
The scaled entropy is minimum for $E(n,0,\dots,0) = 0$ and is is maximum for $E(n/P,\dots,n/P) = 1$. This is clearly what we want, though the behaviour in intermediate configurations depends on that fact that Shannon devised his entropy to quantify the amount of information emitted by a source with P symbols each with probability n_i/n.

Despite this somehow unrelated origin, scaled entropy seems to interpret quite well the *evenness* on which the scope effect hinges. In fact, Figure B.3 shows that the logarithmic worst-case error correlates negatively with scale entropy (and, of course, with n due to the scale effect). Hence, data sets aggregating a substantially equal number of data from each producer yield more utility than equivalent-scale datasets in which most of the data are contributed by few producers.

Finally, Figure B.4 shows the effect of data dimensionality by plotting the logarithm of the worst-case error against the data contribution of $P = 2$ producers working with data of increasing dimensionality $D = 2,3,4$.

Note that, given a certain n_1 and n_2 (and thus fixing the effect of scope and scale), as D increases also the worst-case error increases showing that higher dimensional models are harder to train.

Figure B.4: The logarithmic worst-case error plotted against the contributions of $P = 2$ producers working with data in a D-dimensional space.



Hence, in our simple model, the scale property is confirmed while the positive effect of aggregating a data set from (possibly evenly contributing) sources exhibiting diversity emerges naturally, as well as the effect of using a fixed size data set to train models in high dimensional settings.