WORKING PAPERS

N° 1404

August 2024

"Data and Competition: A Simple Framework"

Alexandre de Cornière and Greg Taylor

Toulouse School of Economics

Data and Competition: A Simple Framework*

Alexandre de Cornière† and Greg Taylor‡

August 2, 2024

Abstract

Does enhanced access to data foster or hinder competition among firms? Using a competition-in-utility framework that encompasses many situations where firms use data, we model data as a revenue-shifter and identify two opposite effects: a mark-up effect according to which data induces firms to compete harder, and a surplus-extraction effect. We provide conditions for data to be pro- or anti-competitive, requiring neither knowledge of demand nor computation of equilibrium. We apply our results to situations where data is used to recommend products, monitor insuree behavior, price-discriminate, or target advertising. We also revisit the issue of data and market structure.

Keywords: competition, data, price discrimination, targeted advertising, market structure.

JEL Classification: L1, L4, L5.

*We are grateful to Mark Armstrong, Paul Belleflamme, Daniele Condorelli, Dianne Coyle, Vincenzo Denicolò, George Georgiadis, Justin P. Johnson, Bruno Jullien, Volker Nocke, Martin Peitz, Yossi Spiegel, Tommaso Valletti, the editor and referees for useful comments and suggestions. Thanks are also due to participants at various seminars and conferences for useful comments and discussions. De Cornière acknowledges funding from ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program), the TSE Digital Center, and the Artificial and Natural Intelligence Toulouse Institute (ANITI). Taylor acknowledges the financial support of the Digital Economics Research Network and the Oxford Internet Institute's Research Programme on AI, Government, and Policy, funded by the Dieter Schwarz Stiftung gGmbH.

†Toulouse School of Economics, University of Toulouse Capitole; alexandre.de-corniere@tse-fr.eu

‡Oxford Internet Institute, University of Oxford; greg.taylor@oii.ox.ac.uk

1 Introduction

Data has become one of the most important issues in the debate about competition and regulation in the digital economy.[1] But does the use of data by firms make markets more or less competitive? On the one hand, data is a source of efficiencies. It enables firms to offer new or better products, to identify what products are worth stocking, to make personalized recommendations to consumers, or to improve monetization opportunities. On the other hand, observers have raised many concerns. One class of concerns reflects fears of exploitative behavior such as privacy violations, price-discrimination, and more generally excessive surplus extraction.[2] A second set of concerns encompass adverse implications for market structure, such as raising barriers to entry or creating winner-take-all situations (see, e.g., Furman et al., 2019, 1.71 to 1.79).

One challenge in studying the competitive effects of data lies in the variety of its uses, from targeted advertising to customized product recommendations to personalized pricing. Surprisingly, although many recent articles study markets in which firms can collect, trade, or use consumer data in various ways (see our literature review below), we are not aware of any attempt at systematically categorizing situations depending on whether data plays a pro- or an anti-competitive role.[3] Our first contribution in this article is to provide such a characterization. To do so, we use a simple model of competition-in-utility à la Armstrong and Vickers (2001), where each firm chooses the mean utility u it provides to consumers, resulting in per-consumer revenue $r(u, \delta)$ when the firm has a dataset of quality δ. This approach is flexible enough to encompass various business models, such as price competition (with uniform or personalized prices), ad-supported business models, or competition in quality. We model data as a factor that increases firms' ability to generate revenue for a given level of utility provided, a natural property across many uses of data

[1] For reports dealing with this issue, see Crémer et al. (e.g., 2019), Furman et al. (2019), and Scott Morton et al. (2019). An example hearing on the topic is the FTC's recent Hearing on Privacy, Big Data, and Competition, see `https://www.ftc.gov/news-events/events-calendar/ftc-hearing-6-competition-consumer-protection-21st-century`, accessed 1 May 2019.

[2] E.g., Scott Morton et al. (2019), p.37: "[Big Data] enables firms to charge higher prices (for goods purchased and for advertising) and engage in behavioral discrimination, allowing platforms to extract more value from users where they are weak."

[3] This statement does not apply to the literature on competitive price-discrimination, as reviewed for instance by Stole (2007).

(we provide several microfoundations in Section 5). This might be because data can be used to increase the surplus created by a product (e.g., through better personalization) or because the data can be used to extract a bigger share of the surplus (e.g., through price discrimination) or both.

As a first step, we provide a result that characterizes environments where data is *unilaterally pro-competitive*, in the sense that a better dataset induces a firm to offer more utility to consumers, keeping its rivals' offers fixed (i.e., the firm's best-response in the utility space shifts upwards). Data is *unilaterally anti-competitive* when it shifts the best response downwards. We highlight a potential trade-off between two effects. The first is the *mark-up effect*: because data increases firms' mark-ups, it also induces them to compete more fiercely to attract consumers. The second effect, which we call the *surplus extraction effect*, is more ambiguous: depending on the way it is used, data may enable firms to extract or on the contrary to provide consumer surplus more efficiently. We then show that, in many cases, the overall competitive effect of data can be determined without having to compute the equilibrium or to make functional form assumptions about demand, as it depends only on the shape of the per-consumer revenue function $r(u, \delta)$, and in particular on whether it is (log-)supermodular.

Up to this point, the model is relatively abstract and δ could be interpreted as any factor that shifts the per-consumer revenue or mark-up.[4] But our analysis reveals that the competitive effect of increasing δ depends on *how* it shifts r. Our second contribution is to show how four simple but canonical models of data use (namely, to improve products, target advertisements, mitigate moral hazard, and price discriminate) can be recast into our framework. Crucially, these applications each imply a different per-consumer revenue function, reflecting their different technologies of data use. The overall takeaway is therefore that different uses of data produce starkly different predictions about its competitive effects. Nevertheless, in each of these cases those effects can be decomposed into mark-up and surplus extraction effects, explaining the differences in terms of firms'

[4]For example, δ could be the firm's stock of cost-reducing innovations. It is well-known that a lower marginal cost is passed through with a lower price when firms choose prices in the face of a (residual) demand curve satisfying basic regularity conditions. Our framework would accordingly show δ to be pro-competitive when applied to such a situation.

underlying strategic incentives, and are easily characterized using the simple conditions from our baseline analysis.

Our third contribution is to discuss how the results can be applied to the dynamic implications of data for market structure. The same conditions that can be used to identify whether data is (statically) pro- or anti-competitive also reveal whether data is a barrier to entry and are instructive about whether data will lead to long-run concentration.

The organization of the article is as follows: after discussing the related literature, we present the basic framework in Section 2. In Section 3 we derive conditions for data to be unilaterally pro- or anti-competitive and Section 4 characterizes the equilibrium effects. We apply these results to four microfounded models of markets with data use in Section 5 to show how the effects of data can be determined. Section 6 discusses the model and shows how the analysis can be extended to incorporate dynamic issues related to market structure, consumer privacy concerns and data externalities. We conclude in Section 7.

Related Literature

Data takes many forms and has many different users and uses (Acquisti et al., 2016). Much of the literature has therefore focused on the study of particular applications of data (see Pino, 2022, for a survey). For example, one sizable literature considers the consequences of allowing firms to use data for personalized pricing (e.g., Thisse and Vives, 1988; Fudenberg and Tirole, 2000; Taylor, 2004; Acquisti and Varian, 2005; Calzolari and Pavan, 2006; Anderson et al., 2022; Belleflamme and Vergote, 2016; Kim et al., 2018; Montes et al., 2018; Bonatti and Cisternas, 2019; Gu et al., 2019; Chen et al., 2020; Ichihashi, 2020; Bounie et al., 2021). Another literature studies targeted advertising (e.g., Roy, 2000; Iyer et al., 2005; Galeotti and Moraga-González, 2008; Athey and Gans, 2010; Bergemann and Bonatti, 2011; Rutt, 2012; Johnson, 2013; Bergemann and Bonatti, 2015; de Cornière and de Nijs, 2016). These articles provide a rich picture of how data affects market outcomes in particular institutional environments. However, that picture is complex, with data sometimes being pro-competitive, but reducing consumer surplus on other occasions. Our contribution is to develop a framework that allows us to systematically characterize the

competitive effects of data while remaining agnostic about how the data is used. We stress that we do not aim to nest all extant models—the variety of modelling approaches is too great—but we do offer a model that reflects some of the most important trade-offs and shows how they play out in different contexts.

One important theme in the policy debate concerns the relationship between data use or accumulation and market structure. Recent articles such as Farboodi et al. (2019), Prüfer and Schottmüller (2021) and Hagiu and Wright (2023) study long-run market dynamics when data-enabled learning helps firms improve their products, and emphasize the potential for data to lead to increased concentration (this is related to earlier work on learning-by-doing, e.g., Dasgupta and Stiglitz, 1988; Cabral and Riordan, 1994).[5] In Section 6 we discuss how our framework can shed light on this question. On a related note, some commentators have argued that data may create a barrier to entry (e.g., Grunes and Stucke, 2016). Building on the classic analysis of Fudenberg and Tirole (1984) (see also Bulow et al., 1985), we can use our framework to show that the viability of an entry-deterrence strategy also depends on how the data is used.

The burgeoning literature on the topic has also considered specific institutional arrangements related to data that are not our main focus here. Fainmesser et al. (2023) and de Cornière and Taylor (2022) study how firms' business models affect their incentive to protect users' data from misuse. Several articles look at situations where data is sold by data brokers (e.g., Gu et al., 2021; Ichihashi, 2021a; Bounie et al., 2023; Delbono et al., 2023). Kastl et al. (2020) consider the collection of data by a platform that competes with third parties in its own marketplace. Chen et al. (2022), Herresthal et al. (2022), and de Cornière and Taylor (forthcoming) study the welfare effects of mergers that transfer data between firms, the last article building on the approach developed here. Like us, Condorelli and Padilla (2022) model data as an input that increases revenues. But their focus is on pre-emptive data acquisition as a foreclosure strategy rather than characterizing

[5]See, also, Campbell et al. (2015), Lam and Liu (2020) for theoretical studies of how data regulations may affect market structure, and Johnson et al. (2023) for a related empirical study on the effects of European privacy regulations. Also related is Eeckhout and Veldkamp (2022) who study the link between data and market power from a different perspective: they model data as a way to reduce risk, and show that data leads firms to invest and produce more, but that the effect on mark-ups is ambiguous.

the within-market competitive effects of data.

2 Model

Demand We consider a market with $n \geq 1$ firms. As in Armstrong and Vickers (2001), each firm chooses a mean utility level u_i, resulting in demand $D(u_i, \mathbf{u}_{-i})$, where $\mathbf{u}_{-i}$ are the mean utilities available from other firms (in the case of monopoly we simply have $D(u_i, \mathbf{u}_{-i}) = D(u_i)$). Depending on the context, u_i may depend on firm i's price, on its quality, or on any of its strategic choices, such as the "ad load" that a media firm imposes on viewers for instance. Demand is assumed to be continuously differentiable, and such that $\frac{\partial D(u_i, \mathbf{u}_{-i})}{\partial u_i} \geq 0$ and $\frac{\partial D(u_i, \mathbf{u}_{-i})}{\partial u_j} \leq 0$ for $j \neq i$. [6]. In the logit model, for instance, each consumer gets final utility $u_i + \xi_i = \mathbf{x}_i\boldsymbol{\beta} - \alpha p_i + \xi_i$ where $\mathbf{x}_i$ is a vector of product characteristics, p_i is the price, and ξ_i an unobservable (to the econometrician) shock. Such a model can also be interpreted as one with a representative consumer with taste for diversity (Anderson et al., 1988).

Per-consumer revenue and fixed costs Firms' marginal cost is constant and, with no further loss of generality, normalized to zero. The choice of a mean utility u_i determines firm i's *per-consumer* revenue, $r(u_i)$, which we assume is continuously differentiable. The model accommodates various situations, such as that where r is simply the price, where the price is zero and the firm earns r per-consumer from ads (see Section 5.2), or where r is the revenue from the optimal effort-inducing contract in an environment with moral hazard (see Section 5.3). The fixed cost of choosing u_i is $C(u_i)$, with $C'(u_i) \geq 0$ and $C''(u_i) \geq 0$.[7]

Data Each firm has access to data containing strategically relevant information about the market. The quality of the data may vary with the number of variables or observations

[6]Such a formulation is consistent with discrete choice models such that the utility that consumer l obtains from firm i is of the form $u_{il} = u_i + \epsilon_{il}$, where ϵ_{il} is a random taste shock that is orthogonal to any data the firms might have

[7]In Armstrong and Vickers (2001) and many natural examples, $C(u_i) = 0$, which holds when u_i depends on firm i's price only. With investments in quality, one may have $C'(u_i) > 0$.

it contains, or with the relevance, accuracy or recency of those observations. To reflect the differing qualities of datasets, we assume that they can be ranked such that a better (e.g., more informative) dataset allows the firm to generate more revenue per-consumer for any level of utility. This guarantees that it is possible to represent each dataset by a score, $\delta_i \in \mathbb{R}$, such that the associated mark-up, $r(u_i, \delta_i)$, is increasing in δ_i.[8] Given this representation, we take r as a primitive and perform our analysis using δ_i rather than the dataset it represents.

Assumption 1. A firm with a better dataset (i.e., a higher δ_i) achieves a higher mark-up for any given utility level provided to consumers: $\frac{\partial r(u_i,\delta_i)}{\partial \delta_i} > 0$.

We often say a firm with a higher δ_i has 'more' data, even though a larger δ_i might actually correspond to a more informative dataset of equal size. We can think of r as capturing the technology of data use, and we will see that different ways of using data—such as targeted advertising or price discrimination—generate differences in the shape of r, and this implies different effects of data.[9] Thus, although in our baseline setting δ_i can be reinterpreted as any factor that monotonically shifts the mark-up, the application to data is particularly interesting for its ability to naturally generate both pro- or anti-competitive effects depending on how the data is used.

To give a simple example, consider a situation where a multi-product firm makes recommendations to consumers based on their and other consumers' consumption history (for instance, think of Netflix using a collaborative filtering system to recommend shows). Such recommendations improve the value of the service, so that consumers' mean utility has the form $u_i = V(\delta_i) - p_i$, where the value of the service $V(\delta_i)$ is increasing in the quantity of data held by i, and the price of the service is p_i. If we normalize the marginal cost to zero the per-consumer mark-up is $r(u_i, \delta_i) = p_i = V(\delta_i) - u_i$. We keep this example reduced-form for brevity, but Appendix B.1 gives it a Bayesian microfoundation where

[8] Let the set of all datasets be Ω and the per-consumer revenue associated with $\omega \in \Omega$ be $\tilde{r}(u, \omega)$. Then, so long as better datasets are associated with higher revenue, one example of a valid representation, $\delta : \Omega \to \mathbb{R}$, is $\delta(\omega) = \tilde{r}(0, \omega)$. For any valid representation we have $r(u_i, \delta_i) := \tilde{r}(u_i, \delta^{-1}(\delta_i))$.

[9] Data might also lower the fixed cost. If data reduces the incremental fixed cost of providing utility, $\frac{\partial^2 C_i}{\partial u_i \partial \delta_i} \leq 0$, then this effect in isolation unambiguously leads the firm to offer higher utility so data would more often be pro-competitive. Both parts i and ii of Proposition 1 below, though, would remain unchanged.

correlated signals about consumers' tastes are used to learn which products to recommend.

There are two ways to interpret δ_i. Firstly, it might measure the aggregate data held by i about the overall population of consumers. Having such data might enable the firm to provide a better offer to all consumers as, for example, when a search engine provides better results for queries it has seen before. Alternatively, δ_i might measure the amount of data the firm has about a single specific consumer, in which case u_i is interpreted as a personalized offer to that consumer and each consumer is treated as a separate market, buying from i with probability $D(u_i, \mathbf{u}_{-i})$. Of course, the data used by firms is often personal data, raising potential concerns around privacy or data externalities between consumers. We abstract away from intrinsic privacy concerns in the main model, but discuss how these issues can easily be incorporated into the analysis in Section 6.

Consider a hypothetical game where, in the first stage, firms take actions that determine δ_i. For instance, firms could invest in data collection, harvest data from their existing customer base, or buy data from data-brokers. In the second stage, firms simultaneously choose their u_i to maximize profit, $\pi(u_i, \mathbf{u}_{-i}, \delta_i) = r(u_i, \delta_i)D(u_i, \mathbf{u}_{-i}) - C(u_i)$.

In the next three sections, we focus on the second stage of this game, and study whether better data induces firms to offer more utility to consumers. We discuss the implications for the dynamics of data accumulation in Section 6.

Throughout the article we maintain the following assumption:

Assumption 2. (i) $\pi(u_i, \mathbf{u}_{-i}, \delta_i)$ is differentiable and quasi-concave in u_i, chosen from a compact set.

(ii) For any $(\delta_1, ..., \delta_n)$, there exists a unique equilibrium,[10] given by the first-order conditions $\frac{\partial \pi(u_i^*, \mathbf{u^*}_{-i}, \delta_i)}{\partial u_i} = 0$.

(iii) $\frac{\partial r(u_i, \delta_i)}{\partial u_i}$ is non-negative everywhere or non-positive everywhere.

Sufficient conditions for the quasi-concavity of profit are that C is sufficiently convex, or that both r and D are log-concave in u_i. One sufficient condition for existence and uniqueness is that $\frac{\partial^2 \pi_i}{\partial u_i^2} + \sum_{j \neq i} |\frac{\partial^2 \pi_i}{\partial u_i \partial u_j}| < 0$ (see Vives, 2001, p47).

[10] Because the strategy set is compact, the unique equilibrium is stable for any n in the case of strategic complements, and for $n = 2$ with strategic substitutes (Vives, 2001).

Our main results would be unchanged if we added an i subscript to $r_i(\cdot,\cdot)$, $D_i(\cdot,\cdot)$, and $C_i(\cdot)$ and allowed these to be firm-specific functions. Because this would clutter the notation without yielding any additional insights, we assume these functions are symmetric. Occasionally, and where no confusion results, we instead write $D_i \equiv D(u_i, \mathbf{u}_{-i})$, $r_i \equiv r(u_i, \delta_i)$ and $\pi_i \equiv \pi(u_i, \mathbf{u}_{-i}, \delta_i)$ for conciseness.

3 Unilateral effects of data and monopolists' incentives

We begin by studying how the quantity of data held by firm i affects its incentives to offer utility, taking as given the utility offered by any rivals it may face. Let $\hat{u}_i(\mathbf{u}_{-i}, \delta_i)$ be firm i's best-response function. We use the following definition.

Definition 1. We say that data is *unilaterally pro-competitive* (UPC) for firm i for a given $\mathbf{u}_{-i}$ if $\frac{\partial \hat{u}_i(\mathbf{u}_{-i},\delta_i)}{\partial \delta_i} > 0$. We say that data is *unilaterally anti-competitive* (UAC) when the inequality is reversed.

This notion of pro- or anti-competitiveness of data captures the "unilateral" effect of data: data is UPC if better data induces a firm to offer more utility to consumers, keeping any rivals' utility offers constant. It therefore fully characterises how a monopolist responds to a change in the data available, as well as being an important ingredient in the competitive equilibrium analysis to follow.

Firm i's best response function, $\hat{u}_i(\mathbf{u}_{-i}, \delta_i)$, is found as the solution to its first-order condition:

$$\frac{\partial \pi(u_i, \mathbf{u}_{-i}, \delta_i)}{\partial u_i} = \frac{\partial r(u_i, \delta_i)}{\partial u_i} D(u_i, \mathbf{u}_{-i}) + \frac{\partial D(u_i, \mathbf{u}_{-i})}{\partial u_i} r(u_i, \delta_i) - \frac{\partial C(u_i)}{\partial u_i} = 0. \qquad (1)$$

By standard arguments, firm i's best-response is increasing in δ_i if and only if $\frac{\partial^2 \pi_i}{\partial u_i \partial \delta_i} > 0$.

Differentiating (1) with respect to δ_i, the condition $\frac{\partial^2 \pi_i}{\partial u_i \partial \delta_i} > 0$ can be rewritten as:

$$\underbrace{\frac{\partial D(u_i, \mathbf{u}_{-i})}{\partial u_i} \frac{\partial r(u_i, \delta_i)}{\partial \delta_i}}_{\text{mark-up effect}} + \underbrace{\frac{\partial^2 r(u_i, \delta_i)}{\partial u_i \partial \delta_i} D(u_i, \mathbf{u}_{-i})}_{\text{surplus extraction effect}} > 0. \tag{2}$$

Data affects the incentive to provide utility in two ways. Firstly, an extra unit of data increases the mark-up earned from an additional consumer and therefore the incentive to attract consumers with high utility offers. This *mark-up effect* corresponds to the first term in (2), which is always positive.

To see why we call the second term a *surplus extraction effect*, note that $-\frac{\partial r(u_i, \delta_i)}{\partial u_i}$ measures how efficient the firm is at extracting surplus from consumers (i.e., at generating revenue by reducing the utility provided).[11] If data makes the firm more efficient at extracting surplus, we have $\frac{\partial}{\partial \delta_i}\left(-\frac{\partial r(u_i, \delta_i)}{\partial u_i}\right) > 0$, i.e. $\frac{\partial^2 r(u_i, \delta_i)}{\partial u_i \partial \delta_i} < 0$. In that case the second term in (2) is negative, so that the overall sign of (2) is ambiguous. We provide examples in Section 5.

Equation (2) thus reveals that a sufficient condition for data to be UPC is that r be supermodular, $\frac{\partial^2 r(u_i, \delta_i)}{\partial u_i \partial \delta_i} \geq 0$, i.e., that data doesn't make the firm more efficient at extracting surplus from consumers. One way to make further progress is to consider the case where the fixed cost is constant, i.e. $C'(u_i) = 0$ (see Section 5 for several natural examples). Then we can substitute the first-order condition, $r_i \frac{\partial D_i}{\partial u_i} + \frac{\partial r_i}{\partial u_i} D_i = 0$, into (2) and obtain that data is UPC if and only if $r_i \frac{\partial^2 r_i}{\partial u_i \partial \delta_i} > \frac{\partial r_i}{\partial u_i} \frac{\partial r_i}{\partial \delta_i}$, which is equivalent to $\frac{\partial^2 \ln(r_i)}{\partial u_i \partial \delta_i} > 0$. We summarize this discussion in the following proposition (whose proof is in Appendix A):

Proposition 1. *Data is unilaterally pro-competitive if and only if the utility-elasticity of demand is larger than the utility-elasticity of the marginal value of data:* $\frac{\partial D(u_i, \mathbf{u}_{-i})}{\partial u_i} \frac{u_i}{D(u_i, \mathbf{u}_{-i})} > -\frac{\partial^2 r(u_i, \delta_i)}{\partial u_i \partial \delta_i} \frac{u_i}{\partial r(u_i, \delta_i)/\partial \delta_i}$. *In particular,*

[11] Here, we have in mind situations where actions that increase utility, such as a price cut, reduce per-consumer revenue. Though less standard, the model also admits the case where $\frac{\partial r_i}{\partial u_i} > 0$, meaning the firm generates revenue by providing rather than extracting utility (Section 5.2 offers an example). Ultimately, the underlying logic of the analysis is the same: what matters for the second term in (2) is whether data makes it more or less attractive for a firm to provide utility to each consumer.

(i) If r is supermodular ($\frac{\partial^2 r(u_i,\delta_i)}{\partial u_i \partial \delta_i} \geq 0$) then data is unilaterally pro-competitive for all $\mathbf{u}_{-i}$.

(ii) When fixed costs are constant ($C' = 0$), data is unilaterally pro-competitive for all $\mathbf{u}_{-i}$ if and only if r is log-supermodular ($\frac{\partial^2 \ln(r(u_i,\delta_i))}{\partial u_i \partial \delta_i} > 0$).

An interesting feature of Proposition 1 is that conditions (i) and (ii) do not depend on the demand function D. Instead, what is most important is the economic technology, $r(u_i, \delta_i)$, that connects data, utility, and revenue. Whether data is UPC or UAC is a local property of this technology. If r is (log-)supermodular in some parts of its domain and (log-)submodular in others then varying u_i can cause data to switch from being pro- to anti-competitive. Thus, although D does not directly appear in conditions (i) and (ii), it can still play an indirect role through the determination of the equilibrium u_i. However, we will later see that in many natural applications of the model r is globally (log-)supermodular or (log-)submodular and the competitive effect of data does not depend even indirectly on demand. Thus, in the case of monopoly, the equilibrium effects of data can often be characterized without computing equilibrium or knowing demand.

4 Equilibrium competitive effects of data

We now turn from the unilateral effects of data to its equilibrium effects under competition.

Increase in general quality of data We first consider the case where the firms are symmetric ($\delta_1 = \ldots = \delta_n = \delta$). Such a situation could, for instance, correspond to one where firms obtain data from third party data brokers who sell data non-exclusively to each of them. Then, one could think of a strengthening of privacy laws as a decrease in δ, or of improvements in analytics technology as an increase in δ. In this setup, the equilibrium is given by the fixed point of $\hat{u}$. An increase in δ causes this fixed point to shift in the same direction as the best responses (see Figure 1). It is immediate that the equilibrium effects of data are the same as the unilateral ones and can be determined using the conditions in Proposition 1 without computing the equilibrium:

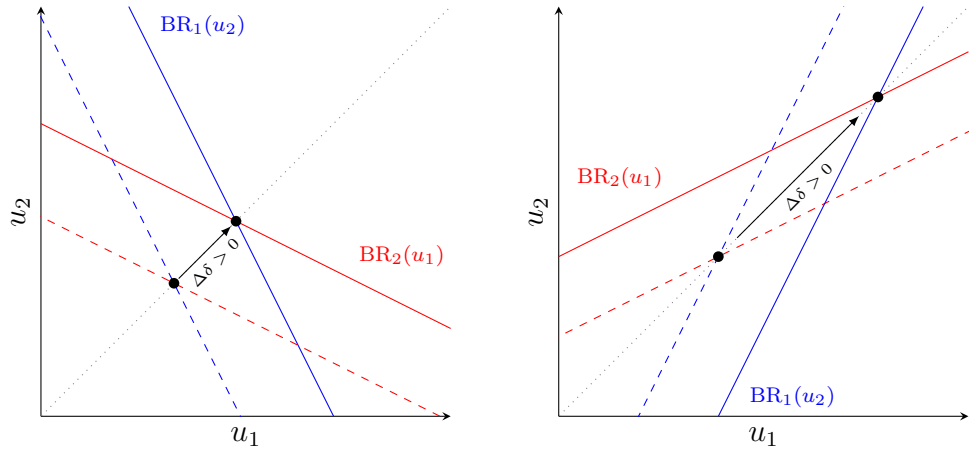

Figure 1: An increase in $\delta(=\delta_1=\delta_2)$ causes equilibrium utility offers to increase when data is pro-competitive. For $n=2$, the left panel shows the case of strategic substitutes, the right strategic complements.

Proposition 2. *Suppose that firms are symmetric ($\delta_1 = \ldots = \delta_n = \delta$). Then an increase in δ increases consumer surplus in equilibrium if and only if data is UPC.*

Asymmetric data We now consider a scenario where firms may have different δs and only some firms enjoy an increase in the quality of their data. Such an exercise could be motivated by policy proposals aimed at reducing an incumbent's data advantage. For example, Article 6(11) of the EU's Digital Markets Act imposes obligations for incumbent "gatekeeper" platforms to share search query and other types of data with rival firms. Formally, this intervention amounts to an increase in δ_i, starting from $\delta_i < \delta_j$. Our results provide guidance on when such a policy would be effective, and sounds a note of warning about cases where it might be counter-productive.

Giving firm i more data has both a direct (unilateral) effect and an indirect (strategic) effect. The direct effect comes from the unilateral shift in i's best response. This is exactly the effect we saw in Section 3 and its sign is characterized in Proposition 1 (e.g., is given by the log-supermodularity of r if fixed costs are constant).

The indirect effect comes as all firms strategically adjust their utility offers to restore equilibrium, given i's new best-response function. The direction of this strategic effect depends on whether firms' actions are strategic complements or substitutes. If payoffs are strategic complements (arguably the more natural case; we provide numerous examples

below) then the indirect and unilateral effects work in the same direction, meaning an increase in δ_i leads all firms to increase their equilibrium utility offer if and only if data is UPC. Proposition 1 can then be used to characterize the equilibrium competitive effect directly from r.

One advantage of the competition-in-utilities approach is that it can readily accommodate both strategic complements and substitutes. But this leaves open the question of how to determine which is the relevant case in any given market. Here, we can usefully invoke the concepts of congruence and conflict from de Cornière and Taylor (2019) if we put more structure on competition, assuming that demand is linear: $D(u_i, \mathbf{u}_{-i}) = \alpha_i^0 + \alpha_i^i u_i - \sum_{j \neq i} \alpha_i^j u_j$, with $\alpha_i^i, \alpha_i^j > 0$. For example, this nests Hotelling duopoly.[12]

Definition 2. Payoffs are *congruent* whenever $\frac{\partial r(u_i, \delta_i)}{\partial u_i} > 0$. When the inequality is reversed, we say that payoffs are *conflicting*.

Whereas the example in Section 2 had $r(u_i, \delta_i) = V(\delta_i) - u_i$ and therefore features conflicting payoffs, a simple model with congruent payoffs would be one where media firms' per-consumer advertising revenue increases with the quality of their content, either because consumers consume more content or because advertisers are willing to pay a premium to be associated with quality content. See Section 5.2 for an example.

The congruence/conflict property suffices to characterize the strategic effect that is the missing ingredient in our equilibrium analysis:

Proposition 3. *Suppose demand is linear. Then (i) u_i and u_j are strategic complements if payoffs are conflicting and strategic substitutes if payoffs are congruent.*

(ii) In the case of duopoly, the effect of an increase in δ_i, on u_i^* and u_j^* is given in the following table:

[12] To obtain the Hotelling model, set $n = 2$, $\alpha_i^0 = \frac{1}{2}$, and $\alpha_i^i = \alpha_i^j = \frac{1}{2t}$, where t is the transport cost.

	Data	
Payoffs	*UAC*	*UPC*
Conflicting	$\downarrow u_i^*, \downarrow u_j^*$	$\uparrow u_i^*, \uparrow u_j^*$
Congruent	$\downarrow u_i^*, \uparrow u_j^*$	$\uparrow u_i^*, \downarrow u_j^*$

(iii) More generally, for any $n \geq 2$, if payoffs are conflicting then all firms' equilibrium utility offers increase in δ_i when data is UPC and decrease when data is UAC.

The proof of Proposition 3 is in Appendix A. Propositions 1–3 together allow us to reduce the problem of signing the unilateral and equilibrium effects of data to the much simpler one of signing at most two derivatives of r_i, namely $\frac{\partial r_i}{\partial u_i}$ along with one of either $\frac{\partial^2 r_i}{\partial u_i \partial \delta_i}$ or $\frac{\partial^2 \ln(r_i)}{\partial u_i \partial \delta_i}$. This obviates, in particular, the need to fully compute equilibrium in order to obtain comparative statics. Instead, one need only identify enough parameters of r_i to sign the two derivatives of interest. This represents a substantial simplification when firms can be asymmetric or r has a functional form that would make solving for equilibrium difficult (see the next section for some examples). Although it assumed linear demand, Proposition 3 continues to hold for other demand specifications so long as either (i) $\frac{\partial^2 D_i}{\partial u_i \partial u_j}$ is small enough or (ii) the congruence or conflict property is sufficiently strong (i.e., $|\frac{\partial r_i}{\partial u_i}|$ is large).[13]

5 Applications

Even though the model presented above is more general than "just" a model of data, we believe that data is a particularly interesting application because we can draw a correspondence between the properties of r and the underlying uses of data and business models that r represents. To see this, we now discuss how some established models of product improvement, targeted advertising, moral hazard and price discrimination can be cast into the competition-in-utility framework. In each case we provide an informational microfoundation to the parameter δ_i, and derive the implied per-consumer revenue $r(u_i, \delta_i)$.

[13] In particular, we have strategic complementarity if $\frac{\partial^2 \pi_i}{\partial u_i \partial u_j} = \frac{\partial r_i}{\partial u_i}\frac{\partial D_i}{\partial u_j} + r_i \frac{\partial^2 D_i}{\partial u_i \partial u_j} > 0$, and strategic substitutability if the inequality is reversed.

We can therefore use the results above to characterize the mark-up and surplus extraction effects and study how the effects of data vary across these applications and why. Indeed, these four examples cover a case where the surplus extraction effect is inactive, ambiguous, positive, and negative, with correspondingly different competitive effects.

5.1 Product improvement

An important use of data is to improve the quality of the products or services offered by firms based on the feedback or choices of past customers. For instance, search engine algorithms use data about past queries to improve their results. This improvement can also take the form of more personalized recommendations without affecting the quality of the underlying products.

The simplest way to model this situation within our framework, as already discussed in Section 2, is to write the mean utility as $u_i = V(\delta_i) - p_i$, where $V(\delta_i)$ is the quality of the product (increasing in δ_i) and p_i its price.[14] This is essentially the formulation of Hagiu and Wright (2023).[15] We provide one possible microfoundation in Appendix B.1. The per-consumer revenue is equal to p_i, meaning we can invert the utility function to write $r(u_i, \delta_i) = V(\delta_i) - u_i$.

It is immediate that $\frac{\partial^2 r_i}{\partial u_i \partial \delta_i} = 0$. The surplus extraction effect is inactive here because the firm can extract surplus via the price, independent of δ_i. Because only the markup effect remains, data is UPC by Proposition 1.

Proposition 4. *In the model of product improvement the surplus extraction effect is inactive ($\frac{\partial^2 r(u_i,\delta_i)}{\partial u_i \partial \delta_i} = 0$). Data is therefore UPC.*

Intuitively, data increases the quality of the product, allowing the firm to hold u_i constant while charging a higher price. This makes the marginal consumer more valuable

[14] Another approach would be to suppose that consumers have multi-unit demand and buy $Q(p_i, \delta_i)$ units of the product, increasing in δ_i. We can use the demand-shifting framework of Cowan (2004). If we equate u_i with the standard measure of consumer surplus then Cowan's model can be recast in our framework to generate insights on the competitive effects of a higher δ_i that shifts demand outwards.

[15] Guembel and Hege (2021) also study a related model where consumers observe the realisation of the firm's signal before purchasing. One could also cast that model in a competition in utility framework, with a small extra notational burden. Note that one substantial difference between our model and Hagiu and Wright (2023) and Guembel and Hege (2021) is that they do not have horizontal differentiation so that the equilibrium is not always given by the first-order condition.

at any given u_i so the firm wants to increase utility to attract more consumers. This is the logic of the markup effect at work.

Note that in this application $r(u_i, \delta_i)$ is decreasing in u_i, meaning that payoffs are conflicting. If we further assume a linear demand, this implies that utilities are strategic complements, so that the equilibrium effect of data is the same as the unilateral one, and consumer surplus increases with data.

5.2 Targeted advertising

Now consider a free-to-access media platform that uses data to facilitate the targeting of advertisements. If the platform sells a_i slots of advertising then its per-consumer revenue is $a_i P(a_i, \delta_i)$, where the inverse demand for ads, $P(a_i, \delta_i)$, is decreasing and log-concave in a_i, and increasing in δ_i.[16]

The relationship between the quantities of ads sold and consumers' utility is ambiguous in general. On the one hand, if consumers view ads as a nuisance (as in Anderson and Coate, 2005) then a higher a_i corresponds to a lower u_i. On the other hand, in a model with an intensive consumption margin, showing more ads may require keeping consumers on the platform for longer, for instance by investing more in the quality of the content. In that case a higher a_i would correspond to a higher u_i. In Appendix B.2 we provide microfoundations for an *ad-nuisance* and an *intensive margin* model, with $C'(u_i) = 0$ in the former model and $C'(u_i) > 0$ in the latter. In each model one can write firm i's revenue as $r(u_i, \delta_i) D(u_i, \mathbf{u}_{-i})$, with $r(u_i, \delta_i) = A(u_i) P\left(A(u_i), \delta_i\right)$.[17] The function $A(u_i)$ represents the number of ads as a function of the utility provided. $A' \leq 0$ in the ad nuisance model, and $A' \geq 0$ in the intensive margin model.

To better understand the competing forces at play here, let us analyze the sign of the

[16] In the literature on targeted advertising (e.g. Johnson and Myatt, 2006), an improvement in information is often modelled as inducing a clockwise rotation in demand. Under that framework we would need the additional assumption that the relevant range is located to the left of the rotation point, so that P is increasing in δ_i.

[17] We have $\frac{\partial r_i}{\partial u_i} = A'(u_i)[P_i + A(u_i)\frac{\partial P_i}{\partial a}]$. The term in square brackets is the marginal revenue, which i will never choose to make negative. Thus, part (iii) of Assumption 2 is satisfied over the relevant range.

surplus extraction effect:

$$\frac{\partial^2 r(u_i, \delta_i)}{\partial u_i \partial \delta_i} = A'(u_i) \left[\frac{\partial P(A(u_i), \delta_i)}{\partial \delta_i} + A(u_i) \frac{\partial^2 P(A(u_i), \delta_i)}{\partial a_i \partial \delta_i} \right]. \tag{3}$$

The term between square brackets captures how the equilibrium number of ads changes with more data. The first term ($\frac{\partial P(A(u_i),\delta_i)}{\partial \delta_i}$) corresponds to the idea that ad slots sell for a higher price with more data, which encourages the firm to sell more slots. But data could also change the slope of the demand for ad slots, as captured by $\frac{\partial^2 P(A(u_i),\delta_i)}{\partial a_i \partial \delta_i}$. If that term is negative, data may lead the firm to reduce its ad load so as to extract more value from the advertisers with relatively high willingness to pay. Notice that switching from the ad nuisance to the intensive margin model, ceteris paribus, reverses the sign of the surplus extraction effect.

The following proposition is proven in Appendix B.2.1:

Proposition 5. *In the intensive margin model of targeted advertising (with $A'(u_i) > 0$ and $C'(u_i) > 0$), data is UPC if the marginal advertising revenue ($P(a_i, \delta_i) + a_i \frac{\partial P(a_i,\delta_i)}{\partial a_i}$) is increasing in δ_i.*

In the ad nuisance model of targeted advertising (with $A'(u_i) < 0$ and $C'(u_i) = 0$), data is UPC if and only if data makes demand for advertising less elastic (i.e., $\frac{\partial^2 \ln(P(a_i,\delta_i))}{\partial a_i \partial \delta_i} < 0$).

Note that payoffs are conflicting in the ad nuisance model but congruent in the intensive margin one. When demand is linear this corresponds to strategic complementarity and substitutability respectively.

Using Proposition 5, if data induces a vertical shift in the ad demand (i.e., $P(a_i, \delta_i) = \phi(a_i) + \delta_i$), then data is UPC in the intensive margin model but UAC in the ad nuisance one. If data induces a rotation of demand such that $P(a_i, \delta_i) = \delta_i \phi(a_i) + (1 - \delta_i)m$, the effect of data is ambiguous in the intensive margin model, and data is UPC in the ad nuisance model. Given that the same assumption on the effect of data on advertisers' demand leads to different results in the two models above, let us emphasize that our point here is not to argue that data is per se UPC or UAC when used to target ads, but rather to stress the modelling assumptions that drive such a result, namely, consumers' attitude

to ads and the manner in which data affects advertisers' willingness to pay.

Other considerations As a final remark, ad markets are complex and this section presents a fairly simple model of the role of targeting. In Appendix B we extend this approach to incorporate some other features relevant to ad markets: consumer multihoming and situations where firms can charge a subscription fee as well as showing ads.

5.3 Moral hazard

Data can also be used to alleviate problems of asymmetric information in insurance markets and other situations of moral hazard. For example, insurers like Geico and UnitedHealth Group use vehicle telematics or personal fitness trackers to log customers' behavior and condition insurance contracts on the data recorded.

Consider a simple model of insurance under moral hazard with binary effort level. A risk-averse consumer who exerts no protection effort incurs a loss L with probability 1. If he exerts effort, he avoids the loss with probability α. The utility function is separable in money and effort: if the final wealth is W then utility is $V(W) - ke$, where $e \in \{0, 1\}$ is the level of effort and $k > 0$ the cost of effort. V is increasing and concave, and we normalize consumers' initial wealth to zero. When a consumer suffers a loss even though he exerted effort, his insurer i observes with probability δ_i data that proves that the consumer exerted effort. With probability $1 - \delta_i$ the data is inconclusive and the insurer learns nothing from it.[18]

Each risk-neutral insurer offers a contract $\mathcal{C}_i \equiv \{p_i, X_{Hi}, X_{Li}\}$, where p_i is the insurance premium that consumers pay irrespective of whether they incur the loss, X_{Hi} is the amount to be reimbursed in case of a loss if the insurer's data proves the consumer exerted effort, and X_{Li} is the amount to be reimbursed in case of a loss if the data is inconclusive.[19] Write $U(\mathcal{C}_i)$ for the utility of a consumer who picks insurer i and exerts effort, and $\tilde{U}(\mathcal{C}_i)$

[18] We choose such a stylized technology for analytical tractability, but the main insights do not depend on it. For instance, a technology where the insurer receives a signal when the consumer does not exert the effort would deliver similar results. The important point is that a more precise signal will lead the insurer to offer more insurance, as we discuss below.

[19] We assume that the insurer cannot pretend not to have received a signal.

if he exerts no effort.

Suppose that insurer i wishes to offer a level of expected utility equal to u_i, which would generate a demand $D(u_i, \mathbf{u}_{-i})$.[20] The optimal way to provide such a utility level is the solution to the following program:

$$r(u_i, \delta_i) \equiv \max_{\mathcal{C}_i} \; p_i - (1-\alpha)\left(\delta_i X_{Hi} + (1-\delta_i) X_{Li}\right) \tag{4}$$
$$\text{s.t.} \quad U(\mathcal{C}_i) \geq \tilde{U}(\mathcal{C}_i) \quad \text{and} \quad U(\mathcal{C}_i) = u_i$$

We solve this problem in Appendix B.3, where we show that the incentive constraint is always binding and that the insurer fully reimburses the loss if the effort is observed ($X^*_{Hi} = L$). We also show that $r(u_i, \delta_i)$ is increasing in δ_i: having more data allows the insurer to provide more insurance without violating the incentive constraint, which in turns allows it to increase the premium to keep utility constant and obtain more profit. Applying Proposition 1 to the implied revenue function yields the following result.

Proposition 6. *In the model of insurance with moral hazard, the surplus extraction effect is positive ($\frac{\partial^2 r(u_i,\delta_i)}{\partial u_i \partial \delta_i} > 0$) if consumers have constant absolute risk aversion or a constant relative risk-aversion above* $1/2$*. Data is then UPC.*[21]

As with product improvement, data is UPC. But unlike product improvement, here the surplus extraction effect is active and positive. In other words, data makes it cheaper for a firm to offer utility to consumers. Intuitively, data mitigates the hidden action problem and thereby allows higher levels of insurance to be offered. More insurance means less risk for the consumer who, because he is risk averse, therefore requires less wealth to reach the same utility u. Lower wealth in turn means the consumer's marginal utility of wealth is higher and it is cheaper to give the consumer additional utility: $\frac{\partial r(u_i,\delta_i)}{\partial u_i \partial \delta_i} > 0$.

One can also check that payoffs are conflicting, so that utilities are strategic complements in the case of linear demand. The equilibrium effect of data is then of the same

[20] Firm i's program is equivalent to the design of an insurance contract with random participation. Roger (2016) provides a general treatment of moral hazard with random participation, though without looking at the situation we are interested in.

[21] When the constant relative risk aversion is below $1/2$, the surplus extraction effect is negative, and must be compared with the mark-up effect. Numerical methods have not delivered a single example where data is UAC.

direction as the unilateral effect.

5.4 Price discrimination with one-stop shopping

Armstrong and Vickers (2001) use the competition-in-utility framework to study price-discrimination.[22] We can easily adapt their framework to study the effects of data about consumers' willingness to pay for products. Consider an environment with several retailers, who each offer a continuum of products. For each product, a consumer's willingness to pay is distributed according to a cumulative distribution F. Consumers are one-stop shoppers: they can only visit one retailer, but can buy all the products whose price is below their willingness to pay.

Retailer i's data allows it to identify, for each consumer, the willingness to pay for a share δ_i of the products. The retailer sets a uniform list price p_i for all its products, and can send personalized discounts to each consumer.[23] For example, one-stop shopping is common in the grocery retail industry, and grocery store loyalty schemes collect data on purchase patterns that is used to generate personalized discount coupons. Formally, if a product is such that $v < p_i$, and if the retailer observes v, it can offer a discount d such that $p_i - d \leq v$ and induce the consumer to buy.

Denote by $Q(p) = 1 - F(p)$ the expected quantity demanded for each product at a list price p.[24] We assume that the price elasticity of Q, $|pQ'(p)/p|$, is increasing in p (Marshall's second law of demand).

Denote the standard measure of consumer surplus associated with price p by $S(p) = \int_p^\infty Q(p)\,dp$, and write $\rho(u)$ for the price such that $S(\rho(u)) = u$. In words, $\rho(u)$ is the uniform price that would induce consumer surplus (utility) of u. Define $\tilde{r}(u) = \rho(u)Q(\rho(u))$

[22] Although most of the analysis in Armstrong and Vickers (2001) takes place in an environment of intense competition (so that the equilibrium is close to marginal cost-pricing), they provide a condition analogous to $\frac{\partial^2 \ln[r(u_i,\delta_i)]}{\partial u_i \partial \delta_i} > 0$ for discrimination to benefit consumers (their Lemma 3), and apply it to compare uniform pricing and two-part tariffs (Corollary 1). By explicitly incorporating data in the model we are able to study marginal improvements in the ability to price-discriminate, as well as asymmetric situations.

[23] Because products are symmetric, the list price is uniform. Our analysis would also work if retailers could *increase* certain prices based on their information, but the case of discounts is easier to present and more realistic.

[24] Q is the quantity bought from the chosen store, and should not be confused with the probability of choosing a given store, D.

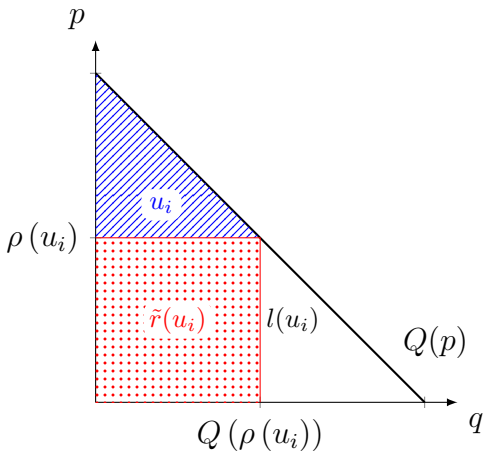

Figure 2: Surplus sharing. $r(u_i, \delta_i) = \tilde{r}(u_i) + \delta_i l(u_i)$

and $l(u) = \int_0^{\rho(u)} (Q(x) - Q(\rho(u)))dx$ for the associated deadweight loss (see Figure 2). For simplicity we assume that there are no fixed costs.

To provide utility level u, it is optimal for a firm to set the list price $p_i = \rho(u)$ and extract all of the surplus from products with $v \leq p_i$ for which it can make a personalized offer.[25] Its per-consumer revenue can then be written $r(u_i, \delta_i) = \tilde{r}(u_i) + \delta_i l(u_i)$.

It is immediate that the surplus extraction effect is negative ($\frac{\partial^2 r(u_i,\delta_i)}{\partial u_i \partial \delta_i} = l'(u_i) < 0$), raising the possibility that data might be UAC. Intuitively, data makes the firm more efficient at extracting consumer surplus because it is able to price discriminate away the deadweight loss. The opportunity cost of providing consumers with utility is therefore higher the more data the firm has. The next result goes one step further and shows that this effect dominates:

Proposition 7. *In the price discrimination model with one-stop shopping, data is UAC.*

For any utility over the monopoly level,[26] payoffs are conflicting, which implies that under linear demand utilities would be strategic complements. In that case, the equilibrium effect of data would be the same as the unilateral one.

[25] To see this, suppose that for a given price schedule a consumer buys all the products $x \in X$. Some of these products are bought at the list price p_i, others are bought at a discount, $p_i - d_i(x)$. Now, for all discounted products such that $v_i \geq p_i - d_i(x) + \epsilon$, reduce the discount $d_i(x)$ by ϵ, and decrease the list price (for all products) by an amount such that the total expense of the consumer over products in X remains the same. The firm's profit over these products is thus also unchanged. But because p_i is reduced the consumer will now buy products not in X, thereby increasing profit. Thus it cannot be optimal to offer discounts that leave some surplus to the consumer.

[26] Utility levels below the monopoly level are never optimal and can be disregarded.

Note that other models of price discrimination may not fit with our competition-in-utility approach. This is in particular the case for spatial models of price discrimination between single product firms, where price discrimination can often intensify competition. We discuss this in Section 6.

6 Discussion and extensions

Dynamics of data accumulation and market structure The framework immediately yields implications for market dynamics. First, a recurrent policy question has been whether data constitutes a barrier to entry (e.g., Grunes and Stucke, 2016; Sokol and Comerford, 2016). Consider a dynamic extension to the model in which a monopolist incumbent serves consumers in the first period and accumulates a dataset whose size is proportional to the number of consumers it serves. In the second period, a potential rival can enter the market at some cost, in which case the two firms compete as in Section 2. Otherwise, the incumbent remains a monopoly. Using the Fudenberg and Tirole (1984) terminology, UPC data makes the incumbent look *tough*: an incumbent with more data will offer a larger utility in the second stage, which reduces the entrant's profit. Over-collecting data in the first period can then be a way to deter entry. Conversely, more data makes the incumbent look *soft* when it is UAC. We can therefore use the conditions in Proposition 1 to say when data constitutes a barrier to entry.

Secondly, in an infinite-horizon setting, several articles have studied environments where data leads to market tipping because (i) more data leads firms to offer better products, which (ii) attracts more consumers and leads to the accumulation of more data in a self-reinforcing cycle (e.g., Prüfer and Schottmüller, 2021; Farboodi et al., 2019; Hagiu and Wright, 2023). The first step in this cycle embeds a pro-competitive logic,[27] which can be overturned if data is UAC. Indeed, in the limit when firms are myopic, more data leads firms to make better offers to consumers if and only if data is UPC, so UPC data is

[27]As discussed in Section 5, the baseline model of Hagiu and Wright (2023) fits our competition-in-utility framework and data is UPC. For Prüfer and Schottmüller (2021), casting the model in a competition in utility framework would lead to $\frac{\partial^2 \pi_i}{\partial u_i \partial \delta_i} > 0$, so data is also UPC. Farboodi et al. (2019)'s model is one with competition in quantities and cannot be expressed in terms of competition-in-utility, but more data leads to higher quantities, and therefore more consumer surplus.

then a necessary condition for such data-driven market tipping.

Both ways of modelling data accumulation point to a tension between short-run exploitative and long-run exclusionary concerns. When data is UPC there is little concern that it will be used to harm consumers in the short-run, but a bigger risk that it will deter entry in the long-run. The opposite is true of UAC data.

Privacy concerns and externalities Two important themes in the debate around data are the potential for consumer privacy harms (e.g., Acquisti et al., 2016; Bundeskartellamt, 2019) and the idea that one consumer's data can be used to make inferences about another, causing externalities between consumers (Choi et al., 2019; Ichihashi, 2021b; Markovich and Yehezkel, 2021; Bergemann et al., 2022; Acemoglu et al., 2022). Our framework can incorporate both issues.

First, if consumers incur a privacy harm $h(\delta)$ then we can define $U \equiv u - h(\delta)$ as the utility net of this harm and $R(U, \delta) \equiv r(U + h(\delta), \delta)$ as the corresponding markup. We can then proceed using R instead of r throughout the analysis. One thing that changes is that R may be decreasing in δ when privacy concerns are sufficiently strong, in which case supermodularity of R becomes a necessary (rather than sufficient) condition in Proposition 1(i). But Proposition 1(ii) is unchanged.

On externalities: suppose a monopoly firm has data δ_l about consumer l, and δ_{-l} about other consumers. It generates *insights* $I_l(\delta_l, \delta_{-l})$ about consumer l, increasing in both arguments, and earns markup $r(u_l, I_l)$. We can apply the results above to determine when insights are UPC or UAC. Because $\frac{\partial u_l^*}{\partial \delta_{-l}} = \frac{\partial u_l^*}{\partial I_l} \frac{\partial I_l}{\partial \delta_{-l}}$, we can immediately see that other consumers' decision to share more data exerts a negative externality on l if insights are UAC and a positive one if UPC, allowing us to apply Proposition 1 to sign the externality.

Consumer heterogeneity Although we have provided numerous examples where our analysis can be applied, we make no claim that our model nests all possible uses of data. One important restriction we impose is with regard to consumer heterogeneity. First, the competition-in-utility framework requires that actions that increase or decrease the mean utility u_i affect all of i's customers equally. Although standard discrete choice

models such as the logit or nested logit are consistent with this specification, models with random coefficients (Berry et al., 1995) are not, because a decrease in price affects different consumers differently.

Second, our way of modelling data implies that consumers are also homogeneous with respect to how data affects their (expected) utility, meaning D depends on δ only indirectly via its effect on u. The framework is thus ill-suited to study issues related to adverse selection or price-discrimination with spatial differentiation, where different types of consumers might be made better-off or worse-off by an increase in the quality of data (see Armstrong and Vickers, 2001, p584). Specifically, this rules-out the class of spatial differentiation models where data lets a firm personalize offers based on each consumer's location (as in Thisse and Vives, 1988), or reduces the differentiation between firms. This feature is shared by many articles on the economics of data (to name a few, Prüfer and Schottmüller, 2021; Hagiu and Wright, 2023; Choi et al., 2019; Acemoglu et al., 2022).

Restrictions on r Our analysis assumes that r_i does not depend on a rival's utility offer (u_j) or data (δ_j). This is not to say that i's *profit* is independent of them. Indeed, firm i's profit depends on u_j, which itself depends on δ_j, [28] and the analysis has accounted for the way this affects i's demand and equilibrium choice of u_i. Meanwhile, r_i is the firm's share of the gains to trade with a consumer (the consumer's share is u_i). In many applications, there is no reason to suppose that these gains from trade—e.g., the value the consumer places on i's product—should depend on u_j or δ_j.

However, there are some cases where data is used in a way that does induce r_i to depend on u_j or δ_j. In Appendix B.1.2 we show that if firms simultaneously choose quality and price then r_i will typically depend on u_j. An example where r_i is likely to depend on both u_j and δ_j is in advertising markets where consumers multihome. A common theme in the literature is that advertisers have a lower willingness to pay to reach a given consumer a second time (e.g. Ambrus et al., 2016; Anderson et al., 2022). The revenue i can generate by showing ads to a multihoming consumer therefore depends on the number

[28] For instance, in the product improvement application an increase in δ_j allows j to offer superior products. This is exploited by j offering higher utility to attract consumers away from i.

of ads shown by j ($A(u_j)$) and the precision of their targeting (δ_j). See Appendix B.2.4 for a formalization of this situation.

More generally, if r_i depends on u_j (but not on δ_j) then the unilateral and equilibrium effects of data are still given by Propositions 1 and 2 so the main substance of our results goes through. What changes is that we can no longer determine whether strategies are complements or substitutes via the congruence or conflict of payoffs.

If r_i depends on δ_j then we can, again, still use Proposition 1 to sign the unilateral effects. But the equilibrium effects are now more complicated because an increase in δ_i causes a shift in both i's and j's best-responses (possibly in opposite directions). The overall equilibrium effect can therefore be unambiguously signed only in some cases. More precisely, we can sign the effect of δ_i on u_i^* if i's and j's best responses shift in the same direction and we have strategic complements, or if they shift in opposite directions and we have strategic substitutes.

7 Conclusion

We have presented a framework for studying the competitive effects of data. Although we do not claim to nest all possible situations where firms use data, we show how key trade-offs are resolved across a wide range of different scenarios. Understanding these general trade-offs is important as policy makers are working to implement economy-wide regulations for data. Data makes each consumer more valuable to a firm (a pro-competitive *mark-up effect*) but may also make the firm better at surplus extraction (a potentially anti-competitive effect). We show that the trade-off between these effects can often be resolved using a simple condition on the firm's per-consumer revenue function. We illustrate the usefulness of this approach through four applications to different uses of data—product improvement, targeted advertising, moral hazard mitigation, and price discrimination. Our results can be used to show how and why the competitive effect of data differs across these cases. The model also sheds light on the efficacy of data sharing policies, the dynamic implications of data, and the sign of data externalities.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar (2022). "Too Much Data: Prices and Inefficiencies in Data Markets". *American Economic Journal: Microeconomics* 14.4, pp. 218–256.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman (2016). "The economics of privacy". *Journal of Economic Literature* 54.2, pp. 442–92.

Acquisti, Alessandro and Hal R. Varian (2005). "Conditioning Prices on Purchase History". *Marketing Science* 24.3, pp. 305–523.

Ambrus, Attila, Emilio Calvano, and Markus Reisinger (2016). "Either or both competition: A" two-sided" theory of advertising with overlapping viewerships". *American Economic Journal: Microeconomics* 8.3, pp. 189–222.

Anderson, Simon P. and Stephen Coate (2005). "Market Provision of Broadcasting: A Welfare Analysis". *The Review of Economic Studies* 72.4, pp. 947–972.

Anderson, Simon P, André De Palma, and J-F Thisse (1988). "A representative consumer theory of the logit model". *International Economic Review*, pp. 461–466.

Anderson, Simon, Alicia Baik, and Nathan Larson (2022). "Price Discrimination in the Information Age: Prices, Poaching, and Privacy with Personalized Targeted Discounts". *Review of Economic Studies* 90.5, pp. 2085–2115.

Armstrong, Mark and John Vickers (2001). "Competitive price discrimination". *RAND Journal of Economics* 32.4, pp. 1–27.

Athey, Susan and Joshua S. Gans (2010). "The Impact of Targeting Technology on Advertising Markets and Media Competition". *American Economic Review* 100.2, pp. 608–13.

Belleflamme, Paul and Wouter Vergote (2016). "Monopoly price discrimination and privacy: The hidden cost of hiding". *Economics Letters* 149, pp. 141–144.

Bergemann, Dirk and Alessandro Bonatti (2011). "Targeting in advertising markets: implications for offline versus online media". *RAND Journal of Economics* 42.3, pp. 417–443.

Bergemann, Dirk and Alessandro Bonatti (2015). "Selling cookies". *American Economic Journal: Microeconomics* 7.3, pp. 259–294.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2022). "The Economics of Social Data". *RAND Journal of Economics* 53.2, pp. 263–296.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995). "Automobile prices in market equilibrium". *Econometrica: Journal of the Econometric Society*, pp. 841–890.

Bonatti, Alessandro and Gonzalo Cisternas (Sept. 2019). "Consumer Scores and Price Discrimination". *The Review of Economic Studies* 87.2, pp. 750–791.

Bounie, David, Antoine Dubus, and Patrick Waelbroeck (2021). "Selling Strategic Information in Digital Competitive Markets". *RAND Journal of Economics* 52.2, pp. 283–313.

— (2023). "Competition Between Strategic Data Intermediaries with Implications for Merger Policy". *CER-ETH Working Paper 23/383*.

Bulow, Jeremy I, John D Geanakoplos, and Paul D Klemperer (1985). "Multimarket oligopoly: Strategic substitutes and complements". *Journal of Political economy* 93.3, pp. 488–511.

Bundeskartellamt (2019). *Case Summary: Facebook, Exploitative business terms pursuant to Section 19(1) GWB for inadequate data processing*. Germany: Bundeskartellamt.

Cabral, Luis M. B. and Michael H. Riordan (1994). "The Learning Curve, Market Dominance, and Predatory Pricing". *Econometrica* 62.5, p. 1115.

Calzolari, Giacomo and Alessandro Pavan (2006). "On the optimality of privacy in sequential contracting". *Journal of Economic theory* 130.1, pp. 168–204.

Campbell, James, Avi Goldfarb, and Catherine Tucker (2015). "Privacy regulation and market structure". *Journal of Economics and Management Strategy* 24.1, pp. 47–73.

Chen, Zhijun, Chongwoo Choe, Jiajia Cong, and Noriaki Matsushima (2022). "Data-Driven Mergers and Personalization". *RAND Journal of Economics* 51.1, pp. 3–31.

Chen, Zhijun, Chongwoo Choe, and Noriaki Matsushima (2020). "Competitive Personalized Pricing". *Management Science* 66, pp. 4003–4023.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim (2019). "Privacy and personal data collection with information externalities". *Journal of Public Economics* 173, pp. 113–124.

Condorelli, Danielle and Jorge Padilla (2022). "Data-driven Predatory Entry with Privacy-Policy Tying". *The Economic Journal* 134.658, pp. 515–537.

Cowan, Simon (2004). "Demand shifts and imperfect competition". *Working Paper*.

Crémer, Jacques, Yves-Alexandre de Montjoye, and Heike Schweitzer (2019). *Competition Policy for the Digital Era*. Report. European Commission.

Dasgupta, Partha and Joseph Stiglitz (1988). "Learning-by-doing, market structure and industrial and trade policies". *Oxford Economic Papers* 40.2, pp. 246–268.

De Cornière, Alexandre and Romain de Nijs (2016). "Online Advertising and Privacy". *RAND Journal of Economics* 47.1, pp. 48–72.

De Cornière, Alexandre and Greg Taylor (2019). "A Model of Biased Intermediation". *RAND Journal of Economics* 50.4, pp. 854–882.

— (2022). "A Model of Information Security and Competition". Working Paper.

— (forthcoming). "Data-driven Mergers". *Management Science*.

Delbono, Flavio, Carlo Reggiani, and Luca Sandrini (2023). "Strategic Data Sales With Partial Segment Profiling". JRC Digital Economy Working Paper 2021-05.

Eeckhout, Jan and Laura Veldkamp (2022). "Data and Market Power". NBER Working Paper No. 30022.

Fainmesser, Itay P., Andrea Galeotti, and Ruslan Momot (2023). "Digital Privacy". *Management Science* 69.6, pp. 3157–3173.

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

Fudenberg, Drew and Jean Tirole (1984). "The fat-cat effect, the puppy-dog ploy, and the lean and hungry look". *American Economic Review* 74.2, pp. 361–366.

— (2000). "Customer poaching and brand switching". *RAND Journal of Economics*, pp. 634–657.

Furman, Jason, Dianne Coyle, Amelia Fletcher, Derek McAuley, and Philip Marsden (2019). *Unlocking Digital Competition*. Report of the Digital Competition Expert Panel. HM Government.

Galeotti, Andrea and José Luis Moraga-González (2008). "Segmentation, advertising and prices". *International Journal of Industrial Organization* 26.5, pp. 1106 –1119.

Grunes, Alan and Maurice Stucke (2016). *Big data and competition policy*. Oxford University Press.

Gu, Yiquan, Leonardo Madio, and Carlo Reggiani (2019). "Exclusive Data, Price Manipulation and Market Leadership". CESifo Working Paper No. 7853.

— (2021). "Data brokers co-opetition". *Oxford Economic Papers* 74.3, pp. 820–839.

Guembel, Alexander and Ulrich Hege (2021). "Data, Product Targeting and Competition". *Working Paper*.

Hagiu, Andrei and Julian Wright (2023). "Data-enabled learning, network effects and competitive advantage". *RAND Journal of Economics* 54 (4), pp. 638–667.

Herresthal, Claudia, Tatiana Mayskaya, and Arina Nikandrova (2022). "Data Linkage between Markets: Does the Emergence of an Informed Insurer Cause Consumer Harm?" *CEPR Discussion Paper 17947*.

Ichihashi, Shota (2020). "Online Privacy and Information Disclosure by Consumers". *American Economic Review*.

— (2021a). "Competing data intermediaries". *The RAND Journal of Economics* 52.3, pp. 515–537.

— (2021b). "The Economics of Data Externalities". *Journal of Economic Theory* 196.

Iyer, Ganesh, David Soberman, and J. Miguel Villas-Boas (2005). "The Targeting of Advertising". *Marketing Science* 24.3.

Johnson, Garrett, Scott K. Shriver, and Samuel Goldberg (2023). "Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR". *Management Science* 69.10, pp. 5695–6415.

Johnson, Justin P. (2013). "Targeted advertising and advertising avoidance". *The RAND Journal of Economics* 44.1, pp. 128–144.

Johnson, Justin P and David P Myatt (2006). "On the simple economics of advertising, marketing, and product design". *American Economic Review* 96.3, pp. 756–784.

Kastl, Jakub, Jorge Padilla, Salvatore Piccolo, and Helder Vasconcelos (2020). "On the Private and Social Value of Consumer Data in Vertically-Integrated Platform Markets". Working Paper.

Kim, Jin-Hyuk, Liad Wagman, and Abraham L. Wickelgren (2018). "The impact of access to consumer data on the competitive effects of horizontal mergers and exclusive dealing". *Journal of Economics and Management Strategy*.

Lam, Wing Man Wynne and Xingyi Liu (2020). "Does data portability facilitate entry?" *International Journal of Industrial Organization* 69.102564.

Markovich, Sarit and Yaron Yehezkel (2021). "Data regulation: who should control our data?" *Available at SSRN 3801314*.

Montes, Rodrigo, Wilfried Sand-Zantman, and Tommaso Valletti (2018). "The Value of Personal Information in Online Markets with Endogenous Privacy". *Management Science*.

Pino, Flavio (2022). "The Microeconomics of Data – A Survey". *Journal of Industrial and Business Economics* 49, pp. 635–665.

Prüfer, Jens and Christoph Schottmüller (2021). "Competing with big data". *Jornal of Industrial Economics* 69.4, pp. 967–1008.

Roger, Guillaume (2016). "Participation in moral hazard problems". *Games and Economic Behavior* 95, pp. 10–24.

Roy, S (2000). "Strategic segmentation of a market". *International Journal of Industrial Organization* 18.8, pp. 1279–1290.

Rutt, James (2012). "Targeted Advertising and Media Market Competition". Working Paper. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2103061.

Scott Morton, Fiona, Theodore Nierenberg, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A Douglas Melamed, and Jamie Morgenstern (2019). "Report: Committee for the Study of Digital Platforms-Market Structure and Antitrust

Subcommittee". *George J. Stigler Center for the Study of the Economy and the State, The University of Chicago Booth School of Business*.

Shapiro, Carl (2006). "Prior user rights". *American Economic Review* 96.2, pp. 92–96.

Sokol, D. Daniel and Roisin E. Comerford (2016). "Antitrust and Regulating Big Data". *George Mason Law Review* 23.5, pp. 1129–1162.

Stole, Lars A (2007). "Price discrimination and competition". *Handbook of industrial organization* 3, pp. 2221–2299.

Taylor, Curtis R. (2004). "Consumer Privacy and the Market for Customer Information". *RAND Journal of Economics* 35.4, pp. 631 –650.

Thisse, Jacques-Francois and Xavier Vives (1988). "On The Strategic Choice of Spatial Price Policy". *The American Economic Review* 78.1, pp. 122–137.

Vives, Xavier (2001). *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press.

A Omitted proofs

Proof of Proposition 1. The condition $\frac{\partial D(u_i,\mathbf{u}_{-i})}{\partial u_i}\frac{1}{D(u_i,\mathbf{u}_{-i})} > -\frac{\partial^2 r(u_i,\delta_i)}{\partial u_i \partial \delta_i}\frac{1}{\partial r(u_i,\delta_i)/\partial \delta_i}$ is found by rearranging (2).

Part i: The first two terms on the right-hand side of (2) are positive: the demand for firm i is increasing in u_i, and its revenue is increasing in δ_i. The sign of $\frac{\partial^2 r_i}{\partial u_i \partial \delta_i}$ is ambiguous but when it is non-negative, we have $\frac{\partial^2 \pi_i}{\partial u_i \partial \delta_i} > 0$, i.e. data is pro-competitive.

Part ii: When $C'(u) = 0$, we have $\frac{\partial D_i}{\partial u_i}/D_i = -\frac{\partial r_i}{\partial u_i}/r_i$ by (1). We thus have

$$\frac{\partial D_i}{\partial u_i}\frac{\partial r_i}{\partial \delta_i} + \frac{\partial^2 r_i}{\partial u_i \partial \delta_i}D_i > 0 \iff -\frac{\partial r_i}{\partial u_i}\frac{\partial r_i}{\partial \delta_i} + \frac{\partial^2 r_i}{\partial u_i \partial \delta_i}r_i > 0$$

$$\iff \frac{1}{r_i^2}\left(-\frac{\partial r_i}{\partial u_i}\frac{\partial r_i}{\partial \delta_i} + \frac{\partial^2 r_i}{\partial u_i \partial \delta_i}r_i\right) > 0 \iff \frac{\partial}{\partial \delta_i}\left(\frac{\frac{\partial r_i}{\partial u_i}}{r_i}\right) > 0$$

$$\iff \frac{\partial^2 \ln(r_i)}{\partial u_i \partial \delta_i} > 0. \qquad \blacksquare$$

Proof of Proposition 2. The symmetric equilibrium, u^*, is given by the fixed point

$$u^* - \tilde{u}(u^*, \delta) = 0, \tag{5}$$

where $\tilde{u}(u, \delta) \equiv \hat{u}((u, \ldots, u), \delta)$ is the best response conditional on all rivals offering utility u. Consider an increase in δ. Totally differentiating (5), we have

$$du^*\left[1 - \frac{\partial \tilde{u}(u^*, \delta)}{\partial u}\right] - d\delta\frac{\partial \tilde{u}(u^*, \delta)}{\partial \delta} = 0 \iff \frac{du^*}{d\delta} = \frac{\frac{\partial \tilde{u}(u^*,\delta)}{\partial \delta}}{1 - \frac{\partial \tilde{u}(u^*,\delta)}{\partial u}}.$$

Now, $\frac{\partial^2 \pi_i}{\partial u_i^2} + \sum_{j \neq i}|\frac{\partial^2 \pi_i}{\partial u_i \partial u_j}| < 0$ implies that $1 - \frac{\partial \tilde{u}(u^*,\delta)}{\partial u} > 0$. The sign of $\frac{du^*}{d\delta}$ is then given by that of $\frac{\partial \tilde{u}(u^*,\delta)}{\partial \delta}$: positive if and only if data is UPC. $\blacksquare$

Proof of Proposition 3. Part (i): By definition, payoffs are strategic complements if $\frac{\partial^2 \pi_i}{\partial u_i \partial u_j} > 0$, i.e. if $\frac{\partial D_i}{\partial u_j}\frac{\partial r_i}{\partial u_i} + r_i\frac{\partial^2 D_i}{\partial u_i \partial u_j} > 0$. With linear demand, $\frac{\partial^2 D_i}{\partial u_i \partial u_j} = 0$, meaning that $\frac{\partial^2 \pi_i}{\partial u_i \partial u_j}$ has the opposite sign to $\frac{\partial r_i}{\partial u_i}$.

Part (ii): u_i^* is the solution to

$$\hat{u}_i(\hat{u}_j(u_i^*, \delta_j), \delta_i) - u_i^* = 0 \tag{6}$$

By totally differentiating this expression, we get

$$\frac{du_i^*}{d\delta_i} = \frac{\frac{\partial \hat{u}_i}{\partial \delta_i}}{1 - \frac{\partial \hat{u}_i}{\partial u_j}\frac{\partial \hat{u}_j}{\partial u_i}} \quad \text{and} \quad \frac{du_i^*}{d\delta_j} = \frac{\frac{\partial \hat{u}_i}{\partial u_j}\frac{\partial \hat{u}_j}{\partial \delta_j}}{1 - \frac{\partial \hat{u}_i}{\partial u_j}\frac{\partial \hat{u}_j}{\partial u_i}}$$

Stability implies that the common denominator in the expressions above is positive, and the result then follows from the definitions of strategic complementarity or substitutability and UPC/UAC data.

Part (iii): by part (i), conflicting payoffs yield strategic complementarity. Standard monotone comparative statics results then imply that an increase in δ_i leads to an increase in each firm's choice of u if an only if $\partial^2 \pi(u_i, \mathbf{u}_{-i}, \delta_i)/\partial u_i \partial \delta_i > 0$—i.e., if data is UPC (see Vives, 2001). ■

B Proofs and analysis for Section 5

B.1 Product improvement

B.1.1 Microfoundation for the model in Section 5.1

Here we present a Bayesian microfoundation for the model of product improvement. Consider a situation where a multiproduct firm with zero marginal cost offers to recommend an experience good (e.g., a movie) to a consumer in return for a subscription price, p. The set of available products is represented by the real line and the consumer's ideal product, which the firm cannot directly observe, is θ_0. The consumer's gross utility from consuming product x is $V - (\theta_0 - x)^2$.

To help it choose which good to recommend, the firm has a dataset, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$, about m past customers' tastes. Each $\hat{\theta}_l = \theta_l + \varepsilon_l$ is a signal about consumer l's true taste, observed with noise $\varepsilon_l \sim N(0, \sigma_\varepsilon^2)$. For example, consumer l may have rated the movies

they watched and these ratings are informative about their underlying taste. The true tastes of consumers, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_m)$, are jointly normally distributed with means zero and variances σ^2, while the covariance between any two consumers' tastes is $\chi > 0$.[29]

Because this is an experience good, the consumer can't observe $|\theta_0 - x|$ before consumption. But the firm has an incentive to develop a reputation for making the best recommendations it can because this increases consumers' expected match quality, and consumers anticipate this when determining their willingness to pay. Our general aim is to understand how an improvement in the firm's access to data affects the offer it makes to the consumer.

We can compute the firm's posterior belief about the consumer's true taste, θ_0, given its data $\hat{\boldsymbol{\theta}}$. Using standard results from probability theory, this posterior is a normal distribution, $N(\mu, \frac{1}{\delta})$, where the posterior mean and variance are calculated as

$$\mu = \frac{\chi \sum_{l=1}^{m} \hat{\theta}_l}{(m-1)\chi + \sigma^2 + \sigma_\varepsilon^2}, \qquad \frac{1}{\delta} = \sigma^2 - \frac{m\chi^2}{(m-1)\chi + \sigma^2 + \sigma_\varepsilon^2}.$$

Given this posterior, the strategy that maximizes the expected value of the product is to recommend product μ. The consumer's expected utility from subscribing is therefore $u = V - E((\theta_0 - \mu)^2) - p = V - \frac{1}{\delta} - p$, which has the form $u = V(\delta) - p$ postulated in Section 5.1. Note that a dataset's value depends on three properties, namely its size (m), relevance (χ) and the signals' accuracy (σ_ε^2), all of which are summarized by δ.

B.1.2 Model where data reduces the cost of quality

Suppose that firms simultaneously choose their price, p_i, and their quality q_i. Data reduces the cost of quality (as in Prüfer and Schottmüller, 2021): the cost is $k(q_i - \delta_i)$, increasing and convex. Let $u_i = q_i - p_i$. For a given $\mathbf{u}_{-i}$, the profit of firm i is

$$p_i D(u_i, \mathbf{u}_{-i}) - k(q_i - \delta_i) = (q_i - u_i) D(u_i, \mathbf{u}_{-i}) - k(q_i - \delta_i).$$

[29] We could easily incorporate the case where consumers are heterogeneous (with a general covariance matrix for $\boldsymbol{\theta}$) at the cost of additional notational complexity, provided we treat each consumer as a separate market. An alternative interpretation is that $\boldsymbol{\theta}$ are realizations of a single consumer's tastes at different points of time.

For a given $(u_i, \mathbf{u}_{-i})$, the optimal quality choice of i satisfies $D(u_i, \mathbf{u}_{-i}) - k'(q_i - \delta_i) = 0$, i.e. $(k')^{-1}(D(u_i, \mathbf{u}_{-i})) + \delta_i = q_i$. The profit of firm i is thus

$$\pi(u_i, \mathbf{u}_{-i}, \delta_i) = \underbrace{\left((k')^{-1}(D(u_i, \mathbf{u}_{-i})) + \delta_i - u_i\right)}_{\equiv r(u_i, \mathbf{u}_{-i}, \delta_i)} D(u_i, \mathbf{u}_{-i}) - \underbrace{k\left((k')^{-1}(D(u_i, \mathbf{u}_{-i}))\right)}_{\equiv C(u_i, \mathbf{u}_{-i})}.$$

Although r_i is now a function of $\mathbf{u}_{-i}$, we can still apply Propositions 1 and 2. In this case we have $\frac{\partial^2 r_i}{\partial u_i \partial \delta_i} = 0$, so that data is UPC.[30]

B.2 Targeted advertising

B.2.1 Proof of Proposition 5

The per-consumer revenue is $r(u_i, \delta_i) = A(u_i)P[A(u_i), \delta_i]$. In the intensive margin model we have $A'(u_i) > 0$. Applying the sufficient condition from Proposition 1, it is immediate from (3) that $\frac{\partial^2 r_i}{\partial u_i \partial \delta_i} > 0$ if and only if $P(a_i, \delta_i) + a_i \frac{\partial P(a_i, \delta_i)}{\partial a_i}$ is increasing in δ_i.

For the ad nuisance case, because $C'(u_i) = 0$, we can apply the log-supermodularity condition from Proposition 1 to yield

$$\frac{\partial^2 \ln[r(u_i, \delta_i)]}{\partial u_i \partial \delta_i} = \frac{\partial^2 \ln[A(u_i)]}{\partial u_i \partial \delta_i} + \frac{\partial^2 \ln[P(A(u_i), \delta_i)]}{\partial u_i \partial \delta_i} = \frac{\partial^2 \ln[P(A(u_i), \delta_i)]}{\partial A(u_i) \partial \delta_i} A'(u_i). \tag{7}$$

whose sign is the opposite of $\frac{\partial^2 \ln[P(A(u_i), \delta_i)]}{\partial A(u_i) \partial \delta_i}$ because $A'(u_i) < 0$. We therefore have

$$\text{data is UPC} \iff \frac{\partial^2 \ln[P(a_i, \delta_i)]}{\partial a_i \partial \delta_i} < 0 \iff \frac{\partial}{\delta_i}\left(\frac{\partial P(a_i, \delta_i)}{\partial a_i} \frac{a_i}{P(a_i, \delta_i)}\right) < 0.$$

Thus, data is UPC if it makes the inverse demand, P, more elastic with respect to a_i, or, equivalently, if it makes the direct demand less price-elastic.

B.2.2 Microfoundation for the model in Section 5.2

In this section we provide microfoundations for the demand for ad slots $P(a_i, \delta_i)$ and for the function $A(u_i)$. Regarding the latter, we discuss two alternative models that result in

[30] Note that a different specification, e.g. $k(q_i/\delta_i)$, would not lead to a cost function $C(u_i, u_{-i})$ but to some $C(u_i, u_{-i}, \delta_i)$, thus changing the unilateral analysis.

a different sign for $A'(u_i)$.

Demand for ad slots Suppose that there are an infinite number of product categories, each with a continuum of advertisers, and that each consumer is interested in a finite number K of categories. A consumer interested in a category is prepared to buy the product of advertiser l with latent probability $\theta_l \sim U[0,1]$. The platform uses its data to target the ads it sells. With probability $\lambda(\delta_i)$, each relevant category is identified as such. Conditional on observing such a "relevance" signal, each advertiser l also receives a signal s about θ_l that is equal to θ_l with probability $\mu(\delta_i)$ and is pure noise with probability $1-\mu(\delta_i)$. The functions λ and μ respectively capture how informative data is about category- and brand match, and are non-decreasing. The willingness to pay of an advertiser who receives a signal s about θ_l is $\mu(\delta_i)s + \frac{(1-\mu(\delta_i))}{2}$. Therefore, the inverse demand for advertising slots is such that

$$K\lambda(\delta_i)\Pr\left[\mu(\delta_i)s + \frac{(1-\mu(\delta_i))}{2} \geq P(a_i,\delta_i)\right] = a_i \iff P(a_i,\delta_i) = \frac{1+\mu(\delta_i)}{2} - \frac{\mu(\delta_i)}{K\lambda(\delta_i)}a_i.$$

If $\mu(\delta_i) = \lambda(\delta_i)$ then demand for ads takes the form $P(a_i,\delta_i) = \phi(a_i) + \psi(\delta_i)$: data induces a vertical shift to demand for ads.

More generally, data can lead to a rotation of the demand for ads, either clockwise (as in Johnson and Myatt (2006)) or counter-clockwise, depending on the relative slopes of λ and μ.

Intensive margin model In this application, we build a model where firms invest in the quality of their content, consumers choose how much time to spend on the website, and the quantity of ads each consumer sees increases with the time spent on the website.

Suppose that each firm invests $k(q_i)$ in the content it displays, resulting in content of quality q_i. Consumers have a fixed time budget of T, from which they choose to spend t units of time on consumption of content and $T-t$ on an outside activity. The resulting utility is $v(q_i,t) + (T-t)$, with $\frac{\partial v}{\partial q_i} > 0$, $\frac{\partial v}{\partial t} > 0$, $\frac{\partial^2 v}{\partial t^2} < 0$, and $\frac{\partial^2 v}{\partial q_i \partial t} > 0$. For each unit of time spent consuming content, the firm can show at most one ad. Ads are sold through a

uniform auction at a price $P(a_i, \delta_i)$.[31]

Let $t(q_i) \equiv \arg\max_t \ v(q_i, t) + (T - t)$. We have $t'(q_i) \geq 0$. Let $u(q_i) \equiv v(q_i, t(q_i)) + (T - t(q_i))$. Define $q(u_i)$ such that $u(q(u_i)) = u_i$. We have $q'(u_i) > 0$. The number of ads seen as a function of the utility provided is then $A(u_i) = \min\{t(q(u_i)), \arg\max_a \ aP(a, \delta_i)\}$, with $A'(u_i) \geq 0$.

The per-consumer revenue is $r(u_i, \delta_i) = A(u_i)P\left(A(u_i), \delta_i\right)$, and is non-decreasing in u_i: payoffs are congruent. Intuitively, providing more utility to consumers keeps them longer on the website, which generates more revenue from advertising.

Because there is a fixed cost of producing quality we can only provide a sufficient condition for data to be UPC, namely if $\frac{\partial^2 r(u_i, \delta_i)}{\partial u \partial \delta} \geq 0$. Using (3), this is the case if $P(a, \delta) + a\frac{\partial P(a,\delta)}{\partial a}$ is increasing in δ.

Ad nuisance model We now present a different model of targeted advertising, building on Anderson and Coate (2005). In this model, firms choose the "ad load", that is the quantity of ads they show alongside their content, and consumers view ads as a nuisance. Suppose that the mean utility is $u_i = V - \gamma a_i$, where γ is the per-ad nuisance. The ad-load associated with utility u_i is thus $A(u_i) = \frac{V - u_i}{\gamma}$, and is decreasing in u_i. The profit of firm i is $A(u_i)P\left(A(u_i), \delta_i\right) D(u_i, u_{-i})$. Using Proposition 1 (ii) and the fact that $A' \leq 0$, a necessary and sufficient condition for data to be UPC is $\frac{\partial^2 \ln(P(a,\delta))}{\partial a \partial \delta} < 0$.

B.2.3 Targeted advertising with positive prices

Here we show how our analysis of targeted advertising can be extended to the case where firms can use two instruments: a price and a quantity of ads. The utility of a consumer is $u_i = v - p_i - \gamma a_i$, while the per-consumer revenue is $p_i + a_i P(a_i, \delta_i)$. In order to compute $r(u_i, \delta_i)$, let us first solve the following problem:

$$\max_{p_i, a_i} p_i + a_i P(a_i, \delta_i) \quad \text{s.t. } v - p_i - \gamma a_i = u_i$$

[31] The firm will show ads as long as $a_i \mapsto a_i P(a_i, \delta_i)$ is increasing

Substituting p_i by $v - u_i - \gamma a_i$ into the objective, we find that the optimal number of ads is given by $P(a_i^*, \delta_i) + a_i^* \frac{\partial P}{\partial a_i} - \gamma = 0$: firm i equalizes the marginal revenue and the advertising nuisance. Indeed, in order to maintain the utility level u_i, an additional ad must be accompanied by a price decrease of γ, and the latter is thus the effective marginal cost of advertising. Firm i's per-consumer revenue is then

$$r(u_i, \delta_i) = v - u_i - \gamma a_i^* + a_i^* P(a_i^*, \delta_i)$$

We have $\frac{\partial^2 r(u_i, \delta_i)}{\partial u_i \partial \delta_i} = 0$. By Proposition 1, we conclude that data is UPC.

Note that this analysis ignores the possible non-negativity constraint on prices. Indeed, if competition is very strong, firms might want to subsidize participation by setting $p_i < 0$. If we restrict p_i and a_i to be non-negative, then, whenever the constraint $p_i \geq 0$ binds, firm i generates all its revenue through advertising and the UPC/UAC condition is that given in the main text in the case without prices.

B.2.4 A model of advertising with multihoming

Suppose that two firms are advertising supported websites, and that consumers can multi-home. Following Ambrus et al. (2016), let us assume that participation decisions are independent, i.e. that consumer l visits website i if $u_i + \epsilon_{il} \geq 0$, where ϵ_{il} is the consumer-specific taste shock.

We use the ad nuisance model from Appendix B.2.2, in which $A'(u_i) < 0$, with the difference that advertisers attach more value the first impression than to the second one. Let $X_i(\delta_i, \delta_j, n(u_j))$ be the probability that a randomly chosen ad at i is exclusive in the sense that the ad matches the consumers tastes and j does not successfully show the same consumer a matching ad. X is increasing in its first argument and decreasing in the second two. We assume that we can write $r(\mathbf{u}, \delta) = A(u_i) P_i(A(u_i), X_i)$, where P_i is the price of an ad.

We know from Proposition 1 that an increase in δ_i causes i's best response function to

shift in a direction given by the sign of

$$\frac{\partial \ln(r_i)}{\partial u_i \partial \delta_i} = \underbrace{\frac{\partial X_i}{\partial \delta_i}}_{>0} A'(u_i) \frac{\frac{\partial^2 P_i}{\partial u_i \partial X_i} P_i - \frac{\partial P_i}{\partial u_i}\frac{\partial P_i}{\partial X_i}}{P_i^2}.$$

By a similar reasoning, the effect of δ_i on j's best response function is given by the sign of

$$\frac{\partial \ln(r_j)}{\partial u_j \partial \delta_i} = \underbrace{\frac{\partial X_j}{\partial \delta_i}}_{<0} A'(u_j) \frac{\frac{\partial^2 P_j}{\partial u_j \partial X_j} P_j - \frac{\partial P_j}{\partial u_j}\frac{\partial P_j}{\partial X_j}}{P_j^2}.$$

We therefore see that i's and j's best responses shift in opposite directions. In particular, if $\frac{\partial^2 P_j}{\partial u_j \partial X_j} \geq 0$ then an increase in δ_i causes a unilaterally anti-competitive response from i and a unilaterally pro-competitive response from j.

B.3 Moral hazard: proof of Proposition 6

Suppose that a consumer picks insurer i. His expected utility if he exerts effort is

$$U(\mathcal{C}_i) = \alpha V(-p_i) + (1-\alpha)\left[\delta_i V(-p_i + X_{Hi} - L) + (1-\delta_i)V(-p_i + X_{Li} - L)\right] - k$$

His expected utility if he does not exert effort is $\tilde{U}(\mathcal{C}_i) = V(-p_i + X_{Li} - L)$.

The incentive compatibility and target utility constraints are respectively

$$\alpha V(-p) + (1-\alpha)\left[\delta V(-p + X_H - L) + (1-\delta)V(-p + X_L - L)\right] - k \geq V(-p + X_L - L) \quad (8)$$

and

$$\alpha V(-p) + (1-\alpha)\left[\delta V(-p + X_H - L) + (1-\delta)V(-p + X_L - L)\right] - k = u. \quad (9)$$

It is fairly easy to prove that (8) must bind in equilibrium: the insurer could improve upon a non-binding constraint by offering slightly more insurance in exchange for a higher

premium, until the constraint binds. Combining the two constraints therefore implies that $V(-p+X_L-L)=u$, i.e. $X_L=L+p+V^{-1}(u)$. We then substitute X_L in the objective (4) and in (9), and write the Lagrangian

$$\begin{aligned}\mathcal{L} = p-(1-\alpha)\left[\delta X_H+(1-\delta)(L+p+V^{-1}(u))\right] \\ +\lambda\left\{\alpha V(-p)+(1-\alpha)\left[\delta V(-p+X_H-L)+(1-\delta)u\right]-k-u\right\}. \quad (10)\end{aligned}$$

By combining the first-order conditions with respect to p and X_H, we obtain $V'(-p)=V'(-p-L+X_H)$, i.e. $X_H=L$: it is optimal for the insurer to fully compensate consumers when it can prove that they exerted effort. Replacing X_H by L in the constraint, we obtain $V(-p)=u+k/(\alpha+\delta-\alpha\delta)$, i.e $p(u,\delta)=-V^{-1}\left(u+k/(\alpha+\delta-\alpha\delta)\right)$. This also implies that $X_L(u,\delta)=L+p(u,\delta)+V^{-1}(u)=L-V^{-1}\left(u+k/(\alpha+\delta-\alpha\delta)\right)+V^{-1}(u)$.

We can now rewrite the per-consumer profit as a function of u:

$$\begin{aligned}r(u,\delta) &= p(u,\delta)-(1-\alpha)\left[\delta X_H+(1-\delta)X_L(u,\delta)\right] \\ &= (\alpha+\delta-\alpha\delta)\left[V^{-1}(u)-V^{-1}\left(u+\frac{k}{\alpha+\delta-\alpha\delta}\right)\right]-V^{-1}(u)-(1-\alpha)L.\end{aligned} \quad (11)$$

The cross-derivative of $r(u,\delta)$ is

$$\begin{aligned}\frac{\partial r(u,\delta)}{\partial u\partial\delta} = (1-\alpha)\Bigg[(V^{-1})'(u)-(V^{-1})'\left(u+\frac{k}{\alpha+\delta-\alpha\delta}\right) \\ +\frac{k}{\alpha+\delta-\alpha\delta}(V^{-1})''\left(u+\frac{k}{\alpha+\delta-\alpha\delta}\right)\Bigg]. \quad (12)\end{aligned}$$

This is positive whenever $(V^{-1})'$ is convex. To see this, notice that (12) is positive if

$$(V^{-1})''(u+x)>\frac{(V^{-1})'(x+u)-(V^{-1})'(u)}{(x+u)-u},$$

where $x=k/(\alpha+\delta-\alpha\delta)$.

For constant absolute risk aversion we have

$$V(W)=c-e^{-\beta W} \iff (V^{-1})'(v)=\frac{1}{(c-v)\beta} \iff (V^{-1})'''(v)=\frac{2}{(c-v)^3\beta}>0.$$

With constant relative risk aversion, the utility takes the form $V(W) = \frac{W^{1-\theta}-1}{1-\theta}$ (and the initial wealth is high enough that wealth is never negative), so that $V^{-1}(v) = ((1-\theta)v+1)^{\frac{1}{1-\theta}}$. We then have $\left(V^{-1}\right)'(v) = ((1-\theta)v+1)^{\frac{\theta}{1-\theta}}$. $\left(V^{-1}\right)'$ is convex if $\theta > 1/2$.

B.4 Price discrimination: proof of Proposition 7

Applying the log-supermodularity condition to $r(u_i, \delta_i) = \tilde{r}(u_i) + \delta_i l(u_i)$ reveals data to be UAC if and only if

$$\frac{\partial^2 \ln r(u_i, \delta_i)}{\partial u_i \partial \delta_i} = \frac{\tilde{r}(u_i)l'(u_i) - l(u_i)\tilde{r}'(u_i)}{[\tilde{r}(u_i) + \delta_i l(u_i)]^2} < 0 \iff \frac{\tilde{r}'(u_i)}{\tilde{r}(u_i)} > \frac{l'(u_i)}{l(u_i)}.$$

In other words, data is UAC if and only if the ratio of deadweight loss ($l(u_i)$) to "monopoly" profit ($\tilde{r}(u_i)$) is decreasing in u_i, i.e. increasing in the price $\rho(u_i)$. Shapiro (2006) shows that a sufficient condition for this is that the elasticity of Q be increasing in p.