NBER WORKING PAPER SERIES

DATA UNION AND REGULATION IN A DATA ECONOMY

Lin William Cong
Simon Mayer

Working Paper 30881
http://www.nber.org/papers/w30881

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2023, revised December 2023

This paper was previously circulated under the name "Antitrust, Regulation, and User Union in the Era of Digital Platforms and Big Data." We are grateful to Matthieu Bouvard (discussant), Sebastian Gryglewicz, Shota Ichishashi, Anthony Lee Zhang (discussant), Evgeny Lyandres, Michael Sockin (discussant), Leonard Treuren, and Laura Veldkamp for detailed comments. We also thank Greg Buchak, Jacelly Caspedes, Emilio Calvano, Christopher Clayton, Hans Degryse, Thierry Foucault, Denis Gromb, Andrei Hagiu, Zhigeng He, Jing Huang, Shiyang Huang, Laurence Lescourret, Dan Luo, Roxana Mihet, Aniko Oery Christine Parlour, James Poterba, Raghuram Rajan, Huan Tang, Glen Weyl, Liyan Yang, and conference and seminar participants at AI and Big Data in Finance Research Forum Webinar, University of Calgary Haskayne School of Business, ESSEC Business School, 2022 Artificial Intelligence and Financial Markets Conference in Oxford, CUHK Shenzhen, the 4th Greater China Area Finance Conference, HKUST Guangzhou, IGP Annual Conference, 2022 International Conference on FinTech and Digital Finance, KU Leuven, Peking University National School of Development Institute of Digital Finance FinTech Workshop, and the 2022 Yale Junior Finance Conference for constructive feedback. Zhiheng He and Qihong Ruan provided excellent research assistance. There is no special funding for the project. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Data Union and Regulation in a Data Economy
Lin William Cong and Simon Mayer
NBER Working Paper No. 30881
January 2023, revised December 2023
JEL No. L10,L41,L50,O30

ABSTRACT

In a model of data-driven firm competition, data are jointly produced by users, buying firms' services and contributing data, and firms, investing in data infrastructure and collection. Data collection improves services, benefiting users, but may reduce competition, harming users. Dispersed users do not internalize the impact of their data contribution on (i) service quality, (ii) competition, and (iii) firms' investment incentives, causing inefficient data over- or underinvestment. Unlike data sharing, user privacy protection policies, or data markets, a data union — which coordinates users' data contributions — or data trust— which intermediates data sales — can address these inefficiencies.

Lin William Cong
SC Johnson College of Business
Cornell University
Sage Hall
Ithaca, NY 14853
and NBER
will.cong@cornell.edu

Simon Mayer
1 Rue de la Liberation
Jouy-en-Josas, Ile- 78350
France
simonmay@andrew.cmu.edu

An online appendix is available at http://www.nber.org/data-appendix/w30881

In a digital economy, product users actively contribute to firm production through their data, while firms (platforms included) undertake costly investments to collect and process user data to improve their products. These developments have raised concerns regarding the anti-competitive effects of data and user privacy (Acquisti, Taylor, and Wagman, 2016; Farboodi, Mihet, Philippon, and Veldkamp, 2019; Acemoglu, Makhdoumi, Malekian, and Ozdaglar, 2022), and motivated recent data-related regulation, including privacy protection policies (e.g., GDPR or CCPA) or data-sharing initiatives (e.g., Open Banking). Our study aims to highlight core inefficiencies in such a data economy, to evaluate existing data-related policies to detect potential pitfalls, and to show how a *data union* can address inefficiencies and enhance user welfare. To this end, we develop a simple model of data-driven firm competition with the novel feature that data are jointly produced by users' economic activities and firms' investments in data infrastructure.

In our model, two firms sell products ("services") to a mass of identical customers ("users") over two periods under price competition. Firms potentially differ in baseline service quality and abilities in processing data to improve services. Users' payoff from buying a firm's service increases with intrinsic service quality and data-related quality improvements, but decreases as firms collect user data and cause privacy-related costs (disutility). The stock of data depends on a firm's costly infrastructure investment for data generation/collection/processing and users' buying and data contribution decisions.

In period one, firms choose their data investment, the *data price* they pay to users for sharing/contributing data, and the service price. Users then each choose at which firm to buy the service and how much data to contribute, trading off privacy costs against compensation for data contribution. Importantly, users are compensated for their data contribution and thus their privacy costs either directly through higher data prices or indirectly through lower service prices. In period two, firms process the data to improve services, and compete again for users. In both periods, service quality is subject to random shocks, giving one firm a competitive advantage upon experiencing a positive shock. Firms set prices in asymmetric Bertrand competition with one firm serving ("winning") the entire market. Prices are set simultaneously after the quality shock realizes, while investments are chosen pre-shock.

Data collection improves services, benefiting users, but induces dynamic feedback effects, potentially harming competition and users. The firm, which wins the market in period one, accumulates data and improves its service, leading to a competitive advantage in the second period. When this data advantage is sufficiently large, the winning firm becomes monopolist as its competitor exits, causing *market tipping* to the detriment of users.

Dispersed users do not take into account the broader effects of their data contribution, focusing only on their own privacy costs and the compensation they receive. Specifically, they do not internalize their impact on (i) service quality affecting all users, (ii) competition, and (iii) firms' incentives to invest in data collection. Equilibrium data collection decreases in users' privacy costs, inversely capturing their willingness to contribute data, and generally does not maximize user welfare or total surplus. The laissez-faire equilibrium therefore features inefficient data *undercollection* (underinvestment), compromising service quality, or data *overcollection* (overinvestment), causing market tipping and lack of competition. Unlike existing data-related policies, a data union, coordinating users' data contributions, can address these inefficiencies and enhance user welfare.

We study user privacy protection policies that allow users to opt out from contributing data to firms, such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). To incentivize data contribution, firms pay users a direct compensation for their privacy cost. However, since firms transfer these additional costs of data collection to users through higher service prices, these privacy protection policies have no effect on equilibrium payoffs and data collection in our setting. Moreover, a reduction in users' privacy costs (due to better technology or regulation enhancing user privacy) can increase or decrease user welfare. Indeed, lower privacy costs stimulate data collection for service improvements; however, excessively low privacy costs trigger data overcollection and market tipping.

We evaluate a data sharing policy that requires firms to share a fraction of their data with competitors. Data sharing may address data overcollection and enhance competition, but exacerbates data undercollection. Because individual firms incur the full cost of data collection while sharing the benefits with their competitors, data sharing introduces a free-rider problem, which curbs data investments. On the bright side, data sharing reduces the competitive advantage that a firm can gain from data, enhancing competition. The optimal data sharing policy, maximizing user welfare, minimizes the extent of data sharing subject to preventing market tipping. However, data sharing might not be effective at preventing market tipping, and it can be optimal to have no data sharing at all.

Given the limitations of existing data-related policies, we show how a *data union*, coordinating users' data contributions, enhances user welfare by addressing the inefficiencies in data collection. The data union effectively determines the compensation (i.e., the data price) that users receive for contributing data, influencing the costs and incentives to invest in data collection. This can be achieved by taxing (subsidizing) users' data contributions to lower (raise) their willingness to contribute data and thus to raise (lower) the compensation they require for contributing data. Interestingly, the data union and optimal data price can also be implemented in a market-based approach by introducing a novel type of data intermediary, a *data trust*, which buys data from users and sells them to firms.

The data union maximizes user welfare, but, as we show, the analysis can be generalized to other objective functions (such as total surplus). Because more data generally benefits users, the data union sets the lowest data price to enable maximum data collection subject to preventing market tipping. In particular, the data union raises (lowers) the data price relative to the laissez-faire benchmark to curb (stimulate) data collection, when data overcollection (undercollection) ensues.

Our analysis sheds light on how the optimal data price depends on firm and market characteristics and, more generally, on when regulation should stimulate or curb firms' data collection. The data price should be higher, when (i) data are highly productive in improving services, (ii) the differences in data productivity are large across competing firms, or (iii) services are highly valuable to users. When data are highly productive and can create significant competitive advantages, a high data price is necessary to limit data collection to prevent market tipping, especially when there are large differences in data productivity between firms. When services are highly valuable to users, monopoly rents are high and firms have strong incentives to strategically overinvest in data collection to cause market tipping, making it necessary to curb data collection.

We also show that the optimal data price is lower when firms (must) share data with each other, highlighting substitutionary interactions between data sharing and data union. Intuitively, data sharing limits the competitive advantage from data collection, which enhances competition and allows the data union to lower the data price without triggering market tipping. Interestingly, our results suggest that the combination of data sharing policies and a data union effectively addresses inefficiencies and enhances user welfare in a data economy. When the data price is optimally determined by the data union, more data sharing (i) introduces a free-rider problem, curbing data collection, but (ii) reduces the optimal data price, stimulating data collection. Crucially, the second effect dominates the first one. Consequently, when the data price is optimally determined by the data union, data sharing policies increase firms' data collection investments and so service quality, thereby also improving user welfare. This contrasts the outcome from the baseline model without data union, where data sharing reduced data collection and user welfare.

We extend our baseline setting to study user-centric data markets, where users own and sell their data to firms. As users do not internalize the broader consequences of their data sales (on firms' investments or competition), they sell their data to both firms, de-facto implementing data sharing among firms. Accordingly, data generated by one firm is shared with the competitor, reducing firms' incentives to collect data. Like data sharing, data markets exacerbate data undercollection, but potentially enhance competition. A data trust, which intermediates data sales from users to firms considering their impact on competition and firms' investments, improves the efficiency of data collection and user welfare.

Finally, we consider firm-centric data markets, where firms own and trade user data. If one firm sells data to its competitor, it enhances its competitor's service quality and therefore competition, which benefits users but decreases firms' joint payoff. Notably, firms optimally trade data only when trading improves their joint payoff, which happens if and only if trading reduces competition by causing market tipping. Consequently, data trades among firms and the introduction of firm-centric data markets harm users by reducing competition.

Following Jones and Tonetti (2020); Farboodi and Veldkamp (2021); Veldkamp and Chung (2022), we model data as by-products of economic activities. Akin to dynamic network effects (Biglaiser, Calvano, and Crémer, 2019; Halaburda, Jullien, and Yehezkel, 2020) or learning-by-doing (Dasgupta and Stiglitz, 1988; Cabral and Riordan, 1994), data collection causes dynamic feedback effects that affect competition. Several papers study the link between data and competition (Biglaiser et al., 2019; Calvano and Polo, 2021; Eeckhout and Veldkamp, 2021; Prüfer and Schottmüller, 2021; De Corniere and Taylor, 2020; Hagiu and Wright, 2021; Bergemann and Bonatti, 2023). Unlike existing papers, our model of data-driven firm competition endogenizes data collection as depending on both users' data contributions and firms' investments, with novel implications for data-related regulation.

We highlight how data sharing, e.g., in Open Banking (He, Huang, and Zhou, 2023; Goldstein, Huang, and Yang, 2022), or data sales (Parlour, Rajan, and Zhu, 2022) undermine firms' incentives to collect data. We show how a data union, coordinating users' data contributions, addresses inefficiencies in data collection and can enhance user welfare, especially when combined with data sharing policies. This data union admits a market-based implementation via a data intermediary (*data trust*), which, different from the ones studied in Ichihashi (2021); Bergemann, Bonatti, and Gan (2022); Liu, Ma, and Veldkamp (2023), maximizes users' payoff. Our analysis also speaks to data markets, their regulation (Bergemann, Crémer, Dinielli, Groh, Heidhues, Schafer, Schnitzer, Morton, Seim, and Sullivan, 2023), the discussion of data trusts in other academic fields (Delacroix and Lawrence, 2019; Houser and Bagby, 2022), recent policies on centralizing data ownership (such as China's national data bureau[1]), or the emergence of data DAOs (akin to a data union).[2]

Several papers focus on user privacy (Campbell, Goldfarb, and Tucker, 2015; Dosis and Sand-Zantman, 2019; Liu, Sockin, and Xiong, 2020; Choi, Jeon, and Kim, 2019; Fainmesser,

[1] See this article on Reuters.
[2] See this article.

Galeotti, and Momot, 2022). In our model, users incur (reduced form) privacy-related costs when contributing data (Ichihashi, 2020), for instance, arising from data-driven price discrimination (Bergemann, Brooks, and Morris, 2015; Brunnermeier, Lamba, and Segura-Rodriguez, 2021; Ichihashi and Smolin, 2022). We contribute by showing how users' privacy concerns (or the lack thereof) distort equilibrium outcomes, which can be corrected by introducing a data union (Arrieta-Ibarra, Goff, Jiménez-Hernández, Lanier, and Weyl, 2018).

1 Model Setup

We develop a purposefully stylized model that illustrates the key mechanisms, while admitting analytical closed-form solutions. Appendix K discusses modeling assumptions and robustness.

Firms and products. Two firms $x \in \{A, B\}$ compete in price for a unit measure of homogeneous customers ("users") over two periods, $t = 1, 2$, without discounting. We denote x's competitor by $-x$. In each period, firms produce services (i.e., products) at cost normalized to zero and sell them at price p_t^x, while a user buys at most one unit of service from either $x = A$ or $x = B$, or no unit at all. $N_t^x \in [0, 1]$ is the mass of users purchasing service from x in t. The value or quality of firm x's service is

$$Y_t^x = K^x + Z_t^x + \phi^x D^x \quad \mathbb{I}_{\{t=2\}}, \tag{1}$$

where $\mathbb{I}_{\{\cdot\}}$ is an indicator function. $K^x \geq 0$ represents the base service quality. $Z_t^x \in \{0, Z\}$ is a quality/demand/technology shock. The constant Z satisfies $Z > \max\{\Delta_K^A, \Delta_K^B\} \geq 0$ where $\Delta_K^x := K^x - K^{-x}$. For simplicity, $Z_t := (Z_t^A, Z_t^B)$ takes values $(Z, 0)$ and $(0, Z)$ with equal probability 1/2, randomizing firms' relative competitive advantage. Finally, $\phi^x D^x \mathbb{I}_{\{t=2\}}$ captures period-2 service improvements from data D^x collected in period 1, where $\phi^x \geq 0$ reflects x's ability to process/utilize/analyze data. Our formulation broadly describes different forms of data-enabled learning that improve services and products, such as within-user and across-user learning (Hagiu and Wright, 2021).[3]

Data collection. Following Jones and Tonetti (2020), data are generated as by-products of economic activities when a user buys/consumes a firm's services in $t = 1$. For simplicity, there is no more data collection in the terminal period $t = 2$. The key novelty of our setting is that the amount of data generated through firm x, that is, $\hat{D}^x := I^x \theta^x N_1^x$, depends on (i) the number of services sold to users in $t = 1$ (i.e., N_1^x), (ii) users' data contribution decisions $\theta^x \in [0, 1]$, and (iii) the firm's (publicly) observable data infrastructure investment, $I^x \in [0, \overline{I}]$ with upper bound $\overline{I} > 0$. Firm x chooses investment I^x at the onset of period 1 against a (private) cost $\frac{1}{2}\lambda(I^x)^2$ with $\lambda \geq 0$. We impose a parameter condition (presented in Appendix A) to ensure that $I^x \leq \overline{I}$ does not bind in optimum.

Intuitively, when a user buys firm x's service in $t = 1$, I^x units of data are generated/collected, and the user decides on the fraction $\theta^x \in [\underline{\theta}, 1]$ of these data that it contributes or sells to (i.e., shares with) firm x. $\underline{\theta} \in \{0, 1\}$ captures whether users or firms govern data; users can (cannot) opt out from contributing data when $\underline{\theta} = 0$ ($\underline{\theta} = 1$).[4] Any user, buying from firm x in $t = 1$, sets the same θ^x to maximize its payoff $(d^x - c^x)\theta^x I^x$ from contributing data, leading to $\theta^x = \underline{\theta}$ when $d^x < c^x$ and $\theta^x = 1$ when $d^x \geq c^x$.

Here, $d^x I^x \theta^x$ is the *direct* compensation for data contribution paid by x, with a (per unit) *data price* d^x. Next, $c^x I^x \theta^x$ is a privacy-related cost (disutility) of contributing data with privacy cost parameter $c^x (\leq \phi^x)$, inversely capturing users' willingness to contribute data.[5] Users pay an *effective price* of $P_t^x := p_t^x - \theta^x d^x I^x \mathbb{I}_{\{t=1\}}$ for x's service, which accounts for users' compensation for contributing data. As shown later, data and service prices (d^x, p_t^x) affect payoffs only via P_t^x. Hence, firms have two pricing instruments (d^x, p_t^x) for one product and effectively compete in setting P_t^x. To preclude degenerate pricing equilibria (see Appendix

[3] All (identical) users act symmetrically and can be aggregated to one representative user under the assumption that this representative user takes product prices and quality as given. The stipulation in (1) then captures that an individual user's data may improve both (i) its own service quality, e.g., due to better customization, ("within-user learning") and (ii) the service quality for others ("across-user learning"). See Hagiu and Wright (2021) for a discussion and examples of these two types of data-enabled learning.

[4] Users (partially) govern their data under privacy protection regulation (e.g., GDPR and CCPA).

[5] $c^x \theta^x I^x$ may also reflect potential positive effects of data contribution, such as improved customization; possibly, $c^x \leq 0$.

K), we assume $P_t^x \geq 0$ and firms cannot make a loss by selling to users.

Because data as a resource are *non-rival*, both firms can in principle use the same data simultaneously. Thus, firm x's amount of data D^x used for service improvements in $t = 2$ may differ from the amount of data generated by x in $t = 1$, $\hat{D}^x$. We consider that firm x shares (exogenous) fraction $\eta^x \in [0, 1]$ of its data with its competitor $-x$ (at no cost), which could be the outcome of regulation (e.g., Open Banking) or, as we show in Section 3.1, of data markets.[6] That is, $D^x = \hat{D}^x + \eta^{-x}\hat{D}^{-x}$; potentially, $\eta^x = 0$. An additional unit of data increases x's and $-x$'s product quality by ϕ^x and $\eta^x\phi^{-x}$ respectively, generating an overall advantage of $\phi_\eta^x := \phi^x - \eta^x\phi^{-x}$ for x. We assume $\phi_\eta^x > 0$ (more data benefits x).

Payoffs. A user's payoff from buying from x (while contributing fraction θ^x of data) equals $u_t^x := Y_t^x - P_t^x - c^x\theta^x I^x \mathbb{I}_{\{t=1\}}$. Firm x's payoff in $t = 2$ equals $\pi_2^x := N_2^x p_2^x$, while its (post-investment) payoff in $t = 1$ is $\pi_1^x := N_1^x p_1^x - d^x N_1^x I^x \theta^x + N_2^x p_2^x = N_1^x P_1^x + \pi_2^x$. We assume that firm $x \in \{A, B\}$ operates in period 2 if and only if $\mathbb{E}_2[\pi_2^x] > 0$ where the expectation is over Z_2 at onset of $t = 2$. When firm x exits, the remaining firm $-x$ becomes monopolist, an outcome we refer to as *market tipping*. This assumption allows us to tractably capture the tension between data and competition. Indeed, as we show, too much data collection can trigger market tipping and lack of competition. To abstract from trivial cases, we assume that in $t = 1$, both firms operate (with $\mathbb{E}_1[\pi_1^x] > 0$) and there is meaningful competition; see Appendix A for parameter conditions.

Equilibrium. We focus on a subgame perfect equilibrium in pure strategies where all users act symmetrically (taking prices and investments as given) and firms maximize their (expected) payoffs. When firm x is monopolist, users buy from x if and only if $u_t^x \geq 0$. When both firms operate, users buy from firm x that delivers higher u_t^x (subject to breaking even). To ensure that users break even, firms charge maximally $Y_t^x - c^x\theta^x I^x \mathbb{I}_{\{t=1\}}$.

The timing within period $t = 2$ is as follows. First, firms simultaneously decide whether to

[6] c^x could depend on η^x to capture that users incur larger privacy cost for sharing data with x when x shares their data with $-x$. In practice, not all data generated by x is useful or can be ported to $-x$, so $\eta^x < 1$ generally.

operate or exit. Second, the shock $Z_2 = (Z_2^A, Z_2^B)$ realizes. Third, firms choose their prices p_2^x simultaneously to maximize π_2^x (taking the choice of $-x$ as given); users make buying decisions given prices. In period $t = 1$, firms first simultaneously choose their investment I^x to maximize $\mathbb{E}_1[\pi_1^x] - \frac{\lambda(I^x)^2}{2}$. Second, the shock $Z_1 = (Z_1^A, Z_1^B)$ realizes. Third, firms x choose d^x and p_1^x to maximize π_1^x, taking the choices of the other firm as given; users make buying and data contribution decisions given prices $(p_1^x, d^x)_{x=A,B}$.

2 Solving the Equilibrium

User payoff u_t^x and buying decisions depend on (d^x, p_t^x) only via P_t^x. Therefore, firms price-compete by setting $P_t^x \geq 0$. Since users are identical, firms set prices in asymmetric Bertrand competition (provided they both operate), with one firm serving (winning) the entire market in that $N_t^x, N_t^{-x} \in \{0, 1\}$ and $N_t^x + N_t^{-x} = 1$. Determining the (unique) pricing equilibrium in asymmetric Bertrand competition is straightforward. Firm x wins the market in period t (i.e., $N_t^x = 1$), whenever the quality of its service (minus the privacy cost) is higher than its competitor's, that is, whenever $Y_t^x - c^x I^x \theta^x \mathbb{I}_{\{t=1\}} > Y_t^{-x} - c^{-x} I^{-x} \theta^{-x} \mathbb{I}_{\{t=1\}}$.[7] The losing firm $-x$ charges price $P_t^{-x} = 0$. By assumption, it cannot further lower the price to undercut x. The winning firm x charges $P_t^x = Y_t^x - c^x I^x \theta^x \mathbb{I}_{\{t=1\}} - \left[Y_t^{-x} - c^{-x} I^{-x} \theta^{-x} \mathbb{I}_{\{t=1\}}\right] \geq 0$, so users are indifferent between at x or $-x$. When firm x wins in period 1 and $N_1^x = 1$, which happens if and only if $Z_1^x = Z$, the price P_1^x and thus period-1 revenue decrease with $c^x I^x \theta^x$. That is, users are compensated for their data contribution and privacy costs either *directly* through higher data prices d^x or *indirectly* through lower service prices p_1^x.[8]

Firm x optimally incentivizes data contribution $\theta^x = 1$, i.e., x sets $d^x \geq c^x$ when $\underline{\theta} = 0$. Otherwise, when $\theta^x < 1$, firm x could profitably reduce I^x (to save the investment cost) and raise d^x and lower p_1^x in a way that P_1^x and $\hat{D}_1^x$ remain unchanged. Since p_t^x and d^x affect payoffs and equilibrium outcomes only through $P_t^x := p_t^x + \theta^x d^x I^x \mathbb{I}_{\{t=1\}}$, i.e., firms

[7]We break ties in favor of $x = A$, so $N_t^A = 1$ when $Y_t^A + Z_t^A - c^A I^A \theta^A \mathbb{I}_{\{t=1\}} = Y_t^B + Z_t^B - c^B I^B \theta^B \mathbb{I}_{\{t=1\}}$.

[8]Parameter condition (A.2) in Appendix A ensures that firm x wins and realizes positive payoff in $t = 1$ if and only if $Z_1^x = Z$.

have two pricing instruments (p_1^x, d^x) for one product, we consider in what follows without loss of generality $d^x = c^x$.[9]

When $-x$ exits in $t = 2$ and the market tips, x becomes the monopolist in $t = 2$ (so $N_2^x = 1$) and charges users the price $P_2^x = p_2^x = Y_2^x$, leaving them with zero payoff in $t = 2$. The indicator $T^x \in \{0, 1\}$ denotes this outcome and is equal to one if and only if the market tips upon $Z_1^x = Z \iff N_1^x = 1$. When $T^x = 0$, $N_2^x = 1$ if and only if $Z_2^x = Z$. The following Proposition summarizes equilibrium outcomes.

Proposition 1. *When the market tips upon $Z_1^x = Z$ and $T^x = 1$, data investment satisfies $I^x = I_T^x := \max\left\{\frac{\phi^x - c^x}{2\lambda}, \frac{Z - \Delta_K^x}{\phi_\eta^x}\right\}$. Otherwise, when $T^x = 0$, data investment satisfies $I^x = I_{NT}^x := \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda}$ where $[\cdot]^+ = \max\{\cdot, 0\}$.*

The market tips ($T^x = 1$), if and only if $G(c^x) > 0$ for

$$G(c^x) := \frac{K^A + K^B}{2} + (\phi^x - c^x) I_T^x - \lambda (I_T^x)^2 - \lambda (I_{NT}^x)^2. \tag{2}$$

A sufficient condition for $T^x = 1$ is $\frac{\phi^x - c^x}{2\lambda} \geq \frac{Z - \Delta_K^x}{\phi_\eta^x}$.

Given I^x, T^x, expected payoff for firm x (before Z_1 is realized) reads:

$$\begin{aligned}\mathbb{E}_1[\pi_1^x] = \frac{1}{2}\Bigg[&\Delta_K^x + Z - (c^x I^x - c^{-x} I^{-x}) + T^x \left(\frac{Z}{2} + K^x + \phi^x I^x\right) \\ &+ \frac{(1 - T^x)}{2}\Big(I^x \phi_\eta^x + \Delta_K^x + Z\Big)\Bigg] + \frac{(1 - T^{-x})}{4}\left[\Delta_K^x + Z - I^{-x} \phi_\eta^{-x}\right],\end{aligned}$$

User welfare $U := \mathbb{E}_1\left[\sum_{x=A,B}(N_1^x u_1^x + N_2^x u_2^x)\right]$ equals

$$\begin{aligned}U :=& (K^A + K^B)\left(1 - \frac{T^A + T^B}{4}\right) + \frac{1}{2}\left[-c^{-x} I^{-x} + \frac{(1 - T^x)}{2}\Big\{\phi^{-x} \eta^x I^x + \phi^x I^x\Big\}\right] \\ &+ \frac{1}{2}\left[-c^x I^x + \frac{(1 - T^{-x})}{2}\Big\{\phi^{-x} I^{-x} + \eta^{-x} \phi^x I^{-x}\Big\}\right].\end{aligned} \tag{3}$$

The Proposition follows by direct calculation; Appendix B provides details. Absent

[9] Indeed, p_t^x and d^x affect payoffs and equilibrium outcomes only through $P_t^x := p_t^x + \theta^x d^x I^x \mathbb{I}_{\{t=1\}}$. As only the level of P_t^x (instead of p_t^x and d^x separately) is payoff-relevant, p_t^x and d^x are not uniquely determined in equilibrium. We therefore follow the convention that $d^x = c^x$ — which is without loss of generality.

market tipping ($T^x = 0$), data investments I^x — solving first-order condition $\frac{\partial \mathbb{E}_1[\pi_1^x]}{\partial I^x} - \lambda I^x = 0$ (when $I^x > 0$) — increase with data productivity ϕ^x and decrease with privacy costs c^x. Increasing I^x, while incentivizing users to buy from firm x, requires to raise users' compensation for contributing data either directly through higher data prices d^x or indirectly through lower service prices p_1^x. Hence, the effective price P_1^x and x's period-1 profits decrease with $I^x c^x$, making data collection costly.

The market tips upon $Z_1^x = Z$ (i.e., $T^x = 1$) and x becomes monopolist in $t = 2$, when $-x$ cannot win in $t = 2$ and $N_2^{-x} = 0$ even upon $Z_2^x = Z$. Thus, $T^x = 1$ if $Z - \Delta_K^x + \phi^{-x} D^{-x} \leq \phi^x D^x$. Due to $Z_1^x = Z \iff N_1^x = 1$ and $\theta^x = 1$, we have $D^x = I^x$ and $D^{-x} = \eta^x I^x$. Consequently, market tipping $T^x = 1$ occurs (upon $Z_1^x = Z$), if x's data investment I^x exceeds $\underline{I}^x := \frac{Z - \Delta_K^x}{\phi_\eta^x}$. Because market tipping allows x to enjoy monopoly rents, firm x might strategically "overinvest" in data collection to precipitate market tipping by raising I^x above $\underline{I}^x := \frac{Z - \Delta_K^x}{\phi_\eta^x}$. The function $G(c^x)$ captures firm x's (scaled) payoff gain from doing so. Thus, the market tips and $T^x = 1$ if and only if $G(c^x) > 0$.[10]

Conditional on $T^x = 1$, $I^x = I_T^x$ solves the first-order condition $\frac{\partial \mathbb{E}[\pi_1^x]}{\partial I^x} - \lambda I^x = 0$ if and only if $\frac{\phi^x - c^x}{2\lambda} \geq \frac{Z - \Delta_K^x}{\phi_\eta^x}$ (in which case $G(c^x) > 0$). Otherwise, when $\frac{\phi^x - c^x}{2\lambda} < \frac{Z - \Delta_K^x}{\phi_\eta^x}$, overinvestment to precipitate market tipping is costly and $I_T^x = \underline{I}^x$, possibly leading to $G(c^x) < 0$. High investment cost λ, high privacy costs c^x, or low data productivity ϕ^x render data investment costly and tend to avert market tipping.

Dispersed users do not consider the broader (period-2) consequences of their data contribution, focusing only on their own privacy costs and compensation they receive. In particular, they do not internalize their impact on (i) service quality, (ii) firms' investment incentives, and (iii) competition. Thus, equilibrium data collection does not maximize user welfare, resulting in inefficient data *undercollection* (underinvestment) or *overcollection* (overinvestment). Data *undercollection* means that user welfare would increase, if c^x were lower and firms collected more data to improve services. Data *overcollection* ensues, when too much data collection $I^x \geq \underline{I}^x$ (due to excessively low c^x), causes the market to tip, harming compe-

[10] We assume that when indifferent ($G(c^x) = 0$), a firm does not precipitate market tipping.

tition in $t = 2$ and user welfare. Note that similar results and inefficiencies in data collection arise if users are myopic (not necessarily atomistic) and therefore ignore the period-2 consequences of their data contribution.

More data collection can increase or decrease user welfare. Provided $\phi^x > 2c^x$ and $T^x = 0$ (no market tipping), user welfare increases as more data is collected.[11] Under these circumstances, more data generates gains from service improvements, and users can capture part of these gains, since competition is fierce. However, too much data investment, specifically $I^x \geq \underline{I}^x$, harms user welfare by causing market tipping $T^x = 1$. Relatedly, an increase in the productivity of data ϕ^x has ambiguous effects on user welfare. When $T^x = 0$, higher ϕ^x enhances data collection investments and service quality to the benefit of users. However, increasing ϕ^x beyond a certain point causes data overcollection and market tipping (by pushing I^x above $\underline{I}^x$), reducing user welfare.[12]

We evaluate efficiency in terms of user welfare, with more efficient outcomes generating higher user welfare. Section 2.4 and Appendix H discuss efficiency in terms of total surplus, i.e., the sum of firms' payoffs and user welfare. Our key findings remain *qualitatively similar* with total surplus as "objective," with both inefficient data under- and overcollection arising in equilibrium.

2.1 Existing Data-Related Policies

We show that existing data-related policies fail to address the inefficiencies in data collection.

Privacy protection. Privacy protection policies, such as CCPA or GDPR, allow users to opt out from contributing data, which, in our model, is captured by $\underline{\theta} = 0$. However, as relevant equilibrium outcomes (e.g., payoffs and investments) do not depend on $\underline{\theta}$, such

[11] With some abuse of notation, treat I^x as "parameter," and calculate for $T^x = 0$: $\frac{\partial U}{\partial I^x} = \frac{\phi^x + \eta^x \phi^{-x} - 2c^x}{4} > 0$ when $\phi^x > 2c^x$, i.e., $I^x > 0$.

[12] For $T^x = 0$ and $I^x > 0$:

$$\frac{\partial U}{\partial \phi^x} = \frac{1}{4}\left(I^x + (\phi^x + \phi^{-x}\eta^x)\frac{\partial I^x}{\partial \phi^x}\right) - \frac{c^x}{2}\frac{\partial I^x}{\partial \phi^x}.$$

$\frac{\partial U}{\partial \phi^x} > 0$ when $\phi^x > 2c^x$, i.e., $I^x = \frac{[\phi^x_\eta - 2c^x]^+}{4\lambda} > 0$.

privacy regulation has no effects (on user welfare). The reason is that firms have two pricing instruments (p_1^x, d^x) for one product, while only P_1^x matters. Thus, restricting one price (via $\underline{\theta}$) does not affect equilibrium outcomes. For some intuition, suppose that $\underline{\theta} = 1$, and firms set $d^x = 0$ and compete via p_1^x; users are then indirectly compensated for their data contribution through lower service prices. Under these circumstances, more data collection, i.e., higher I^x, causes more privacy-related disutility to users and so requires a lower service price p_1^x to entice users to buy from x. When $\underline{\theta}$ becomes zero and users can opt out from contributing data, firms raise d^x above c^x to incentivize data contribution $\theta^x = 1$ and increase p_1^x by $d^x I^x$, leaving P_1^x and payoffs unchanged. Intuitively, firms pass the additional cost of data collection caused by privacy protection policies on to users via higher service prices, leaving the equilibrium outcomes unchanged.

Interestingly, a reduction in c^x (due to privacy protection regulation or investments (Fainmesser et al., 2022)) may increase or decrease user welfare. When $T^x = 0$, a decrease in c^x increases data investment and contributions I^x, boosting service quality and user welfare. However, excessively low privacy costs c^x imply data overcollection (i.e., $I^x \geq \underline{I}^x$), triggering market tipping to the detriment of users.[13]

Data sharing. Data sharing, such as in Open Banking or Finance initiatives, allows firms to access the data of their competitors; Section 3.1 shows that data sharing can be the outcome of data markets, with users owning and selling data to firms. Overall, data sharing may address data overcollection, but exacerbates data undercollection.

Data sharing $\eta^x > 0$ introduces a free-rider problem regarding data collection, curbing firms' incentives to collect data. While individual firms incur its full cost, the benefits of data collection are shared with competitor firms and eventually also users, since data sharing enhances price competition in $t = 2$. When $T^x = 0$, an increase in η^x reduces data

[13] When $T^x = 0$ and $I^x > 0$, calculate $\frac{\partial U}{\partial c^x} = \frac{\phi^x + \eta^x \phi^{-x} - 2c^x}{4} \frac{\partial I^x}{\partial c^x} - \frac{I^x}{2} < 0$ for $I^x > 0$. When c^x is sufficiently low, then $\frac{\phi^x - c^x}{2\lambda} \geq \frac{Z - \Delta_K^x}{\phi_\eta^x}$, which implies $T^x = 1$.

investments I^x and user welfare due to this free-rider problem.[14] However, data sharing may benefit users by enhancing competition and preventing market tipping. Indeed, the market tips upon $N_1^x = 1$ and $T^x = 1$ when $\phi_\eta^x I^x \geq Z - \Delta_K^x$, which is harder to satisfy when ϕ_η^x is small and η^x is large.

The optimal data sharing policy $\eta_*^x := \arg\max_{\eta \in [0,\overline{\eta}]} U$, maximizing user welfare (subject to an exogenous upper bound $\eta^x \leq \overline{\eta} \leq 1$), therefore minimizes the extent of data sharing subject to preventing market tipping. That is, $\eta_*^x = \inf\{\eta^x \in [0, \overline{\eta}] : T^x = 0\}$.[15] However, data sharing might not be effective at preventing market tipping (i.e., $T^x = 1$ for all $\eta^x \in [0, \overline{\eta}]$), and it can be optimal to have no data sharing at all (i.e., $\eta_*^x = 0$), highlighting the limits of data sharing policies.

2.2 Data Union and Data Trust

We introduce a *data union* which, given any η^x, enhances user welfare by addressing the inefficiencies in data collection. Intuitively, the data union coordinates users' data contribution decisions θ^x, determining users' compensation for data contribution and firms' data investments. This can be achieved by taxing (subsidizing) users' data contributions to decrease (increase) their willigness to share data and thus to increase (decrease) the compensation they require for contributing data. Later, we discuss an alternative implementation of the data union, which is market-based and involves a data intermediary.

The data union requires users to pay a tax $\tau^x \gtrless 0$ to the union per unit of data they contribute to firm x, amounting to total tax revenue $\mathcal{T} := \sum_{x=A,B} \tau^x \theta^x I^x N_1^x$; data contributions are subsidized when $\tau^x < 0$. For now, we consider that the data union is *budget neutral*, which could be relaxed. That is, taxes/subsidies $\mathcal{T}$ are rebated/financed to/from

[14] Take $T^x = 0$ and $I^x = \frac{\phi_\eta^x - 2c^x}{4\lambda} > 0$, and calculate using (3) and $\frac{\partial I^x}{\partial \eta^x} = \frac{-\phi^{-x}}{4\lambda}$:

$$\frac{\partial U}{\partial \eta^x} = \frac{1}{4}\left(\phi^{-x} I^x - (\phi^x + \eta^x \phi^{-x})\frac{\phi^{-x}}{4\lambda}\right) + \frac{c^x \phi^{-x}}{8\lambda}.$$

For $\phi^{-x}, \lambda, \eta^x > 0$, $\frac{\partial U}{\partial \eta^x}$ has the same sign as $2c^x + \phi_\eta^x - 2c^x - (\phi^x + \phi^{-x}\eta^x) \leq \phi_\eta^x - \phi^x < 0$, so $\frac{\partial U}{\partial \eta^x} < 0$.

[15] Provided there exists feasible data sharing rule η^x preventing market tipping, η_*^x is such that $G(c^x) = 0$. The upper bound $\overline{\eta} < 1$ captures in reduced form the costs/limits of data sharing as well as that data generated by x might be of limited use for the other firm $-x$.

users equally through lump-sum transfers, whereby any user receives/pays the same amount $\mathcal{T}$ regardless of it choice of θ^x. These lump-sum transfers do not affect users' data contribution or buying decisions, and their exact timing is immaterial.[16]

We assume a privacy cost of $c < \frac{\phi^x}{2}$ absent taxation.[17] When $\underline{\theta} = 0$, a user, buying from x, sets $\theta^x = 1$ if and only if $d^x \geq c + \tau^x$. With a slight abuse of notation, we define $c^x = c + \tau^x$ and note that, given c^x, Proposition 1 applies. The tax τ^x determines users' willingness to contribute data (i.e., c^x) and therefore their (required) compensation for contributing data, thereby influencing firms' costs and incentives to collect data. Thus, the data union controls how much user data $I^x N_1^x$ is collected, but does not affect the choice of θ^x — which is one whenever $I^x > 0$. Without any loss, we consider that the data price d^x exactly compensates users for their costs of contributing data, i.e., $d^x = c^x = c + \tau^x$, noting that compensation for data contribution could also take the form of a lower service price p_1^x rather than higher data price. The data union chooses tax τ^x or equivalently data price $d^x = c^x = c + \tau^x$ to maximize sum of (expected) user welfare and tax proceeds:

$$U^* := \max_{\tau^A, \tau^B \in [\underline{\tau}, \overline{\tau}]} \mathbb{E}\left[U + \sum_{x=A,B} \tau^x \theta^x I^x N_1^x \right]. \tag{4}$$

As the parameter conditions in Appendix A require c^x to be bounded, we impose $\tau^x \in [\underline{\tau}, \overline{\tau}]$. Without further specifying the bounds $\underline{\tau} < \overline{\tau}$, we focus on the case $\tau^x \in (\underline{\tau}, \overline{\tau})$, i.e., parameters are such that this regularity constraint does not bind. Below Proposition characterizes the solution to (4), yielding optimal compensation for data contribution $c_*^x = c + \tau^x$. In what follows, we refer to c_*^x as the *optimal data price.*

Proposition 2. *Suppose that $\tau^x \in (\underline{\tau}, \overline{\tau})$ in optimum. The (optimal) data price c_*^x $(= d^x)$*

[16] If $\tau^x > 0$, the union could pay out $\mathcal{T}$ to every user at the end of $t = 1$. If $\tau^x < 0$, the union could raise $\mathcal{T}$ from every user at the onset of $t = 1$. The exact timing of the lump-sum transfer $\mathcal{T}$ is irrelevant, so we do not specify it further.

[17] If $c > \frac{\phi^x}{2}$, user welfare is maximized for $I^x = 0$, which, trivially, is achieved by prohibitively high tax.

is the unique root of $G(c^x)$ for $c^x \geq \frac{\phi^x_\eta}{2} - \frac{2\lambda(Z-\Delta^x_K)}{\phi^x_\eta}$. When $c^x_* \leq \frac{\phi^x_\eta}{2}$:

$$c^x_* = \frac{\phi^x_\eta}{2} - \frac{2\lambda(Z-\Delta^x_K)}{\phi^x_\eta} + \left[2\lambda\left(\frac{2(Z-\Delta^x_K)\phi^x - (Z-\Delta^x_K)\phi^x_\eta + (K^A+K^B)\phi^x_\eta}{\phi^x_\eta}\right)\right]^{\frac{1}{2}}. \quad (5)$$

When $c^x_* > \frac{\phi^x_\eta}{2}$:

$$c^x_* = \phi^x - \frac{\lambda(Z-\Delta^x_K)}{\phi^x_\eta} + \frac{\phi^x_\eta(K^A+K^B)}{2(Z-\Delta^x_K)}.$$

The market does not tip ($T^x = 0$) and data investment is $I^x = \frac{[\phi^x_\eta - 2c^x_*]^+}{4\lambda}$.

We sketch heuristically the arguments leading to Proposition 2, with more formal arguments deferred to Appendix C.

The data union resolves the trade-off between improving services through stimulating data collection and preventing market tipping through limiting data collection. Because more data benefits users and user welfare increases with data collection (absent market tipping), the data union implements the lowest data price c^x and maximum level of data investment $I^x = \frac{[\phi^x - 2c^x]^+}{4\lambda}$ subject to preventing market tipping. The optimal data price $c^x = c^x_*$ puts firm x indifferent between strategically triggering market tipping (by raising I^x above $\underline{I}^x$) and not doing so. Thus, $G(c^x_*) = 0$ holds, which we can solve for closed-form expressions for c^x_*. Depending on whether the laissez-faire equilibrium features data undercollection or overcollection (market tipping), the data union sets $c^x_* \gtrless c$ and $\tau^x \gtrless 0$. When $c^x_* > c$ ($c^x_* < c$), the data union taxes (subsidizes) users' data contributions, increasing (decreasing) the data price and curbing (boosting) data collection.

By construction, data union dominates any data sharing policy, as it improves user welfare given any data sharing rule η^x. The next Section 2.3 highlights the interaction between data sharing policies and data union by calculating comparative statics of c^x_* with respect to η^x.

Market-Based Implementation of Data Union and Data Trust. In the proposed implementation of the data union with taxes/subsidies, users manage their data and sell their data at price $d^x = c^x_*$. Firm x pays $d^x = c^x_* = c + \tau^x$ to users who pay taxes (receive subsidies) to (from) the data union. The optimal data price can also be implemented in a market-based approach by introducing a *data trust* as intermediary. To see this, consider a user contributing data to firm x. Firm x pays $d^x I^x = (c + \tau^x)I^x$ to this user who then pays $\tau^x I^x$ to the data union (or receives $-\tau^x I^x$ if $\tau^x < 0$). Without any loss, we can rearrange the chain of transfers such that firm x pays $d^x I^x$ to the data union; the data union pays cI^x to the user and keeps the remainder. Hence, the data union acts as an intermediary which buys data from users at price c and sells them to firm x at price $d^x = c + \tau^x$, with $d^x = c^x$ maximizing (4).[18] To prevent data overcollection, data trust might set $c^x_* > c$, realizing profits which are redistributed to users. When the data trust stimulates data collection via $c^x_* < c$, it realizes a loss which is financed by users. The interpretation is that the data trust acts as data intermediary which intermediates data contributions from users to firms.[19]

Non-discriminatory Pricing. When data union cannot apply discriminatory taxation or pricing of data — that is, when (τ^A, c^A) and (τ^B, c^B) must be equal — it must set $\tau := \max\{\tau^A, \tau^B\}$ or $c_* := \max\{c^A_*, c^B_*\}$ to prevent market tipping. Then, under reasonable parameter configurations, $\max\{c^A_*, c^B_*\}$ becomes the optimal data price. In this case, the optimal data price also reflects the trade-off between stimulating data collection for service improvements and curbing data collection for enhancing competition. See Appendix I for additional details.

Union Membership. While we characterize the effects of the data union and optimal data price, we do not model users' decision to join the data union, assuming it represents all users. Appendix J discusses the incentives to join the union. As discussed in greater detail in Appendix J, an individual user is better off joining the union, when the data union taxes data contributions and thus pays out tax proceeds to its "members." However,

[18] We can reverse above logic too, starting out with a data trust. Consider a user contributing data to x. Then, firm x pays $d^x I^x = (c + \tau^x)I^x$ dollars to the data trust, which pays cI^x dollars to the user and keeps $\tau^x I^x$. Without changing outcomes, we can rearrange this chain of transfers such that the firm pays $d^x I^x$ to the user; the user keeps cI^x and pays the data trust $\tau^x I^x$, akin to the tax-based implementation.

[19] Here, the data trust sells data only to the firm that has generated the data: When $N^x_1 = 1$, data trust only sells data to x but not to $-x$, while x shares η^x of the data with $-x$ at zero cost. We relax this assumption in Section 3.1, studying data markets.

when the data union subsidizes data contributions and finances these subsidies from users joining the union, an individual user could be better off not joining the union to avoid financing the subsidy (while free-riding on the data union). While our analysis highlights the effects of the data union and characterizes the optimal data price (abstracting from incentive compatibility frictions), we leave the further study of the exact union design and its incentive compatibility for future research. We note, however, that this incentive problem could be resolved by (i) requiring users to be represented by data union or (ii) having a third party (e.g., the government) cover the subsidy of data collection, in that the data union does not have to be budget neutral. Indeed, in practice, initiatives that share similarities with a data union could be government-led, as, for instance, is the case for China's national bureau of data.

2.3 Data Union: Analysis and Comparative Statics

We now study how the optimal data price depends on market and firm characteristics. More broadly, our results of this Section highlight when data-related regulation should curb data collection (i.e., $c_*^x > c$) or stimulate data collection (i.e., $c_*^x < c$). Corollary 1 (proven in Appendix D) presents comparative statics of c_*^x.

Corollary 1. *When c_*^x satisfies (5), the following holds:*

1. *Holding Δ_K^x fixed, c_*^x increases in K^A and K^B.*
2. *c_*^x is hump-shaped in λ, i.e., increases (decreases) for low (high) values of λ.*
3. *When ϕ^x is not too large, in that $(\phi^x)^2 \leq 2\lambda\big[(Z - \Delta_K^x) + K^A + K^B\big]$, then $\frac{\partial c_*^x}{\partial \phi^x} > 0$, $\frac{\partial c_*^x}{\partial \eta^x} < 0$, $\frac{\partial c_*^x}{\partial Z} < 0$, and $\frac{\partial c_*^x}{\partial \phi^{-x}} \leq 0$ (with the inequality being strict for $\eta^x > 0$).*

When services are highly valuable (high K^x), monopoly rents are large, which incentivizes firms to strategically precipitate market tipping and requires a high data price to prevent this outcome. Likewise, when data are highly productive (large ϕ^x), a high data price is needed to prevent market tipping, especially when there are large differences in data productivity between firms. Thus, c_*^x tends to increase in ϕ^x and decrease in ϕ^{-x}. More in detail, an increase in ϕ^x increases x's competitive advantage from data collection, necessitating a higher data price to limit x's data collection to prevent market tipping. Overall, data price should be higher when the data union's main concern is lack of competition (data overcollection) and lower when the main concern is data undercollection. The comparative statics of c_*^x with respect to λ reflect this trade-off, since low (high) λ causes data overcollection (undercollection), making c_*^x hump-shaped in λ.

Finally, the data price c_*^x decreases with the extent of data sharing as captured by η^x. More data sharing (higher η^x) reduces the competitive advantage that a firm can gain from data collection and therefore also the incentives to collect data. As a result, more data sharing (higher η^x) enhances competition and allows the data union to lower the data price to stimulate data collection while preventing market tipping. Hence, a high price for data and data sharing are substitutes in preserving competition, so c_*^x decreases with η^x.

Can the combination of data sharing policies and a data union enhance user welfare? What is the optimal data sharing policy in the presence of a data union? The following Corollary (proven in Appendix E) addresses these questions by showing that data investments and user welfare increase with the extent of data sharing η^x, when the data price is optimally determined by the data union (i.e., $c^x = c_*^x$).

Corollary 2. *Suppose ϕ^x is not too large, i.e., $(\phi^x)^2 \leq 2\lambda\big[(Z - \Delta_K^x) + K^A + K^B\big]$, $\phi^x > 2c$, and $\tau^x \in (\underline{\tau}, \overline{\tau})$ under the optimal data union policy. Then, under the optimal data price determined by the data union (i.e., $c^x = c_*^x$), data investments I^x and user welfare U^* from (4) increase with η^x. That is, $\frac{\partial Y}{\partial \eta^x} \geq 0$ for $Y \in \{I^x, U^*\}$ when $c^x = c_*^x$, where the inequality is strict if and only if $I^x > 0$.*

Recall that absent data union intervention (i.e., for constant c^x), more data sharing (higher η^x) reduces data collection investment I^x and user welfare (when $T^x = 0$). Notably, the opposite result prevails when the data price is optimally chosen by the data union, in that $c^x = c_*^x$. Under these circumstances, more data sharing stimulates data collection and enhances user welfare.

When the data price is optimally determined by the data union, more data sharing (higher η^x) has two effects. First, as in the baseline, data sharing reduces firms' data investments by introducing a free-rider problem regarding data collection. Second, data sharing tends to enhance competition by preventing market tipping in $t = 2$. Consequently, more data sharing allows the data union to reduce the data price c^x_* while preserving competition, which in turn increases firms' data investments. Interestingly, the decrease in data price due to data sharing outweighs the associated free-rider problem, so that data investment I^x increases with η^x under the optimal data price. Because, in addition to boosting data investments I^x, data sharing $\eta^x > 0$ also improves services for a given level of investment I^x, we obtain that more data sharing enhances user welfare U^* under the optimal data price.

Interpreted broadly, our results indicate that the combination of data sharing policies and data union can effectively address inefficiencies in data collection and enhance user welfare in a data economy. Provided that the data price is optimally determined by the data union and the conditions of Corollary 2 are met, the optimal data sharing policy η^x implements maximum data sharing (subject to constraints).[20] In contrast, in our baseline model absent data union, the optimal data sharing policy minimizes the extent of data sharing subject to preventing market tipping.

Overall, our findings suggest that data sharing policies are in general beneficial only when combined with a data union. The intuition is that the data union can alleviate the free-rider problem in data collection arising from data sharing by appropriately subsidizing data contributions.

[20] For regularity purposes and to ensure that the conditions of Corollary 2 are met, we impose the constraint $\eta^x \in [0, \overline{\eta}]$ for $\overline{\eta} < 1$. This constraint captures in reduced technological/practical limits and costs of data sharing or that data generated by one firm x is less useful for $-x$. The optimal data sharing policy with data union is then $\eta^x = \overline{\eta}$.

2.4 Data Union: Objective Function and Total Surplus

We assumed the data union maximizes user welfare. We now (heuristically) analyze the case that (given any η^x) the data union maximizes *total surplus* and show that our key findings remain qualitatively similar. Details are provided in Appendix H. As a first step, we calculate total surplus — i.e., sum of user welfare and firms' payoffs — as

$$\mathcal{S} := \sum_{x=A,B} \left[\mathcal{S}^x - \frac{\lambda (I^x)^2}{2} \right],$$

with

$$\mathcal{S}^x := \frac{1}{2}\Bigg[K^x - c^x I^x + Z + T^x \left(K^x + \phi^x I^x + \frac{Z}{2} \right) \\ + \frac{(1 - T^x)}{2} \left(K^x + K^{-x} + 2Z + \phi^x I^x + \phi^{-x} \eta^x I^x \right) \Bigg]$$

and investments I^x characterized in Proposition 1. Total surplus is simply the expected sum of service quality minus the privacy costs and investment costs; it does not directly depend on service and data prices which are merely transfers.

As we show and formalize in Appendix H, market tipping is inefficient in terms of total surplus when the sufficient parameter condition $\frac{Z-\Delta_K^x}{\phi^x - c^x} \geq \frac{\phi^x - c^x}{2\lambda}$ holds.[21] In this case, our key findings remain qualitatively similar to the baseline, featuring inefficient data overcollection (market tipping) or undercollection in laissez-faire equilibrium. Data overcollection ensues when too much data collection (i.e., $I^x \geq \underline{I}^x$) harms total surplus by precipitating market tipping. Otherwise, provided there is no market tipping, data undercollection ensues and investment I^x lies below the efficient surplus-maximizing investment level $\mathcal{I}^x = \mathcal{I}^x_{NT} := \frac{\phi^x + \eta^x \phi^{-x} - 2c^x}{4\lambda} \geq I^x$. Individual firms underinvest in data collection, because they do not take into account that their data improves the other firm's services too (due to data sharing or

[21] Market tipping $T^x = 1$ can be inefficient in terms of total surplus, as it precludes that the firm with a positive productivity shock serves the market in $t = 2$. When $-x$ exits upon $Z_1^x = Z$ ($T^x = 1$), x serves the market in $t = 2$ regardless of the realization Z_t^x, while it could be efficient in terms of total surplus to have $-x$ serve the market in case $Z_t^{-x} = Z > 0 = Z_t^x$. Absent market tipping, it is precisely the firm with the positive shock that serves the market, i.e., $N_2^x = 1 \iff Z_2^x = Z$.

technological spillovers captured by $\eta^x > 0$).

Appendix H then also analyzes a data union that maximizes total surplus, when the privacy cost (absent data union) is c. The data union implements a tax τ^x to set the price of data $c^x = c + \tau^x$. The data union maximizes $\mathcal{S} + \mathbb{E}[\sum_{x=A,B} \tau^x I^x N_1^x]$, i.e., the sum of surplus (given c^x) and tax revenues which are redistributed to users. As we show, the optimal data price $c^x = c_S^x$ satisfies (provided $\tau^x \in (\underline{\tau}, \overline{\tau})$):

$$c_S^x = \max\{c_*^x, c - \eta^x \phi^{-x}\},$$

with c_*^x from Proposition 2. When $c_S^x = c_*^x$, the data union, maximizing total surplus, implements the lowest data price and maximum data collection subject to preventing market tipping. When $c_S^x = c - \eta^x \phi^{-x}$, the data union sets the data price to implement $I^x = \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda} = \frac{\phi^x + \eta^x \phi^{-x} - 2c}{4\lambda}$ — where $\frac{\phi^x + \eta^x \phi^{-x} - 2c}{4\lambda}$ is the efficient surplus-maximizing investment when $T^x = 0$ and the privacy cost is c.

Provided $\eta^x > 0$, the data price c_S^x ($= c^x$) lies above (below) c and $\tau^x > 0$ ($\tau^x < 0$) if and only if $c_*^x > c$ ($c_*^x < c$). The data union addresses the inefficiencies in data collection by taxing (subsidizing) data contributions to curb (stimulate) I^x when $c_*^x > c$ ($c_*^x < c$). By construction, the data union, maximizing total surplus given η^x, achieves higher total surplus than data sharing policies alone or user privacy protection policies (as quantified by $\underline{\theta}$). We conclude that our findings remain qualitatively similar relative to the baseline (with user welfare as the objective), when evaluating efficiency in terms of total surplus and considering a surplus-maximizing data union.

3 Data Markets

We study markets for data in two forms: (i) users own and sell data ("user-centric data markets") or (ii) firms own and sell data ("firm-centric data markets").

3.1 User-Centric Data Markets

We extend our baseline to study user-centric data markets. In user-centric data markets, users can contribute/sell data (generated in $t = 1$) to both firms, unlike in the baseline where users only contribute data to the firm whose service they bought in $t = 1$. Akin to such data markets, Open Banking initiatives allow users to share their banking data with many banks and firms (e.g., FinTechs).

The data market, allowing users to sell data to the firm that sells the service in $t = 1$, is already modeled in the baseline when $\underline{\theta} = 0$. Buying the service from x, users sell fraction $\theta^x \in [0, 1]$ of their data to x at price $d^x = c^x$.

Next, we introduce a market that allows users to sell data to the firm at which they did *not* buy the service in $t = 1$. Suppose fraction $\tilde{\eta}^x := \tilde{\theta}^x \eta^x$ of the data generated by firm x in $t = 1$ (i.e., $\hat{D}^x$) is shared with $-x$ at the end of $t = 1$, with exogenous $\eta^x \in [0, 1]$ and users' choice $\tilde{\theta}^x \in [\underline{\tilde{\theta}}, 1]$ for $\underline{\tilde{\theta}} \in \{0, 1\}$.[22] Formally, $D^x = \hat{D}^x + \tilde{\eta}^{-x}\hat{D}^{-x}$, where $\hat{D}^x = \theta^x I^x N_1^x$ (and optimally $\theta^x = 1$). When $\underline{\tilde{\theta}} = 1$, $-x$ obtains data from x without having to compensate users, yielding the baseline. Otherwise, when $\underline{\tilde{\theta}} = 0$, users sell their data to $-x$ at endogenous (per unit) price $\tilde{d}^{-x}$, while incurring (per unit) disutility of c^{-x}. Therefore, users set $\tilde{\theta}^x = 1$ if and only if $\tilde{d}^{-x} \geq c^{-x}$. As it would be suboptimal for $-x$ to pay more than c^{-x}, data price satisfies $\tilde{d}^{-x} = d^{-x} = c^{-x}$. Users do not internalize the broader effects of their data sales to either firm, focusing only on their own privacy costs and compensation.

When $\underline{\tilde{\theta}} = 0$, data is traded at the end of $t = 1$ (or alternatively at the beginning of period 2). Suppose $N_1^x = \hat{D}^x = 0 < \hat{D}^{-x}$ and $N_1^{-x} = 1$, i.e., firm $-x$ wins the market in $t = 1$. Then, firm x decides how much data $D^x - \hat{D}^x = D^x$ to acquire data from users at price $\tilde{d}^x = c^x$. At the end of $t = 1$, x solves $\max_{D^x \in [\hat{D}^x, \hat{D}^x + \eta^{-x}\hat{D}^{-x}]} \mathbb{E}_2[\pi_2^x] - c^x[D^x - \hat{D}^x]$.

If $\max_{D^x \in [\hat{D}^x, \hat{D}^x + \eta^{-x}\hat{D}^{-x}]} \mathbb{E}[\pi_2^x] - c^x[D^x - \hat{D}^x] \leq 0$, firm x exits, causing market tipping $T^{-x} = 1$. Conditional on operating, $\frac{\partial \mathbb{E}[\pi_2^x]}{\partial D^x} = \frac{\phi^x}{2}$.[23] Thus, firm x buys all available data when $\phi^x \geq 2c^x$, while buying no data otherwise. There is no mandatory data sharing; everything

[22] Generally, $\eta^x < 1$ capturing that (i) not all data can be ported from one firm to another or (ii) not all data of firm x is useful to $-x$.

[23] When both firms compete, then $\mathbb{E}_2[\pi_2^x] = \frac{\Delta_K^x + \phi^x D^x - \phi^{-x} D^{-x} + Z}{2}$.

else remains as in the baseline. Proposition 3 (proven in Appendix F) summarizes some equilibrium outcomes.

Proposition 3. *Define* $\phi^x_\eta = \phi^x - \eta^x \phi^{-x} \mathbb{I}_{\{\phi^{-x} \geq 2c^{-x}\}}$ *and consider* $\underline{\tilde{\theta}} = \underline{\theta} = 0$*. When* $T^x = 1$*, data investment satisfies* $I^x = I^x_T := \max\left\{\frac{\phi^x - c^x}{2\lambda}, \frac{Z - \Delta^x_K}{\phi^x - \eta^x[\phi^{-x} - 2c^{-x}]^+}\right\}$*. When* $T^x = 0$*, data investment satisfies* $I^x = I^x_{NT} := \frac{[\phi^x_\eta - 2c^x]^+}{4\lambda}$*. The market tips in* $t = 2$ *upon* $Z^x_1 = Z$ *and* $T^x = 1$*, if and only if* $G(c^x) > 0$ *with* $G(c^x)$ *from* (2)*. A sufficient condition for market tipping is* $\frac{\phi^x - c^x}{2\lambda} \geq \frac{(Z - \Delta^x_K)}{\phi^x - \eta^x[\phi^{-x} - 2c^{-x}]^+}$*.*

User-centric data markets lead to similar outcomes in terms of investment and market tipping as the baseline with (mandatory) data sharing. Because users do not internalize the effects of their data sales on firms' investment incentives and competition, they sell their data to all firms, leading to $D^x = \hat{D}^x + \eta^{-x}\hat{D}^{-x}$ akin to data sharing.

When x wins the market in $t = 1$, the market tips in $t = 2$ (i.e., $T^x = 1$) whenever $-x$ cannot win in $t = 2$ even upon $Z^x_2 = Z$. That is, $T^x = 1$ if and only if $I^x \geq \underline{I}^x := \frac{Z - \Delta^x_K}{\phi^x - \eta[\phi^{-x} - 2c^{-x}]^+}$. Different to the baseline, the firm, losing period 1, now buys data from the competing firms' users instead of receiving it for free. This makes operating in period 2 more costly and therefore market tipping more likely, relative to the baseline. Formally, $\underline{I}^x$ is larger in the baseline (where we had $\underline{I}^x = \frac{Z - \Delta^x_K}{\phi^x - \eta\phi^{-x}}$) than in the extension with user-centric data markets.

Because user-centric data markets are similar to our baseline with mandatory data sharing, our main conclusions regarding the optimal data union and data price remain valid too. We stipulate $c^x = c + \tau^x$, where $c < 2\phi^x$ is the intrinsic privacy cost and $\tau^x \gtreqless 0$ a tax chosen by the data union according to (4). The data union effectively determines the data price $c^x_* = d^x = \tilde{d}^x$ and thus firms' data collection investments. See Appendix F.4 for details and an expression for c^x_*. The optimal data price may lie above or below the laissez-faire price, depending on whether the data union needs to address data overcollection or undercollection.

In the tax-based implementation, firm x pays any user the data price $d^x = c + \tau^x$ per unit of its data, while this user pays τ^x (per unit) to the data union; the tax proceeds/subsidies are then rebated/financed to/by all users equally. To obtain a market-based implementation, we can rearrange these transfers and have the data union act as intermediary (data trust). When x buys one unit of a specific user's data, then x pays (per unit) data price d^x to the data union, which pays c to the user and captures τ^x. In other words, the data trust buys data from users at price c and sells them to firm x at price d^x (and to firm $-x$ at price d^{-x}). The data trust intermediates data sales from users to firms taking into account the impact of data sales on competition and firms' investment incentives, thereby improving user welfare.

Finally, if the data union cannot price-discriminate in that $c^A = c^B = d$ must hold, it must set a data price of $\max\{c_*^A, c_*^B\}$ to prevent market tipping.

3.2 Firm-Centric Data Markets

We evaluate firm-centric data markets, where firms own and trade user data at the beginning of period $t = 2$. The key finding is that data trades among firms and therefore the introduction of firm-centric data markets cause market tipping and lack of competition, harming users. Intuitively, if one firm sells data to its competitor, it enhances its competitor's service quality and therefore competition, which benefits users but decreases firms' joint payoff. Firms optimally trade data only when trading improves their joint payoff, which happens if and only if trading reduces competition by causing market tipping in period $t = 2$.

Formally, we consider that at the end of period 1, before deciding whether to operate in $t = 2$, firms can trade the stock of data gathered in period $t = 1$ (i.e., $\hat{D}^x$) — which we take as given. We do not restrict attention to a specific trading protocol, but consider that trade occurs if and only if trade puts both firms at least weakly and one firm strictly better off.

Formally, let $\overline{\pi}_2^x(D^x, D^{-x})$ denote firm x's expected payoff in $t = 2$ given data (D^x, D^{-x}). Consider that $N_1^x = 1$ and $\hat{D}^x = I^x > \hat{D}^{-x} = 0$. Absent data trade, $D^x = I^x > 0 = D^{-x}$ (there is no mandatory data sharing nor a user-centric data market). However, if firm x trades (and thus shares) data with $-x$, we obtain $D^x = I^x$ and $D^{-x} \in (0, \eta I^x]$ where we impose that maximally (exogenous) fraction $\eta \in [0, 1]$ of x's data can be sold to $-x$. Data is traded if and only if both firms' joint (expected) payoff strictly increases upon trade, that is, if and only if there exists (feasible) $D^{-x} \in (0, \eta I^{-x})$ such that[24]

$$\overline{\pi}_2^x(I^x, D^{-x}) + \overline{\pi}_2^{-x}(D^{-x}, I^x) > \overline{\pi}_2^x(I^x, 0) + \overline{\pi}_2^{-x}(0, I^x). \tag{6}$$

Provided (6) is satisfied, there an appropriate transfer between firms implements the desired split of surplus created from trade, which pins down the terms of trade (not further characterized).[25]

The following Proposition (proven in Appendix G) summarizes our key result.

Proposition 4. *Consider $N_1^x = 1$ and $\hat{D}^x = I^x$. Data trade occurs, i.e., (6) holds for admissible $\hat{D}^x > 0$, only if the market tips post-trade and firm x exits the market. A necessary condition for data trade to occur is $\eta\phi^{-x} > \phi^x + (\Delta_K^x + Z)/\overline{I}$.*

When both firms operate in $t = 2$ and trade data, they enhance product quality and competition in period $t = 2$, which solely benefits the users. That is, firms cannot capture the gains from trade when competition is fierce. However, firms benefit from trading data when this weakens competition by causing market tipping. Therefore, data is traded only if it precipitates market tipping in $t = 2$. A trade in which x sells data to $-x$ precipitates market tipping if the data productivity ϕ^{-x} of $-x$ is sufficiently large. Intuitively, data trades make data flow toward companies that can gain a significant competitive edge due to high data productivity. Consequently, data trades among firms and therefore the introduction of firm-centric data markets weaken competition by triggering market tipping. We therefore conclude (without further formal analysis) that the introduction of firm-centric data markets tends to harm users.

[24]The equilibrium in the period-2 subgame is defined as in the baseline, given stocks of data D^x and D^{-x}.

[25]Characterizing the terms of trade would require further assumptions on the trading protocol — which we omit here.

4 Conclusion

We develop a model of data-driven firm competition in which data are jointly produced by dispersed users' economic activities and firms' investments in data infrastructure. Data collection improves service quality, benefiting users, but may reduce competition, harming users. Dispersed users do not internalize the impact of their data contributions on (i) service quality, (ii) firm competition, and (iii) firms' incentives to invest in data generation. The laissez-faire equilibrium features inefficient data overcollection, reducing competition, or undercollection, compromising service quality. Data privacy protection policies, open data initiatives, and data markets all have limitations. We show how a data union, coordinating users' data contributions, enhances user welfare by addressing inefficiencies in data collection. The data union can be implemented in a market-based approach by introducing a data trust, intermediating data sales from users to firms. Moreover, our results highlight that the combination of a data union and data sharing policies can enhance user welfare in a data economy. We also shed light on the determinants of the optimal data price, with implications for data-related regulation.

Analyzing a purposefully stylized model, we formalize the concept of a data union but leave several key questions unaddressed. It would be interesting to characterize optimal data union in a more general and potentially quantitative framework. As our framework is tractable and admits closed-form solutions, it could be extended in various ways, for instance, by allowing for preference heterogeneity. Future research should also explore whether a data union can be implemented through private entities (market-based approach) markets or requires regulatory intervention. Furthermore, it is important to explore the legal or technological implementation of a data union.

Finally, the concept of a data union — which we introduce and formalize in this paper — has its counterparts in practice. For instance, legal entities in the form of data trusts start to emerge (Houser and Bagby, 2022). Practitioners have also explored the concepts of data pool or cooperative, where users contribute and share data. Examples include Driver's Seat and MIDATA.coop. More recently, the concept of DataDAO provides an implementation of an entity akin to a data union based on distributed ledgers and blockchain technology. In addition, China's newly introduced national data bureau constitutes an example of a goverment-led initiative to centralizing data ownership.

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics 14*(4), 218–256.

Acquisti, A., C. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of economic Literature 54*(2), 442–92.

Arrieta-Ibarra, I., L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl (2018). Should we treat data as labor? moving beyond" free". In *aea Papers and Proceedings*, Volume 108, pp. 38–42.

Bergemann, D. and A. Bonatti (2023). Data, competition, and digital platforms. *arXiv preprint arXiv:2304.07653*.

Bergemann, D., A. Bonatti, and T. Gan (2022). The economics of social data. *The RAND Journal of Economics*.

Bergemann, D., B. Brooks, and S. Morris (2015). The limits of price discrimination. *American Economic Review 105*(3), 921–57.

Bergemann, D., J. Crémer, D. Dinielli, C.-C. Groh, P. Heidhues, M. Schafer, M. Schnitzer, F. M. S. Morton, K. Seim, and M. Sullivan (2023). Market design for personal data. *Yale J. on Reg. 40*, 1056.

Biglaiser, G., E. Calvano, and J. Crémer (2019). Incumbency advantage and its value. *Journal of Economics & Management Strategy 28*(1), 41–48.

Brunnermeier, M. K., R. Lamba, and C. Segura-Rodriguez (2021). Inverse selection. *Available at SSRN 3584331*.

Cabral, L. M. and M. H. Riordan (1994). The learning curve, market dominance, and predatory pricing. *Econometrica: Journal of the Econometric Society*, 1115–1140.

Calvano, E. and M. Polo (2021). Market power, competition and innovation in digital markets: A survey. *Information Economics and Policy 54*, 100853.

Campbell, J., A. Goldfarb, and C. Tucker (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy 24*(1), 47–73.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics 173*, 113–124.

Cong, L. W. and S. Mayer (2023). Antitrust, regulation, and user union in the era of digital platforms and big data. Technical report, National Bureau of Economic Research.

Dasgupta, P. and J. Stiglitz (1988). Learning-by-doing, market structure and industrial and trade policies. *Oxford Economic Papers 40*(2), 246–268.

De Corniere, A. and G. Taylor (2020). Data and competition: a general framework with applications to mergers, market structure, and privacy policy.

Delacroix, S. and N. D. Lawrence (2019). Bottom-up data trusts: disturbing the 'one size fits all'approach to data governance. *International data privacy law 9*(4), 236–252.

Dosis, A. and W. Sand-Zantman (2019). The ownership of data. *Available at SSRN 3420680*.

Eeckhout, J. and L. Veldkamp (2021). Data and market power. Technical report, Columbia University Working Paper Sep.

Fainmesser, I. P., A. Galeotti, and R. Momot (2022). Digital privacy. *Management Science Forthcoming*.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp (2019). Big data and firm dynamics. In *AEA papers and proceedings*, Volume 109, pp. 38–42.

Farboodi, M. and L. Veldkamp (2021). A growth model of the data economy. Technical report, National Bureau of Economic Research.

Goldstein, I., C. Huang, and L. Yang (2022). Open banking with depositor monitoring. Technical report, Working paper.

Hagiu, A. and J. Wright (2021). Data-enabled learning, network effects and competitive advantage. *Working Paper*.

Halaburda, H., B. Jullien, and Y. Yehezkel (2020). Dynamic competition with network externalities: how history matters. *The RAND Journal of Economics 51*(1), 3–31.

He, Z., J. Huang, and J. Zhou (2023). Open banking: Credit market competition when borrowers own the data. *Journal of financial economics 147*(2), 449–474.

Houser, K. and J. W. Bagby (2022). The data trust solution to data sharing problems. *Vanderbilt Journal of Entertainment & Technology Law, Forthcoming*.

Ichihashi, S. (2020). Online privacy and information disclosure by consumers. *American Economic Review 110*(2), 569–95.

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

Ichihashi, S. and A. Smolin (2022). Data collection by an informed seller. *arXiv preprint arXiv:2204.08723*.

Jones, C. I. and C. Tonetti (2020). Nonrivalry and the economics of data. *American Economic Review 110*(9), 2819–58.

Liu, E., S. Ma, and L. Veldkamp (2023). Data sales and data dilution.

Liu, Z., M. Sockin, and W. Xiong (2020). Data privacy and temptation. Technical report, National Bureau of Economic Research.

Parlour, C. A., U. Rajan, and H. Zhu (2022). When fintech competes for payment flows. *The Review of Financial Studies 35*(11), 4985–5024.

Prüfer, J. and C. Schottmüller (2021). Competing with big data. *The Journal of Industrial Economics 69*(4), 967–1008.

Veldkamp, L. and C. Chung (2022). Data and the aggregate economy. *Journal of Economic Literature*.

Appendix

A Regularity Conditions

To ensure $I^x < \overline{I}$ in equilibrium, we assume

$$\overline{I} > \max\left\{\frac{\phi^x - c^x}{2\lambda}, \frac{(Z - \Delta_K^x)}{\phi_\eta^x}\right\}. \tag{A.1}$$

To ensure that there exists $P_t^x > 0$ so that $N_1^x = 1$ upon $Z_1^x = Z$ (and thus $\mathbb{E}_1[\pi_1^x] > 0$), we assume

$$\Delta_K^x + Z - (c^x I^x - c^{-x} I^{-x}) > 0. \tag{A.2}$$

A sufficient condition for (A.2) is

$$\Delta_K^x + Z > \overline{I}\left[\max\{c^x, 0\} - \min\{c^{-x}, 0\}\right]. \tag{A.3}$$

Condition (A.2) ensures that $N_1^x = 1$ (i.e., x wins the market in $t = 1$) if and only if $Z_1^x = Z$.

Finally, we make assumptions how to break ties in asymmetric Bertrand pricing competition and regarding market tipping. First, in determining the pricing equilibrium, we break ties in favor of $x = A$, so $N_t^A = 1$ when $Y_t^A + Z_t^A - c^A I^A \theta^A \mathbb{I}_{\{t=1\}} = Y_t^{-x} + Z_t^B - c^B I^B \theta^B \mathbb{I}_{\{t=1\}}$. Second, we assume in Proposition 1 that when indifferent, a firm does not precipitate market tipping. That is, when $G(c^x) = 0$, the market does not tip upon $Z_1^x + Z$, so $T^x = 0$. When we call an equilibrium *unique*, we mean that is unique subject to these assumptions on how to break ties.

B Proof of Proposition 1

We solve the model (i.e., the equilibrium) separately for the two cases, $T^x = 0$ and $T^x = 1$, where $x = A$ ($-x = B$) or $x = B$ ($-x = A$). Having solved the model for the two cases $T^x \in \{0, 1\}$, we then determine which case prevails in the (unique) equilibrium. In what follows, $\mathbb{E}_t$ denotes the expectation taken at the beginning of period t before the shock $Z_t = (Z_t^A, Z_t^B)$ realizes. Moreover, as can be easily verified, first-order conditions are sufficient for optimality in our setting, so we do not always separately check the second-order conditions for the sake of brevity.

B.1 No Market Tipping $T^x = 0$

Consider that $T^x = 0$, i.e., the market does not tip and both firms operate when $Z_1^x = Z \iff N_1^x = 1$ realizes in period 1. Consider $Z_1^x = Z$, i.e., x wins the market in $t = 1$. Then, $Z_2^x = Z$ implies $N_2^x = 1 > 0 = N_2^{-x}$, $P_2^{-x} = \pi_2^{-x} = 0$, while $Z_2^x = 0 < Z_2^{-x} = Z$ implies $N_2^x = 0$. The (realized) period-2 payoff of x upon $Z_2^x = Z$ becomes:

$$\pi_2^x = P_2^x = Y_2^x - Y_2^{-x} = \Delta_K^x + \phi^x D^x - \phi^{-x} D^{-x} + Z,$$

with $D^x = N_1^x I^x + \eta^{-x} N_1^{-x} I^{-x}$ and $D^{-x} = N_1^{-x} I^{-x} + \eta^x N_1^x I^x$. As firm x wins the market with probability 1/2 (and otherwise has zero revenue), firm x's expected payoff at beginning of period 2 reads

$$\mathbb{E}_2[\pi_2^x] = \frac{\Delta_K^x + \phi^x D^x - \phi^{-x} D^{-x} + Z}{2}.$$

In period $t = 1$, we have $N_1^x = 1$ if and only if $Z_1^x = Z$ — which happens with probability 1/2. Then, $Z_1^x = Z$ implies that x wins the market, in which case $P_1^{-x} = 0$ and

$$P_1^x = Y_1^x - Y_1^{-x} - (c^x I^x - c^{-x} I^{-x}) = \Delta_K^x - (c^x I^x - c^{-x} I^{-x}) + Z.$$

Condition (A.2) implies $P_1^x > 0$, so $\mathbb{E}_1[\pi_1^x] > 0$.

We calculate the expected payoff at the beginning of period 1:

$$\begin{aligned}\mathbb{E}_1[\pi_1^x] = &\frac{1}{2}\Big(\Delta_K^x - (c^x I^x - c^{-x} I^{-x}) + Z + \frac{1}{2}\Big(I^x(\phi^x - \eta^x \phi^{-x}) + \Delta_K^x + Z\Big)\Big)\\ &+ \frac{1}{4}\left[\Delta_K^x + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})\right]^+,\end{aligned}$$

where $[\Delta_K^x + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})]^+ = [\Delta_K^x + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})]\,(1 - T^{-x})$ is zero if and only if firm x exits upon $Z_1^{-x} = Z > Z_1^x = 0$, i.e., $T^{-x} = 1$.

Firm x maximizes $\mathbb{E}_1[\pi_1^x] - \frac{\lambda (I^x)^2}{2}$ over $I^x \in [0, \overline{I}]$ at the beginning of period $t = 1$ (taking I^{-x} as given). If $I^x \in (0, \overline{I})$, this leads to the first-order condition (FOC):

$$\frac{1}{2}\Big[-c^x + \frac{1}{2}\big(\phi^x - \eta\phi^{-x}\big)\Big] - \lambda I^x = 0 \iff I^x = \frac{\phi_\eta^x - 2c^x}{4\lambda}.$$

Condition (A.1) ensures $I^x < \overline{I}$. The FOC is clearly sufficient, i.e., the second order condition is satisfied as $\frac{\partial^2}{\partial (I^x)^2}\left(\mathbb{E}_1[\pi_1^x] - \frac{\lambda (I^x)^2}{2}\right) < 0$.

Thus,

$$I^x = I_{NT}^x = \max\left\{0, \frac{\phi_\eta^x - 2c^x}{4\lambda}\right\} = \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda}.$$

In this scenario, total expected payoff for firm x becomes

$$\begin{aligned}
\Pi^x_{NT} &:= \max_{I^x \geq 0} \left(\mathbb{E}_1[\pi^x_1] - \frac{\lambda (I^x)^2}{2} \right) \\
&= \frac{1}{2} \left(\Delta^x_K + c^{-x} I^{-x} + Z + \frac{1}{2} \Big(\Delta^x_K + Z \Big) \right) + \frac{\lambda (I^2_{NT})^2}{2} \\
&\quad + \frac{1}{4} \left[\Delta^x_K + Z + I^{-x} (\eta^{-x} \phi^x - \phi^{-x}) \right]^+ .
\end{aligned}$$

B.2 Market Tipping $T^x = 1$

Next, consider that $T^x = 1$, i.e., the market tips upon $Z^x_1 = Z$, which occurs if and only if firm $-x$ cannot "win" the market upon $Z^x_2 = 0 < Z^{-x}_2 = Z$. That is, $T^x = 1$ if and only if

$$\Delta^x_K - Z + I^x (\phi^x - \eta^x \phi^{-x}) \geq 0 \quad \iff \quad I^x \geq \underline{I}^x := \frac{Z - \Delta^x_K}{\phi^x - \eta^x \phi^{-x}} = \frac{Z - \Delta^x_K}{\phi^x_\eta}. \tag{B.1}$$

Condition (A.1) implies $\overline{I} > \underline{I}^x$. As $Z > \Delta^x_K$ and $\phi^x_\eta > 0$, $\underline{I}^x > 0$.

We consider $T^x = 1$, i.e., $I^x \geq \underline{I}$. Then, conditional on $Z^x_1 = Z$, firm x becomes monopolist in period $t = 2$ and charges optimally the monopoly $P^x_2 = Y^x_2 + Z^x_2$, so that users just break even. Thus, x realizes period-2 profits of $\pi^x_2 = P^x_2 = K^x + Z^x_2 + \phi^x I^x$.

If firm x wins in period 1, i.e., $Z^x_1 = Z$ (which happens with probability 1/2), it charges price $P^x_1 = \Delta^x_K - (c^x I^x - c^{-x} I^{-x})$, while firm $-x$ charges price $P^{-x}_1 = 0$. Condition (A.2) ensures $P^x_1 > 0$.

Then, period-1 expected payoff for firm x reads

$$\begin{aligned}
\mathbb{E}_1[\pi^x_1] &= \frac{1}{2} \left(\Delta^x_K - (c^x I^x - c^{-x} I^{-x}) + \frac{3Z}{2} + K^x + \phi^x I^x \right) \\
&\quad + \frac{1}{4} \left[\Delta^x_K + Z + I^{-x} (\eta^{-x} \phi^x - \phi^{-x}) \right]^+ ,
\end{aligned}$$

where $[\Delta^x_K + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})]^+$ is zero if and only if firm x exits upon $Z^{-x}_1 = Z > Z^x_1 = Z$ (i.e., if and only if $T^{-x} = 1$).

Conditional on $T^x = 1$ and hence $I^x \geq \underline{I}^x$, firm x maximizes $\mathbb{E}_1[\pi^x_1] - \frac{\lambda (I^x)^2}{2}$ over $I^x \geq \underline{I}^x$ at the beginning of period $t = 1$ (taking I^{-x} as given). If $I^x \in (\underline{I}^x, \overline{I})$, this leads to the FOC

$$\frac{1}{2}(-c^x + \phi^x) - \lambda I^x = 0 \quad \iff \quad I^x = \frac{\phi^x - c^x}{2\lambda}.$$

(A.1) ensures $I^x < \overline{I}$. The second-order condition is satisfied as $\frac{\partial^2}{\partial (I^x)^2} \left(\mathbb{E}_1[\pi^x_1] - \frac{\lambda (I^x)^2}{2} \right) < 0$.

Thus,

$$I^x = I^x_T = \max\left\{\frac{\phi^x - c^x}{2\lambda}, \underline{I}^x\right\},$$

with $I^x_T < \overline{I}$ due to (A.1). When $T^x = 1$, total expected payoff becomes

$$\Pi^x_T := \max_{I^x \geq \underline{I}^x}\left(\mathbb{E}_1[\pi^x_1] - \frac{\lambda(I^x)^2}{2}\right).$$

When $I^x_T > \underline{I}^x$, then

$$\Pi^x_T = \frac{1}{2}\left(\Delta^x_K + c^{-x}I^{-x} + \frac{3Z}{2} + K^x\right) + \frac{\lambda(I^x_T)^2}{2} + \frac{1}{4}\left[\Delta^x_K + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})\right]^+.$$

Otherwise, when $I^x_T = \underline{I}^x$, we have

$$\begin{aligned}\Pi^x_T = \frac{1}{2}&\left(\Delta^x_K - (c^x\underline{I}^x - c^{-x}I^{-x}) + \frac{3Z}{2} + K^x + \phi^x\underline{I}^x\right) - \frac{\lambda(\underline{I}^x)^2}{2}\\ &+ \frac{1}{4}\left[\Delta^x_K + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})\right]^+.\end{aligned}$$

B.3 Optimal Choice

Recall $\phi^x_\eta = \phi^x - \eta^x\phi^{-x}$. Combining the two cases $T^x \in \{0, 1\}$, we can write the firm x' expected payoff at the beginning of period 1 as follows:

$$\begin{aligned}\mathbb{E}_1[\pi^x_1] = \frac{1}{2}\Bigg[&\Delta^x_K + Z - (c^xI^x - c^{-x}I^{-x}) + T^x\left(\frac{Z}{2} + K^x + \phi^xI^x\right)\\ &+ \frac{(1-T^x)}{2}\Big(I^x(\phi^x - \eta^x\phi^{-x}) + \Delta^x_K + Z\Big)\Bigg]\\ &+ \frac{(1-T^{-x})}{4}\left[\Delta^x_K + Z + I^{-x}(\eta^{-x}\phi^x - \phi^{-x})\right].\end{aligned}$$

Taking the choice of I^{-x} as given, firm x's total payoff (pre-investment) becomes

$$\Pi^x_1 = \max\{\Pi^x_{NT}, \Pi^x_T\}.$$

Effectively, firm x chooses (taking the choice of the other firm as given) at the beginning of the first period the investment $I^x \in \{I^x_T, I^x_{NT}\}$ to maximize

$$\Pi^x_1 = \Pi^x_{NT}\mathbb{I}_{\{I^x < \underline{I}^x\}} + \Pi^x_T\mathbb{I}_{\{I^x \geq \underline{I}^x\}}$$

. In other words, x chooses whether it precipitates market tipping upon $Z^x_1 = Z$ by raising I^x above $\underline{I}^x$.

When $\Pi^x_{NT} \geq \Pi^x_T$, then the market does not tip upon $Z^x_1 = Z$ and x chooses $I^x = I^x_{NT}$ so $T^x = 0$; ties are broken in favor of no tipping. Otherwise, when $\Pi^x_{NT} < \Pi^x_T$, the market tips upon $Z^x_1 = Z$ (i.e., $T^x = 1$) and $I^x = I^x_T$.

Given I^{-x}, we obtain that $\Pi^x_{NT} \gtreqless \Pi^x_T$ is equivalent to

$$\begin{aligned}
&\frac{1}{2}\left(-c^x I^x_{NT} + \frac{1}{2}\Big(I^x_{NT}\phi^x_\eta + \Delta^x_K + Z\Big)\right) - \frac{\lambda(I^x_{NT})^2}{2} = \frac{1}{4}\Big(\Delta^x_K + Z\Big) + \frac{\lambda(I^x_{NT})^2}{2} \\
\gtreqless &\frac{1}{2}\left(-c^x I^x_T + K^x + \frac{Z}{2} + \phi^x I^x_T\right) - \frac{\lambda(I^x_T)^2}{2},
\end{aligned}$$

which simplifies to

$$\lambda(I^x_{NT})^2 \gtreqless \frac{K^A + K^B}{2} + (\phi^x - c^x)I^x_T - \lambda(I^x_T)^2. \tag{B.2}$$

Thus, the market tips if and only if

$$G(c^x) := \frac{K^A + K^B}{2} + (\phi^x - c^x)I^x_T - \lambda(I^x_T)^2 - \lambda(I^x_{NT})^2 > 0.$$

Note that when $I^x_T > \underline{I}^x$, i.e., $\frac{\phi^x - c^x}{2\lambda} > \frac{Z - \Delta^x_K}{\phi^x_\eta}$, then

$$G(c^x) = \frac{K^A + K^B}{2} + \lambda(I^x_T)^2 - \lambda(I^x_{NT})^2 > 0,$$

which triggers market tipping $T^x = 1$. As a result, a sufficient condition for market tipping $T^x = 1$ reads $\frac{\phi^x - c^x}{2\lambda} > \frac{Z - \Delta^x_K}{\phi^x_\eta}$.

B.4 User Welfare

When there is market tipping $t = 2$ and x becomes monopolist (i.e., $T^x = 1$), user welfare (i.e., users' per-period payoff) in $t = 2$ is zero, since users are charged their reservation value (i.e., $P^x_2 = Y^x_2$) and merely break even.

Otherwise, when there is no market tipping (i.e., $T^x = 0$) and $Z^x_2 = Z$, i.e., x wins the market, we have $P^{-x}_2 = 0$ and $P^x_2 = Y^x_2 - Y^{-x}_2 + Z$, so users' utility/payoff/welfare in this case becomes $Y^x_2 - P^x_2 = Y^{-x}_2 = K^{-x} + \phi^{-x}D^{-x}$.

Therefore, if both firms operate in $t = 2$, expected user utility at the beginning of period 2 equals

$$u^{NT}_2 = \frac{Y^{-x}_2 + Y^x_2}{2} = \frac{1}{2}(K^{-x} + \phi^{-x}D^{-x}) + \frac{1}{2}(K^x + \phi^x D^x).$$

This payoff depends on which firm has won period $t = 1$ through D^x, D^{-x}. When x has won period 1, then $D^x = I^x$ and $D^{-x} = \eta^x I^x$ and $u^{NT}_2 = \frac{1}{2}(K^{-x} + \phi^{-x}\eta I^x) + \frac{1}{2}(K^x + \phi^x I^x)$.

Next, consider period $t = 1$. When $Z^x_1 = Z$, i.e., x wins the market, we have $P^{-x}_1 = 0$ as well as $P_1^x = Y_1^x - Y_1^{-x} - (c^x I^x - c^{-x} I^{-x})$. Thus, $Y_1^x - P_1^x = Y_1^{-x} + (c^x I^x - c^{-x} I^{-x})$. The per-period utility for users in period 1 is then $Y_1^x - P_1^x - c^x I^x = Y_1^{-x} - c^{-x} I^{-x} = K^{-x} - c^{-x} I^{-x}$.

Next, we can write user welfare — users' expected (lifetime) payoff/utility at the beginning of period $t = 1$ — as

$$\begin{aligned} U =& \frac{1}{2}\Big[K^{-x} - c^{-x} I^{-x} + (1 - T^x)\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}\eta^x I^x) + \frac{1}{2}(K^x + \phi^x I^x)\Big\}\Big] \\ &+ \frac{1}{2}\Big[K^x - c^x I^x + (1 - T^{-x})\Big\{\frac{1}{2}(K^{-x} + \phi^{-x} I^{-x}) + \frac{1}{2}(K^x + \eta^{-x}\phi^x I^{-x})\Big\}\Big], \end{aligned}$$

where we account for the fact that period-2 user payoff becomes zero upon market tipping — i.e., when $T^x = 1$ and $Z_1^x = Z \iff N_1^x = 1$.

One can also write:

$$U = \frac{1}{2}\sum_{x=A,B}\Big[K^{-x} - c^{-x} I^{-x} + (1 - T^x)\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}\eta^x I^x) + \frac{1}{2}(K^x + \phi^x I^x)\Big\}\Big].$$

We can rewrite this expression for U as

$$\begin{aligned} U :=& (K^A + K^B)\left(1 - \frac{T^A + T^B}{4}\right) + \frac{1}{2}\Big[-c^{-x} I^{-x} + \frac{(1 - T^x)}{2}\Big\{\phi^{-x}\eta^x I^x + \phi^x I^x\Big\}\Big] \\ &+ \frac{1}{2}\Big[-c^x I^x + \frac{(1 - T^{-x})}{2}\Big\{\phi^{-x} I^{-x} + \eta^{-x}\phi^x I^{-x}\Big\}\Big], \end{aligned}$$

which coincides with (3). One can insert $(x, -x) \in \{(A, B), (B, A)\}$ to obtain a more explicit representation.

C Data Union — Proof of Proposition 2

C.1 Part I — Auxiliary Result

Consider parameters such that

$$\frac{Z - \Delta_K^x}{\phi_x^\eta} > \frac{\phi_\eta^x - 2c^x}{4\lambda} \quad \iff \quad c^x > \frac{\phi_\eta^x}{2} - \frac{2\lambda(Z - \Delta_K^x)}{\phi_x^\eta}.$$

Thus, $I_T^x = \frac{Z - \Delta_K^x}{\phi_x^\eta}$ and $G(c^x)$ can be rewritten as

$$G(c^x) := \frac{K^A + K^B}{2} + \frac{(\phi^x - c^x)(Z - \Delta_K^x)}{\phi_\eta^x} - \frac{\lambda(Z - \Delta_K^x)^2}{(\phi_\eta^x)^2} - \frac{([\phi_\eta^x - 2c^x]^+)^2}{16\lambda} > 0,$$

Clearly, when $c^x > \frac{\phi_\eta^x}{2}$, then $G'(c^x) < 0$ (owing to $Z > \Delta_K^x$).

Next, consider $c^x < \frac{\phi_\eta^x}{2}$. Then, we calculate

$$G'(c^x) = -\frac{Z - \Delta_K^x}{\phi_x^\eta} + \frac{\phi_\eta^x - 2c^x}{4\lambda} = I^x - \underline{I}^x,$$

which is strictly smaller than zero if and only if

$$\frac{Z - \Delta_K^x}{\phi_x^\eta} > \frac{\phi_\eta^x - 2c^x}{4\lambda} \quad \iff \quad c^x > \frac{\phi_\eta^x}{2} - \frac{2\lambda(Z - \Delta_K^x)}{\phi_x^\eta}.$$

In particular, we have that $G(c^x)$ strictly decreases when $\frac{(Z-\Delta_K^x)}{\phi_\eta^x} \geq \frac{\phi^x - c^x}{2\lambda}$, i.e., $c^x \geq \phi^x - \frac{2\lambda(Z-\Delta_K^x)}{\phi_\eta^x} > \frac{\phi_\eta^x - 2c^x}{4\lambda}$, with $G'(c^x) < 0$ whenever $G(c^x)$ is differentiable.

Next, we evaluate

$$\begin{aligned} G\left(\phi^x - \frac{2\lambda(Z - \Delta_K^x)}{\phi_\eta^x}\right) &= \frac{K^A + K^B}{2} + \frac{\lambda(Z - \Delta_K^x)^2}{(\phi_\eta^x)^2} - \frac{([\phi_\eta^x - 2c^x]^+)^2}{16\lambda} \\ &= \frac{K^A + K^B}{2} + \frac{(\phi^x - c^x)^2}{4\lambda} - \frac{([\phi_\eta^x - 2c^x]^+)^2}{16\lambda} \\ &> 0, \end{aligned}$$

with $c^x = \phi^x - \frac{2\lambda(Z-\Delta_K^x)}{\phi_\eta^x} < \phi^x$.

As $G(c^x)$ is continuous on $\mathcal{S} := \left(\phi^x - \frac{2\lambda(Z-\Delta_K^x)}{\phi_\eta^x}, \infty\right)$, it follows that $G(c^x)$ has a unique root on $\mathcal{S}$ denoted c_*^x with $G(c_*^x) = 0$.

We can solve $G(c_*^x) = 0$ in closed form. When $c_*^x \leq \frac{\phi_\eta^x}{2}$, then

$$c_*^x = \frac{\phi_\eta^x}{2} - \frac{2\lambda(Z - \Delta_K^x)}{\phi_\eta^x} + \left[2\lambda\left(\frac{2(Z - \Delta_K^x)\phi^x - (Z - \Delta_K^x)\phi_\eta^x + (K^A + K^B)\phi_\eta^x}{\phi_\eta^x}\right)\right]^{\frac{1}{2}}.$$

Otherwise, when $c_*^x > \frac{\phi_x^\eta}{2}$, we have

$$c_*^x = \frac{\phi_\eta^x}{Z - \Delta_K^x}\left(\frac{K^A + K^B}{2} + \frac{\phi^x(Z - \Delta_K^x}{\phi_\eta^x} - \frac{\lambda(Z - \Delta_K^x)^2}{(\phi_x^\eta)^2}\right),$$

i.e.,

$$c_*^x = \phi^x - \frac{\lambda(Z - \Delta_K^x)}{\phi_\eta^x} + \frac{\phi_\eta^x(K^A + K^B)}{2(Z - \Delta_K^x)}.$$

C.2 Part II — Data Union Optimization

For $c^x = c + \tau^x$, choosing c^x is equivalent to choosing τ^x, subject to ensuring $\tau^x \in [\underline{\tau}, \overline{\tau}]$. It is assumed that this constraint does not bind in optimum, i.e., $\tau^x \in (\underline{\tau}, \overline{\tau})$ in optimum.

As such, the data union maximizes

$$\max_{c^A,c^B}\Big(\underbrace{U+\frac{1}{2}\tau^x I^x+\frac{1}{2}\tau^{-x}I^{-x}}_{=:\overline{U}}\Big),$$

with U characterized in Proposition 1 for a given level of $c^x = c+\tau^x$. This is akin to solving $\max_{c^A,c^B}\overline{U}$ with

$$\overline{U} = \frac{1}{2}\bigg[K^{-x}-cI^{-x}+(1-T^x)\Big\{\frac{1}{2}(K^{-x}+\phi^{-x}\eta^x I^x)+\frac{1}{2}(K^x+\phi^x I^x)\Big\}\bigg] \tag{C.1}$$

$$+\frac{1}{2}\bigg[K^{x}-cI^{x}+(1-T^{-x})\Big\{\frac{1}{2}(K^{-x}+\phi^{-x}I^{-x})+\frac{1}{2}(K^x+\eta^{-x}\phi^x I^{-x})\Big\}\bigg]. \tag{C.2}$$

The objective function $\overline{U}$ depends on c^x only via I^x. As such, the data union's optimization becomes $\max_{I^A,I^B}\overline{U}$, subject to all relevant constraints — where one can treat I^x (with a slight abuse of notation) as a parameter.

Clearly, market tipping $T^x = 1$ or T^{-x} reduces $\overline{U}$ (heuristically, one can calculate $\frac{\partial \overline{U}}{\partial T^x} < 0$). As such, the data union sets (c^A, c^B) so that there is no market tipping whenever feasible, i.e., $T^x = T^{-x} = 0$. Thus, whenever $\tau^x \in (\underline{\tau}, \overline{\tau})$, then $T^x = 0$. By Proposition 1, this requires that

$$\frac{Z-\Delta_K^x}{\phi_\eta^x} > \frac{\phi^x - c^x}{2\lambda} \quad\iff\quad c^x > \phi^x - \left(\frac{2\lambda(Z-\Delta_K^x)}{\phi_\eta^x}\right).$$

Furthermore, for no market tipping to occur, it must be that $G(c^x) \leq 0$. As we know from the previous part, $G(c^x)$ strictly decreases in c^x for $c^x > \phi^x - \left(\frac{2\lambda(Z-\Delta_K^x)}{\phi_\eta^x}\right)$ and has a unique root c_*^x. As such, market tipping does not occur if $c^x \geq c_*^x$; data union therefore implements $c^x = c_*^x$.

The partial derivative with respect to I^x becomes (holding all else equal):

$$\frac{\partial \overline{U}}{\partial I^x} = \frac{1}{2}\left[-c+\frac{(1-T^x)}{2}\big(\phi^x+\eta^x\phi^{-x}\big)\right]$$

Absent market tipping ($T^x = 0$) and under $\phi^x > 2c$, $\frac{\partial \overline{U}}{\partial I^x} > 0$.

Thus, the data union optimally induces maximum I^x subject to ensuring the market does not tip. As $I^x = \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda}$ decreases with c^x, the union sets the lowest price of data that prevents market tipping. Thus, provided $\tau^x \in (\underline{\tau}, \overline{\tau})$, optimal data price equals $c^x = c_*^x$, with c_*^x solving $G(c_*^x) = 0$ and characterized in Part I of the proof.

D Proof of Corollary 1

To begin with, define

$$\mathcal{K} := 2\lambda\left(\frac{2(Z-\Delta_K^x)\phi^x - (Z-\Delta_K^x)\phi_\eta^x + (K^A+K^B)\phi_\eta^x}{\phi_\eta^x}\right) > 0.$$

Thus,

$$c_*^x = \frac{\phi_\eta^x}{2} - \frac{2\lambda(Z-\Delta_K^x)}{\phi_\eta^x} + \sqrt{\mathcal{K}}.$$

As $\phi^x \geq \phi_\eta^x$, we have $\mathcal{K} \leq 2\lambda[Z-\Delta_K^x+K^A+K^B]$. As an auxiliary result, we note that a sufficient condition for $\phi^x \leq \sqrt{\mathcal{K}} \iff (\phi^x)^2 \leq \mathcal{K}$ is given by

$$(\phi^x)^2 \leq 2\lambda\big[(Z-\Delta_K^x)+K^A+K^B\big]. \tag{D.1}$$

We now perform comparative statics.

Holding Δ_K^x fixed, c_*^x clearly decreases with K^A and K^B.

Next,

$$\frac{\partial c_*^x}{\partial \lambda} = -\frac{2(Z-\Delta_K^x)}{\phi_\eta^x} + \frac{1}{2\sqrt{\mathcal{K}}}\frac{\mathcal{K}}{\lambda} = -\frac{2(Z-\Delta_K^x)}{\phi_\eta^x} + \frac{\sqrt{\mathcal{K}}}{2\lambda}.$$

Thus,

$$\frac{\partial c_*^x}{\partial \lambda} \gtrless 0 \iff \sqrt{\mathcal{K}}\phi_\eta^x \gtrless 4\lambda(Z-\Delta_K^x) \iff \frac{\sqrt{\mathcal{K}}\phi_\eta^x}{4(Z-\Delta_K^x)} \gtrless \lambda.$$

Thus, the expression for c_*^x in (5) is hump-shaped (inverted U-shaped) in λ, i.e., increases in λ for low levels of λ and decreases in λ for high levels of λ.

We now calculate (using $\phi_\eta^x = \phi^x - \eta^x\phi^{-x} > 0$):

$$\begin{aligned}\frac{\partial c_*^x}{\partial \phi^x} &= \frac{1}{2} + \frac{2\lambda(Z-\Delta_K^x)}{(\phi_\eta^x)^2} + \frac{2\lambda(Z-\Delta_K^x)}{\sqrt{\mathcal{K}}}\frac{\phi_\eta^x-\phi^x}{(\phi_\eta^x)^2} \\ &= \frac{1}{2} + \frac{2\lambda(Z-\Delta_K^x)}{(\phi_\eta^x)^2}\left(1-\frac{\phi^x-\phi_\eta^x}{\sqrt{\mathcal{K}}}\right).\end{aligned}$$

This expression is positive when $\phi^x - \phi_\eta^x \leq \sqrt{\mathcal{K}}$. In particular, $\frac{\partial c_*^x}{\partial \phi^x} > 0$ when (D.1) holds or when η^x is small (e.g., $\eta^x = 0$).

Next, we calculate (using $\phi_\eta^x = \phi^x - \eta^x\phi^{-x} > 0$):

$$\begin{aligned}\frac{\partial c_*^x}{\partial \eta^x} &= -\frac{\phi^{-x}}{2} - \phi^{-x}\left(\frac{2\lambda(Z-\Delta_K^x)}{(\phi_\eta^x)^2}\right) + \frac{2\lambda\phi^{-x}}{\sqrt{\mathcal{K}}}\,\frac{(Z-\Delta_K^x)\phi^x}{(\phi_\eta^x)^2}\\ &= \phi^{-x}\left[-\frac{1}{2} - \frac{2\lambda(Z-\Delta_K^x)}{(\phi_\eta^x)^2}\left(1-\frac{\phi^x}{\sqrt{\mathcal{K}}}\right)\right].\end{aligned}$$

Thus, $\frac{\partial c_*^x}{\partial \eta^x} < 0$ when (D.1) holds and $\phi^{-x} > 0$.

Due to $\phi_\eta^x = \phi^x - \eta^x\phi^{-x}$, $\frac{\partial c_*^x}{\partial \phi^{-x}}$ has the same sign as $\frac{\partial c_*^x}{\partial \eta^x}$, provided $\eta^x > 0$. Thus, when $\eta^x > 0$ and (D.1) holds, then $\frac{\partial c_*^x}{\partial \phi^{-x}} < 0$.

Calculate

$$\frac{\partial c_*^x}{\partial Z} = -\frac{2\lambda}{\phi_\eta^x} + \frac{\lambda}{\sqrt{\mathcal{K}}}\frac{2\phi^x - \phi_\eta^x}{\phi_\eta^x} = \frac{2\lambda}{\phi_\eta^x}\left(\frac{\phi^x - 0.5\phi_\eta^x}{\sqrt{\mathcal{K}}} - 1\right),$$

which is negative under condition (D.1).

E Proof of Corollary 2

From (C.1) and under $T^A = T^B = 0$ (under the optimal data union policy), we obtain

$$\begin{aligned}\overline{U} =& \frac{1}{2}\left[K^{-x} - cI^{-x} + \left\{\frac{1}{2}(K^{-x} + \phi^{-x}\eta^x I^x) + \frac{1}{2}(K^x + \phi^x I^x)\right\}\right]\\ &+ \frac{1}{2}\left[K^x - cI^x + \left\{\frac{1}{2}(K^{-x} + \phi^{-x} I^{-x}) + \frac{1}{2}(K^x + \eta^{-x}\phi^x I^{-x})\right\}\right],\end{aligned}$$

where $I^x = \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda}$. Note that $U^* = \max_{c^A, c^B}\overline{U}$. As we consider $\tau^A, \tau^B \in (\underline{\tau}, \overline{\tau})$, optimal data prices c_*^x are characterized in Proposition 2. That is, $U^* = \overline{U}$ under $c^x = c_*^x$.

It is apparent that — holding (I^A, I^B) fixed — U^* increases with η^x and η^{-x}. Further, due to $\phi^x, \phi^{-x} > 2c$, U^* increases with I^x and I^{-x}, holding all else equal. Thus, if I^x increases with η^x, then U^* increases with η^x too. Formally, $\frac{\partial I^x}{\partial \eta^x} > 0$ implies $\frac{\partial U^*}{\partial \eta^x} > 0$.

In case $I^x = 0$, η^x does not affect U^*, and $\frac{\partial I^x}{\partial \eta^x} = \frac{\partial U^*}{\partial \eta^x} = 0$ (provided differentiability). It therefore suffices to analyze $I^x > 0$. Suppose that $I^x > 0$ under $c^x = c_*^x$, with the optimal data price c_*^x characterized in Proposition 2. As $I^x = \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda}$ under the optimal data union policy $c^x = c_*^x$, strictly positive investment $I^x > 0$ requires $c^x = c_*^x < \frac{\phi_\eta^x}{2}$, so that c_*^x is characterized in (5).

Next, for $I^x > 0$ and $c^x = c_*^x$ from (5), we calculate that $I^x = \frac{\mathcal{C}}{4\lambda}$, with

$$\mathcal{C} := \frac{2\lambda(Z - \Delta_K^x)}{\phi_\eta^x} - \left[2\lambda\left(\frac{2(Z - \Delta_K^x)\phi^x - (Z - \Delta_K^x)\phi_\eta^x + (K^A + K^B)\phi_\eta^x}{\phi_\eta^x}\right)\right]^{\frac{1}{2}}.$$

Define

$$\mathcal{K} := 2\lambda\left(\frac{2(Z - \Delta_K^x)\phi^x - (Z - \Delta_K^x)\phi_\eta^x + (K^A + K^B)\phi_\eta^x}{\phi_\eta^x}\right) > 0.$$

Finally, we calculate (using $\phi_\eta^x = \phi^x - \eta^x\phi^{-x} > 0$):

$$\frac{\partial \mathcal{C}}{\partial \eta^x} = \phi^{-x}\left(\frac{2\lambda(Z - \Delta_K^x)}{(\phi_\eta^x)^2}\right) - \frac{2\lambda\phi^{-x}}{\sqrt{\mathcal{K}}}\,\frac{(Z - \Delta_K^x)\phi^x}{(\phi_\eta^x)^2} = \frac{2\phi^{-x}\lambda(Z - \Delta_K^x)}{(\phi_\eta^x)^2}\left(1 - \frac{\phi^x}{\sqrt{\mathcal{K}}}\right).$$

Thus, $\frac{\partial \mathcal{C}}{\partial \eta^x} > 0$ when (D.1) holds and $\phi^{-x} > 0$.

Consequently, when $I^x > 0$ and $c^x = c_*^x$, we have $\frac{\partial I^x}{\partial \eta^x} > 0$ and $\frac{\partial U^*}{\partial \eta^x} > 0$ as well as $\frac{\partial c_*^x}{\partial \eta^x} < 0$ (according to Corollary 1). This concludes the proof.

F User-Centric Data Markets: Proof of Proposition 3

The proof is analogous to the proof of Proposition 1, modulo slight changes. We solve the model separately for the two cases $T^x \in \{0, 1\}$, and finally determine which case prevails in equilibrium.

F.1 No Market Tipping $T^x = 0$

First, suppose that both firms operate in $t = 2$ upon $Z_1^x = Z$, i.e., $T^x = 0$. Let us analyze firms' decisions whether to acquire data in the end of $t = 1$. For this sake, consider that $N_1^x = 1$, i.e. $Z_1^x = Z$, and investment $I^x > 0$; that is, firm x has won period $t = 1$. Thus, $\hat{D}^x = D^x = I^x$ and $\hat{D}^{-x} = 0$. The expected period-2 payoff of firm $-x$, given data D^{-x}, becomes

$$\mathbb{E}_2[\pi_2^{-x}] = \frac{1}{2}\Big(Z - \Delta_K^x + \phi^{-x}D^{-x} - \phi^x D^x\Big).$$

At the end of period $t = 1$, firm $-x$ solves

$$\max_{D^{-x} \in [0, \eta^x I^x]} \mathbb{E}_2[\pi_2^{-x}] - c^{-x}D^{-x}$$

Due to $\frac{\partial \mathbb{E}[\pi_2^{-x}]}{\partial D^{-x}} = \frac{\phi^{-x}}{2}$, we get that in optimum/equilibrium, $-x$ buys all available data so that $D^{-x} = \eta^x D^x = \eta^x I^x$ when $\phi^{-x} \geq 2c^{-x}$ and otherwise buys no data, i.e., $D^{-x} = 0$; we break ties in favor of buying all data, if indifferent.

Conditional on $T^x = 0$ and $Z_1^x = Z$, we then have under optimal data acquisition policy:

$$\mathbb{E}_2[\pi_2^{-x}] - c^{-x}D^{-x} = \frac{1}{2}\Big(Z - \Delta_K^x + [\phi^{-x} - 2c^{-x}]^+\eta^x I^x - \phi^x I^x\Big),$$

where $[\cdot]^+ = \max\{\cdot, 0\}$.

Building on our insights so far, we can derive for $x = A, B$, $T^x = 0$ and $\phi_\eta^x = \phi^x - \eta^x\phi^{-x}\mathbb{I}_{\{\phi^{-x}\geq 2c^{-x}\}}$ the expected period-1 payoff of firm x:

$$\begin{aligned}\mathbb{E}_1[\pi_1^x] = &\frac{1}{2}\left(\Delta_K^x - (c^x I^x - c^{-x}I^{-x}) + Z + \frac{1}{2}\Big(I^x\phi_\eta^x + \Delta_K^x + Z\Big)\right) \\ &+ \frac{1}{4}\left[\Delta_K^x + Z + I^{-x}(\eta^{-x}[\phi^x - 2c^x]^+ - \phi^{-x})\right](1 - T^{-x}).\end{aligned}$$

Note that according to (A.2), $\Delta_K^x - (c^x I^x - c^{-x}I^{-x}) > 0$.

Firm x maximizes at the beginning of $t = 1$ the objective $\mathbb{E}_1[\pi_1^x] - \frac{\lambda(I^x)^2}{2}$. If interior, optimal investment therefore solves the first order condition:

$$\frac{\phi_\eta^x}{4} - \frac{c^x}{2} - \lambda I^x = 0 \quad \Longleftrightarrow \quad I^x = \frac{\phi_\eta^x - 2c^x}{4\lambda}.$$

As a consequence,

$$I^x = I_{NT}^x = \left[\frac{\phi_\eta^x - 2c^x}{4\lambda}\right]^+.$$

F.2 Market Tipping $T^x = 1$

Consider $T^x = 1$. The market tips in $t = 2$, i.e. $T^x = 1$, and $-x$ exits if and only if

$$\max_{D^{-x}\in[0,\eta^x I^x]}\left(\frac{1}{2}\Big(Z - \Delta_K^x + \phi^{-x}D^{-x} - \phi^x I^x\Big) - c^{-x}D^{-x}\right) \leq 0.$$

This condition is equivalent

$$\phi^x I^x \geq Z - \Delta_K^x + \eta^x I^x[\phi^{-x} - 2c^{-x}]^+ \quad \Longleftrightarrow \quad I^x \geq \frac{Z - \Delta_K^x}{\phi^x - \eta^x[\phi^{-x} - 2c^{-x}]^+} =: \underline{I}^x,$$

with $\Phi_\eta^x := \phi^x - \eta^x[\phi^{-x} - 2c^{-x}]^+$.

Suppose that indeed the market tips upon $Z_1^x = Z$, i.e., $I^x \geq \underline{I}^x$. Then, conditional on $Z_1^x = Z \iff N_1^x = 1$, firm x becomes monopolist in period $t = 2$ and charges optimally $P_2^x = Y_2^x + Z_2^x$, realizing time-2 profits $\pi_2^x = P_2^x = K^x + Z_2^x + \phi^x I^x$.

Conditional on market tipping in $t = 2$ upon $Z_1^x = Z$ (i.e., $T^x = 1$), time-1 expected payoff for firm x reads

$$\mathbb{E}_1[\pi_1^x] = \frac{1}{2}\left(\Delta_K^x - (c^x I^x - c^{-x} I^{-x}) + \frac{3Z}{2} + K^x + \phi^x I^x\right)\Big)$$
$$+ \frac{1}{4}\left[\Delta_K^x + Z + I^{-x}(\eta^x[\phi^x - 2c^x]^+ - \phi^{-x})\right](1 - T^{-x}).$$

Note that according to (A.2), $\Delta_K^x - (c^x I^x - c^{-x} I^{-x}) > 0$.

Firm x maximizes $\mathbb{E}_1[\pi_1^x] - \frac{\lambda(I^x)^2}{2}$ over $I^x \geq \underline{I}^x$ at the beginning of period $t = 1$ (taking I^{-x} as given). If $I^x \in (\underline{I}^x, \overline{I})$, this leads to the first-order-condition (FOC):

$$\frac{1}{2}(-c^x + \phi^x) - \lambda I^x = 0 \quad \Longleftrightarrow \quad I^x = \frac{\phi^x - c^x}{2\lambda}.$$

Note that (A.1) ensures $I^x < \overline{I}$. Further, $\underline{I}^x < \overline{I}$ by (A.1).

As a result,

$$I^x = I_T^x = \max\left\{\frac{\phi^x - c^x}{2\lambda}, \underline{I}^x\right\},$$

with $I_T^x < \overline{I}$. Under market tipping upon $Z_1^x = Z$, the payoff becomes

$$\Pi_T^x := \max_{I^x \geq \underline{I}^x}\left(\mathbb{E}_1[\pi_1^x] - \frac{\lambda(I^x)^2}{2}\right).$$

F.3 Optimal Choice

Taking the choice of I^{-x} by the other firm $-x$ as given, firm x's payoff becomes

$$\Pi_1^x = \max\{\Pi_{NT}^x, \Pi_T^x\}$$

When $\Pi_{NT}^x \geq \Pi_T^x$, then the market does not tip upon $Z_1^x = Z$; ties are broken in favor of no tipping. Otherwise, the market tips.

Given I^{-x}, we obtain that $\Pi_{NT}^x \gtrless \Pi_T^x$ is equivalent to

$$\frac{1}{2}\left(-c^x I_{NT}^x + \frac{1}{2}\Big(I_{NT}^x \phi_\eta^x + \Delta_K^x + Z\Big)\right) - \frac{\lambda(I_{NT}^x)^2}{2}$$
$$\gtrless \frac{1}{2}\left(-c^x I_T^x + K^x + \frac{Z}{2} + \phi^x I_T^x\right) - \frac{\lambda(I_T^x)^2}{2}$$

which simplifies to

$$\lambda(I_{NT}^x)^2 \gtrless \frac{K^A + K^B}{2} + (\phi^x - c^x) I_T^x - \lambda(I_T^x)^2.$$

Thus, the market tips if and only if

$$G(c^x) = \frac{K^A + K^B}{2} + (\phi^x - c^x) I_T^x - \lambda (I_T^x)^2 - \lambda (I_{NT}^x)^2 > 0.$$

F.4 Data Union and Data Trust

F.4.1 User Welfare

We first calculate user welfare as

$$\begin{aligned} U :=& \frac{1}{2}\Bigg[K^{-x} - c^{-x} I^{-x} + (1 - T^x)\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}\eta^x I^x \mathbb{I}_{\{\phi^{-x} \geq 2c^{-x}\}}) + \frac{1}{2}(K^x + \phi^x I^x)\Big\}\Bigg] \\ &+ \frac{1}{2}\Bigg[K^x - c^x I^x + (1 - T^{-x})\Big\{\frac{1}{2}(K^{-x} + \phi^{-x} I^{-x}) + \frac{1}{2}(K^x + \eta^{-x}\phi^x I^{-x}\mathbb{I}_{\{\phi^x \geq 2c^x\}})\Big\}\Bigg]. \end{aligned}$$

We can rewrite this as

$$\begin{aligned} U :=& (K^A + K^B)\left(1 - \frac{T^A + T^B}{4}\right) \\ &+ \frac{1}{2}\Bigg[- c^{-x} I^{-x} + \frac{(1 - T^x)}{2}\Big\{\phi^{-x}\eta^x I^x \mathbb{I}_{\{\phi^{-x} \geq 2c^{-x}\}} + \phi^x I^x\Big\}\Bigg] \\ &+ \frac{1}{2}\Bigg[- c^x I^x + \frac{(1 - T^{-x})}{2}\Big\{\phi^{-x} I^{-x} + \eta^{-x}\phi^x I^{-x}\mathbb{I}_{\{\phi^x \geq 2c^x\}}\Big\}\Bigg]. \end{aligned} \tag{F.1}$$

F.4.2 Auxiliary Result

Define $\phi_\eta^x = \phi^x - \eta^x \phi^{-x} \mathbb{I}_{\{\phi^{-x} \geq 2c^{-x}\}}$ and $\Phi_x^\eta := \phi^x - \eta^x[\phi^{-x} - 2c^{-x}]^+$. Recall that when the market tips upon $Z_1^x = Z$ and $T^x = 1$, data investment satisfies $I^x = I_T^x := \max\left\{\frac{\phi^x - c^x}{2\lambda}, \frac{(Z - \Delta_K^x)}{\Phi_\eta^x}\right\}$. When the market does not tip upon $Z_1^x = Z$ and $T^x = 0$, data investment satisfies $I^x = I_{NT}^x := \frac{[\phi_\eta^x - 2c^x]^+}{4\lambda}$.

Consider parameters such that

$$\frac{(Z - \Delta_K^x)}{\phi^x - \eta[\phi^{-x} - 2c^{-x}]^+} \geq \frac{\phi_\eta^x - 2c^x}{4\lambda} \quad \Longleftrightarrow \quad c^x > \frac{\phi_\eta^x}{2} - \frac{2\lambda(Z - \Delta_K^x)}{\Phi_x^\eta}, \tag{F.2}$$

which is met for c^x sufficiently large.

Thus, $I_T^x = \frac{Z - \Delta_K^x}{\Phi_x^\eta}$ and $G(c^x)$ can be rewritten as

$$G(c^x) := \frac{K^A + K^B}{2} + \frac{(\phi^x - c^x)(Z - \Delta_K^x)}{\Phi_\eta^x} - \frac{\lambda((Z - \Delta_K^x))^2}{(\Phi_\eta^x)^2} - \frac{([\phi_\eta^x - 2c^x]^+)^2}{16\lambda} > 0,$$

Clearly, when $c^x > \frac{\phi_\eta^x}{2}$, then $G'(c^x) < 0$.

Next, consider $c^x < \frac{\phi_\eta^x}{2}$. Then, we calculate

$$G'(c^x) = -\frac{Z - \Delta_K^x}{\Phi_x^\eta} + \frac{\phi_\eta^x - 2c^x}{4\lambda} = I^x - \underline{I}^x.$$

In particular, we have that $G(c^x)$ strictly decreases when $\underline{I}^x = \frac{(Z-\Delta_K^x)}{\Phi_\eta^x} \geq \frac{\phi^x - c^x}{2\lambda}$, i.e., $c^x \geq \phi^x - \frac{2\lambda(Z-\Delta_K^x)}{\Phi_\eta^x}$, with $\lim_{y\uparrow c^x} G'(y) < 0$ and $\lim_{y\downarrow c^x} G'(y) > 0$ as well as $G'(c^x) < 0$ whenever $G(c^x)$ is differentiable.

Next, we evaluate

$$\begin{aligned} G\left(\phi_\eta^x - \frac{2\lambda(Z-\Delta_K^x)}{\Phi_\eta^x}\right) &= \frac{K^A + K^B}{2} + \frac{\lambda(Z-\Delta_K^x)^2}{(\Phi_\eta^x)^2} - \frac{([\phi_\eta^x - 2c^x]^+)^2}{16\lambda} \\ &= \frac{K^A + K^B}{2} + \frac{(\phi^x - c^x)^2}{4\lambda} - \frac{([\phi_\eta^x - 2c^x]^+)^2}{16\lambda} \\ &> 0. \end{aligned}$$

As $G(c^x)$ is continuous on $\mathcal{S} := \left(\phi^x - \frac{2\lambda(Z-\Delta_K^x)}{\Phi_\eta^x}, \infty\right)$, it follows that $G(c^x)$ has a unique root on $\mathcal{S}$ denoted c_*^x with $G(c_*^x) = 0$.

We can solve $G(c_*^x) = 0$ in closed form. When $c_*^x \leq \phi_\eta^x/2$, then

$$c_*^x = \frac{\phi_\eta^x}{2} - \frac{2\lambda(Z-\Delta_K^x)}{\Phi_\eta^x} + \left[2\lambda\left(\frac{2(Z-\Delta_K^x)\phi^x - (Z-\Delta_K^x)\phi_\eta^x + (K^A+K^B)\Phi_\eta^x}{\Phi_\eta^x}\right)\right]^{\frac{1}{2}}.$$

Otherwise, when $c_*^x > \phi_x^\eta/2$, we have

$$\begin{aligned} c_*^x &= \frac{\Phi_\eta^x}{Z-\Delta_K^x}\left(\frac{K^A+K^B}{2} + \frac{\phi^x(Z-\Delta_K^x)}{\Phi_\eta^x} - \frac{\lambda(Z-\Delta_K^x)^2}{(\Phi_x^\eta)^2}\right) \\ &= \phi^x + \frac{\Phi_\eta^x(K^A+K^B)}{2(Z-\Delta_K^x)} - \frac{\lambda(Z-\Delta_K^x)}{\Phi_\eta^x}. \end{aligned}$$

F.4.3 Union Optimization

For $c^x = c + \tau^x$, choosing c^x is equivalent to choosing τ^x, subject to ensuring $\tau^x \in [\underline{\tau}, \overline{\tau}]$. Suppose that in optimum $\tau^x \in (\underline{\tau}, \overline{\tau})$. As such, the data union maximizes

$$\max_{c^A, c^B}\left(U + \frac{I^x}{2}\left[\tau^x + (1-T^x)\tau^{-x}\eta^x \mathbb{I}_{\{\phi^{-x}\geq 2c^{-x}\}}\right] + \frac{I^{-x}}{2}\left[\tau^{-x} + (1-T^{-x})\tau^x\eta^{-x}\mathbb{I}_{\{\phi^x \geq 2c^x\}}\right]\right).$$

This is akin to solving $\max_{c^A,c^B} \overline{U}$ with

$$\begin{aligned}\overline{U} :=& \frac{1}{2}\Big[K^{-x} - cI^{-x} + (1-T^x)\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}\eta^x I^x \mathbb{I}_{\{\phi^{-x}\geq 2c^{-x}\}}) + \frac{1}{2}(K^x + \phi^x I^x)\Big\}\Big] \\ &+ \frac{1}{2}\Big[K^x - cI^x + (1-T^{-x})\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}I^{-x}) + \frac{1}{2}(K^x + \eta^{-x}\phi^x I^{-x}\mathbb{I}_{\{\phi^x\geq 2c^x\}})\Big\}\Big] \\ &+ \frac{1}{2}\Big[(1-T^x)(c^{-x}-c)\eta^x I^x \mathbb{I}_{\{\phi^{-x}\geq 2c^{-x}\}} + (1-T^{-x})(c^x - c)\eta^{-x}I^{-x}\mathbb{I}_{\{\phi^x\geq 2c^x\}}\Big]\end{aligned}$$

It is clear that the data union sets (c^A, c^B) so that there is no market tipping, which requires that

$$\frac{Z-\Delta_K^x}{\Phi_\eta^x} > \frac{\phi^x - c^x}{2\lambda} \quad \iff \quad c^x > \phi^x - \left(\frac{2\lambda(Z-\Delta_K^x)}{\Phi_\eta^x}\right).$$

Furthermore, for no tipping to occur, it must be that $G(c^x) \leq 0$. As we now from the previous part, $G(c^x)$ strictly decreases in c^x for $c^x > \phi^x - \left(\frac{2\lambda(Z-\Delta_K^x)}{\Phi_\eta^x}\right)$ and has a unique root c_*^x. As such, market tipping does not occur if $c^x \geq c_*^x$.

When $T^x = 0$ and $I^x = \frac{\phi_\eta^x - 2c^x}{4\lambda} > 0$, then $c^x = 0.5\phi_\eta^x - 2\lambda I^x$ and $\frac{\partial c^x}{\partial I^x} = -2\lambda$ with a slight abuse of notation. Holding all else equal and for $\phi^x > 2c^x$, the partial derivative of $\overline{U}$ with respect to I^x becomes:

$$\begin{aligned}\frac{\partial \overline{U}}{\partial I^x} = & \frac{1}{2}\left[-c + \frac{(1-T^x)}{2}\big(\phi^x + \eta^x\phi^{-x}\mathbb{I}_{\{\phi^{-x}\geq 2c^{-x}\}}\big)\right] \\ &+ \frac{(1-T^x)}{2}(c^{-x}-c)\eta^x \mathbb{I}_{\{\phi^{-x}\geq 2c^{-x}\}} - \frac{(1-T^{-x})}{2}I^{-x}\eta^{-x}\mathbb{I}_{\{\phi^x\geq 2c^x\}}2\lambda.\end{aligned}$$

As there is no market tipping, we have $T^x = T^{-x} = 1$.

We focus on ϕ^x large enough so that investment is sufficiently efficient, and $\frac{\partial \overline{U}}{\partial I^x} > 0$. The baseline also focuses on this scenario. Under these circumstances, we obtain that data union optimally sets $\tau^x = c_*^x - c$ (provided $\tau^x \in (\underline{\tau}, \overline{\tau})$), as in the baseline. The data union therefore implements maximum data collection subject to preventing market tipping, so that the optimal data price c_*^x solves $G(c_*^x) = 0$. The optimal price for data c_*^x is available in closed-form; see Internet Appendix F.4.2.

G Firm-Centric Data Markets: Proof of Proposition 4

Suppose that $N_1^x = 1$, i.e., $Z_1^x = Z$, and the investment level of firm x is $I^x > 0$. Thus, $\hat{D}^x = I^x > 0 = \hat{D}^{-x}$.

To begin with, consider arbitrary D^x, D^{-x} and suppose the market does not tip in $t = 2$. Then,

$$\overline{\pi}_2^x(D^x, D^{-x}) = \frac{1}{2}\left(\Delta_K^x + Z + \phi^x D^x - \phi^{-x}D^{-x}\right)$$

and

$$\pi_2^{-x}(D^{-x}, D^x) = \frac{1}{2}\left(-\Delta_K^x + Z + \phi^{-x}D^{-x} - \phi^x D^x\right)$$

so that

$$\pi_2^x(D^x, D^{-x}) + \pi_2^{-x}(D^{-x}, D^x) = Z.$$

This result holds regardless of whether there is data trade.

First, consider that there is no data trade, in that $D^x = I^x > 0 = D^{-x}$. If the market tips and $-x$ exits in $t = 2$, we have $\pi_2^{-x}(D^{-x}, D^x) = 0$ and

$$\pi_2^x(D^x, D^{-x}) + \pi_2^{-x}(D^{-x}, D^x) = \pi_2^x(D^x, D^{-x}) = K^x + \phi^x I^x + \frac{Z}{2}.$$

In addition, the condition for market tipping must hold in that

$$\phi^x I^x \geq Z - \Delta_K^x. \tag{G.1}$$

This condition implies that

$$K^x + \phi^x I^x + \frac{Z}{2} \geq K^{-x} + \frac{3Z}{2} > Z.$$

If there is no market tipping, we have $\pi_2^x(D^x, D^{-x}) + \pi_2^{-x}(D^{-x}, D^x) = Z$, which is strictly lower than firms' joint payoff under market tipping whenever (G.1) holds.

Second, consider there is a data trade, in that $D^x = I^x$ and $D^{-x} \in (0, \eta I^x]$. A data trade (from x to $-x$) is feasible, in that it strictly increases firms' joint payoff, only if $-x$ operates in $t = 2$. Suppose to the contrary that post-trade, the market does not tip and both firms operate, so that $\pi_2^x(D^x, D^{-x}) + \pi_2^{-x}(D^{-x}, D^x) = Z$. As a data trade occurs only if it strictly increases firms' payoff, it must be that absent a data trade, the market would have tipped in that $-x$ would have exited. That is, (G.1) holds. In addition, the pre-trade surplus $K^x + \phi^x I^x + \frac{Z}{2}$ must be lower than Z. However, we have according to (G.1):

$$K^x + \phi^x I^x + \frac{Z}{2} \geq K^{-x} + \frac{3Z}{2} > Z,$$

a contradiction. It follows that the market tips post-trade, i.e., a data trade occurs if and only if it is followed by market tipping and exit by firm x in $t = 2$.

A data trade (from x to $-x$) is only feasible, in that it strictly increases firms' joint payoff, if $-x$ operates in $t = 2$. Thus, if the market tips post-trade, then it must be that firm x — instead of $-x$ — exits. This requires that

$$\Delta_K^x + Z + \phi^x I^x \leq \phi^{-x} D^{-x},$$

with $D^{-x} \leq \eta I^x$. A necessary condition for market tipping post-trade and data trade is therefore

$$\eta \phi^{-x} \geq \phi^x + \frac{\Delta_K^x + Z}{I^x} > \phi^x + \frac{\Delta_K^x + Z}{\bar{I}}.$$

Furthermore, firms' joint payoff is then $\overline{\pi}_2^{-x}(D^{-x}, I^x) = K^{-x} + \phi^{-x} D^{-x} + \frac{Z}{2}$. This joint payoff increases in D^{-x} and is maximized for $D^{-x} = \eta I^x$.

H Total Surplus

We calculate total surplus. For this sake, define

$$\begin{aligned}\mathcal{S}^x := \frac{1}{2}\Bigg[&K^x - c^x I^x + Z + T^x \left(K^x + \phi^x I^x + \frac{Z}{2}\right) \\ &+ \frac{(1-T^x)}{2}\left(K^x + K^{-x} + 2Z + \phi^x I^x + \phi^{-x}\eta^x I^x\right)\Bigg].\end{aligned}$$

Total surplus reads then

$$\mathcal{S} := \sum_{x=A,B}\left[\mathcal{S}^x - \frac{\lambda (I^x)^2}{2}\right]. \tag{H.1}$$

We can write $\mathcal{S}$ as a function of T^x, T^{-x}, i.e., $\mathcal{S} = \mathcal{S}(T^x, T^{-x})$. When $T^x = 1$ and there is market tipping, we have $I^x = I_{NT}^x$. When $T^x = 0$ and there is no market tipping, we have $I^x = I_T^x$, with I_T^x, I_{NT}^x characterized in Proposition 1.

H.1 Surplus-maximizing Investment

We now look for the surplus-maximizing investment level $\mathcal{I}^x = \arg\max_{I^x \in [0,\bar{I}]} \mathcal{S}$. Suppose $\phi^x > 2c^x$ (data collection is efficient).

We calculate $\mathcal{I}^x$, holding T^x fixed. First, fix $T^x = 0$. Then, the first-order condition $\frac{\partial \mathcal{S}}{\partial I^x} = 0$ becomes

$$-\frac{c^x}{2} + \frac{\phi^x + \phi^{-x}\eta^x}{4} - \lambda I^x = 0,$$

so that

$$\mathcal{I}^x = \mathcal{I}_{NT}^x := \frac{\phi^x + \eta^x \phi^{-x} - 2c^x}{4\lambda}.$$

When $\eta^x > 0$, $\mathcal{I}_{NT}^x$ is larger than the privately optimal investment level under $T^x = 0$ — which equals $I_{NT}^x = \frac{[\phi^x - \eta^x \phi^{-x} - 2c^x]^+}{4\lambda}$ (see Proposition 1). When $\eta^x > 0$ and $\mathcal{I}_{NT}^x > I_{NT}^x$, individual firms do not take into account the surplus created through data sharing, causing data undercollection (underinvestment). Note that, in addition to data sharing, $\eta^x > 0$ could also capture technological spillovers, in that improvements in x's service quality spill over to improves $-x$'s service too. This effect is not necessarily related to data sharing. Thus, even in the absence of data sharing, $\eta^x > 0$ seems empirically plausible, leading to data undercollection (underinvestment).

Second, fix $T^x = 1$. Then, the first-order condition $\frac{\partial \mathcal{S}}{\partial I^x} = 0$ becomes $\frac{\phi^x - c^x}{2} - \lambda I^x = 0$, so

$$\mathcal{I}^x = \mathcal{I}^x_T = \frac{\phi^x - c^x}{2\lambda},$$

which is weakly larger than I^x_T from Proposition 1.

H.2 Total Surplus and Market Tipping

Holding all else equal, market tipping upon $Z^x_1 = Z$, i.e., $T^x = 1$, leads to lower surplus than no market tipping $T^x = 0$ if $\mathcal{S}(0, T^{-x}) > \mathcal{S}(1, T^{-x})$. This is the case if

$$\frac{Z - \Delta^x_K}{2} + \frac{1}{2}(\phi^x + \eta^x \phi^{-x} - 2c^x) I^x_{NT} - \lambda (I^x_{NT})^2 > (\phi^x - c^x) I^x_T - \lambda (I^x_T)^2. \tag{H.2}$$

We can rewrite (H.2) as

$$\frac{Z - \Delta^x_K}{2} + \lambda (I^x_{NT})^2 + \frac{\eta^x \phi^{-x} I^x_{NT}}{2} > (\phi^x - c^x) I^x_T - \lambda (I^x_T)^2. \tag{H.3}$$

Assume $\phi^x - c^x \geq 0$. The right-hand-side is maximized for $I^x_T = \mathcal{I}^x_T = \frac{\phi^x - c^x}{2\lambda}$. Thus, a sufficient condition for (H.3) is given by

$$\frac{Z - \Delta^x_K}{2} + \frac{\left([\phi^x_\eta - 2c^x]^+\right)^2 + 2\eta^x \phi^{-x} [\phi^x_\eta - 2c^x]^+}{16\lambda} > \frac{(\phi^x - c^x)^2}{4\lambda}. \tag{H.4}$$

Finally, a sufficient condition for (H.4) and thus for (H.3) is

$$\frac{Z - \Delta^x_K}{\phi^x - c^x} \geq \frac{\phi^x - c^x}{2\lambda} \geq I^x_T,$$

with I^x_T from (1). That is, market tipping is inefficient in terms of total surplus when Z is large, Δ^x_K is low, c^x is large, or the cost of data collection λ is large.

H.3 Data Under- and Overcollection

We assume that (H.2) is met, in that market tipping is inefficient in terms of total surplus. Then, as in the baseline, data overcollection ensues when too much data collection causes market tipping, i.e., when $I^x \geq \underline{I}^x$.

On the other hand, when the market does not tip and $\eta^x > 0$, we have data undercollection as $I^x < \mathcal{I}^x_{NT}$. That is, data collection investments by firms and data contributions by users are inefficiently low in terms of total surplus.

These takeaways are similar to the ones obtained in the baseline and main text, when we consider user welfare (instead of total surplus) as the "objective," with both inefficient data under- and overcollection arising in equilibrium.

H.4 Data Union and Surplus Maximization

We now analyze a data union that aims to maximize total surplus and show that our results are qualitatively similar to our baseline analysis that considers a data union that maximizes user welfare.

For this sake, consider (as in the baseline) that the privacy cost is constant c, while the data union taxes data contributions at (potentially negative) rate $\tau^x \gtrless 0$. We write for the effective privacy cost $c^x = c + \tau^x$, and assume $2c < \phi^x$ for $x = A, B$.

The union chooses (τ^A, τ^B) or equivalently (c^A, c^B) to maximize

$$\overline{\mathcal{S}} := \mathcal{S} + \mathbb{E}\Big[\tau^A N_1^A I^A + \tau^B N_1^B I^B\Big],$$

where the equilibrium, given union intervention τ^x and c^x, is summarized in Proposition 1. We can rewrite

$$\overline{\mathcal{S}} := \sum_{x=A,B} \left[\overline{\mathcal{S}}^x - \frac{\lambda (I^x)^2}{2}\right], \tag{H.5}$$

where

$$\overline{\mathcal{S}^x} = \frac{1}{2}\Bigg[K^x - cI^x + Z + T^x\left(K^x + \phi^x I^x + \frac{Z}{2}\right) + \frac{(1-T^x)}{2}\left(K^x + K^{-x} + 2Z + \phi^x I^x + \phi^{-x}\eta^x I^x\right)\Bigg].$$

We focus on parameters such that (H.2) is satisfied (under $c^x = c$), i.e., market tipping is inefficient when it comes to total surplus. A conservative, sufficient condition for this to be the case is

$$\frac{Z - \Delta_K^x}{\phi^x - c} \geq \frac{\phi^x - c}{2\lambda}.$$

The surplus-maximizing investment, conditional on no market tipping, is then $\mathcal{I}^* = \arg\max_{I^x} \overline{\mathcal{S}}$, with $I^x = \mathcal{I}^* := \frac{\phi^x + \eta^x \phi^{-x} - 2c}{4\lambda}$.

Recall that there is no market tipping and $T^x = 0$ if $G(c^x) \leq 0$, with $G(c^x)$ from Proposition 1. Denote the root of $G(c^x)$ for $c^x > \phi^x - \frac{2\lambda(Z - \Delta_K^x)}{\phi_\eta^x}$ by c_*^x, which is available in closed-form in Proposition 2.

As market tipping is inefficient (in terms of total surplus), data union implements $I^x = \mathcal{I}^*$ whenever $G(c^x) < 0$ (under optimal c^x), while otherwise $G(c^x) = 0$. Then, the optimal price of data c_S^x solves

$$c_S^x = \max\{c_*^x, c_{**}^x\}$$

where c_{**}^x is such that I_{NT}^x equals $\mathcal{I}^*$. That is, solve

$$\frac{\phi^x + \eta^x \phi^{-x} - 2c}{4\lambda} = \frac{\phi_\eta^x - 2c^x}{4\lambda}$$

for $c^x = c_{**}^x$ to obtain $c_{**}^x = c - \eta^x \phi^{-x}$. Therefore, the optimal *data price* satisfies

$$c_S^x = \max\{c_*^x, c - \eta^x \phi^{-x}\}.$$

As in the baseline, optimal data price $c^x = c_S^x$ may lie above or below c. Thus, the data union may tax or subsidize data contributions to curb or stimulate data collection respectively. To prevent market tipping, it might be necessary to raise $c^x = c_S^x$ above c by setting $\tau^x > 0$ — which happens when $c_*^x > c$. Then, data overcollection and market tipping ensue in laissez faire. On the other hand, when $c_*^x < c$, data union stimulates data collection to alleviate data undercollection. As such, the optimal policy is qualitatively similar, both when the union maximizes total surplus or user welfare. Therefore, we view our key takeaways regarding optimal data union or data trust as robust to the specific objective.

The key difference is that when maximizing user welfare, the data union implements maximum data collection subject to preventing market tipping, leading to the optimal data price c_*^x. When maximizing total surplus, the data union implements *efficient* data collection $\mathcal{I}_{NT}^x$ subject to preventing market tipping, so that $c_S^x > c_*^x$ is possible.

Finally, note that total surplus is the (equally weighted) sum of firms' and users' payoffs. By extension, our results would likely not change significantly, if we considered a different objective function (potentially one that overweights user welfare).

I Data Union and Non-Discriminatory Pricing

We now impose the constraint that data union must set $c^A = c^B$. Thus, data union maximizes $\max_{c^A, c^B} \overline{U}$ subject to $c^A = c^B$ and with

$$\begin{aligned}\overline{U} :=& \frac{1}{2}\Big[K^{-x} - cI^{-x} + (1-T^x)\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}\eta^x I^x) + \frac{1}{2}(K^x + \phi^x I^x)\Big\}\Big] \\ &+ \frac{1}{2}\Big[K^x - cI^x + (1-T^{-x})\Big\{\frac{1}{2}(K^{-x} + \phi^{-x}I^{-x}) + \frac{1}{2}(K^x + \eta^{-x}\phi^x I^{-x})\Big\}\Big].\end{aligned}$$

Denote the maximand by $c_* = c^A = c^B$. Consider for simplicity that the solution to the optimization is well-behaved. Absent the constraint $c^A = c^B$, data union sets $c^x = c_*^x$ (see Proposition 2). Conditional on $T^A = T^B = 0$, lowering $c^A = c^B$ increases I^A and I^B and increases user welfare. Thus, data union would like to maximize data investments, while limiting users' losses from market tipping.

Consider now without loss of generality that $c_*^A \leq c_*^B$; otherwise, we relabel the firms. It is clear that $c_* > c_*^B$ is suboptimal, as the union could lower c_* to strictly increase investments without precipitating market tipping. Thus, $c_* \leq c_*^B$. Next, $c_* < c_*^A$ causes market tipping with probability one and cannot be optimal, so $c_* \geq c_*^A$. Setting $c_* \in (c_*^A, c_*^B)$ causes market tipping upon $Z_1^B = Z$. However, conditional on $T^B = 1$, it is optimal to set $c_* = c_*^A$. In particular, when $c_* \in (c_*^A, c_*^B)$ and $T^B = 1$, union could increase user welfare by lowering c_* to increase I^A without precipitating $T^A = 1$. As such, $c^* \in \{c_*^A, c_*^B\}$.

To determine c_*, one would need to compute welfare in these two scenarios — i.e., $c_* = c_*^A$ or $c_* = c_*^B$ — and check which choice of c_* leads to higher user welfare. As market tipping leads to the (extreme) outcome that users realize zero payoff in period 2, we expect that under reasonable parameter conditions, $c_* = c_*^B \geq c_*^A$, in that market tipping is inefficient.

For the purpose of our analysis, we therefore (implicitly) focus on the case that market tipping is inefficient. Consequently, the price of data becomes $c_* = \max\{c_*^A, c_*^B\}$, when the data union cannot discriminate between firms. In this case, the tradeoffs underlying the determination of optimal data price remain qualitatively similar as in the baseline. In particular, the optimal data price reflects the trade-off between stimulating data collection for service improvements and curbing data collection for enhancing competition.

J Incentive Compatibility of Data Union

To heuristically study an individual user's incentives to join the union, suppose that an individual deviant user decides not to join the union at the onset at time $t = 1$ (before Z_1 is realized and taking I^x, p_1^x, d^x as given) — which does not change equilibrium outcomes as the user is atomistic. We assume that in case of staying out of the union, this user cannot contribute data to firms, e.g., because the data union provides the necessary infrastructure or the firm only gathers data union user data. If a user in the data union contributes a unit of data, this user is paid $d^x = c + \tau^x$ per unit data, but must pay the tax τ^x and incurs the privacy cost c, leaving zero net payoff from contributing data and implying indifference between contributing data or not.

Accordingly, any user makes zero net profits from contributing data (regardless of union membership), also implying that both deviant and union members make the same buying decisions. Provided x wins period 1, the additional payoff from joining the union is then $\tau^x I^x$ — the rebate from the union to users. Essentially, when $\tau^x > 0$ ($\tau^x < 0$), the data union controls firms' investment incentives by implementing a tax (subsidy) from firm x to users, amounting to total proceeds $N_1^x I^x \tau^x$ distributed to users (financed by users if $\tau^x < 0$). As both firms win with equal probability 1/2, the overall additional payoff from joining the union is $\mathcal{U} := \frac{I^A\tau^A + I^B\tau^B}{2}$.

Note now that the deviant is worse off (better off) staying out of the union when $\mathcal{U} \geq 0$ ($\mathcal{U} < 0$). Consequently, users would voluntarily join the data union, if the data union (on average) taxes data collection, e.g., $\mathcal{U} \geq 0$. However, when $\mathcal{U} < 0$, an individual user is better off not joining the union, as this requires to contribute financially to the subsidy to firms.

Thus, users may not find it optimal to join the union when the union subsidizes data collection. This incentive compatibility problem can be resolved by either requiring users to be represented by the data union or by having a (benevolent) third party (e.g., the government) come up with the cost of subsidizing data collection, so that the data union is no more required to be budget-neutral. We leave the further study of this topic to future research.

K Discussion of Modelling Assumptions

To afford maximum theoretical clarity and to obtain analytical closed-form solutions, we purposefully develop a stylized model which tractably captures the key economic mechanisms that we would like to highlight. We now elaborate on several modelling assumptions and model robustness in greater detail.

Shock Structure. Shocks to service quality, randomizing the relative competitive advantage between A and B, are needed to get meaningful competition and trade-offs with homogenous customers (leading to asymmetric Bertrand competition). Without shocks, the outcome of the competition would be deterministic and the "losing" firm has no incentive to invest in data collection, leading to less interesting equilibrium outcomes. However, with uncertainty and shocks to service qualitt, both firms have incentives to invest, allowing us to study how data collection investments are affected by different types of regulation in a tractable setting.

The shock structure assumes that $Z_t = (Z_t^A, Z_t^B) \in \{(Z, 0), (0, Z)\}$, while Z_t^x takes values Z and 0 with equal probability 1/2. This specific structure is merely chosen for simplicity. Similar findings would arise under a different shock structure. For instance, we could assume that Z_t^x and Z_t^{-x} are i.i.d. and take values Z and 0 with equal probability. In this case, we would need to distinguish four cases, namely $(Z_t^A, Z_t^B) \in \{(0, 0), (Z, 0), (0, Z), (Z, Z)\}$, which would not affect the key results but make the analysis more cumbersome. Likewise, we could assume that Z_t^x is a continuous random variable and/or can take negative values.

User Homogeneity. To obtain closed-form solutions and to highlight the key economic mechanisms analytically, we stick to a setting with homogeneous users and stochastic service quality. While such modelling of price competition, to the best of our knowledge, is rather uncommon, we note that similar findings would arise in a setting where customers are located on a Hotelling line. In fact, a subsumed version of this paper relied on such modelling and obtained similar key finding, although some of the analysis relied on numerical solutions and thus was significantly less tractable. See Cong and Mayer (2023) for details. We view our modelling of price competition with user homogeneity and shocks to service quality as an improvement, since it affords more theoretical clarity and tractability.

Market Tipping. We assume that a firm exits in $t = 2$, when it cannot win the market anymore, which happens because the competing firm has gained a large advantage from data generated in $t = 1$. This assumption is a simple way to capture that reduced competition, due to large (data-driven) differences in service quality, can harm users. The assumption that firms exit unless they derive strictly positive (expected) payoffs from operating could be micro-founded by incorporating a fixed cost of operations. Intuitively, we make this assumption to capture the tension between data and competition (Farboodi et al., 2019).

Without this assumption, we would always have (duopolistic) asymmetric Bertrand competition in period 2, with both firms competing. Consider that both firms (always) operate in $t = 2$. Then, even if a firm has zero chance of winning the market in $t = 2$, it would stand by and provide services at zero cost. If firm $x = A$ wins in period 2 — which would occur with probability one if $\Delta_K^A + D^A\phi^A - D^B\phi^B - Z > 0$ — then $x = B$ charges price zero and users' payoff becomes Y_2^B, while firm A charges $P_2^A = Y_2^A - Y_2^B$. User's payoff in this event is Y_2^B. An increase in D^A would leave users' payoff unchanged or increase if $\frac{\partial D^B}{\partial D^A} > 0$ (due to data sharing).

Overall, users' expected payoff in period $t = 2$ could then be written as

$$Y_2^A N_2^B + Y_2^B N_2^A = (K^A + \phi^A D^A + Z_2^A) N_2^B + (K^B + \phi^B + Z_2^B) N_2^A.$$

This payoff (at least weakly) increases in D^A, D^B, so more data and more data investment would unambiguously benefit users. More pressing, under this stipulation when $N_2^A = 1$ with probability one, reduced competition due to an increase in Y_2^A neither affects prices nor user welfare.

The assumption that the market tips is therefore necessary to capture the tension between data and competition (and that lack of competition can harm users). One could likely model such adverse effects of reduced (data-driven) competition in a different way, without changing the key findings.

User Welfare. In the main text, we use user welfare as the objective when evaluating the efficiency of equilibrium allocations. Notably, Internet Appendix H highlights that our findings remain qualitatively similar under this different objective function, namely total surplus (equally weighted sum of firms' and users' payoff). By extension, we expect qualitatively similar results also when considering a differently weighed average of firms' and users' payoffs (tilted toward users).

Non-negativity constraint on effective service price. We impose $P_t^x \geq 0$ to preclude degenerate pricing-equilibria which are supported by the losing firm pricing in a way that results in negative payoff if it instead won the market at the prices charged. A similar assumption is made in Hagiu and Wright (2021). Without the assumption $P_t^x \geq 0$, we would need an assumption of similar nature to preclude degenerate pricing equilibria.

A natural, alternative constraint would be that the losing firm's continuation payoff must be positive, if it won the customers. Let's sketch what happens under this alternative constraint. In period 2, we have $P_2^x = p_2^x$ and $\pi_2^x = N_2^x p_2^x$. Thus, the constraint $\pi_2^x \geq 0$ would simply imply $P_2^x \geq 0$, as is assumed in the main text.

In period 1, the losing firm may find it optimal to strategically underprice in period 1 by setting a negative price. To see this, suppose that $Z_1^A = Z$ and A wins the market in period 1 (in our baseline), with $P_1^A > 0$ and $P_1^B = 0$ while users are indifferent between buying at A or B. Then, in general, B could profitably deviate by further undercutting A through $P_1^B < 0$ — which would imply negative period-1 profits but also more data and higher period 2 (expected) profits.

In particular, denote by $\overline{\pi}_2^x(\hat{D}^A, \hat{D}^B)$ the expected period-2 payoff for firm x, when firm x $(-x)$ generates $\hat{D}^x$ $(\hat{D}^{-x})$ units of data. Clearly, when A wins, then $\hat{D}^A = I^A$ and $\hat{D}^B = 0$.

When A wins period 1, then the period-1 payoff for B is zero and the expected period-2 payoff (which can be zero if B exits) is $\overline{\pi}_2^B(I^A, 0)$. However, if B were to undercut A and win the market, B's period 1 payoff becomes P_1^B and period-2 payoff becomes $\overline{\pi}_2^B(0, I^B) \geq \overline{\pi}_2^B(I^A, 0)$. The lowest price P_1^B that B is willing to set is then

$$P_1^B = -\left[\overline{\pi}_2^B(0, I^B) - \overline{\pi}_2^B(I^A, 0)\right] \leq 0.$$

Under these circumstances, the analysis would become more complicated, while we expect the key findings to remain similar. We therefore impose $P_t^x \geq 0$.

Modeling of Data Sharing. One interpretation of our modeling of (mandatory) data sharing — i.e., $D^{-x} = \hat{D}^{-x} + \eta^x \hat{D}^x$ — is that due to regulation, x *must* share fraction η^x of their data with competitor firm $-x$. In general, $\eta^x < 1$, as not all data generated on x can be ported to or used by $-x$.

In practice, not all data shared by x with $-x$ is (equally) useful for $-x$, so η^x may also reflect such technological constraints. For instance, we could rewrite $D^x = \hat{D}^x + \hat{\eta}^{-x}\alpha^{-x}\hat{D}^{-x}$, where $\hat{\eta}^{-x}$ corresponds to the fraction of the data $\hat{D}^{-x}$ that $-x$ shares with x and α^{-x} captures to what extent these data improve x's service. We could then define $\eta^x = \alpha^x \hat{\eta}^x$. Thus, even if firms must share all data ($\hat{\eta}^x = 1$), the data of $-x$ may not be as useful to x in that $\alpha^x < 1$ and $\eta^x < 1$.

Finally, $\eta^x > 0$ could also capture data-related technological spillovers.

Across- and Within-User Data-Enabled Learning. Our formulation broadly describes different forms of data-enabled learning that improve services and products, such as within-user and across-user learning (Hagiu and Wright, 2021).

Specifically, in equilibrium, all (identical) users act symmetrically and can be aggregated to one representative user under the assumption that this representative user takes product prices and quality as given. The stipulation in (1) then captures that an individual user's data may improve both (i) its own service quality, e.g., due to better customization, ("within-user learning") and (ii) the service quality for others ("across-user learning"). See Hagiu and Wright (2021) for a discussion and examples of these two types of data-enabled learning.

Interpreted broadly, within-user learning could also be reflected in the parameter c^x. That is, $c^x\theta^x I^x$ may not only capture the adverse consequences of sharing data — such as loss of privacy or data-enabled price discrimination — but also potential positive effects (e.g., when $c^x < 0$) — such as improved customization.