DISCUSSION PAPER SERIES

DP17842

MACHINE DATA: MARKET AND ANALYTICS

Giacomo Calzolari, Anatole Cheysson and Riccardo Rovatti

INDUSTRIAL ORGANIZATION

ISSN 0265-8003

MACHINE DATA: MARKET AND ANALYTICS

Giacomo Calzolari, Anatole Cheysson and Riccardo Rovatti

Discussion Paper DP17842
Published 23 January 2023
Submitted 23 January 2023

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

MACHINE DATA: MARKET AND ANALYTICS

Abstract

Recent technological developments in ICT and Artificial Intelligence allow extracting valuable information from data that machines generate with production, machine data (MD). Although possibly more valuable than personal data, the growing market for MD and its analytics may suffer from several issues, such as datasets fragmented into many small data-producers, externalities as with non-rival information and fuzzy property rights. We combine these market elements with critical properties that we derive from actual Machine Learning algorithms for analytics. We explore how and to what extent a data aggregator can operate, contracting with different data producers to share data and analytics. We identify conditions that impact the market organization for MD, such as producers' heterogeneity, their preference for anonymity, and the intensity of competition in final markets.

JEL Classification: N/A

Keywords: N/A

Giacomo Calzolari - giacomo.calzolari@eui.eu
Department Of Economics, European University Institute and CEPR

Anatole Cheysson - anatole.cheysson@eui.eu
Department Of Economics, European University Institute

Riccardo Rovatti - riccardo.rovatti@unibo.it
University Of Bologna

Acknowledgements
We are grateful for detailed comments to Jens Prüfer, Peerawat Samranchit, Xavier Lambin, Vincenzo Denicolò, Nicolas Petit, Philip Hanspach, Natalie Kessler, Alexandre de Cornière, Linus Hoffmann, Luca Tomesani, Giusella Finocchiaro and Enrico Al Mureden. We thank seminars participants at several institutions, the EUI Artificial Intelligence group, the Competition Law Working Group, 2022 MaCCI Conference, Paris 13th Conference on Digital Economics 2022, and EARIE conference 2022.

Machine Data: market and analytics *

Giacomo Calzolari †
European University Institute
CEPR

Anatole Cheysson
European University Institute

Riccardo Rovatti
University of Bologna

January 23, 2023

Abstract

Recent technological developments in ICT and Artificial Intelligence allow extracting valuable information from data that machines generate with production, machine data (MD). Although possibly more valuable than personal data, the growing market for MD and its analytics may suffer from several issues, such as datasets fragmented into many small data-producers, externalities as with non-rival information and fuzzy property rights. We combine these market elements with critical properties that we derive from actual Machine Learning algorithms for analytics. We explore how and to what extent a data aggregator can operate, contracting with different data producers to share data and analytics. We identify conditions that impact the market organization for MD, such as producers' heterogeneity, their preference for anonymity, and the intensity of competition in final markets.

Keywords: Data Analytics, Machine Learning, Artificial Intelligence, Machine Generated Data, Non-Personal Data, IoT, 5G, ICT, Enabling Technology, Market Organization, Externality, Anonymity, Property Rights, Competition.

JEL Classification: L12, L13, M11, M15, D21, D43.

*We are grateful for detailed comments to Jens Prüfer, Peerawat Samranchit, Xavier Lambin, Vincenzo Denicolò, Nicolas Petit, Philip Hanspach, Natalie Kessler, Alexandre de Cornière, Linus Hoffmann, Luca Tomesani, Giusella Finocchiaro and Enrico Al Mureden. We thank seminars participants at several institutions, the EUI Artificial Intelligence group, the Competition Law Working Group, 2022 MaCCI Conference, Paris 13th Conference on Digital Economics 2022, and EARIE conference 2022.

†Corresponding author: Giacomo Calzolari, Department of Economics, European University Institute. E-mail: giacomo.calzolari@eui.eu.

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 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 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

[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.

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 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

[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.

[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.

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 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]

[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.

[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).

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 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

[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).

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 lays out the baseline model. In Section 3 we discuss some benchmarks. The market-based analytics with the organization of an aggregator is in Section 4. In Section 5 we show the difficulties in dealing with anonymity. Section 6 discusses cooperative analytic with producers that compete for consumers and Section 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 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,

$$B_i = \alpha_i \eta(n) - \gamma_i n_i - Q_i, \tag{1}$$

where $\eta(n)$ is the *value of the analytics* that relies on data $n = (n_1, \ldots, n_P)$ 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 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,

$$B_{\text{agg}} = \sum_{j=1}^{P} Q_j - \bar{\delta} - \delta \sum_{j=1}^{P} n_i - \epsilon \eta(n) \tag{2}$$

where $\bar{\delta} + \delta \sum_{j=1}^{P} n_i$ is the fixed and variable cost to produce the analytics and $\epsilon \eta(n)$ is a cost related to managing the analytics of a given value $\eta(n)$. 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.
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,

$$\eta\left(n_1, \ldots, n_P\right) = \bigoplus_{j=1}^{P} \upsilon\left(n_i\right) = \bigoplus_{j=1}^{P} \eta\left(n_i\right) \tag{3}$$

[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.

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 $n = (0, \ldots, 0)$. 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 $\epsilon\eta(n)$. 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 of the analytics (e.g. if they do not provide data), and each joining producer benefits from the analytics without degrading its 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 $\epsilon < 0$, for example, when the aggregator manufactures and sells machines that produce the data.

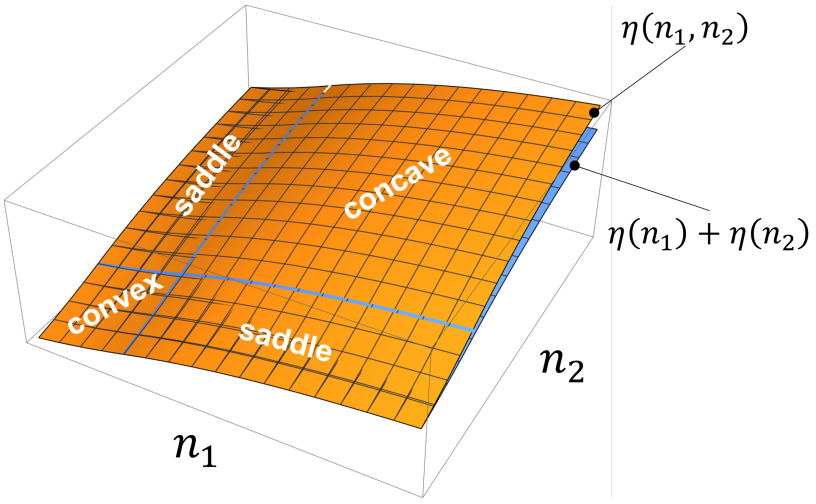


Figure 1: Value of analytics $\eta(n_1, n_2)$ in the Example (details in Appendix A2.1) showing Scale and Scope.

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 1 illustrates the associated value of the analytics $\eta(n_1, n_2)$.[13]

[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.

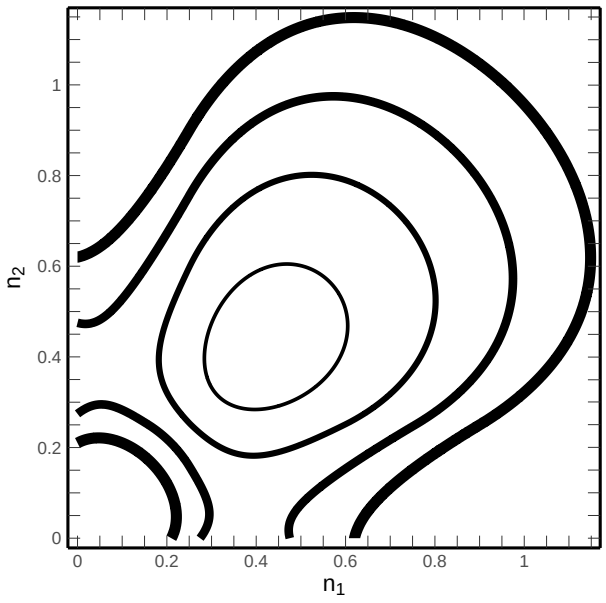


Figure 2: Feasible analytics (i.e. combinations of n_1, n_2 s.t. $W(n) \geq B_{\text{agg}}^{\min} + \sum_{j=1}^{P} B_j^{\min}$) with thicker lines associated with lower (symmetric) marginal costs.

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:

$$W(n) = B_{\text{agg}} + \sum_{j=1}^{P} B_j = \eta(n) \left(\sum_{j=1}^{P} \alpha_j - \epsilon\right) - \bar{\delta} - \sum_{j=1}^{P} (\delta + \gamma_j) n_j \tag{4}$$

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 shows the combinations of n such that the surplus is 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)

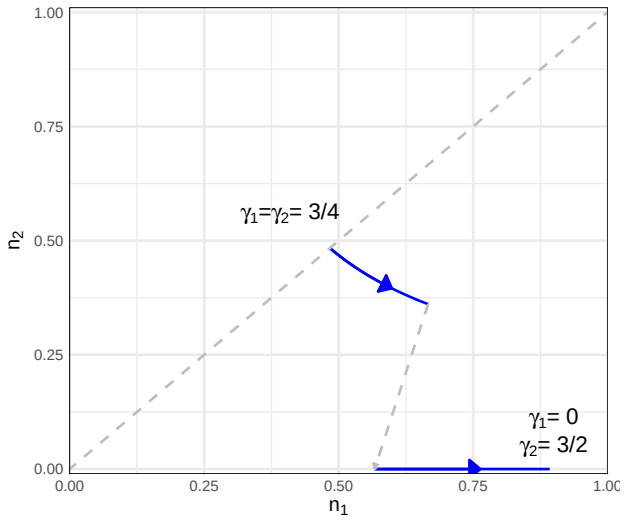


Figure 3: Efficient data when costs γ_1 and γ_2 respectively decrease and increase by a same amount.

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]

$$\frac{\partial \eta(n)}{\partial n_i}\left(\sum_{i=1}^{P}\alpha_i - \epsilon\right) = \delta + \gamma_i, \quad i = 1, \ldots, P. \tag{5}$$

The cost of procuring and handling the marginal unit of data of producer i (the r.h.s. in condition (5)) 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 (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 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*

[14] Given the properties of $\eta(n)$, condition (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.

independently of their value of the analytics.

Figure 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 (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).

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,

$$\max_{n_i} \quad B_i \tag{6a}$$

$$\text{s.t.} \quad B_i - B_i^{\min} \geq 0 \tag{6b}$$

(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,

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} \alpha_i - \gamma_i = 0. \tag{7}$$

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, (7) is satisfied at $n_i = n_i^0$ and $\hat{n}_{-i} = n_{-i}^0$.

Comparing the optimality conditions (7) with those that guarantee the efficient analytics (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

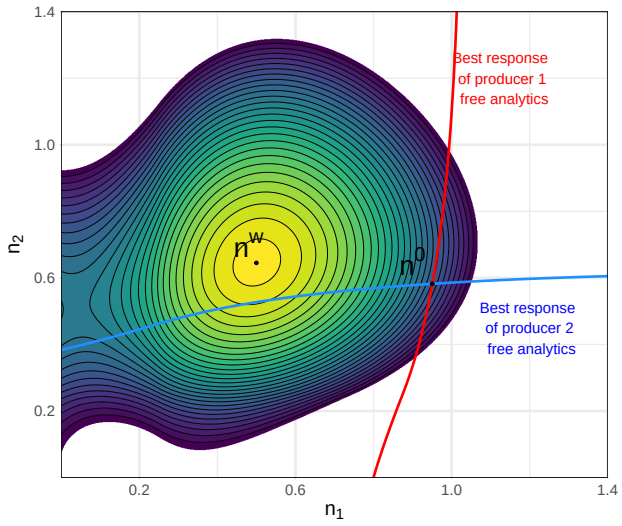


Figure 4: Combinations of data such that $W^{\max} \geq B^{\min}_{\text{agg}} + B^{\min}_1 + B^{\min}_2$ (and heat-map) with the two best response curves for free analytics (asymmetric case).

analytics $\frac{\partial \eta(n)}{\partial n_i}\epsilon + \delta$. 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

$$\Delta_i^0 = \frac{\partial \eta(n^0)}{\partial n_i} \sum_{j \neq i}^{P} \alpha_j - \left(\delta + \frac{\partial \eta(n^0)}{\partial n_i} \epsilon \right) \tag{8}$$

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 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]

[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.

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.*

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 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,

$$\max_{(Q_1,n_1),\ldots,(Q_P,n_P)} \quad B_{\text{agg}} = \sum_{j=1}^{P} Q_j - \bar{\delta} - \delta \sum_{j=1}^{P} n_j - \epsilon\eta\,(n) \tag{9a}$$

$$\text{s.t.} \qquad B_{\text{agg}} - B_{\text{agg}}^{\min} \geq 0, \tag{9b}$$

$$B_i - B_i^{\min} \geq 0 \qquad i = 1, \ldots, P \tag{9c}$$

Since B_{agg} increases and B_i decreases in Q_i, at the optimum the transfer Q_i is set so that constraint (9c) binds, i.e.,

$$Q_i = \alpha_i \eta(n) - \gamma_i n_i - B_i^{\min}. \tag{10}$$

Substituting this payment into B_{agg}, the aggregator's profit rewrites as the right-hand side of (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 $B_{\text{agg}} \geq B_{\text{agg}}^{\min}$ 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, $B_i \geq B_i^{\min}$.[16]

[16]Given the optimal analytics $n^{\text{a}} = n^{\text{w}}$, since $W^{\max} \geq B_{\text{agg}}^{\min} + \sum_{j=1}^{P} B_j^{\min}$, there exist payments $(Q_i, \ldots, Q_P)$ that guarantee all producers'

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

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 (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:

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i}\alpha_i - \gamma_i - q_i = 0 \tag{11}$$

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 (11) becomes equivalent to the condition for efficient data (5) whenever,

$$q_i = \frac{\alpha_i(\delta + \gamma_i)}{\sum_{j=1}^{P} \alpha_j - \epsilon} - \gamma_i. \tag{12}$$

The fixed fee $\bar{q}_i$ can then be set so that $B_i = \alpha_i \eta(n^*) - (\gamma_i + q_i) n_i^* - \bar{q}_i = B_i^{\min}$. 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 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 5 shows how with variable parts as in (12), the aggregator modifies the best responses of producers and

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 $B_i^{\min}$.

[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 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 (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.

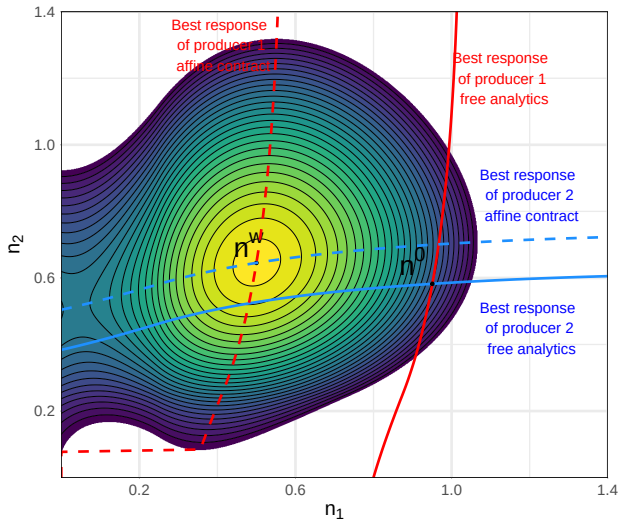


Figure 5: A for-profit aggregator modifies producers' best responses to maximize their own profit and replicates the efficient outcome.

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 (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 other producers. This case is reported in Figure 6, 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 (5), producer 2 is remunerated for its data. Instead, producer 1 provides data but pays for the analytics.

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

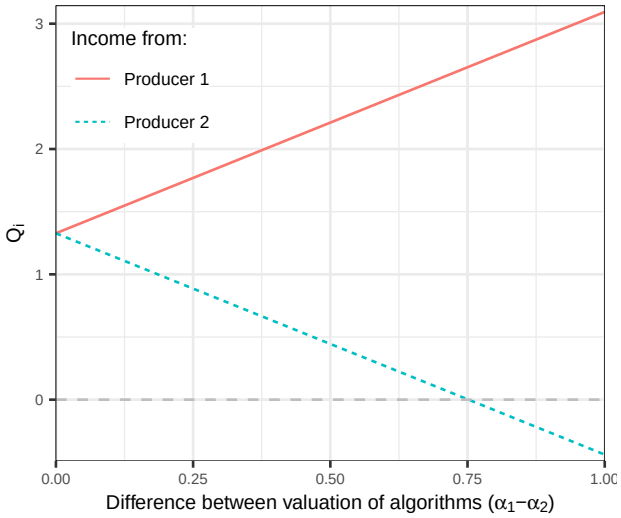


Figure 6: Transfer as a function of the difference in the abilities to gain from the analytics (with identical costs).

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

$$B_i^{\text{self}} = \max_{n_i} \alpha_i \eta \left(n_i, 0, ..., 0\right) - \bar{\delta} - \delta n_i \tag{13}$$

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 $\epsilon = 0$). 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}.

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 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]

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

$$\alpha_i \eta \left(n_i^*, n_{-i}^*\right) - \gamma_i n_i^* - Q_i \geq \alpha_i \eta \left(n_j^*, n_{-i}^*\right) - \gamma_i n_j^* - Q_j. \tag{14}$$

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.

[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.

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 (14) for $i = 1$, $j = 2$ as well as for $i = 2$ and $j = 1$ to yield the conditions

$$\alpha_1\eta'(n_1^*, n_2^*) - \gamma_1 n_1^* - Q_1 \geq \alpha_1\eta'(n_2^*, n_2^*) - \gamma_1 n_2^* - Q_2 \tag{15}$$

$$\alpha_2\eta'(n_1^*, n_2^*) - \gamma_2 n_2^* - Q_2 \geq \alpha_2\eta'(n_1^*, n_1^*) - \gamma_2 n_1^* - Q_1 \tag{16}$$

that can be satisfied if and only if,

$$\alpha_1\left[\eta'(n_2^*, n_2^*) - \eta'(n_1^*, n_2^*)\right] - \gamma_1\left(n_2^* - n_1^*\right) \leq Q_2 - Q_1 \leq \alpha_2\left[\eta'(n_1^*, n_2^*) - \eta'(n_1^*, n_1^*)\right] - \gamma_2\left(n_2^* - n_1^*\right).$$

A necessary condition for this is,

$$C(\gamma_1, \gamma_2) := \alpha_2\left[\eta'(n_1^*, n_2^*) - \eta'(n_1^*, n_1^*)\right] - \alpha_1\left[\eta'(n_2^*, n_2^*) - \eta'(n_1^*, n_2^*)\right] - (\gamma_2 - \gamma_1)\left(n_2^* - n_1^*\right) \geq 0. \tag{17}$$

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,

$$C(\bar{\gamma} + d\gamma, \bar{\gamma}) \simeq -\Psi(\alpha_1 - \alpha_2)d\gamma \tag{18}$$

where $\Psi > 0$. Equation (18) 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 (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.

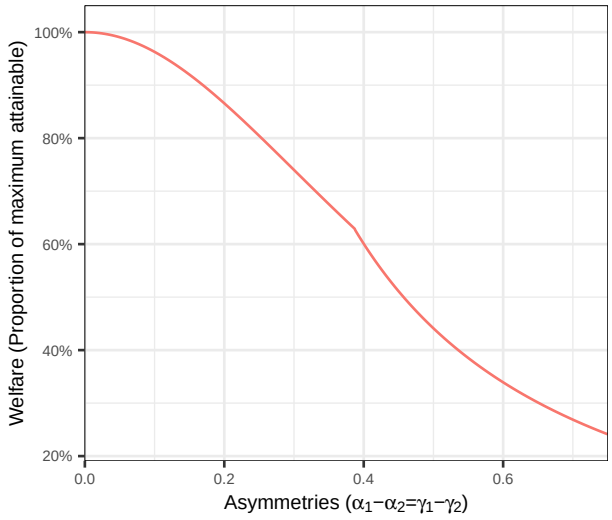


Figure 7: The cost of anonymity: welfare reduction with different producers (parameterized with $\Delta = \alpha_1 - \alpha_2 = \gamma_1 - \gamma_2 \geq 0$), with anonymity preserved with a unique contract for all producers.

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 (17) 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 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.

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. $\epsilon = 0$, 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

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

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,

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} = \frac{\gamma_i + q_i}{\alpha_i}, \tag{19}$$

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 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,

$$\max_{\{\bar{q}_i, q_i\}_i} \quad B_{\text{agg}} = \sum_{i=1} [(q_i - \delta) n_i(q_i, \hat{n}_{-i}) + \bar{q}_i] - \bar{\delta} \tag{20a}$$

$$\text{s.t.} \quad B_{\text{agg}} \geq 0, \tag{20b}$$

$$B_i(n_i(q_i, \hat{n}_{-i}), \hat{n}_{-i}) \geq B_i^{\min} \quad \text{for any } i. \tag{20c}$$

As in section 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 the (optimally) binding participation constraints (20c), the program becomes,

$$\max_{\{n_i\}_i} \quad \sum_{i=1} [\alpha_i \eta(n_i, \hat{n}_{-i}) - (\gamma_i + \delta) n_i] - \bar{\delta} \tag{21a}$$

$$\text{s.t.} \quad B_{\text{agg}} \geq 0 \tag{21b}$$

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]

$$\frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} \alpha_i = \gamma_i + \delta. \tag{22}$$

The difference between (22) and condition (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 (22) does not

[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.

[25] One can then recover the price per-unit of data for each producer by combining equations (19) and (22).

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 (22) are, in fact, very similar to the case of free analytics (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 (22) 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 $\epsilon > 0$ to manage the analytics.

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

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

This shows that when $\epsilon > 0$, 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 $\epsilon \eta(n)$. 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.

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

[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).

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:

$$c_i(n_i, n_{-i}) = \bar{c} - \eta(n_i, n_{-i}), \tag{23}$$

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]

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

$$max_{(x_1,...,x_P)} U(x_1, ..., x_P) - \sum_i p_i x_i. \tag{24}$$

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

$$U(x_1, x_2) = \theta(x_1 + x_2) - \frac{1-\rho}{2}(x_1^2 + x_2^2) - \rho x_1 x_2 \tag{25}$$

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

[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.

[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 25 guarantees that when increasing ρ the market size is kept constant. This implies that, in general, the consumer' surplus increases in ρ.

positive quantity of both goods), the demand function for each producer $i = 1, 2$ is,

$$x_i(p_i, p_{-i}) = \frac{\theta(1-2\rho) + \rho p_{-i} - (1-\rho)p_i}{1-2\rho}. \tag{26}$$

Free analytics.

For given data $(n_1, ..., n_P)$ and analytics $\eta(n_1, ..., n_P)$, when active in the final consumer markets competing producers independently set prices according to the following optimality conditions,[33]

$$x_i(p_i, p_{-i}) + (p_i - \bar{c} + \eta(n_i, n_{-i}))\frac{\partial x_i(p_i, p_{-i})}{\partial p_i} = 0, \ i = 1, ..., P \tag{27}$$

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]

$$\begin{aligned} x_i(n_i, n_{-i})\frac{\partial \eta}{\partial n_i} - \gamma_i + \\ + [p_i(\eta(n_i, n_{-i})) - \bar{c} + \eta(n_i, n_{-i})] \times \frac{\partial \eta}{\partial n_i} \sum_{j \neq i} \frac{\partial p_j(\eta(n_i, n_{-i}))}{\partial \eta} \frac{\partial x_i(\eta(n_i, n_{-i}))}{\partial p_j} = 0 \end{aligned} \tag{28}$$

The first line is equivalent to the optimality condition for the free analytics and non-competing producers, i.e. (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 (28) 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

[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 (27). We discuss the case of non-interior solutions later on.

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 (28) 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 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 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 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 (9)-(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:

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

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,

$$\begin{aligned}\left[\frac{\partial\eta(n)}{\partial n_i}\left(\sum_{j=1}^{P} x_j - \epsilon\right) - (\delta + \gamma_i)\right] + \frac{\partial\eta(n)}{\partial n_i}\sum_{j=1}\sum_{k\neq j}[p_k - \bar{c} + \eta(n)]\frac{\partial x_k}{\partial p_j}\frac{\partial p_j}{\partial \eta} = \\ = \sum_{j\neq i}\sum_{k\neq j}\left(p_j^r - \bar{c}\right)\frac{\partial x_j^r}{\partial p_k^r}\frac{\partial p_k}{\partial \eta}\frac{\partial \eta_{-j}}{\partial n_i}\end{aligned} \tag{29}$$

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 (5) with non-competing producers and accounts for the internalization of the analytics' positive externality across all producers. All other terms in (29) 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 (28). 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. Interestingly, 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)

[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.

and that our analysis substantiates.[36] The last term, on the right-hand side of (29), 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 (27). Comparing with this social optimum, the solid lines of panel (a) in Figure 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 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;

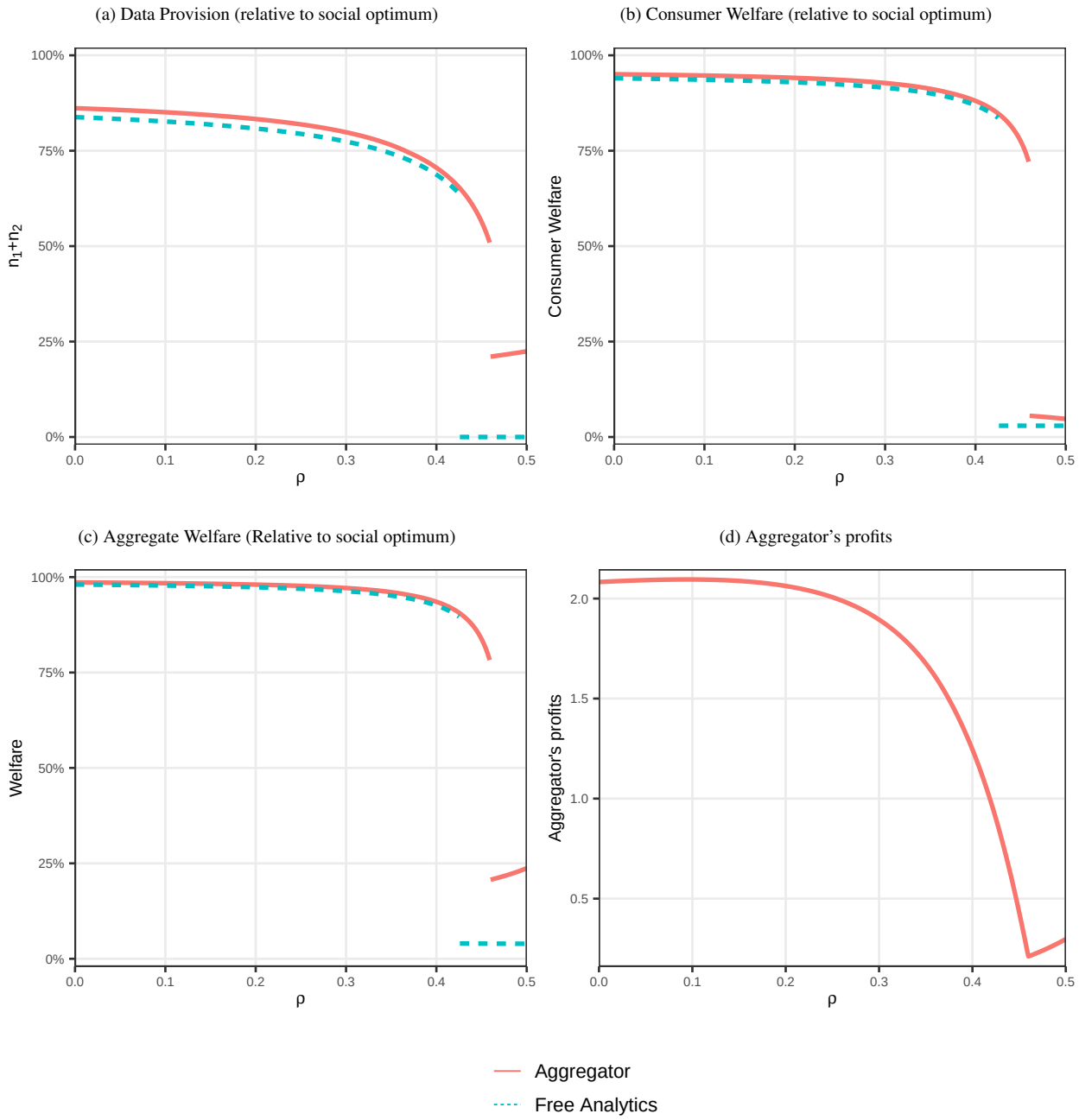
(iii) the aggregator optimally combines data of producers and products associated with an intensity of competition that (generically) differs from that yielding maximal welfare or consumers' surplus.

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. 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 8 shows that, although the aggregator relies on more data than with free analytics, the difference can be quite small.

[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.

Figure 8: Competition between data producers: Market-based Aggregator (solid line) and Free-analytics (dashed line) on the the intensity of competition ρ (horizontal axis).



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 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.

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 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.

[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.

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. 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.

[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.

References

Alam, F., Mehmood, R., Katib, I., Albogami, N. N., and Albeshri, A. (2017). Data Fusion and IoT for Smart Ubiquitous Environments: A Survey. *IEEE Access*, 5:9533–9554.

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

Amir, R., Liu, H., Machowska, D., and Resende, J. (2019). Spillovers, subsidies, and second-best socially optimal R&D. *Journal of Public Economic Theory*, 21(6):1200–1220.

Anderson, G. M., Shughart, W. F., and Tollison, R. D. (2004). The Economic Theory of Clubs. In *The Encyclopedia of Public Choice*, pages 499–504. Springer US.

Atik, C. and Martens, B. (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*, page 40.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483.

Bergemann, D., Alessandro, and Gan, T. (2019). The economics of social data. *Discussion Paper No. 2203R, Cowles Foundation, New Haven, CT*.

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

Calvano, E., Calzolari, G., Denicolo', V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267—-3297.

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

Dosis, A. and Sand-Zantman, W. (2019). The Ownership of Data. *SSRN Electronic Journal*.

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

Duch-Brown, N., Martens, B., and Mueller-Langer, F. (2017). The Economics of Ownership, Access and Trade in Digital Data. *SSRN Electronic Journal*.

European Commission (2017). Building a European Data Economy.

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.

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.

Farboodi, M., Mihet, R., Philippon, T., and Veldkamp, L. (2019). Big Data and Firm Dynamics. *AEA Papers and Proceedings*, 109:38–42.

Gambardella, A., Heaton, S., Novelli, E., and Teece, D. J. (2021). Profiting from Enabling Technologies? *Strategy Science*, 6(1):75–90.

Graef, I., Gellert, R., Purtova, N., and Husovec, M. (2018). Feedback to the Commission's Proposal on a Framework for the Free Flow of Non-Personal Data. *SSRN Electronic Journal*.

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

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. *arXiv:1712.00409 [cs, stat]*. arXiv: 1712.00409.

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

Johnson, J., Rhodes, A., and Wildenbeest, M. R. (2023). Platform design when sellers use pricing algorithms. *forthcoming Econometrica*, pages 1–55.

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

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

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

Meng, T., Jing, X., Yan, Z., and Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion*, 57:115–129.

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

Mumuni, A. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, page 100258.

Prono, L., Bich, P., Mangia, M., Pareschi, F., Rovatti, R., and Setti, G. (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, L., Mangia, M., Pareschi, F., Rovatti, R., and Setti, G. (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, J. and Schottmüller, C. (2020). Competing with Big Data. *Journal of Industrial Economics*.

Rey, P. and Verge, T. (2004). Bilateral Control with Vertical Contracts. *The RAND Journal of Economics*, 35(4):728.

Schaefer, M. and Sapi, G. (2022). Complementarities in Learning from Data: Insights from General Search. page 43.

Zech, H. (2016). Data as a Tradeable Commodity. In De Franceschi, A., editor, *European Contract Law and the Digital Single Market*, pages 51–80. Intersentia, 1 edition.

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

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,

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

in which case condition (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, $\frac{\partial \eta(0,n_{-i})}{\partial n_i} \left(\sum_{i=1}^{P} \alpha_i - \epsilon \right) < \delta + \gamma_i$, in which case condition (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

$$B^*_{agg} = \sum_{j=1}^{P} \left[\alpha_j \eta(n^*) - (\gamma_i + \delta) n^*_i - B^{self}_i \right] - \epsilon \eta(n^*), \tag{30}$$

be the maximum payoff the aggregator can obtain. If one ore more B^{self}_i 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^{self}_i.[41] Result (i) then follows.

To see why the aggregator may prefer to exclude the producer with the highest B^{self}_i 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 $W(n^*) - (B^{\min}_{\text{agg}} + \sum_{j=1}^{P} B^{\min}_j) < 0$ (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 $B^{\text{self}}_1 > B^{\text{self}}_2 = B^{\text{self}}_3$, 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 (5) implies that these data depend (also) on the costs, $n^{\text{w}}_i = n^{\text{w}}_i(\gamma_i)$ for $i = 1, \ldots, P$, so that $C(.)$ in (17) is a function of γ_1 and γ_2. For producers with identical costs $\gamma_1 = \gamma_2 = \bar{\gamma}$, we would have $n^{\text{w}}_1(\gamma_1) = n^{\text{w}}_2(\gamma_2) = \bar{n}$. Clearly $C(\bar{\gamma}, \bar{\gamma}) = 0$ so that the sign of $C(.)$ in that neighborhood depends on its derivative

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

with respect to the parameter that changes. In the following we will assume that γ_1 varies and thus we have to compute

$$\begin{aligned}
\frac{\partial C(\bar{\gamma},\bar{\gamma})}{\partial \gamma_1} &= \alpha_2 \left[\frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} + \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \\
&\quad -\alpha_1 \left[\frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} + \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_1} \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial \eta'(\bar{n})}{\partial n_2} \frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \\
&= \alpha_2 \frac{\partial \eta'(\bar{n})}{\partial n_2} \left[\frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} \right] - \alpha_1 \frac{\partial \eta'(\bar{n})}{\partial n_1} \left[\frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \\
&= (\alpha_2 - \alpha_1) \frac{\partial \eta'(\bar{n})}{\partial n_1} \left[\frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} \right] \qquad (31)
\end{aligned}$$

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 (5) and derive them with respect to γ_1 to obtain,

$$\sum_{j=1}^{P} \frac{\partial^2 \eta(\bar{n})}{\partial n_1 \partial n_j} \frac{\partial n_j^{\mathrm{w}}(\bar{n})}{\partial \gamma_1} = 1 \qquad (32)$$

$$\sum_{j=1}^{P} \frac{\partial^2 \eta(\bar{n})}{\partial n_i \partial n_j} \frac{\partial n_j^{\mathrm{w}}(\bar{n})}{\partial \gamma_1} = 0, \quad i = 2, \ldots, P \qquad (33)$$

The variations of n_i^{w} with respect to γ_1 can be derived by considering (32) and (33) as a linear system and solve it for $\frac{\partial n_j^{\mathrm{w}}(\bar{n})}{\partial \gamma_1}$ for $j = 1, \ldots, P$. Due to the symmetry of η we have $\frac{\partial^2 \eta(\bar{n})}{\partial n_i^2} = a \leq 0$ (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 (32) and (33) 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

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

whose inverse is

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

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

$$\frac{\partial n_1^{\mathrm{w}}(\bar{n})}{\partial \gamma_1} = \frac{(P-2)b+a}{(a-b)[(P-1)b+a]} \tag{34}$$

$$\frac{\partial n_i^{\mathrm{w}}(\bar{n})}{\partial \gamma_1} = \frac{b}{(b-a)[(P-1)b+a]} \quad i = 2, \dots, P \tag{35}$$

and thus

$$\frac{\partial n_2^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} - \frac{\partial n_1^{\mathrm{w}}(\bar{\gamma})}{\partial \gamma_1} = \frac{1}{b-a} \tag{36}$$

Finally, this can be plugged into (31) to obtain (37) with (38), as in the main text:

$$\frac{\partial C(\bar{\gamma}, \bar{\gamma})}{\partial \gamma_1} = K\left(\alpha_2 - \alpha_1\right) \tag{37}$$

with

$$K = \frac{\frac{\partial \eta'(\bar{n})}{\partial n_1}}{\frac{\partial^2 \eta'(\bar{n})}{\partial n_1 \partial n_2} - \frac{\partial^2 \eta'(\bar{n})}{\partial n_1^2}} \geq 0. \tag{38}$$

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 (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

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

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,

$$\alpha_i \frac{\partial \eta(n_i, \hat{n}_{-i})}{\partial n_i} - \epsilon \frac{\partial \eta(n_i, n_{-i})}{\partial n_i} = \gamma_i + \delta. \tag{39}$$

The second order conditions can be rewritten as

$$(\alpha_i - \epsilon)\eta_{ii} \leq 0 \tag{40}$$

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

$$\frac{(\alpha_1 - \epsilon)(\alpha_2 - \epsilon)}{\epsilon^2} > \bar{\sigma} \tag{41}$$

where

$$\bar{\sigma} = \frac{\eta_{12}^2}{\eta_{11}\eta_{22}} \tag{42}$$

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 (41) fails. In fact, it must be $\epsilon < \alpha_i$ 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,

$$\begin{aligned} &\max_{n_1, n_2} 2\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta - \bar{c} + \eta(n_1, n_2))^2 - (\gamma + \delta)(n_1 + n_2) - \epsilon\eta(n_1, n_2) \\ &+ \sum_{i=1}^{2} \left[\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)} \eta(n_i, 0)(\theta - \bar{c}) - \left(\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)} \right)^2 (\eta(n_i, 0))^2 \right], \end{aligned} \tag{43}$$

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,

$$\begin{aligned} &\left[4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta - \gamma + \eta(n)) - \epsilon \right] \frac{\partial \eta(n^*_{\text{agg}})}{\partial n_i} \\ &-2\left(\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)} \right)^2 \frac{\partial \eta(n^*_{\text{i,agg}}, 0)}{\partial n_i} \eta(n^*_{\text{i,agg}}, 0) \\ &\qquad + \frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)} \frac{\partial \eta(n^*_{\text{i,agg}}, 0)}{\partial n_i} (\theta - \gamma) = \gamma + \delta, \quad \forall i, \end{aligned} \tag{44}$$

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,

$$\left[4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta-\gamma+\eta(n))-\epsilon\right]\frac{\partial\eta(n^{**})}{\partial n_i} + \frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\frac{\partial\eta(n_i^{**},0)}{\partial n_i}(\theta-\gamma)=\gamma+\delta, \tag{45}$$

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

$$\left[4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}(\theta-\gamma+\eta(n))-\epsilon\right]\frac{\partial\eta(n^{***})}{\partial n_i} + \frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}\frac{\partial\eta(n^{***})}{\partial n_i}(\theta-\gamma+\eta(n^{***}))=\gamma+\delta, \tag{46}$$

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,

$$\max_{n_1,n_2}\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}(\theta-\bar{c}+\eta(n))^2-(\gamma+\delta)(n_1+n_2)-\epsilon\eta(n) \tag{47}$$

are determined by the following condition,

$$\left[\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}(\theta-\gamma+\eta(n^*))-\epsilon\right]\frac{\partial\eta(n^*)}{\partial n_i}=\gamma+\delta. \tag{48}$$

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

$$\frac{(1-\rho)(3-5\rho)}{(2-3\rho)^2}\geq 4\frac{(1-2\rho)(1-\rho)}{(2-3\rho)}+\frac{\rho(1-\rho)}{(2-3\rho)(2-\rho)}, \tag{49}$$

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:

$$\max_{n_1}\frac{1}{4(1-\rho)}(\theta-c+\eta(n_1,0))^2-(\gamma+\delta)n_1-\epsilon\eta(n_1,0) \tag{50}$$

Which always yields a lower amount of data than the resolution of problem (47).

(ii) Let $n^*(\rho)$ be the solution of problem (29). Maximized profits of the aggregator when contracting with both producers write:

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

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

$$\begin{aligned}\frac{\partial \pi^{agg}(n^*(\rho),\rho)}{\partial \rho} =& \sum_{i=1}^{2}(\theta-c)\eta(n_i,0)\frac{4(2-9\rho+9\rho^2-3\rho^3)}{(2-\rho)^2(2-3\rho)^3} \\ &+\sum_{i=1}^{2}\eta(n_i,0)^2\frac{\rho(-24+132\rho-270\rho^2+259\rho^3-124\rho^4+25\rho^5)}{(1-2\rho)^2(2-3\rho)^3(2-\rho)^3} \\ &-\frac{2\rho}{(2-3\rho)^3}\left(2(\theta-c)\eta_{12}+2\sum\eta(n_i,0)\eta_{12}+2\eta(n_1,0)\eta(n_2,0)+\eta_{12}^2\right)\end{aligned} \tag{51}$$

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:

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

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:

$$-2\frac{\partial \eta(n_1,n_2)}{\partial n_i}\frac{\frac{\partial \Pi_i}{\partial \rho}}{\frac{\partial \Pi_i}{\partial n_i}+\frac{\partial \Pi_i}{\partial n_{-i}}} = -\frac{(1-2\rho)}{(1-\rho)(3-5\rho)}(\theta-c+\eta(n_1,n_2)) \tag{52}$$

where:

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

Substituting for these expressions, the solution to equation (52) 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,

$$\eta\left(n_1, \ldots, n_P\right) = \bigoplus_{j=1}^{P} \upsilon\left(n_i\right) = \bigoplus_{j=1}^{P} \eta\left(n_i\right) \tag{53}$$

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

$$\eta\left(n_1, \ldots, n_P\right) \geq \eta\left(n_1, \ldots, n_{p-1}\right) + \eta\left(n_p, n_{p+1}, \ldots, n_P\right) \tag{54}$$

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 $a = \frac{v_0'}{v^{\max}}$, $b = \frac{v_{\bar{n}} - a\bar{n}}{\bar{n}^2}$, $c = 1 - v_{\bar{n}}$, $d = \frac{a\bar{n} - 2v_{\bar{n}}}{\bar{n}(v_{\bar{n}} - 1)}$.

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

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

in which $\sigma = 0$ corresponds to no synergy while increasing $\sigma > 0$ generates the super-additive effect in (54). In fact, such an Υ implies

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

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

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

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

In the Example, we use the following parameters: $v_0' = 3/4$, $\bar{n} = 1/4$, $v^{\max} = 1$, $v_{\bar{n}} = 1/4$ and $\sigma = 2$. For the aggregator, we set $\bar{\delta} = 0$, $\delta = 1/2$, $\epsilon = 1/3$, 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:

$$\max_{n_i} \alpha_i \eta(n) - \gamma_i n_i \tag{55}$$

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*

$$B_i(0, n_{-i}) \geq B_i(Br_i(n_{-i}), n_{-i}) \tag{56}$$

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

$$B_i(0, n_{-i}^{alt}) > B_i(n_i, n_{-i}^{alt}) \tag{57}$$

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

$$\alpha_i \eta(0, n_{-i}) \geq \alpha_i \eta(Br_i(n_{-i}), n_{-i}) - \gamma_i n_i^* > \alpha_i \eta(Br_i(n_{-i}^{alt}), n_{-i}) - \gamma_i Br_i(n_{-i}^{alt}) \implies \tag{58}$$

$$\alpha_i \eta(0, n_{-i}^{alt}) > \alpha_i \eta(Br_i(n_{-i}^{alt}), n_{-i}) - \gamma_i Br_i(n_{-i}^{alt}). \tag{59}$$

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 $\mathcal{I}$ is the set of producers not providing any data. Set $\mathcal{J}$ 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 $\mathcal{J}$ move to set $\mathcal{I}$. Data provision in set $\mathcal{J}$ 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 $\mathcal{J}$ move to set $\mathcal{I}$. Importantly, by continuity of the best response function when not providing any data (Lemma 4), the movement of producers from set $\mathcal{J}$ to set $\mathcal{I}$ does not change the best response of producers in set $\mathcal{I}$.

This method can be iterated until either (ii) all producers are in set $\mathcal{I}$, (ii) no producer in set $\mathcal{J}$ 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 $\mathcal{I}$ are playing their best responses. In the former case it ensures a Nash equilibrium is reached since all players are in set $\mathcal{I}$. In the latter case, as no producer in set $\mathcal{J}$ 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:

$$\frac{\partial^2 B_{\text{agg}}(n_1, n_2)}{\partial n_1^2} = \alpha_1 \frac{\partial^2 \eta(n_1, n_2^e)}{\partial n_1^2} - \epsilon \frac{\partial^2 \eta(n_1, n_2)}{\partial n_1^2} \tag{60}$$

It is negative in solution if:

$$(\alpha_1 - \epsilon) \frac{\partial^2 \eta(n_1, n_2)}{\partial n_1^2} \leq 0 \tag{61}$$

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

$$\frac{\partial^2 B_{\text{agg}}(n_1, n_2)}{\partial n_1 \partial n_2} = -\epsilon \frac{\partial^2 \eta(n_1, n_2)}{\partial n_1 \partial n_2} \tag{62}$$

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

$$B_{\text{agg}_{11}} B_{\text{agg}_{22}} > B_{\text{agg}_{12}}^{\ 2} \tag{63}$$

Which corresponds to:

$$\frac{(\alpha_1 - \epsilon)(\alpha_2 - \epsilon)}{\epsilon^2} > \frac{\eta_{12}^2}{\eta_{11}\eta_{22}} \tag{64}$$

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 $x_1, \ldots, x_D$ 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

$$\hat{\tau}_i = \max_x \left\{ f(x) \middle| x \in \hat{X}_i \wedge y > 0 \right\} \leq \tau \tag{65}$$

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

$$\hat{\tau} = \max_x \left\{ f(x) \middle| x \in \bigcup_{j=1}^{P} \hat{X}_j \wedge y > 0 \right\} = \max_{j=1,\dots,P} \{\hat{\tau}_j\} \leq \tau \tag{66}$$

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

$$\sum_{i=1}^{P} V(X_i \cap \hat{H}) \geq \sum_{i=1}^{P} V(X_i \cap \hat{H}_i) \tag{67}$$

that is equivalent to say that the value of the analytics based on the aggregated data (left-hand side of (67)) is larger than the sum of the values of the analytics based on separate datasets (right-hand side of (67)).

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 (65) and (66) 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

$$\mathbf{E}[\hat{\tau}(n+1)] = \int_0^\infty Z(n,\xi)\mathrm{d}\,\xi = \frac{D}{D+1}\frac{\tau^{D+1} - \hat{\tau}(n)^{D+1}}{\tau^D - \hat{\tau}(n)^D} \tag{68}$$

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, (68) 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 $\epsilon(n) = 1 - \frac{\hat{\tau}(n+1)}{\tau}$ that is always non-negative and pushed to zero by the training. The above

relationship yields

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

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 $y = \mathrm{sgn}\,(x_D - 1/2)$. 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

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

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

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

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 $1/2 \le x_D \le g_{\alpha,\beta}(x_1,\ldots,x_{D-1})$ are false negatives, while all the data points such that $g_{\alpha,\beta}(x_1,\ldots,x_{D-1}) \le x_D \le 1/2$ 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

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

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

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

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

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

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)

B1.2.1 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 $1/\Delta$

$$X_i = \left[\bigtimes_{k=1}^{D-1}\left[(\xi_{i,k}-1)\,\Delta, \xi_{i,k}\Delta\right]\right] \times [0,1]$$

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 9 shows an example of the setting for $D = 3$ and $P = 3$ producers.

B1.2.2 Empirical evidence and emerging properties

Consider $D = 2, 3, 4$, $n = 10, 15, \ldots, 300$, $v = 1/64$, $P = 4$ 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 $i = 1,\ldots,P$.

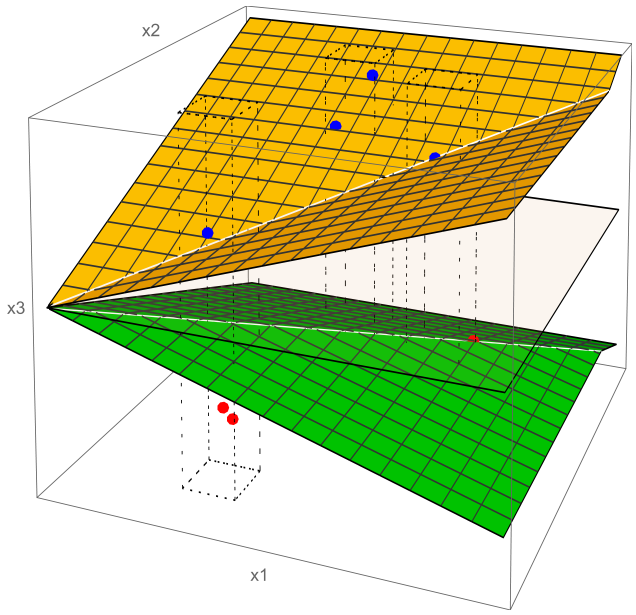


Figure 9: An example of the simplified setting with $D = 3$, $v = 1/64$ and $P = 3$ producers of data. Positive (blue) data points and negative (red) data points determine the separation manifold (in yellow) of the worst-false-negative classifier and the separation manifold (in green) of the worst-false-positive 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.

For each of the resulting P-tuple, $n_1, \ldots, n_P$, the training and performance evaluation of our worst-case classifier is repeated for 10^5 trials. In each trial, P random sets of indexes $0 \leq \xi_{i,1}, \ldots \xi_{i,D-1} \leq M$ 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 10 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 $n_1 + n_2 = n = \text{constant}$, the worst-case performance consistently improves as one approaches the even distribution of the data set between the two producers $n_1 = n_2 = n/2$.

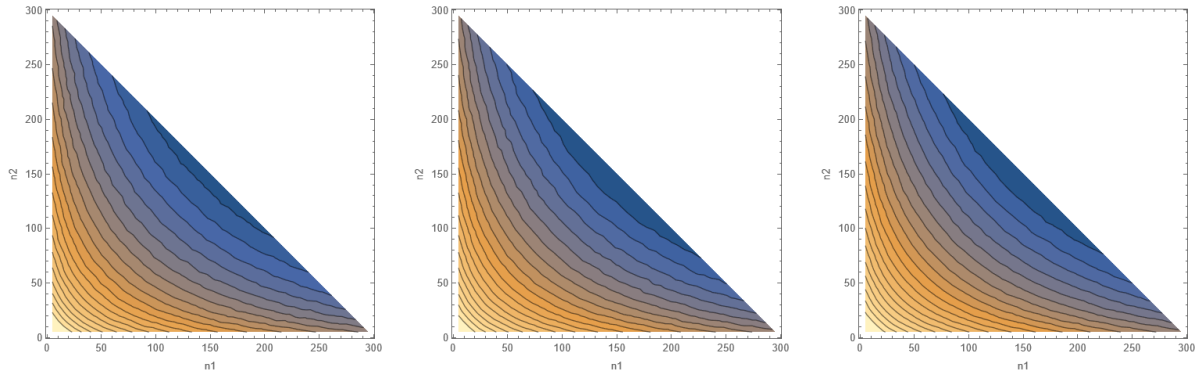
To assess whether this scope effect holds with $P > 2$ we should agree on how to measure the *evenness* of a partition of n among more than 2 producer. Among the many ways of measuring *evenness* we choose scaled Shannon entropy, i.e., in the case of P producers

$$E(n_1, \ldots, n_P) = -\frac{1}{\log P} \sum_{i=1}^{P} \frac{n_i}{n} \log \frac{n_i}{n}$$

whatever the basis of the logarithm.

The scaled entropy is minimum for $E(n, 0, \ldots, 0) = 0$ and is is maximum for $E(n/P, \ldots, n/P) = 1$. 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



a) b) c)

Figure 10: 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.

scope effect hinges. In fact, Figure 11 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 12 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.

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.

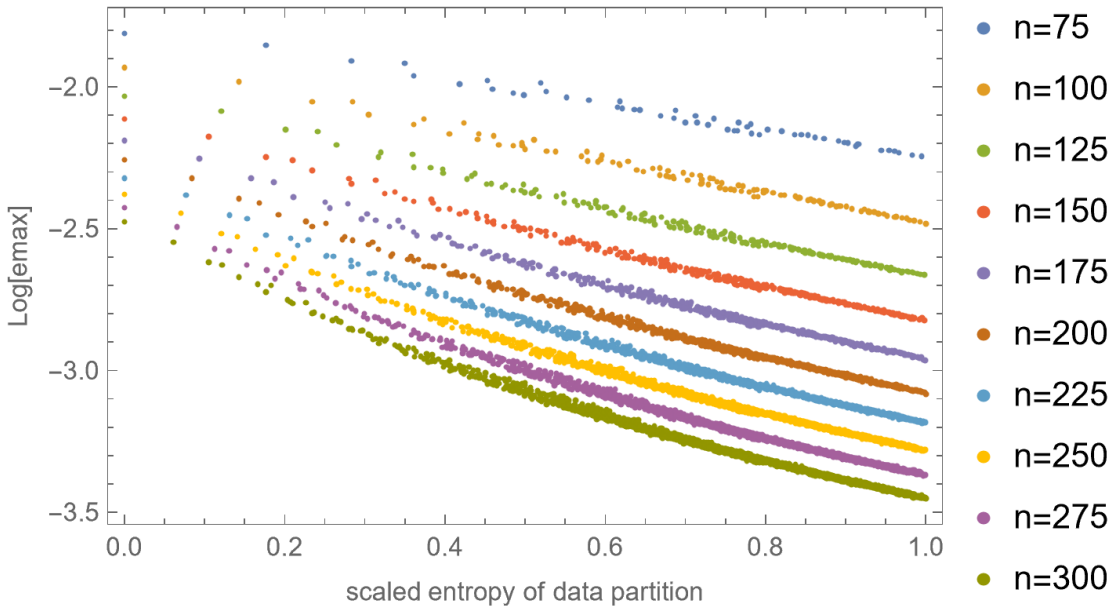


Figure 11: Logarithmic worst-case performance plotted against the scaled entropy of the distribution of n data points among $P = 4$ producers.

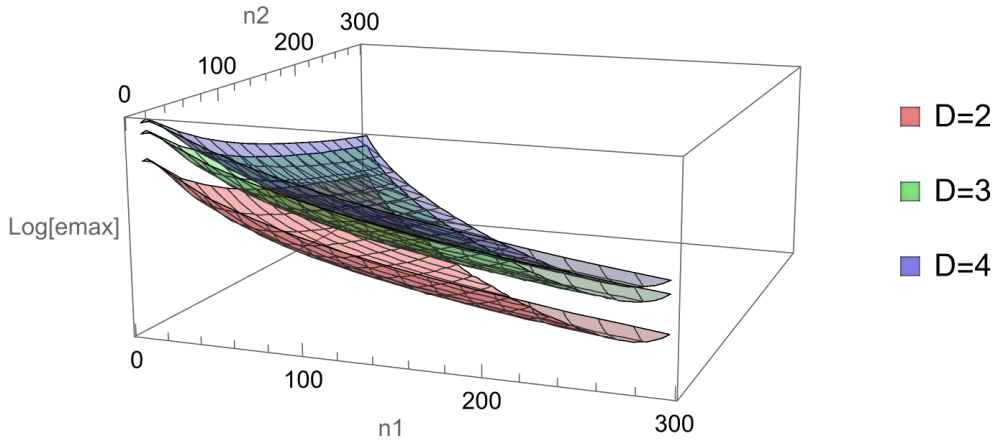


Figure 12: The logarithmic worst-case error plotted against the contributions of $P = 2$ producers working with data in a D-dimensional space.