Antitrust, Regulation, and User Union in the Era of Digital Platforms and Big Data[*]

Lin William Cong[†] Simon Mayer[§]

Initial draft: April 2022; current draft: November 2022.

Abstract

We model platform competition with endogenous data generation, collection, and sharing, thereby providing a unifying framework to evaluate data-related regulation and antitrust policies. Data are jointly produced from users' economic activities and platforms' investments in data infrastructure. Data improves service quality, causing a feedback loop that tends to concentrate market power. Dispersed users do not internalize the impact of their data contribution on (i) service quality for other users, (ii) market concentration, and (iii) platforms' incentives to invest in data infrastructure, causing inefficient over- or under-collection of data. Data sharing proposals, user privacy protections, platform commitments, and markets for data cannot fully address these inefficiencies. We introduce and analyze user union, which represents and coordinates users, as a potential solution for antitrust and consumer protection in the digital era.

JEL Classification: L10, L41, L50, O30

Keywords: Data Sharing, Data Trust, Digital Economy, Open Banking, Privacy.

[*] We are grateful to Zhiheng He, Shota Ichishashi, 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, Sebastian Gryglewicz, Andrei Hagiu, Jing Huang, Shiyang Huang, Laurence Lescourret, Dan Luo, Roxana Mihet, Christine Parlour, James Poterba, Raghuram Rajan, Huan Tang, Glen Weyl, Liyan Yang, Anthony Lee Zhang (discussant), 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 (Cergy-Pontoise), 2022 Artificial Intelligence and Financial Markets Conference in Oxford, CUHK Shenzhen, the 4th Greater China Area Finance Conference, HKUST Guangzhou, IGP Annual Conference on "From Internet Governance to Digital Political Economy," 2022 International Conference on FinTech and Digital Finance, KU Leuven, and the 2022 Yale Junior Finance Conference for constructive feedback. Zhiheng He and Qihong Ruan provided excellent research assistance.

[†] Cornell University Johnson Graduate School of Management and NBER. E-mail: will.cong@cornell.edu

[§] HEC Paris. Email: mayer@hec.fr

In the era of digital platforms and big data, customers and their data actively contribute to firms' production of goods and services. For instance, the data users generate affect future product innovation and improvements, and one user's adoption of a platform or product affects the other users' service utility (through content creation and network externalities), leading to dynamic data feedback and network effects that interact. These phenomena pose challenges to traditional antitrust laws (e.g., the Sherman Act) and recent policies targeting privacy protection and data sharing.[1] Meanwhile, computer and data scientists are actively exploring how privacy-perserving computation and user-centric data ownership can empower people and communities (e.g., Zyskind, Nathan, and Pentland, 2015; Pentland, Lipton, and Hardjono, 2021). To formulate appropriate antitrust policies, regulate data-driven platforms, and unleash the potential of big data, it is crucial to understand the economic forces underlying the new economy.

To this end, we build a tractable two-period model of firm/platform competition incorporating the defining features of a data economy and key channels through which users contribute to digital platforms or business ecosystems. Naturally, a platform offering better products (potentially due to existing data advantage) increases user adoption, spurs data collection, and enjoys a bigger advantage in future by utilizing data to improve its products. We further innovate by modeling the joint data projection by users and platforms as well as the dispersed nature of users relative to platforms, which allow us to capture for the first time the externalities of user participation and data contribution on platform service quality, market power, and incentives at the same time. We thus identify novel sources of inefficiencies that can lead to both under- and over-collection of data, and offer a unifying framework to analyze the effectiveness of a wide array of data-related antitrust and regulatory policies, including privacy protection regulation (e.g., GDPR or CCPA), data sharing initiatives (e.g., open banking), and approaches involving data markets and intermediaries. We introduce user union as a potential solution for antitrust and consumer protection, which is novel in the literature and, as we show, can better mitigate the inefficiencies than existing approaches.

In our model, two firms/platforms compete for users through price and potentially differ in

[1] Many companies increasingly rely on the platforms or ecosystems they foster and the world's largest publicly-traded companies (Amazon, Alphabet, Meta, Apple) are all (digital) platforms. Even conventional (non-platform) firms and manufacturers such as Tesla increasingly rely on userbase scale (e.g., more users leading to more charging stations) and user-generated data (e.g., for training self-driving car algorithms). Though many believed around the turn of the century that extant antitrust laws were sufficient to deal with the new economy (Posner, 2017), lawyers and policy-makers are increasingly aware of how they largely fail to address the challenges posed by digital markets which facilitate uniquely durable market power and digital inequality (Newman, 2019) through. e.g., suboptimal allocation of data control (Fisher and Streinz, 2021). Platforms such as Alibaba and Google have a staggering capability for tracking users' behavior across a wide range of their activities both online and offline (Varian, 2010), subjecting users to data risks and exploitation (Economist, 2018). Economists have also the antitrust implications of networks (Elliott and Galeotti, 2019), learning algorithms on digital platforms (Vaitilingam, 2020), and new forms of algorithmic collusion (Johnson and Sokol, 2020). Recent initiatives concerning data sharing include open banking that has received significant attention (He, Huang, and Zhou, 2022; Goldstein, Huang, and Yang, 2022).

baseline quality and abilities in processing data to improve service quality. They sell services to a continuum of users with privacy concerns whose payoff in each period depends on the intrinsic product quality, heterogeneous user preference, network adoption, and quality improvements from analyses of historical data. The stock of data, by-products of users' historical activities on each platform, depends on a platform's infrastructure investment for data generation and collection, user adoption, and endogenous data sharing by users and potentially also platforms.

In the first period, platforms decide how much to invest in data generation and collection, what compensation or perks they provide users who contribute data, as well as what price they set for their goods (i.e. platform services). Users then each choose a platform to adopt and exercise their discretion, if any, over how much of the data generated to share with the platforms, trading off privacy costs for compensation for data contribution. In the second period, platforms process the data collected to improve service quality, set prices, and compete again for the users. Importantly, users can switch between the platforms without friction or costs, which reflects the rising platform interoperability and ease of multi-homing in practice.

In equilibrium, data-driven platforms initially underprice (i.e., set low service prices) and subsidize adoption to gather data, which improves service quality and leads to a competitive advantage in the second period allowing to charge high prices. Data accumulation and quality improvement through data induce a dynamic feedback loop which tends to amplify an initial advantage one platform has over another (e.g., in terms of user base, market share, or data processing technology). For instance, a greater market share of a platform in the first period leads to accumulating more data, improving product quality, and gaining a large market share in the second period. This data feedback effect allows the stronger platform to gain high market power and, whilst improving service quality, weakens price competition in the second period, which hurts users.

Because users are dispersed and atomistic—a key assumption in our model, they do not internalize the broader impact of their actions (e.g., data contribution and sharing) on (i) future service or product quality which affects all users, (ii) concentration of market power, and (iii) platforms' incentives to invest in data infrastructure and to collect data. Since platforms compete with price, service prices as well as the platforms' profits need not internalize these effects either. As such, the equilibrium level of data collection generally does not maximize user welfare and hence is inefficient. We analyze the resulting inefficiencies in the presence of both data feedback and network effects, and provide a simple framework to assess various policy interventions including the General Data Protection Regulation (GDPR), California Consumer Privacy Act of 2018 (CCPA), Payment Services Directive Two (PSD2), and Open Banking Standard. Whereas extant policies cannot fully address these issues, we show how a simple setup of user union or data trust can more effectively coordinate users to contribute and share data to improve consumer welfare.

Several defining features of a data economy have been well-studied in the literature. For example, data are by-products of economic activities and non-rival (e.g., Jones and Tonetti, 2020; Veldkamp and Chung, 2022), and data are associated with data subjects (users) who have privacy and other costs when contributing data (e.g., Ichihashi, 2020; Liu, Sockin, and Xiong, 2020). Our model adds by underscoring new inefficiencies once we consider how data are jointly generated by dispersed users, who decide how much data to share or contribute, and by platforms, which undertake costly investment to collect, generate, or process data.[2] In our model, both inefficient under-collection of data (due to low platform investment in data collection), which compromises service quality, and over-collection of data, which concentrates market power and reduces price competition, can arise. When the two platforms are relatively symmetric, price competition is fierce and users' failure to internalize their effect on platform incentives and quality improvement from data dominates. Under-collection and under-sharing of data ensue. In contrast, if the asymmetry (e.g., in terms of data processing technology) between platforms is sufficiently large, users' failure to internalize the impact of data contribution on market power can dominate, leading to over-collection of data, high market power of the dominant platform, and thus high service prices.[3]

Our model enables a general qualitative evaluation of data regulation and antitrust policies that include privacy protection proposals, open data initiatives, and even decentralized or market-based innovations which facilitate trading of data among platforms and users, in terms of how they help or fail to address inefficient under- or over-collection/contribution of data.[4]

First, we find that privacy protection proposals, such as the GDPR or CCPA, have no material effect on economic outcomes, as the additional costs they cause for platforms are passed on to consumers via higher service/product prices. Next, we show that mandatory data sharing, such as

[2] The Data Freedom Act of RadicalXChange aptly puts it, "Data about people is always the output of a network of social activity. Even apparently "individual" data, such as a particular consumer's shopping habits or travel itinerary, is a product of the social world in which that person lives." Unlike other products, data or information is dispersed at inception and, if siloed, is of very limited value. This implies that aggregation, exchange, and sharing of data from individuals are key. However, dispersed individuals cannot coordinate to efficiently share data nor do they internalize the impact of their actions on market concentration, service quality improvement, or platform incentives. These issues are not restricted to platforms, but digitization and network effects on platforms amplify them.

[3] In our setting, over-collection of data implies high market power of one platform which, in turn, harms users through high service prices. One could easily introduce other channels through which high market power or the over-collection of data harms users without changing the paper's key findings, such as price discrimination by the dominant platform, inverse selection, etc (see, e.g., Ichihashi (2020); Montes, Sand-Zantman, and Valletti (2019); Brunnermeier, Lamba, and Segura-Rodriguez (2021).

[4] Similarly, it helps us understand the consequences of market-wide regulatory changes concerning data even beyond tech firms or banks. One salient example is the ongoing debate regarding consolidated tape on exchanges in Europe under MIFID II.

the open banking initiative in Europe, helps reduce the market power of dominant platforms that arises due to data feedback. However, mandatory data sharing also implies a classical freerider problem regarding data collection and leads to under-collection of data as well as increases in service prices in the first period due to platforms' reduced incentives to attract users for data. Our findings even reveal a striking paradox: Data sharing — often intended to increase platforms' stock of available data to improve services — can backfire and reduce platforms' stock of available data.

We then introduce a market for data in two specifications. First, we consider that users own data generated through their previous interactions with a platform and can sell them. As data are non-rival, users tend to sell their data to all platforms including ones which they have not adopted. This outcome erodes data exclusivity, and undermines platforms' incentives to invest in data infrastructure and to collect data, although it can help reduce market power. Similar to the baseline, the key inefficiency is that dispersed users do not internalize their impact on platform incentives when they sell their data to platforms, thereby causing inefficiently low data collection. Second, we consider that platforms own data and can strategically sell or share them with each other. Interestingly, the stronger platform then buys data from the weaker one, aggravating the concentration of market power and reducing price competition. In other words, a market for data leads to a novel form of anti-competitive "data collusion."

Overall, any of the discussed policy interventions — that is, privacy protection, mandatory data sharing, or markets for data — fail to address both potential inefficiencies, i.e., the under-collection of data curbing quality improvements and the over-collection of data weakening competition, and can backfire and harm users under certain circumstances. We therefore introduce and model the concept of user union, which recognizes users' decentralized nature, as well as their fundamental role in a platform's production or service provision. User union represents and coordinates users in order to maximize their welfare, and unlike other policy interventions and market-based solutions, it unambiguously raises user welfare by addressing both inefficiencies. First, because user union considers the joint payoff of all users, it internalizes the externality of individual data contribution on service quality affecting all users. Second, with coordination, it takes into consideration the impact of the aggregate data contribution on platforms' market power and investment incentives.

We then discuss several ways that the user union can curb excessive (data-induced) market power of large platforms, whilst mitigating the possible under-collection of data compromising service quality. In particular, we show that through appropriately subsidizing (or taxing) users' individual data contribution, the user union can stimulate (or curb) data collection if under-provision (or over-provision) of data ensues. Alternatively, the user union could be organized as a data trust that collects and accesses user data and then sells these data at a price (potentially determined via "collective" bargaining) to the different platforms, taking into account potential inefficiencies when setting the price and inducing an efficient allocation and generation of data.

Literature. Our study adds to what is now a large literature on the data economy. Jones and Tonetti (2020), Farboodi and Veldkamp (2021), Veldkamp and Chung (2022), and Cong, Wei, Xie, and Zhang (2022) study how data affects economic growth. Eeckhout and Veldkamp (2021) examine data and market power to show that firms' data-driven reallocation of production to the goods consumers desire can explain the divergence between product, firm and industry markups. We abstract from issues concerning macroeconomics and informational asymmetry (Ichihashi, 2020; Ichihashi and Smolin, 2022; Brunnermeier et al., 2021), to focus on how endogenous data generation by both users and platforms improves product quality and affects market concentration. The interaction of data feedback effect and product pricing complements earlier studies on how network effects affect platforms' strategic pricing (Fainmesser and Galeotti, 2016, 2020). In addition, we identify data-related inefficiencies and evaluate various protocols for data collection and sharing.

While existing literature has documented inefficient over- or under- supply/sharing of data, we contribute by rationalizing both in one unified model in the presence of dynamic data feedback.[5] The idea of data feedback is not new (see surveys by Biglaiser, Calvano, and Crémer, 2019; Calvano and Polo, 2021) and is related to the learning-by-doing literature (e.g., Dasgupta and Stiglitz, 1988; Cabral and Riordan, 1994). For example, Prüfer and Schottmüller (2021) describe conditions under which a data advantage leads to market tipping with dynamic R&D competition.[6] De Corniere and Taylor (2020) and Hagiu and Wright (2021) study how different types or uses of data affect competition. In particular, Hagiu and Wright (2021) find that data-enabled learning leads to

[5] On that front, a number of studies point to an under-supply of data. Data are non-rival (Jones and Tonetti, 2020; Cong, Xie, and Zhang, 2021) and they improve allocation of online resources such as advertising space (Stigler, 1980; Posner, 1981; Goldfarb and Tucker, 2011; Bergemann, Brooks, and Morris, 2015; Farboodi and Veldkamp, 2021). Empirical evidence on the aggregate value of data is also abundant (e.g., Bajari, Chernozhukov, Hortaçsu, and Suzuki, 2019; Schaefer, Sapi, and Lorincz, 2018). For example, data-driven, efficient decision-making in U.S. manufacturing nearly tripled between 2005 and 2010 (Brynjolfsson and McElheran, 2016). Failure to consider the aggregate net benefit naturally creates under-supply, sharing, and utilization of data. Meanwhile, other studies argue that consumers may over-supply data and data are over-shared in equilibrium, because consumers easily surrender their data or underestimate the costs of data breaches, contrary to what policy regulations such as GDPR implicitly assume (e.g., Taylor, 2004; Carrascal, Riederer, Erramilli, Cherubini, and de Oliveira, 2013; Acquisti, Taylor, and Wagman, 2016; Athey, Catalini, and Tucker, 2017; Agarwal, Ghosh, Ruan, and Zhang, 2020). In addition, gatekeeper and copycat externalities, as well as informational externality of individuals' data sharing due to, e.g., the correlation in user types can create over-supplies or over-sharing of data (e.g., Acemoglu, Makhdoumi, Malekian, and Ozdaglar, 2019; Bergemann, Bonatti, and Gan, 2022; Choi, Jeon, and Kim, 2019).

[6] The authors identify market power spillover in connected markets and positive net effects of data sharing when data feedback is strong. We link data to endogenous pricing, demonstrating inefficiencies under potentially symmetric platforms without tipping.

socially efficient outcomes in an infinite horizon, with a single firm capturing the entire market.[7] Different from these papers, we endogenize data collection, in that it depends both on users' endogenous data sharing and platforms' investments to collect data, and emphasize the inefficiencies in platform investment, data collection/sharing, and pricing. Note that our study differs from the patenting literature in which firms make innovation decisions themselves without users' discretion.

Also closely related to our paper are studies evaluating the various arrangements of data ownership. Among them, Campbell, Goldfarb, and Tucker (2015) show that opt-in privacy regulation can entrench monopolies, and Easley, Huang, Yang, and Zhong (2018) incorporate microfounded use of information in production to show how under-sharing can be corrected through a profit-maximizing data vendor. Instead, we focus on antitrust and regulation and the decentralized nature of data subject (and thus their lack of coordination) in the presence of platform investment and potential network effect. Parlour, Rajan, and Zhu (2022) also discuss how data privacy or sharing policies aimed to give consumers more direct control can have unintended consequences through a negative payment data externality due to information asymmetry in the lending market. We do not consider data spillovers across markets, and focus on non-information-based data externality that can be positive (through product quality improvement) or negative (through increasing platforms' market power or reducing their investment incnetives). We also differ from all these studies by generating both under- and over-provision of data, as well as proposing/analyzing the concept of user union.

Among the earliest studies on privacy protection and data sharing, Bouckaert and Degryse (2013) employ a two-period model of localized competition and find that opt out is the socially preferred consumer default option. More recently, Fainmesser, Galeotti, and Momot (2022) show that a firm may over-collect data and under-invest in data protection, an inefficiency correctable by regulation. Jin et al. (2018) and Acquisti et al. (2016) provide comprehensive surveys highlighting that firms do not internalize data harms to consumers and cannot commit to consumer-friendly data policy, which our model captures. Dosis and Sand-Zantman (2019) study optimal data allocation between a monopolist platform and users to resolve the issues arising from such lack of commitment or incomplete contracts. Garratt and Lee (2021) suggest that privacy-preserving CBDCs would improve consumer welfare without catalyzing data monopolies. Tang (2019) and Liu et al. (2020), among others, provide empirical measurement and theoretical microfoundations for consumer privacy. We contribute by analyzing whether privacy protection policies such as GDPR

[7] Hagiu and Wright (2021) distinguish across-user and within-user data and allow richer dynamics whereas we incorporate data-sharing-dependent privacy costs, platforms' data-related investment incentives, and consumer heterogeneity (which effectively creates horizontal product differentiation). Moreover, whereas Hagiu and Wright (2021) consider asymmetric Bertrand competition resulting in one firm covers the entire market each period, we analyze monopolistic competition and its resulting market concentration which antitrust and regulatory proposals target.

or CCPA mitigate inefficiencies in data sharing and platform competition.

Beyond data privacy policies, we also add to recent studies on open banking and open data initiatives, which are a part of a bigger data/infrastructure neutrality issue (Easley, Guo, and Krämer, 2017). For example, Goldstein et al. (2022) and He et al. (2022) theoretically demonstrate how open banking either hinders efficient resource allocation or hurts borrowers' welfare. We add by highlighting that platforms also contribute to data generation and open data initiatives can distort their incentives.[8] Our model predicts that mandatory data sharing reduces platforms' incentives to collect data via (i) investment in data collection and (ii) low service prices to attract users, potentially harming users. These results are consistent with the empirical findings of Martens, De Streel, Graef, Tombal, and Duch-Brown (2020), who find that user welfare is not maximized under mandatory data sharing due to increases in product price, and of Babina, Buchak, and Gornall (2022), who show that customer-directed data sharing increases entry but can reduce ex-ante information production. Also in line with our findings, Jin and Vasserman (2021) estimate that requiring auto-insurance firms to publicly share monitoring data leads to less monitoring, data elicitation, and lower consumer welfare in equilibrium.[9]

The user union we introduce can be interpreted as a special type of data intermediary which — different from the ones studied in the literature (see, e.g., Ichihashi, 2021a; Bergemann et al., 2022) — maximizes users' payoff instead of its own. It has been recognized that users should be compensated for data contributions and data may need joint management (Posner and Weyl, 2018; Arrieta-Ibarra, Goff, Jiménez-Hernández, Lanier, and Weyl, 2018). Multiple law articles argue that recent regulatory policies fail to resolve the issues and create new challenges, and advocate for alternative legal frameworks (e.g., Delacroix and Lawrence, 2019; Houser and Bagby, 2022), potentially aided by new technologies such as blockchains. Our model exactly provides the economic foundation for returning the power of aggregated data to individuals through such mechanisms.

Finally, our study is broadly related to the literature on multi-sided platforms (Rochet and Tirole, 2003, 2006; Armstrong, 2006) and network effects (Katz and Shapiro, 1985; Becker, 1991). We add dynamic data feedback to network effects (current platform adoption improves current product quality). Unlike prior studies, our findings do not rely on switching cost (Von Weizsäcker,

[8] Complementary to our study concerning the incentives of the platforms, Fang and Kim (2022) introduce a new hypothetical regulation called data neutrality to demonstrate how non-discriminatory and open access to a platform's data does not necessarily make consumers better off because the platform optimally reduces the amount of data provision under the regulation.

[9] Monitoring programs have been introduced in the auto-insurance ("pay-how-you-drive" program in the United States), life insurance (e.g., Vitality program from John Hancock), and lending (e.g., Ant Financial) industries where firms collect consumer data to better assess accident/medical risks and adjust future premiums/interests (Jin and Vasserman, 2021).

1984; Klemperer, 1987, 1995; Farrell and Klemperer, 2007), platform multi-sidedness (Weyl, 2010; Cong, Tang, Xie, and Miao, 2021), focality-based arguments, or higher order beliefs (Akerlof, Holden, and Rayo, 2021; Halaburda and Yehezkel, 2019; Halaburda, Jullien, and Yehezkel, 2020). A future competitive advantage for platforms with higher user base may similarly arise in a model of platform competition with switching cost (sticky customers), yet the implications and mechanisms fundamentally differ: First, data are non-rival whereas user base is generally rival, and users' adoption and data sharing decisions interact to affect other users' product quality and platform incentives. Second, data feedback effects depend on users' and platforms' endogenous data sharing, as well as platforms' data-related investments, whereby users may contribute data to a platform different from their service provider. These elements absent in models of dynamic network effects or switching cost allow us to evaluate a wide array of policies and regulations in practice.

1 Model Setup

We study an economy with two periods, $t = 1, 2$, and no time discounting, in which two digital platforms, indexed by $x \in \{A, B\}$, compete for a unit measure of atomistic users indexed by $z \in [0, 1]$.[10] Data generated in the first period as by-products of economic activities improves platform services in the second period.

Platforms and users. In each period, all platforms produce and sell perishable services ("products") to users under price competition. The role of the platforms and their services or products can be interpreted broadly. For instance, the platforms could represent digital marketplaces — such as Amazon or Alibaba — where the service is to facilitate transactions among their users.[11] In each period t, an individual user buys at most one unit of service from either A or B. Users can switch platforms without friction or cost, which is akin to (intertemporal) multi-homing and captures the rising platform interoperability and ease in multi-homing in practice. The platform produces the service at zero cost, and sells it to users at price p_t^x in terms of the numeraire ("dollars"); we do not model price discrimination. The price p_t^x can also be seen as a fee for using the platform or its services. We denote by N_t^x the measure of users who buy the service from platform x at t, which

[10] The model features only two periods, because it is the minimum number of periods required to model dynamic data collection and usage which lead to the key economic mechanisms we are after. These key mechanisms that our model captures as well as our main results would likely arise in a setting with more time periods too.

[11] The platform could also represent social networks such as Meta or Instagram. Given this broad interpretation, our model describes platforms with diverse structures, including both one-sided and two-sided platforms. For simplicity, we model network effects in reduced form and abstract away from the finer details of multi-sided platforms (e.g., Rochet and Tirole, 2003).

we also refer to as the level of platform adoption.

Platform services and quality. As in a canonical Hotelling model, users are uniformly distributed on the interval $[0, 1]$ with location index z and platforms A and B are located at 0 and 1 respectively. The service or product of platform x at t then gives user z a utility $Y_t^x - \kappa^x(z)$, where Y_t^x is the service quality (discussed below), and $\kappa^x(z)$ is user z's "transport cost" taking the form:

$$\kappa^x(z) = \begin{cases} \hat{\kappa} z & \text{for} \quad x = A \\ \hat{\kappa}(1-z) & \text{for} \quad x = B, \end{cases} \tag{1}$$

with constant $\hat{\kappa} > 0$. As such, any platform x has some local market power regarding the users z that are located "close by." As $\hat{\kappa}$ decreases, competition among the two platforms becomes more fierce, whereby the limit case $\hat{\kappa} \to 0$ results in a Bertrand competition.[12]

We specify the quality of the service that platform x provides to each user in t to be:

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

where K^x, γ^x, and ϕ^x are positive constants and D^x is platform x's endogenous amount of data used in production in period $t = 2$ (which has been generated in period $t = 1$); here, $\mathbb{I}_{\{\cdot\}}$ is an indicator function which equals one if $\{\cdot\}$ is true and zero otherwise.[13] The usefulness of platforms, digital marketplaces, or social networks generally depends on their adoption and user base, giving rise to network effects, which we capture by the service quality being increasing in the adoption level N_t^x. The parameters γ^x for $x = A, B$ quantify the (marginal) strength of these network effects.

In addition, data lie at the core of the business model of many platforms. Indeed, in the digital era, most firms and social networks collect and process an enormous amount of data to improve their products and services. Following the literature (e.g., Farboodi et al., 2019), we assume data are by-product of economic activity, and platforms use accumulated data to improve efficiency

[12] Without local market power and, in particular, in the limit $\hat{\kappa} \to 0$, the outcome of competition between two non-identical platforms would generally feature — due to network effects — "winner takes all," whereby all users adopt one single platform.

[13] In the specification of (2), data exhibits constant returns to scale (i.e., service quality is linear in data). Existing literature considers both increasing and decreasing returns to scale. Eeckhout and Veldkamp (2021) model data's increasing returns to scale and how firms' aversion to risk affects product and firm-level markups. By raising D^x to a power less than 1, our framework can model decreasing returns to the data scale discussed in Farboodi, Mihet, Philippon, and Veldkamp (2019). While there is empirical evidence for the diminishing returns to data when used for a particular task or product (e.g., Chiou and Tucker, 2017; Bajari et al., 2019), there have not been empirical studies on data complementarities and how data diversity enhances prediction accuracy. It is generally taken as given that more data gives a firm or platform advantages (Prüfer and Schottmüller, 2021; Biglaiser et al., 2019). We note that our key findings would remain similar, if we assumed decreasing or increasing returns to scale. That is, our key findings arise as long as data improves service quality.

or product/service quality. In other words, a platform's service quality Y_2^x increases with D^x, i.e., the quantity of data generated in period $t = 1$ from users' activities that platform x uses in production at $t = 2$, a notion that we will make precise shortly. D^x is public knowledge.[14] The parameter ϕ^x quantifies the extent to which data improves platform services.[15] Platforms are potentially heterogeneous in their ability to generate, collect, process, or exploit data, so ϕ^x differs across platforms. Arguably, firms such as Amazon, Google, and Apple are already leaders in data accumulation historically. This unequal landscape can be captured through one platform's having greater K^x (historical data enable it to offer better products) or ϕ^x (experience in analyzing data make them more capable of extracting information from data).

Data collection and generation. Data are by-products of economic activities and, in particular, of platform usage. The data that user z contributes can include personal data (such as sign-up information) or transaction data. For mere simplicity, data are homogeneous in our model and any unit of data has the same effects on platform service quality Y_2^x specified in (2). Data are collected only in period $t = 1$, and improves product quality in period $t = 2$; for simplicity, there is no more data collection in $t = 2$ because it is the terminal period.

One key feature that distinguishes our paper from related works is that it endogenizes data collection in the following way.[16] The "effective" amount of data $I^x\theta^x$, generated through a user z's interaction with platform x at $t = 1$, depends on (i) the user z's willingness to share data with the platform $\theta^x \in [0, 1]$ and (ii) the platform's (publicly) observable investment to collect, to generate, or to process data, $I^x \in [0, 1]$ in period $t = 1$. Investment I^x comes with quadratic (and private) cost $\frac{1}{2}\lambda(I^x)^2$ for the platform for a constant $\lambda \geq 0$, and is bounded from above by 1. Intuitively, when z buys platform x's service in $t = 1$, I^x units of data are generated/collected, and z then decides on the fraction $\theta^x \in [\underline{\theta}, 1]$ of these data that she shares with platform x.[17] Here,

[14] We follow the literature (e.g., Gal-Or, 1985; Vives, 1988) to assume that the amount of data acquired, whether through purchase from vendors or not, is public. In other words, firms do not engage in secret information acquisition (see, e.g., Hauk and Hurkens, 2001).

[15] More broadly, ϕ^x may be related to platform x's existing stock of data or the platform's past experience in processing data, reflecting the idea that any additional unit of data becomes more useful the larger the existing stock of data are, i.e., data have non-decreasing returns to scale. For instance, machine learning or AI algorithms often require a large training data set before they can deliver useful results, giving rise to increasing returns of scale of data. Our framework can be extended to multiple periods with decay of data usefulness and long-run diminishing returns to scale of data on specific tasks, as documented in Farboodi and Veldkamp (2021).

[16] Studies such as Jones and Tonetti (2020); Farboodi and Veldkamp (2021); Prüfer and Schottmüller (2021); Hagiu and Wright (2021) do not consider users' endogenous decisions to share data or firms' investments in data technology/innovation/collection. Our paper crucially differs in this regard because it endogenizes data collection in that way and therefore identifies both inefficient under- and over-collection of data.

[17] Because any user z's privacy cost as well as perks from contributing data (introduced later below) depends on the platform x that she adopts but not on her location on the Hotelling line, it follows that z's choice on whether to contribute data to platform x, i.e., θ^x, only depends on x and not on z.

$\underline{\theta} \in \{0, 1\}$. When the user z must share the data with the platform, e.g., because the platform can require users to share data in order to use platform services (and there is no legal privacy protection in place precluding that), then one sets $\underline{\theta} = 1$ so that mechanically $\theta^x = 1$. When users can opt out of sharing data, then $\underline{\theta} = 0$.[18]

We denote by $\hat{D}^x := \theta^x I^x N_1^x$ the quantity of data generated on platform x up to the beginning of time $t = 2$ (i.e., in time $t = 1$), and by D the total stock of data at the beginning of time $t = 2$, i.e., $D := \hat{D}^A + \hat{D}^B$. Next, D^x denotes the amount of data that platform x uses for production at $t = 2$. These data may also include data that has been generated on a competing platform and can be accessed by platform x, for instance, due to data sharing, so that $D^x \geq \hat{D}^x$. The stock of data D^x that platform x uses in its operations satisfies $D^x \leq D$. However, because the use of data as a resource is non-rival, $D^A + D^B = D$ need not hold. Both platforms could in principle use the entire stock of data for their "production" in which case $D^A = D^B = D$.

At time $t = 1$, any user z collects payoff $(q^x - c^x)\theta^x I^x$ from sharing fraction θ^x of its stock of data I^x, which has been generated through z's consumption of x's service with platform x investing I^x. First, any user gets an endogenous reward or "perk" $q^x I^x \theta^x$ from platform x as compensation for sharing fraction θ^x of its data with x. Second, users have privacy concerns in that z incurs a (privacy) cost or disutility $c^x I^x \theta^x$ for given c^x when platform x has access to her data from period $t = 1$. Importantly, the privacy cost also depends on whether platform x shares or sells user data to other platforms. To capture this feature, we write $c^x = c(1 + \eta^x)$ for a constant c and $\eta^x \in [0, 1]$ denotes the (possibly endogenous) fraction of data that platform x shares with or sells to the other platform $-x$ at the beginning of period $t = 2$. In the baseline, only platform x has access to the data of its users, so $\eta^x = 0$ and $c^x = c$.[19] In other variants of the model, η^x might be non-zero, e.g., due to required data sharing by regulation (e.g., Open Finance), or be an endogenous equilibrium quantity that depends on platforms' optimized decisions on whether to sell or share data.

Note that the "privacy disutility" $c^x \theta^x I^x$ may not only capture the adverse consequences of sharing data — such as the loss of privacy — but also the potential positive effects of sharing (personal) data (e.g., when $c^x < 0$) — such as improved service quality or customization from the use of personal and user-specific data by the platform.[20] Unless otherwise specified, we only

[18] This would be the case under regulatory proposals and privacy protection regulation, like GDPR and CCPA.

[19] Notice that in case z shares data at intensity θ^x with platform x, which exerts investmentI^x, and the competitor platform $-x$ can use fraction η^x of this data, then the total privacy cost of z becomes $(c + \eta^x c) I^x \theta^x$.

[20] That is, negative $c^x < 0$ would capture (in reduced form) the benefits of learning from user-specific data, with the overall benefits $\theta^x I^x c^x$ scaling with platform investment in data infrastructure I^x: A platform might utilize user z's data to tailor the services to z's needs, which benefits z. In contemporaneous work, Hagiu and Wright (2021) refer to this effect as "within-user learning" (as opposed to "across-user learning" capturing that data improves service quality for all users).

consider $c^x, c \geq 0$, i.e., sharing data generates disutility for users. That said, we note that model solution and equilibrium remain the same in case $c^x < 0$.

Payoffs. When user z buys one unit of service from platform x (i.e., when user z adopts platform x) at time t, she derives a net utility payoff

$$u_t^x(z) = Y_t^x - p_t^x - \kappa^x(z) + (q^x - c^x) I^x \theta^x \quad \mathbb{I}_{\{t=1\}}, \tag{3}$$

where Y_t^x is the service quality from (2), p_t^x is the service fee/price, $\kappa^x(z)$ is the cost of consuming platform x's services from (1), and the last term with the indicator function captures users' net payoff from sharing data with the platform. When user z consumes platform x's service at $t = 1$, she also decides on the amount of data she is willing to contribute through her choice of $\theta^x \in [\underline{\theta}, 1]$, so as to maximize $(q^x - c^x) I^x \theta^x$ taking the price ("reward") for contributing data q^x as given. This leads to the optimal choice:

$$\theta^x = \begin{cases} \underline{\theta} & \text{if} \quad q^x < c^x \\ \hat{\theta} \in [\underline{\theta}, 1] & \text{if} \quad q^x = c^x \\ 1 & \text{if} \quad q^x > c^x. \end{cases} \tag{4}$$

Notice that because any user is atomistic, she does not internalize the broader effects of her data sharing (for instance, on period-2 payoffs or market structure) and thus finds it privately optimal to share data if and only if the reward from doing so exceeds the cost, that is, if and only if $q^x \geq c^x$.

Platform x's payoff π_2^x in period $t = 2$ is the revenue from selling N_2^x service units at price p_2^x, i.e., $\pi_2^x := N_2^x p_2^x$. In period $t = 1$, platform x pays a direct transfer q^x to its users for accessing their data. For prices p_t^x and q^x and investment I^x, platform x's payoff at time $t = 1$ reads

$$\pi_1^x := N_1^x p_1^x - q^x N_1^x I^x \theta^x + N_2^x p_2^x - \frac{\lambda (I^x)^2}{2}. \tag{5}$$

We assume that both platforms operate in the market in both periods $t = 1$ and $t = 2$; Section 5.1 studies an extension in which platforms decide on whether to enter the market in period $t = 1$.

Equilibrium concept. Given prices p_t^x, q^x and investment I^x, user z decides at time t whether to consume platform services and, if so, which platform to adopt. Throughout, we assume that parameters are such that all users participate, so that $N_t^A + N_t^B = 1$.[21] Then, user z buys the

[21] As will become clear later, platform competition and price levels depend on the difference between K^A and K^B but not on their level, so we can choose K^A and K^B sufficiently large to ensure full participation, in that

service of platform x if $u_t^x(z) \geq u_t^{-x}(z)$.[22] We look for a subgame perfect equilibrium in pure strategies over the two periods $t = 1, 2$ where platforms x maximize their payoffs. The timing in period $t = 2$ is as follows: Platforms choose their prices p_2^x simultaneously to maximize $\pi_2^x = N_2^x p_2^x$ (taking the choice of the competing platform $-x$ as given) and, given prices p_2^A and p_2^B, users decide which platform to join. Next, we discuss the timing in period 1. First, platforms x choose rewards q^x, service prices p_1^x, and investment I^x simultaneously to maximize π_1^x from (5), taking the choices of the other platform $(q^{-x}, p_1^{-x}, I^{-x})$ as given. Second, users, who observe I^x, p_1^x, and q^x, decide on whether to buy platform services and, if so, which platform they buy at (recall that any user z buys a service from at most one platform). They also decide on their optimal data contribution, which is characterized in equilibrium by θ^x in (4).[23] At time $t = 1$, platform x cannot commit to a future price p_2^x; as shown in Section 5.3, commitment to future prices does not change the outcomes or equilibrium.

2 Equilibrium Characterization and Baseline Solutions

2.1 General Characterization

Platform adoption and user decisions. We characterize the demand for platform services in period t, given p_t^x, q^x, I^x and θ^x for $x = A, B$. First, we conjecture $N_t^x \in (0, 1)$ and solve for the marginal user $\hat{z}_t$ who is indifferent between adopting platform A and adopting platform B. Then $u_t^A(\hat{z}_t) = u_t^B(\hat{z}_t)$, $N_t^A = \hat{z}_t$, and $N_t^B = 1 - \hat{z}_t$. Using $u_t^x(z)$ from (3), Y_t^x from (2), and $\kappa^x(z)$ from (1), we can show (details in Appendix A.1):

$$\hat{z}_t = \frac{1}{2} + \frac{\Delta_K - (p_t^A - p_t^B) + \left[\phi^A D^A - \phi^B D^B\right] \mathbb{I}_{\{t=2\}} + \left[I^A \theta^A (q^A - c^A) - I^B \theta^B (q^B - c^B)\right] \mathbb{I}_{\{t=1\}}}{2\kappa}, \tag{6}$$

where we define, assuming $\hat{\kappa} > \frac{\gamma^A + \gamma^B}{2}$,

$$\kappa := \hat{\kappa} - \frac{\gamma^A + \gamma^B}{2} \quad \text{and} \quad \Delta_K := K^A - K^B + \frac{\gamma^A - \gamma^B}{2}.$$

When the expression for $\hat{z}_t$ from (6) lies outside the interval $[0, 1]$, then the "winner takes it all" outcome prevails with one platform covering the whole market, in that $N_t^A = 1$ or $N_t^B = 1$. As our

$\max\{u_t^A(z), u_t^B(z)\} \geq 0$ at all times $t = 1, 2$ and for all $z \in [0, 1]$.

[22] As there is a continuum of users, it is without loss of generality to assume that user z joins platform A if indifferent between A and B.

[23] It does not matter whether users choose which platform x to adopt and θ^x simultaneously or sequentially, in a way that any user z first decides on which platform x she consumes and then decides on the fraction of data θ^x she contributes.

focus is on platform competition, we focus on parameter configurations under which any platform x has non-trivial market share in at least one period; a sufficient condition is

$$3\kappa > \max\{\Delta_K + \phi^A - \phi^B, -\Delta_K - \phi^A + \phi^B\}, \tag{7}$$

which we assume to hold throughout.[24] Furthermore and without loss of generality, we consider throughout the paper that platform A is weakly stronger than B in $t = 2$, in that $\hat{z}_2 \geq 1/2$.

Notice that the marginal user $\hat{z}_t$ depends on transport cost $\hat{\kappa}$ and network effects γ^x only via κ and Δ_K. Thus, an increase in the strength of network effects (i.e., an increase in γ^A or γ^B) has similar effects to a decrease in κ. When network effects are sufficiently strong (i.e., $\kappa \to 0$), the expression for $\hat{z}_t$ in (6) tends to either $\pm\infty$, when one platform dominates, or to $1/2$, when platforms are symmetric. That is, sufficiently strong network effects precipitate "market tipping" and the "winner takes it all" outcome. In addition, K^A and K^B affect user choice only via Δ_K. For most of the analysis, we only have to keep track of Δ_K and κ instead of K^x, γ^x and $\hat{\kappa}$ separately. Having characterized equilibrium demand for platform services, we now solve for the subgame perfect equilibrium in the baseline solution. To do so, we start by characterizing the (Nash) equilibrium of the subgame in period $t = 2$. Then, we move backward to period $t = 1$.

Solution in the second period. Given a stock of data D^x that is used in production, platforms chooses their prices p_2^x simultaneously maximize period-2 payoff, i.e., $\max_{p_2^x} N_2^x p_2^x$. The trade-offs that determine the optimal price p_2^x are standard, leading to the following equilibrium.

Lemma 1. *The subgame in period $t = 2$ admits a unique Nash equilibrium. In equilibrium, period-2 service prices read $p_2^x = 2\kappa N_2^x$. Period-2 platform payoffs read $\pi_2^A = \frac{(3\kappa + \Delta_K + D^A\phi^A - D^B\phi^B)^2}{18\kappa}$ and $\pi_2^B = \frac{(3\kappa - \Delta_K - D^A\phi^A + D^B\phi^B)^2}{18\kappa}$. The marginal user $\hat{z}_2$ (i.e., platform A's market share) is*

$$\hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + D^A\phi^A - D^B\phi^B}{6\kappa}.$$

Users' payoff in period $t = 2$ reads $u_2 = N_2^A(Y_2^A - p_2^A) + N_2^B(Y_2^B - p_2^B) - \bar{\kappa}_2$, with the total/aggregate transport cost $\bar{\kappa}_2 := \frac{\hat{\kappa}\left((N_2^A)^2 + (N_2^B)^2\right)}{2}$.

Notice that π_2^x may depend on N_1^x for $x = A, B$, as N_1^x affects data accumulation and therefore platform service quality Y_2^x. Moreover, platform A's market share $\hat{z}_2$ captures A's market and price setting power: Holding κ fixed, an increase in market power $\hat{z}_2$ allows platform A to charge a higher price to the detriment of users.

[24] Lemma 1 reveals that (7) implies $\hat{z}_2 \in (0, 1)$, so $N_2^x \in (0, 1)$ for $x = A, B$.

Solution in the first period. In period $t = 1$, the platforms simultaneously solve $\max_{q^x, p_1^x, I^x} \pi_1^x$, where the objective function π_1^x is characterized in (5). In general, the data in period $t = 2$, i.e., D^x for $x = A, B$, is a function of platforms' choices in period $t = 1$, specifically, $D^x = d^x(N_1^A I^A, N_1^B I^B)$ for a function $d^x(\cdot, \cdot)$. The optimal price level p_1^x for platform x's service solves then the first-order condition

$$\frac{\partial \pi_1^x}{\partial p_1^x} = \underbrace{N_1^x + \left(\frac{\partial N_1^x}{\partial p_1^x}\right) p_1^x}_{\text{Static revenue maximization}} + \underbrace{\sum_{x'=A,B} \left(\frac{\partial \pi_2^x}{\partial D^{x'}} \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial p_1^x}\right)}_{\text{Data benefits}} - \underbrace{\left(\frac{\partial N_1^x}{\partial p_1^x}\right) I^x \theta^x q^x}_{\text{Data cost}} = 0. \tag{8}$$

As illustrated by (8), the choice of price p_1^x reflects dynamic considerations: A higher price p_1^x raises revenue per user, but curbs the demand for platform services and thus generates less data, which hampers growth in platform quality. That is, the price p_1^x affects platform adoption N_1^x which may affect the data stock of platforms A and B. In setting its price p_1^x, platform x takes into account the future data-driven competition. Lowering the price p_1^x, platform x not only boosts demand for its own services and thus its data accumulation, but also limits demand for the competitor platform's services and in turn the data accumulation of the competitor platform. All these effects are captured by the "data benefits" term in (8). The "data cost" term reflects the privacy-related cost of data collection, in that privacy-concerned users require compensation q^x for sharing data (if they have the choice).

Next, we study the platforms' incentives to generate and to collect data via I^x. If optimal investment is interior (i.e., $I^x \in (0, 1)$), it solves the first-order condition:

$$\frac{\partial \pi_1^x}{\partial I^x} = \underbrace{\left(\frac{\partial N_1^x}{\partial I^x}\right) p_1^x}_{\text{Static revenue maximization}} + \underbrace{\sum_{x'=A,B} \left(\frac{\partial \pi_2^x}{\partial D^{x'}} \left[\frac{\partial D^{x'}}{\partial I^x} + \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial I^x}\right]\right)}_{\text{Data benefits}}$$
$$\underbrace{- \left(\left(\frac{\partial N_1^x}{\partial I^x}\right) I^x \theta^x q^x + \theta^x N_1^x q^x + \lambda I^x\right)}_{\text{Data cost}} = 0. \tag{9}$$

Similar to price-setting, the choice of investment reflects dynamic and strategic considerations. Higher I^x comes at the expense of users' privacy and additional costs and, all else equal, reduces period-1 payoff, but increases the stock of data and platform x's payoff in period $t = 2$.

Proposition 1 below shows that it is optimal for platform x to implement $\theta^x = 1$.[25] Another

[25] If $\theta^x < 1$, then the platform could profitably reduce investment I^x and increase θ^x, reducing the cost of the effort for generating data whilst holding data collection intensity $\theta^x I^x$ fixed. Also, when $I^x = 0$, then the level of θ^x does not matter.

notable result is that platforms pass the direct monetary costs of collecting data, i.e., the compensation for sharing data paid to users $q^x I^x \theta^x$, one-to-one on to users through prices p_1^x in a way that q^x does not affect user welfare, market concentration, or platforms' payoffs. Specifically, an increase in q^x implies a commensurate increase in p_1^x and the exact value q^x is payoff-irrelevant. The intuition is that users require a compensation for their privacy-related cost when joining a platform and accordingly contributing data to this platform which can come in the form of a direct reward q^x or lower service price p_1^x, i.e., they are substitutes in compensating for privacy costs. User welfare reads $u_1 = \sum_{x=A,B} N_1^x \left[Y_1^x - p_1^x + (q^x - c^x)\theta^x I^x\right] - \bar{\kappa}_1 + u_2$, where u_2 is the period-2 user welfare from Lemma 1 and $\bar{\kappa}_t := \frac{\hat{\kappa}\left((N_t^A)^2+(N_t^B)^2\right)}{2}$ is the aggregate transport cost in t.

Proposition 1. *In a subgame perfect equilibrium, Lemma 1 characterizes the subgame in $t = 2$. Each platform chooses q^x to induce $\theta^x = 1$, and $p_1^x = \bar{p}_1^x + I^x q^x$, where $\bar{p_1}^x$ does not depend on $q^{x'}$ for $x' \in \{A, B\}$, and satisfy $\frac{\partial p_1^x}{\partial q^x} = I^x$ as well as $\frac{\partial p_1^x}{\partial q^{-x}} = 0$.*

Note that the exact level of q^x is payoff-irrelevant (as long as it induces $\theta^x = 1$), in a sense that all other equilibrium quantities (that is, $N_t^{x'}$, $\hat{z}_t$, $p_2^{x'}$, u_t, $\pi_2^{x'}$, $I^{x'}$) do not depend on q^x for $t \in \{1, 2\}$, $x, x' \in \{A, B\}$. The platforms are indifferent between different levels of q^x as long as it induces $\theta^x = 1$. The exact value of $\underline{\theta}$ also does not affect investment I^x, user welfare u_1, π_t^x, p_2^x, $\bar{p}_1^x = p_1^x - I^x q^x$, and N_t^x. We have characterized the essentially unique equilibrium, without making any specific assumptions on data collection, i.e., the relation between D^x and N_1^x. Thus, Proposition 1 remains valid in the following sections with data sharing (which affects the relationship between D^x and N_1^x). The equilibrium choice of q^x is not pinned down as long as it induces $\theta^x = 1$. The price can be rewritten as $p_1^x = \bar{p}_1^x + I^x q^x$, where $\bar{p}_1^x$ does not depend on I^x or q^x, so one could set (without loss of generality) $q^x = c^x$. When λ is sufficiently large in that $I^x \in (0, 1)$, the equilibrium in $t = 1$ is then characterized by four non-linear equations, i.e., (8) and (9) for $x = A, B$, which in general have to be solved numerically for $\bar{p}_1^x := p_1^x - I$ and I^x.[26] In what follows, we assume that a unique solution and thus a unique equilibrium (up to q^x) exists, with unique values for $\hat{z}_t$, I^x, π_1^x, u_1, and $\bar{p}_1^x$ as well as p_2^x.

2.2 Benchmark Under Laissez-Faire Database Directives

We now present the model analysis of the baseline with $\eta^x = 0$, i.e., there is no data sharing, and users cannot opt out from sharing data, i.e., $\underline{\theta} = 1$ implying $\theta^x = 1$. To gain some preliminary

[26] As we shall see, an exception is the symmetric platform case, in which we can solve for the unique symmetric equilibrium in closed-form.

insights, we start by analyzing the symmetric platform case, with $\Delta_K = 0$ and $\phi^A = \phi^B$. This is essentially the laissez-faire situation under the conventional database directives.[27]

Proposition 2. *Consider $\eta = 0$, $\underline{\theta} = 1$, $c^x = c$, and that both platforms are symmetric. In a symmetric equilibrium, $N_t^x = 1/2$, $p_t^A = p_t^B$, $I^A = I^B$, $\theta^A = \theta^B = 1$, and $q^A = q^B$. In period $t = 2$, prices satisfy $p_2^x = \kappa$. In period $t = 1$, prices and investment satisfy:*

$$p_1^x = \kappa + I^x\left(q^x - \frac{2\phi^x}{3}\right) \quad \textit{and} \quad I^x = \min\left\{\left[\frac{\phi^x - 3c}{6\lambda}\right]^+, 1\right\},$$

where $[x]^+ = \max\{x, 0\}$. User welfare reads $u_1 = const + I^x\left(\frac{7\phi^x}{6} - c\right)$ and increases with I^x, where const is a constant that is independent of I^x or p_1^x and only depends on (exogenous) parameters.

Interestingly, in period $t = 1$, both platforms compete fiercely for users and their data via relatively low prices that decrease with ϕ^x, which benefits users: For instance, setting $q^x = 0$ (which is possible due to $\underline{\theta} = 1$) would imply that period-1 prices decrease with $I^x\phi^x$ and are strictly lower than period-2 prices, notably, even though there are no switching costs. Also notice that investment I^x decreases with users' privacy concerns c: When increasing I^x, platform x must pay users higher compensation $q^x\theta^x I^x$ or reduce service prices p_1^x, in order not to lose customers to $-x$ and to maintain market share N_1^x. That is, to induce users to join the platform and accordingly to share data with platform x as $\theta^x = 1$, users require compensation for their privacy cost c either directly via q^x or indirectly via low prices p_1^x, which makes data collection costly for platforms and curbs investment.[28] Individual (atomistic) users do not internalize that their data contribution (i) improves service quality in period $t = 2$, both for themselves and for other users, as well as (ii) affects platform incentives to invest, so they effectively require their "private" cost c as compensation for sharing data. As platforms compete with price, service prices (e.g., $p_2^x = \kappa$ in $t = 2$) do not internalize these effects (i) and (ii) either. The resulting level of investment I^x is therefore inefficiently low from the user perspective, i.e., user welfare increases with I^x and users would in aggregate benefit, if they required less compensation for sharing data and platforms' investment in data collection I^x were higher.

When platforms are not symmetric, data collection also affects market A's market power $\hat{z}_2$ in $t = 2$, which in turn affects period-2 service prices. As we are primarily interested in how data and

[27] Efforts to create exclusive ownership rights in electronic data started with the Database Directive (1996). Data subjects are not the sole creators of data. Platforms also invest in data infrastructure and maintain databases. Database Directive was justified as an incentive for EU firms to invest more in the production of electronic databases and help the EU to catch up with other countries, in particular the United States, in this respect.

[28] Again, recall from Proposition 1 that compensating for privacy cost via low service prices p_1^x or via the reward q^x are (perfect) substitutes, i.e., $\frac{\partial p_1^x}{\partial q^x} = I^x$.

data collection/sharing affect competition, we focus (unless otherwise mentioned) on the case that A's market power in $t = 2$ derives (mainly) from its superior data processing ability, i.e., $\phi^A \geq \phi^B$, rather than from an advantage independent of data (e.g., large $\Delta_K > 0$). The next Proposition presents equilibrium properties when A has superior ability to process data (i.e., $\phi^A > \phi^B$).

Proposition 3. *Consider $\phi^A > \phi^B$, $\eta = 0$, and $\underline{\theta} = 1$. In equilibrium, platform A's market share $\hat{z}_2$ in $t = 2$ increases with its "data advantage" ($\phi^A I^A - \phi^B I^B$). If $\Delta_K \geq 0$ and $c \geq 0$ are sufficiently small, the following holds: (i) $I^A > I^B$, (ii) $p_1^A < p_1^B$ (considering $q^x \leq c^x$), (iii) $p_2^A > p_2^B$, and (iv) $\hat{z}_t > 1/2$ for $t = 1, 2$.*

An interesting case to study is the case in which platforms only differ in their ability to collect, process, or utilize data, i.e., $\Delta_K = 0$ and $\phi^A > \phi^B$, and the privacy-induced costs of data collection (captured by c) are not prohibitively large. Then, platform A benefits more from data in period $t = 2$ than platform B does, so it tends to reduce prices more aggressively to attract users in $t = 1$ ($p_1^A < p_1^B$) and invests more in data collection ($I^A > I^B$). Platform A's data edge translates into high market power in period $t = 2$ which allows A to charge high prices in $t = 2$, i.e., $p_2^A > p_2^B$.

Figure 1 graphically illustrates these findings by plotting equilibrium prices and market shares for different values of ϕ^A, starting from $\phi^A = \phi^B = 1$. An increase in ϕ^A boosts platform A's incentives to collect data in the first period $t = 1$, reducing price p_1^A and increasing investment I^A, whilst crowding out investment I^B from the competitor platform B. The lower price p_1^A puts pressure on B to lower its first-period price as well, so p_1^B decreases with ϕ^A. When ϕ^A and ϕ^B are close to each other, competition among platforms is fierce in both periods, which limits platforms' abilities to benefit from data in period $t = 2$ and curbs A's investment in period $t = 1$. As A collects more data upon an increase in ϕ^A, A's market power, i.e., $\hat{z}_t$ for $t = 1, 2$, as well as its price-setting power and the time-2 price p_2^A, rise. Importantly, users do not internalize that sharing data boosts A's market power in $t = 2$, which may harm them through high prices. The following Lemma characterizes when such an increase in A's market power in $t = 2$ reduces user welfare.

Lemma 2. *In period 2, higher market concentration $\hat{z}_2$ (i.e., platform A's market share) harms users and their welfare, in that $\frac{\partial u_2}{\partial \hat{z}_2} < 0$, if and only if $\gamma^B > \frac{2\kappa(1+\hat{z}_2)+\gamma^A \hat{z}_2}{1-\hat{z}_2}$.*

Unless otherwise mentioned, we focus on the case that γ^B is sufficiently large so that u_2 decreases with A's market share $\hat{z}_2$, i.e., that (ceteris paribus) more intense price competition in $t = 2$ is welfare-improving and high market concentration harms users. One interpretation of why network effect strength γ^B is large for the weaker/smaller platform (and possibly exceeds γ^A) is that the marginal network effect strength decreases with platform adoption, so total network effect

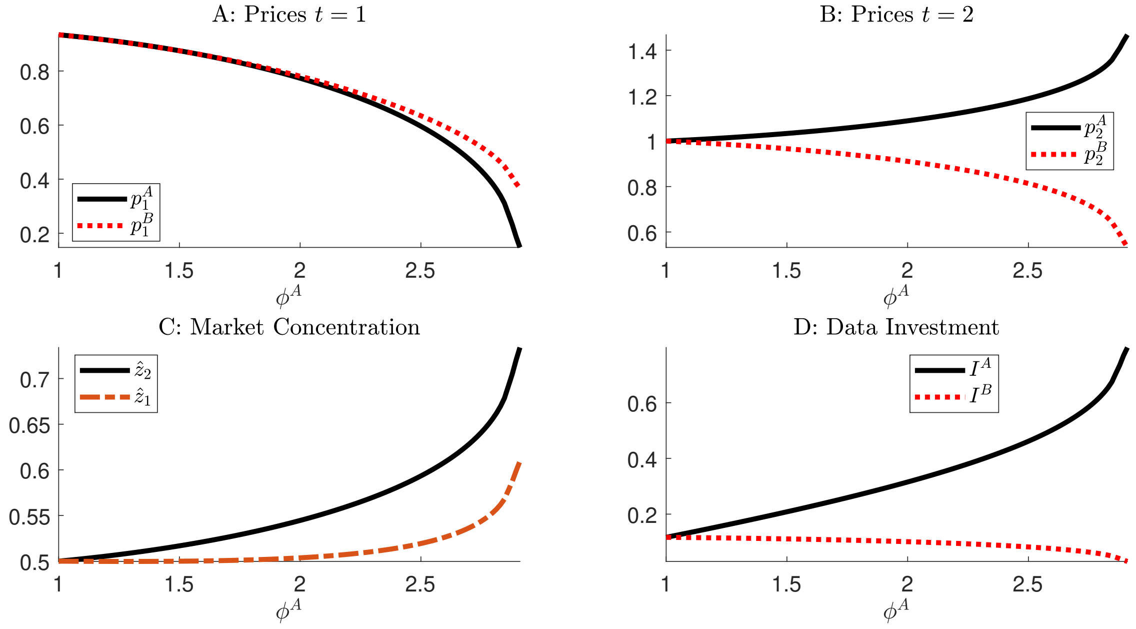

Figure 1: Baseline Solution and Laissez-Faire Equilibrium. Comparative statics with respect to ϕ^A. The relevant parameters are $\lambda = \kappa = \phi^B = 1$ and $\Delta_K = 0$. The model's qualitative outcomes are robust to the choice of these parameters.

strength (i.e., $\gamma^x N_t^x$) is concave in adoption (see, e.g., Cong, Li, and Wang, 2022). This assumption is intuitive and sensible: The incremental value an additional user adds to a platform (e.g., a marketplace or social network) tends to be larger when the number of existing users is small.[29] While in our setting high market power $\hat{z}_2$ (mainly) harms users through high service prices, we note that our key findings would likely not change much, if we introduced different channels through which high market power may harm consumers.[30]

3 Policy Interventions and Market-Based Solutions

In what follows, we study how data-driven competition affects user welfare. We call an equilibrium "inefficient" relative to another one if it leads to lower user welfare.[31] Indeed, the focus on user welfare as the key objective is consistent with data-related antitrust and regulatory proposals which often aim to improve consumer protection and privacy or to break excessive market power and "data

[29] For instance, take an online marketplace where buyers and sellers meet and trade products, such as Amazon or Alibaba. Then, quite likely, buyers (of a specific product) value the arrival of the first seller (of this product), whose arrival is necessary for any trade to occur, more than the arrival of the n'th seller, whose arrival is not necessary for trade to occur but may (slightly) improve the terms of trade (when n is large). A more formal way to capture this would be to assume that $\gamma^x = \gamma(N_t^x)$ is function of N_t^x with $\gamma(\cdot)$ being decreasing in its argument (e.g., $\gamma(x) = \gamma^A + (\gamma^B - \gamma^A)\mathbb{I}_{\{N_t^x < 0.5\}}$), which is similar to the specification in Cong et al. (2022). In this case, total network effect strength, that is, $\gamma^x(N_t^x)N_t^x$, is concave in N_t^x.

[30] For examples such channels, see, for instance, Bergemann et al. (2015) or Brunnermeier et al. (2021).

[31] We thus do not focus on total surplus or platforms' payoff when evaluating different antitrust policies.

monopolies." Section 5.2 provides an extended discussion of our objective.

Our previous analysis highlights two potential inefficiencies regarding data collection. First, there can be over-collection of data when ϕ^A is sufficiently larger than ϕ^B: Data collection, while improving product quality, concentrates market power which harms users via high prices. Second, there can be under-collection of data when the platforms are (approximately) symmetric, because users' privacy concerns and fierce price competition in $t = 2$ reduce platforms' incentives to invest.

We now discuss several potential policy interventions or, more generally, solutions to improve user welfare, which, depending on the parameters, require reducing market dominance by platform A or stimulating data-related investment. We evaluate four different scenarios: (i) Giving data sharing options to users (as, e.g., implemented by GDPR), (ii) data sharing (as, e.g., implemented in open data and finance initiatives), (iii) a market for data when users own data and sell these data to platforms and (iv) a market for data in which platforms trade data. Points (iii) and (iv) suggest potential market solutions that could also be organized by private entities or data intermediaries facilitating trade (see, e.g., Ichihashi, 2021a).

In a nutshell, the remainder of Section 3 shows that none of the approaches (i), (ii), (iii), and (iv) can address both inefficient under- or over-collection and thus, depending on the type of inefficiency ensuing, backfire and reduce user welfare relative to the laissez-faire benchmark ("baseline") from Section 2.2. Accordingly, we introduce in Section 4 user union, which coordinates and represents users, and shows that user union unambiguously improves user welfare.

3.1 Data Privacy Protection

Privacy protection policies such as GDPR (implemented by the European Union) or CCPA in the U.S. strengthen individual ownership rights over personal data by granting rights to access, correct, and delete personal data held by firms. Generally speaking, under privacy protection policies, firms can only process personal data under limited and specific circumstances, such as an individual's explicit opt-in consent.[32] Without other frictions, explicit opt-in or opt-out consents can be easily modeled in our framework through endogenizing $\theta \in [\underline{\theta}, 1]$.

However, as we show in Proposition 1, the equilibrium quantities and, in particular, user welfare is independent of the "data price" q^x or $\underline{\theta}$. Thus, if — by regulation — users own their data and can opt out or opt in of providing data to platforms for free, then $\underline{\theta} = 0$ (instead of $\underline{\theta} = 1$) holds

[32] Additional background information about privacy protection regulation can be found in Appendix F.4. In the U.S., opt out is the most prevalent default option. Examples include the 1999 Gramm-Leach-Bliley Act, the 2000 Fair Credit Repor 2000 Telephone Consumer Protection Act. In the European Union, the opt-in system underlie Parliament and Council Directive 95/46/EC and the European Uni Directive of 1995.

in our framework, and platforms could incentivize users to share their data by setting $q^x = c^x$ (possibly instead of $q^x = 0$). While users now earn additional payoffs from sharing their data, the platforms pass the additional cost of collecting data onto users via higher service prices (relative to a situation with $\underline{\theta} = 1$ and $q^x = 0$), which exactly cancels out the additional user payoff from the compensation. As a result, user welfare and the market share of platform A remain unchanged when regulation transfers data ownership to users and in particular do not depend on the exact value of $\underline{\theta}$ (see Proposition 3). Consequently, within our framework, data privacy protection through opting in or opting out of data contribution has no meaningful effects. This changes drastically when the data can be ported out of the platform where they are generated, which we discuss later when we analyze data markets with user or platform ownerships.

3.2 Data Sharing and Open Data Initiatives

While policies such as GDPR have focused on data ownership rights, open data initiatives, e.g., in the form of data sharing initiatives, have emphasized data access. Recently, there have been many attempts to promote open data access, including Open Banking and Open Finance initiatives (He et al., 2022; Goldstein et al., 2022). For example, the International Data Spaces Association constitutes a private investment for secure data sharing (Richter and Slowinski, 2019).[33] China and South Korea have built open platforms for data sharing to aggregate scattered, isolated, and varied data to help integrate technology and business data to lower information barriers.[34] The EU proposed the Digital Market Act (DMA) in 2020, explicitly emphasizing data sharing for a fair competition. Extended background information on open data initiatives can be found in Appendix F.4. We now study the effect of open data initiatives and (mandatory) data sharing through the lens of our model. The following Proposition analytically characterizes the effects of data sharing when platforms are symmetric, whereby x must share fraction η of its data with its competitor $-x$ "for free," so that $D^x = \hat{D}^x + \eta \hat{D}^{-x}$.

Proposition 4. *Suppose that the platforms are symmetric. Consider a symmetric equilibrium whereby platform x shares fraction η of its data at the beginning of period $t = 2$ with platform $-x$ at zero cost. In period $t = 2$, prices satisfy $p_2^x = \kappa$. In period $t = 1$, prices and investment satisfy*

$$p_1^x = \kappa + I^x \left(q^x - \frac{2(1-\eta)\phi^x}{3} \right) \quad \textit{and} \quad I^x = \min \left\{ \left[\frac{\phi^x(1-\eta) - 3c(1+\eta)}{6\lambda} \right]^+, 1 \right\}.$$

[33] See the Digital Markets Act by the European Commission.

[34] See here.

Price p_1^x increases and investment I^x decreases with η, so that the data generated on platform x (i.e., $\hat{D}^x = I^x/2$) decreases with η. Provided $I^x \in (0,1)$, data sharing decreases the amount of data $D^x = (1+\eta)\hat{D}^x$ that platforms use in period $t = 2$, in that $\frac{\partial D^x}{\partial \eta} < 0$. Under data sharing with $\eta > 0$, user welfare u_1 is (strictly) lower than under the baseline with $\eta = 0$ (when $\phi^x > 3c$). User welfare (strictly) decreases with η (when $\phi^x > 3c$).

Data sharing promises two potential benefits: First, it increases any platform's stock of data and thus overall service quality and, second, it may reduce excessive market power. However, quite strikingly, Proposition 5 shows that data sharing does not succeed in the first pursuit, as it leads to less investment in data collection and even reduces rather than increases a platform's stock of data D^x in period $t = 2$ (i.e., $\frac{\partial D^x}{\partial \eta} < 0$ when $I^x \in (0,1)$). Data sharing requirement (η) undermines platforms' incentives to collect data via investment I^x and low prices in period $t = 1$, in that I^x decreases and p_1^x increases with η. The reason is that data sharing creates a free-rider problem with respect to platforms' incentives to collect data. With data sharing, individual platforms bear the cost of gathering data but the benefits are mutualized. In addition, data sharing makes users' data more widely available and therefore raises the privacy-induced cost of data collection $c^x = c(1+\eta)$. Anticipating that their data are eventually shared with other platforms too, users require then higher compensation (either indirectly via low prices p_1^x or directly via q^x) for sharing data with platform x in the first place.

As a result, when platforms are symmetric and market power is not a concern, data sharing leads to under-collection of data as well as increases period-1 services prices due to platforms' reduced incentive to attract users for data. Data sharing then unambiguously hurts user welfare, in that user welfare u_1 decreases with data sharing intensity η. These findings are consistent with recent empirical studies on open data initiatives. Martens et al. (2020) find that user welfare is not maximized under mandatory data sharing due to increases in product price, and Babina et al. (2022) find that customer-directed data sharing can reduce ex-ante information production.

Next, to illustrate how data sharing can mitigate the concentration of market power in $t = 2$, we consider the general case in which platforms are not necessarily symmetric with full data sharing, $\eta = 1$. That is, $D^x = I^A N_1^A + I^B N_1^B$, and there are no monetary transfers associated with data sharing. The following Proposition characterizes the equilibrium.

Proposition 5. *Consider $\eta = 1$, and $\phi^A \geq \phi^B$. Then, in equilibrium, only platform A collects data, so that $I^A \geq 0 = I^B$ and $D = D^x = N_1^A I^A$. Suppose that $\phi^A > \phi^B$, in which case $I^A > 0 = I^B$. When $\lambda > 0$ and $c > 0$ are sufficiently small, then platform A's market share in $t = 2$ is strictly lower than under the baseline.*

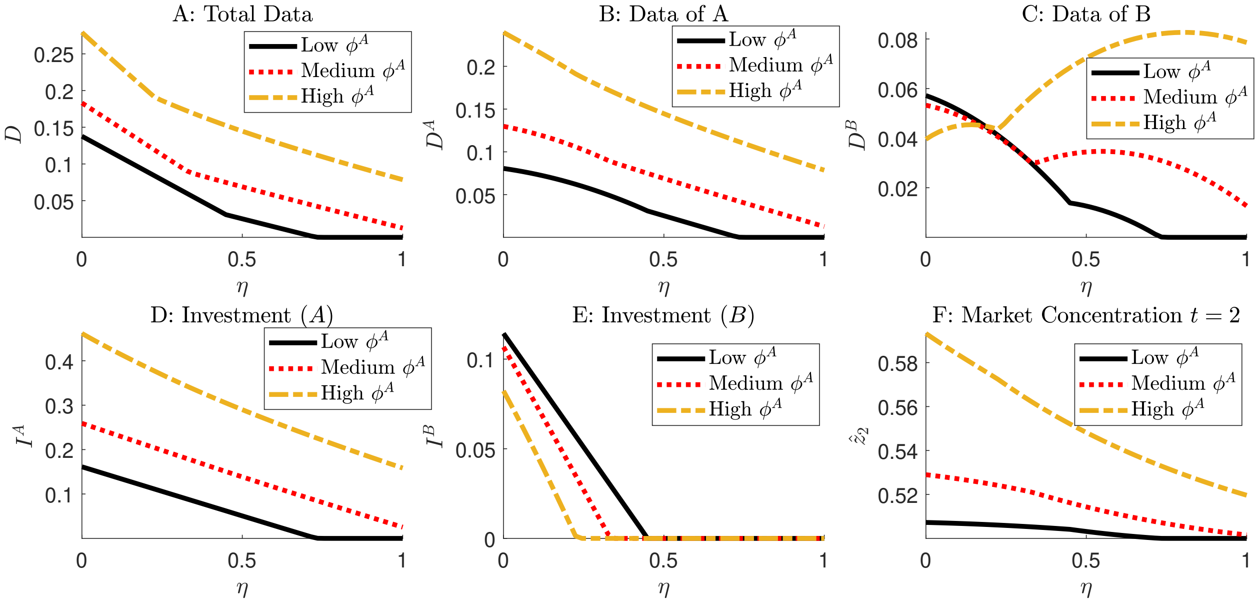

Figure 2: The Effects of Data Sharing. Comparative statics with respect to η. The relevant parameters are $\lambda = \kappa = \phi^B = 1$, and $\Delta_K = 0$. The low level of ϕ^A is 1.25, the medium level is 1.75, and the high level is 2.5.

Similar to Proposition 4, data sharing undermines incentives to collect data, leading to the stark outcome $I^B = 0$. We can show that under certain parameter conditions (i.e., when λ and c are not too large), data sharing reduces the market share and therefore price-setting power of platform A in period $t = 2$. Figure 2 graphically illustrates these results by showing that total data D (Panel A), market power in $t = 2$ (Panel F), as well as platform investments (Panels D and E) decrease with the extent of data sharing η for different levels of ϕ^A. Strikingly, data sharing reduces platform A's stock of data D_2^A in period $t = 2$ (see Panel B), platform B's stock of data D_2^B (see Panel C), except when ϕ^A is sufficiently large, and total data D (see Panel A). As such, the following paradox arises: Data sharing agreements — put in place with the goal of maximizing available data — can actually reduce platforms' overall stock of data.

3.3 Data Market When Users Own Data

As the mandatory data-sharing requirement appears mechanical and inefficient, we now examine a market-based solution to data sharing, potentially operating via data intermediaries (Ichihashi, 2021a).[35] Suppose that users own their data, including data that is generated with their interactions on platforms, and can sell their data to platforms at the beginning of $t = 2$ (before period-2 prices are chosen), while platforms must buy these data to access it. In the baseline and all other model

[35] For simplicity, we do not explicitly model data intermediaries and view them as a perfect passthrough. The IO, Pricing, and design issues in data intermediation are interesting on their own and worthy of separate studies such as those pioneered in Ichihashi (2022) and Bergemann, Bonatti, and Smolin (2018).

variants that we have considered so far, users' consumption on the platform and their contribution to the platform in the form of data were linked or "bundled." A market for data unbundles consumption and contribution. Users can contribute their data to platforms they do not use and, similarly, can consume on platforms on which they do not contribute. Notably, this is different from privacy protection analyzed in Section 3.1, where users could sell their data exclusively to the platform which they adopted in $t = 1$. As such, with the data market, the period-1 marginal user satisfies $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}$, i.e., relative to (6), which holds in the baseline and with mandatory data sharing, the terms related to users' data contribution decision θ^x drop out.[36]

In more detail, consider any user z who has consumed at platform x in $t = 1$ and therefore owns I^x units of data which, notably, depend on the platform's investment. User z is willing to sell data to platform $x' \in \{A, B\}$ as long as the price compensates for her privacy loss, i.e., user z requires at least a payment of cI^x dollars (or a per unit price of c).[37] When buying data, platform x' optimally offers user z per unit price of data of c dollars (i.e., in total $I^x c$ dollars), which is the lowest price possible to induce any user to accept trade at this price.[38]

Given I^x and N_1^x, the total stock of data reads $D = I^A N_1^A + I^B N_1^B$. After the market for data in period $t = 2$ determining D^x, the platforms simultaneous choose prices to maximize $N_2^x p_2^x$, leading to the outcomes in Lemma 1. At the beginning of period $t = 2$, platform x decides on the quantity of data D^x that it wishes to buy to maximize: $\max_{D^x \in [0,D]} \pi_2^x - cD^x$, taking the choice of the other platform D^{-x} as given and taking into account the data-dependent (continuation) payoff π_2^x in Lemma 1. At the beginning at time $t = 1$, platform x optimizes:

$$\max_{p_1^x, I^x} \left(N_1^x p_1^x - \frac{\lambda (I^x)^2}{2} + \max_{D^x \in [0,D]} \left[\pi_2^x - cD^x \right] \right), \tag{10}$$

whereby period-1 demand is characterized by $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}$ and $D = N_1^A I^A + N_1^B I^B$.

Proposition 6 studies the equilibrium under this model variant in the symmetric platform case.

Proposition 6. *Suppose that both platforms are symmetric. When users own data and can sell*

[36] As data contribution and consumption decisions are now unbundled and independent, all terms in (6) that are multiplied by θ^x drop out, which is akin to setting $\theta^x = 0$. More in detail, anticipating that she does not make a gain from selling data in $t = 2$ (i.e., any user breaks even from selling data in $t = 2$), the net utility of user z in $t = 1$ equals $u_1^x(z) = Y_1^x - p_1^x - \kappa^x(z)$. The marginal user satisfies $u_1^A(\hat{z}_1) = u_1^B(\hat{z}_1)$. This leads then to the marginal user $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}$, pinning down market shares in $t = 1$, i.e., $N_1^A = \hat{z}_1$ and $N_1^B = 1 - \hat{z}_1$.

[37] The structure is as follows: platform x' sets a per unit price for data and the atomistic users take this price as given and decide whether to sell their data to x'.

[38] Intuitively, as users are atomistic and the platforms are "large," the platforms can therefore extract the whole surplus from data trading in $t = 2$. Also notice that a Nash bargaining solution to determine the price of a data trade between an individual user and a platform, whereby atomistic users have zero bargaining power, would imply that users just break even and receive per unit price c for data.

these data in a market that opens at the beginning of period $t = 2$ (before platforms set prices p_2^x), then, in the (unique) symmetric equilibrium, prices satisfy $p_1^x = p_2^x = \kappa$. Platforms' investment is $I^x = 0$, so $D^x = D = 0$. User welfare is (strictly) lower than under the baseline (when $\phi^x > 3c$).

Proposition 6 illustrates that — just like data sharing with $\eta = 1$ (compare Proposition 4) — the market for data wipes out platforms' incentives to invest. Notably, key equilibrium quantities, such as market concentration and platform investments, are identical under data sharing with $\eta = 1$ and the model variant with a market for data in which users can sell their data. Paradoxically, the market for data, in which users own and sell their data, harms users and their welfare when platforms are symmetric. The intuition behind this result is as follows. Because data are non-rival, users are tempted to sell their data to both platforms and, in particular, also to platforms that they did not join in $t = 1$, leading to $D^x = D$ as under data sharing with $\eta = 1$. The fact that data generated on platform x is also used by platform $-x$ implies a free-rider problem regarding data collection investment, causing under-collection of data. The key inefficiency here is that individual users do not internalize the adverse effect of their data sales on investment. The next Proposition highlights that the market for data, whilst curbing data collection, can mitigate A's market power.

Proposition 7. *When users own data and can sell these data in a market that opens at the beginning of period $t = 2$ (before platforms set prices p_2^x), then there exists a unique equilibrium with the following properties. First, platforms either buy all available data or no data at all, in that $D^x \in \{0, D\}$. Second, when $\phi^A \geq \phi^B$, only platform A collects data, so that $I^A \geq 0 = I^B$ and $D = D^x = N_1^A I^A$. Suppose that $\phi^A > \phi^B$, in which case $I^A > 0 = I^B$. When λ and c are sufficiently small, then $D^x = D$ and platform A's market share in period $t = 2$ is strictly lower than under the baseline.*

3.4 Data Market When Platforms Own Data

We now consider a market for data when platforms own the stock of data collected in period $t = 1$. Suppose that at the beginning of period $t = 2$, before platforms set prices p_2^x, a market for data opens that allows the two platforms to trade data. We denote the stock of data of platform x before the data trade by $\hat{D}^x = N_1^x I^x$ (assuming $\theta^x = 1$ in optimum) and the stock of data of platform x after the data trade by D^x. As data are non-rival, $D^x \geq \hat{D}^x$, i.e., if platform x sells data to platform $-x$, the stock of data of platform $-x$ increases but the stock of data of platform x does not decrease. Since our setup features only two platforms, we consider that the two platforms determine the optimal allocation of data and the transfer via Nash Bargaining, whereby platforms

A and B have, for simplicity, equal Nash bargaining weight $1/2$.[39] The following Lemma shows how the two platforms determine the optimal allocation of data at the beginning of $t = 2$ to maximize their joint continuation surplus.

Lemma 3. *Suppose that $\hat{D}^A\phi^A + \Delta_K \geq \hat{D}^B\phi^B$ as well as $\phi^A \geq \phi^B$. Then, the two platforms' joint payoff $\pi_2^A + \pi_2^B$ is maximized upon $D^A = \hat{D}^A + \hat{D}^B$ and $D^B = \hat{D}^B$. Nash Bargaining at the beginning of $t = 2$, after data in $t = 1$ has been collected and before prices p_2^x are set, therefore leads to the optimal allocation of data of $D^A = \hat{D}^A + \hat{D}^B$ and $D^B = \hat{D}^B$. When $\hat{D}^A\phi^A + \Delta_K < \hat{D}^B\phi^B$, we relabel the platforms (and interchange Δ_K with $-\Delta_K$) and above statements apply too.*

The lemma illustrates that the stronger platform A buys data from the weaker one (i.e., B) but not the other way around, which also implies $c^A = c < c^B = 2c$.[40] Even if platforms are ex-ante symmetric and enter period $t = 2$ with identical stock of data $\hat{D}^x$, the market for data would inevitably lead to concentration of data ownership. As a result, platform A possesses higher market power after this data trade and users may suffer from this increased market power. Platforms share and trade data to intentionally create a situation in which one platform possesses higher market power, preventing that their period-2 payoffs are dissipated by fierce price competition. Because an appropriate transfer compensates the platform that gives up market share, both platforms are better off. This outcome can be interpreted as "data killer acquisition," whereby the stronger "incumbent" platform acquires data of the weaker "entrant," or as "data collusion."

The detailed description and solution of this model variant, including the first-order conditions with respect to price p_1^x and investment effort I^x, is presented in Appendix C.5. Figure 3 plots key equilibrium quantities (under numerical solution) against ϕ^A both under the baseline (solid black line) and under the model variant with a market for data that is owned and traded by platforms (dotted red line). Panel B shows that indeed the presence of a data market, in which platforms trade data, unambiguously concentrates market power in period $t = 2$. As Panels C and D illustrate, investment I^A by platform A is similar under both model variants (yet weakly higher with a data market), while investment I^B and total stock of data can be higher or lower in either variant.

Finally, to analytically characterize how the market for data affects investment, Appendix F.1 considers ex-ante symmetric platforms A and B (e.g., with $\phi^A = \phi^B$ and Δ_K) and studies a model extension with symmetric equilibrium in the subgame in $t = 1$ and thus $I^A = I^B$ and $p_1^A = p_1^B$.

[39] Alternatively, data intermediaries, as in Ichihashi (2021a), could facilitate this trade too. With perfect competition among intermediaries, the intermediary sector would likely be just a pass-through. We leave this for future research.

[40] When users share data with platform B in $t = 1$, they anticipate that B will sell these data to A in $t = 2$, so they incur privacy-related dis-utility of $c^B = 2c$.

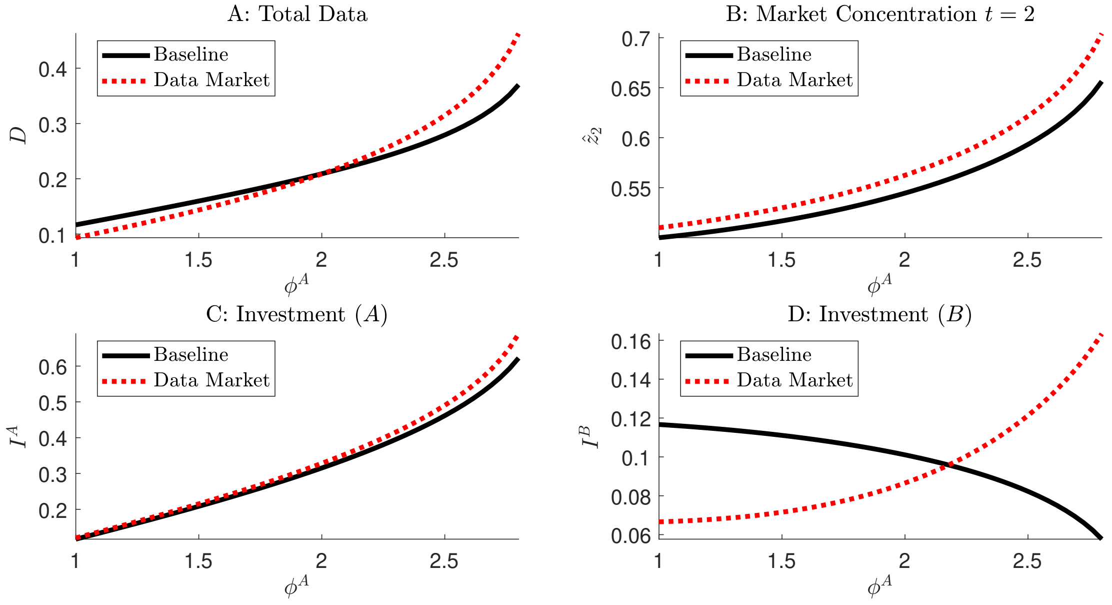

Figure 3: Data Market When platforms Own and Trade Data. Comparative statics with respect to ϕ^A.The relevant baseline parameters are $\lambda = \kappa = \phi^B = 1$ and $\Delta_K = 0$.

Notably, we show in Appendix F.1 that when c is low, the market for data, whilst concentrating market power, stimulates data collection by both platforms, leading to lower period-1 prices p_1^x and higher investment I^x in period $t = 1$ than in the baseline.

3.5 User Welfare Under Policy Interventions and Market-Based Solutions

We now evaluate the welfare effects of the previously discussed policies and market-based solutions. To this end, Figure 4 plots user welfare under the baseline from Section 2.2 (solid black line), data sharing from Section 3.2 with $\eta = 1$ (dashed orange line), market for data with users owning data from Section 3.3 (dotted purple line), and market for data with platforms owning data from Section 3.4 (yellow line) against ϕ^A starting from $\phi^A = \phi^B = 1$.

Recall that data sharing or a market for data with users owning data reduces market concentration but curb investment and increase (period-1) service prices: As such, these approaches tend to increase user welfare (relative to the baseline) when the asymmetry between the two platforms is large and excessive market power is a concern, but reduce it when the platforms are (approximately) symmetric. A market for data with platforms owning the data concentrates market power, which under the chosen parameters (large γ^B) reduces user welfare. Overall, none of the proposed policy interventions and market-based approaches constitutes a robust means to improving user welfare. In particular, any of the policies and market-based approaches can backfire and reduce user welfare (relative to the baseline) under certain parameter configurations. Next, we introduce and model

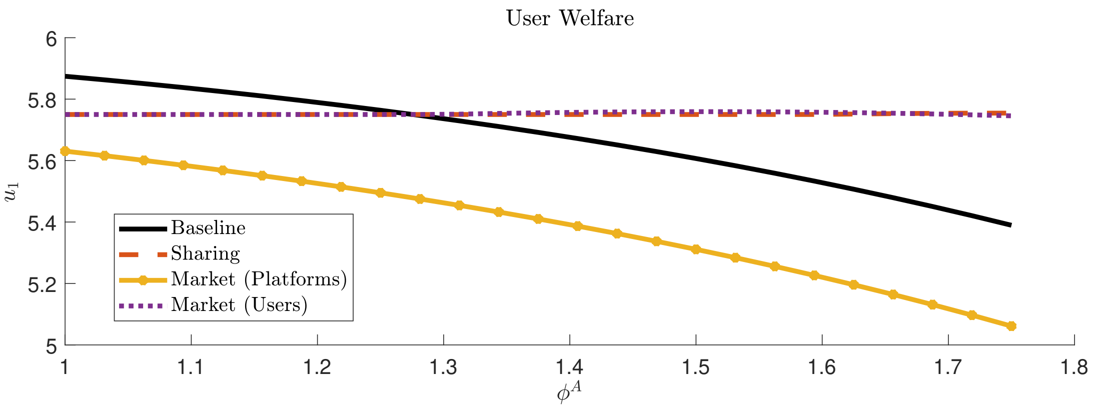

Figure 4: Welfare Effects and Comparison. The relevant baseline parameters are $\lambda = \kappa = \phi^B = 1$ and $\Delta_K = 0$ as well as $\gamma^B = 15 > \gamma^A = 0$. High levels of γ^B (relative to γ^A) imply in our model that high market concentration $\hat{z}_2$ in $t = 2$ has a sufficiently negative impact on user welfare (see Lemma 2).

user union, which, by construction, always increases user welfare relative to the baseline.

4 User Union: Coordinating and Protecting Users

The fundamental inefficiency in our data economy is that users do not internalize the broader effects of their participation and data contribution. We now introduce and evaluate user union and data trust as an alternative solution. Whereas all other policy interventions discussed thus far (e.g, data sharing or a market for data) reduce user welfare under certain circumstances, coordinating users' decisions via a "user union" unambiguously improves user welfare.[41]

4.1 Implementation of User Union

We introduce an implementation of a user union or data trust when there is no data sharing by platforms ($\eta^x = 0$ and $D^x = N_1^x I^x \theta^x$) and users can (but are not required to) contribute data to the platform they joined in $t = 1$, i.e., $\underline{\theta} = 0$. Consider for now that all users $z \in [0, 1]$ are members of this user union.

The user union commits at time $t = 1$, before platforms choose prices and investments, to a reward level of f to induce optimal investment in data collection and to maximize user welfare as follows. The union pays users f dollars per unit of data contributed on any platform x (i.e., in total $f\theta^x I^x$ dollars); when $f < 0$, then users pay the union a fee per unit of data contributed. In a nutshell and as will become clearer later, the user union is analogous to a planner who subsidizes

[41] Note that user union is akin to labor union or firm merger (which has a large literature) in that they all increase the bargaining power of dispersed players against an upstream or larger player. Data sharing and its interaction with network effects distinguish our setting.

or taxes users to incentivize efficient data sharing and collection.

The reward f transforms the privacy cost in a sense that $c^x = c - f$; contributing one unit data, users incur privacy cost c but also receive a reward f from the union, leading to effective cost $c^x = c - f$. For simplicity, the reward f is the same across platforms. Given f, which affects the equilibrium only via $c^x = c - f$, the (continuation) equilibrium is characterized in Proposition 1, yielding user (continuation) welfare u_1. Importantly, because of $\underline{\theta} = 0$, platforms must compensate users in the union for their privacy cost and set $q^x \geq c^x$ for all users in the union.[42]

Total rewards $T := f\theta^A I^A N_1^A + f\theta^B I^B N_1^B$ that the user union distributes for data contributions in $t = 1$ are financed by the users through a membership fee (for being part of the union) which is paid at inception (i.e., before prices and effort are chosen). User z pays, at inception, a membership fee $m(z)$ to the union, where $m(z)$ might vary across z (i.e., type z is observable or contractible for the union) and may become negative negative (membership subsidy). Total rewards T must then equal total membership fees $M := \int_0^1 m(z)dz$, i.e., $T = M$. Importantly, in the current implementation, the user union is completely self-financed and does not make or receive any transfers to or from the platforms.[43]

If $z \leq \hat{z}_1$, user z joins platform A and contributes data to this platform; otherwise, z joins B. We stipulate the membership fee to be $m(z) = fI^A\mathbb{I}\{z \leq \hat{z}_1\} + fI^B\mathbb{I}\{z > \hat{z}_1\}$, which ensures that any user z pays ex-ante the same amount of membership fee that she expects to receive later on as a reward for data contributions.[44] Moreover, this membership fee satisfies by construction $\int_0^1 m(z)dz = T$. We emphasize that the membership $m(z)$ is not contingent on users' decision which platform to adopt in $t = 1$, but rather it is contingent on user type z in anticipation of the equilibrium level of $\hat{z}_1$.

The user union maximizes total (ex-ante) payoff/welfare of its member:

$$\max_f \left(u_1 - f\theta^A I^A N_1^A + f\theta^B I^B N_1^B\right), \tag{11}$$

where u_1, θ^x, and I^x are characterized in Proposition 1 and $M = T = f[\theta^A I^A N_1^A + \theta^B I^B N_1^B]$ is the dollar amount of users' ex-ante membership fee as well as $c^x = c - f$. We note that user union does not take into account platform payoff and that it is assumed that platforms are active. The

[42] In addition to the reward f paid by the user union, users receive compensation q^x from platform x. Thus, user z (joining platform x) chooses the fraction of data θ^x she shares with x according to $\max_{\theta^x \in [0,1]} I^x\theta^x(q^x - c^x)$, so that $\theta^x = 1$ only if $q^x \geq c^x = c - f$.

[43] Such design can prove beneficial in practice when, for instance, bargaining or contracting with platforms is hard or infeasible.

[44] For instance, user $z \leq \hat{z}_1$ joins platform A and shares I^A units of data leading to reward fI^A which equals the membership fee $m(z)$ that she pays.

above implementation of user union does not require any direct interactions (such as negotiations) between user union and platforms; rather, user union resembles a (social) planner or "government" that taxes and subsidizes data contribution by users to induce an efficient allocation of data.

4.2 Analysis

Through adjusting f, the user union makes individual users internalize the broader effects of their data contributions, which helps address the two potential inefficiencies discussed earlier. Setting $f < 0$, the user union can curb data collection by effectively taxing individual users' data contributions, which limits market power and preserves competition in period $t = 2$ at the expense of foregone service quality improvements from data. In contrast, setting $f > 0$, the user union stimulates data collection by effectively subsidizing individual users' data contributions, possibly boosting market power of the dominant platform. When high market power is not an issue (e.g., when platforms are symmetric), the user union subsidizes data collection to boost platform investments relative to the baseline. The following Proposition formalizes this result.

Proposition 8. *Suppose that platforms are symmetric and focus on a symmetric equilibrium. When* $6c \geq 7\phi^x$*, the user union sets* $f = 0$ *and implements* $I^x = 0$*. When* $6c < 7\phi^x$*, the user union induces investment* $I^x = 1$ *which exceeds investment in the baseline equilibrium, data sharing equilibrium, or the equilibria with markets for data. The optimal fee satisfies* $f = f^* = \frac{6\lambda+3c-\phi^x}{3}$*.*

Next, when ϕ^A is large compared to ϕ^B and data collection leads to excessive market concentration $\hat{z}_2$ in period $t = 2$, the more pressing inefficiency harming user welfare is the lack of period-2 competition. Then, the user union effectively taxes data collection through $f < 0$, thereby reducing investments and market concentration relative to the baseline. Figure 5 graphically illustrates this result by plotting total data collected (Panel A), A's investment (Panel B), B's investment (Panel C), the optimal reward f solving (11) (Panel D), and market concentration in $t = 1$ (Panel E) and $t = 2$ (Panel F) against ϕ^A, both under the baseline (solid black line) and with the user union (dotted red line). Indeed, for low values of ϕ^A, the user union stipulates $f > 0$ to boost data investments I^x, raising market concentration and the total stock of data relative to the baseline. For larger values of ϕ^A, the user union stipulates $f < 0$ to curb data investments, reducing market concentration and data collection relative to the baseline. Also note that to the extent a higher ϕ^x corresponds to firms with a large amount of historical data and experience handling data, user union remedies an unequal landscape by reducing overall data contribution and platform investment. The discontinuity in Figure 5 corresponds to the point at which the user union effectively

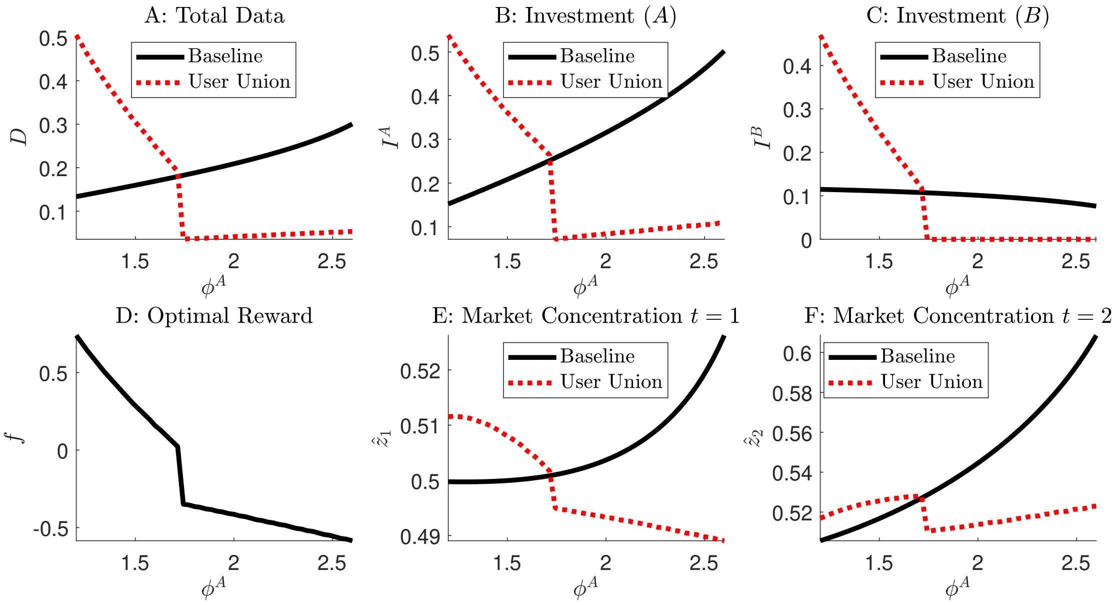

Figure 5: The effects of a with user union. Comparative statics with respect to ϕ^A. The relevant baseline parameters are $\lambda = \kappa = \phi^B = 1$, $\Delta_K = 0$, and $\gamma^B = 15 > \gamma^A = 0$. High levels of γ^B (relative to γ^A) imply in our model that high market concentration $\hat{z}_2$ in $t = 2$ has a sufficiently negative impact on user welfare (Lemma 2).

switches from the objective of boosting data collection to the objective of curbing data collection in favor of competition.

By construction, user union always improves user welfare compared with the baseline. Thus, the implementation of a user union is a robust means to improve user welfare relative to the baseline, whereas other policy interventions or market-based solutions do not lead to unambiguous welfare gains (see Section 3.5). The reason underlying this result is that the user union is the *only* policy intervention that can address both potential inefficiencies (albeit not always simultaneously), i.e., (i) under-collection or (ii) over-collection of data. All other policy interventions target at most one inefficiency. Data sharing or a market for data operated by users may reduce market power by one platform but curbs platforms' investment (see Figure 2). A market for data operated by platforms may increase investment, but leads to concentration of market power in the long run (see Figure 3). Finally, Figure 6, which is similar to Figure 4 but adds user welfare under user union (solid red line), highlights how under our baseline parameters user union dominates all other policy interventions and the baseline.

4.3 Incentives to Join the User Union

Depending on the exact implementation and regulatory environment regarding user union in the future, union membership may be voluntary or (implicitly) required, e.g., when union membership

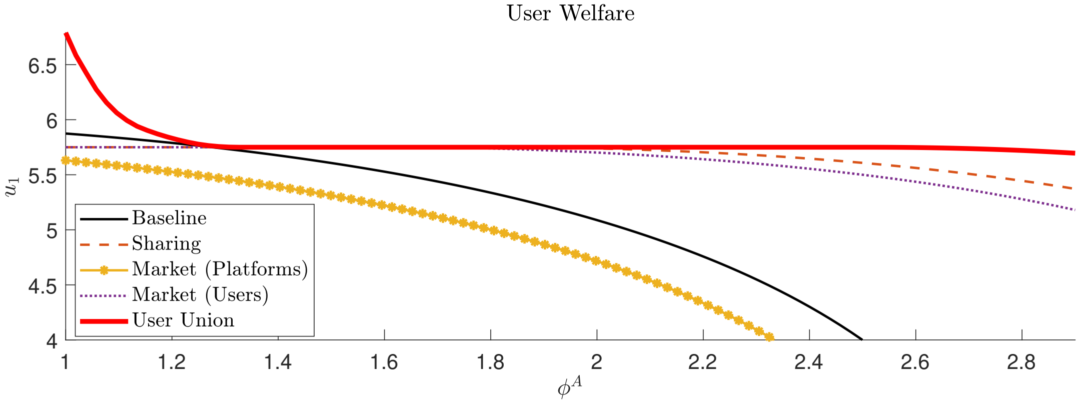

Figure 6: Welfare Effects and Comparison with User Union. The relevant baseline parameters are $\lambda = \kappa = \phi^B = 1$ and $\Delta_K = 0$ as well as $\gamma^B = 15 > \gamma^A = 0$. High levels of γ^B (relative to γ^A) imply in our model that high market concentration $\hat{z}_2$ in $t = 2$ has a sufficiently negative impact on user welfare (see Lemma 2).

is a prerequisite for joining an ecosystem/platform or there are (unmodelled) costs associated with leaving the union.[45] It is interesting to study users' incentives to deviate by not joining the user union through the lens of our model (whenever this is possible), starting from the equilibrium with full union membership characterized before. We assume that, after user z joins platform x, platform x can observe whether z is member of the union (i.e., a user cannot fake membership as is the case, e.g., for labor union). If user z is not member of the union, platform x can offer z a potentially different data price $\hat{q}^x$ (e.g., $\hat{q}^x < q^x$) than the data price q^x that it offers to union members. This reflects the idea that members of the "data union" receive different and potentially better compensation for contributing data than non-members, which has its natural analogue in the context of labor unions where non-members might be paid different wages than members. Appendix E.3 formally analyzes users' incentives to join user union under these circumstances and our implementation of user union. In particular, Appendix E.3 shows that when $I^A > 0$ and $I^B > 0$, there exists an equilibrium (unique up to q^x) in which users find it privately optimal to join user union (i.e., union membership is incentive compatible) under the optimal choice f solving (11). Then, all equilibrium quantities (given f) follow from Proposition 1.

4.4 An Alternative Implementation

Another implementation of user union is a "data trust" which collects data from all users (generated by their interactions with platforms) in period $t = 1$ and sells these data to platforms on users'

[45] We do not impose constraints on incentive compatibility of membership in the formal optimization (11) but rather assume that all users are members of the union, which could be because it is incentive compatible to join (see conditions in Appendix E.3), because membership is required to join the platforms, or because there are unmodelled costs of not joining the union.

behalf at endogenous (per unit) price q, where we assume that both platforms can buy data at the same price (i.e., there is no price discrimination by the trust). The proceeds from data sales are then distributed to users via payouts. Specifically, suppose that user union collects all user data generated on platform x through users' interactions, that is, $N_1^x I^x$ units of data. The amount of data depends on platform investment I^x, but the data are owned by the trust and can be bought by platforms. To prevent a free-rider problem regarding data collection, we stipulate that the data trust limits the amount of data platform x can buy, so that $D^x \leq \overline{D}^x$, with the limit $\overline{D}^x$ being increasing in the amount of data $\hat{D}^x$ that has been generated on platform x. For simplicity, we consider $\overline{D}^x = \hat{D}^x$, so that $D^x \leq \hat{D}^x$ and x cannot acquire more data than it has generated.[46] We expect that our results remain similar under a (slightly) different constraint.

With data trust implementation, x decides at the beginning of period $t = 2$ how much data to buy from the trust at price q. That is, x solves $\max_{D^x \in [0, N_1^x I^x]} \pi_2^x - qD^x$, with π_2^x from Lemma 1. It is clear that $D^x = N_1^x I^x = \hat{D}^x$ in optimum; if $D^x < N_1^x I^x$, platform x could profitably reduce costly investment I^x at time $t = 1$. In period $t = 1$, no data is yet shared with platforms, so, to account for that, we set $\theta^x = 0$ (e.g., in (6)) when calculating market shares in period $t = 1$. Instead, data are allocated according to the market that opens at the beginning of $t = 2$. Users do not incur direct privacy costs from sharing data with the union/data trust, but only incur privacy losses if the trust shares these data with platforms. The data trust mandates pro-rata payouts to any user (irrespective of which platform she attends) of $N_1^A I^A q + N_1^B I^B q$ dollars which are the total proceeds from data sales, and maximizes

$$\max_q \left[u_1 + N_1^A I^A (q - c) + N_2^B I^B (q - c)\right], \tag{12}$$

which is the sum of user utility from consumption u_1 and the proceeds from data sales net of privacy cost. As payouts are independent of adoption, adoption decisions do not directly depend on I^x. The following Proposition summarizes the equilibrium when platforms are symmetric.

Proposition 9. *Suppose that both platforms are symmetric and focus on a symmetric equilibrium. When $6c \geq 7\phi^x$, data trust sets $q = 0$ and implements $I^x = 0$. When $6c < 7\phi^x$, the data trust induces investment $I^x = 1$ which exceeds investment in the baseline equilibrium, data sharing*

[46] Equivalently, we could assume that platform x has the exclusive right to buy data, which has been generated on platform x, from the data trust at per-unit price q. That is, platform x $(-x)$ cannot buy data generated on the competing platform $-x$ (x). The assumption $D^x \leq \hat{D}^x$ is for simplicity, but could be easily relaxed or dealt with by using a more sophisticated implementation. In principle, we could allow user union to optimally price-discriminate, leading to two different data prices $q^A \neq q^B$ and thereby controlling data allocation to platforms, or to implement a pricing schedule $q^x(D^x)$, whereby price $q^x(D^x)$ increases with the amount of data D^x sold to x, so as to induce $\hat{D}^x \geq D^x$.

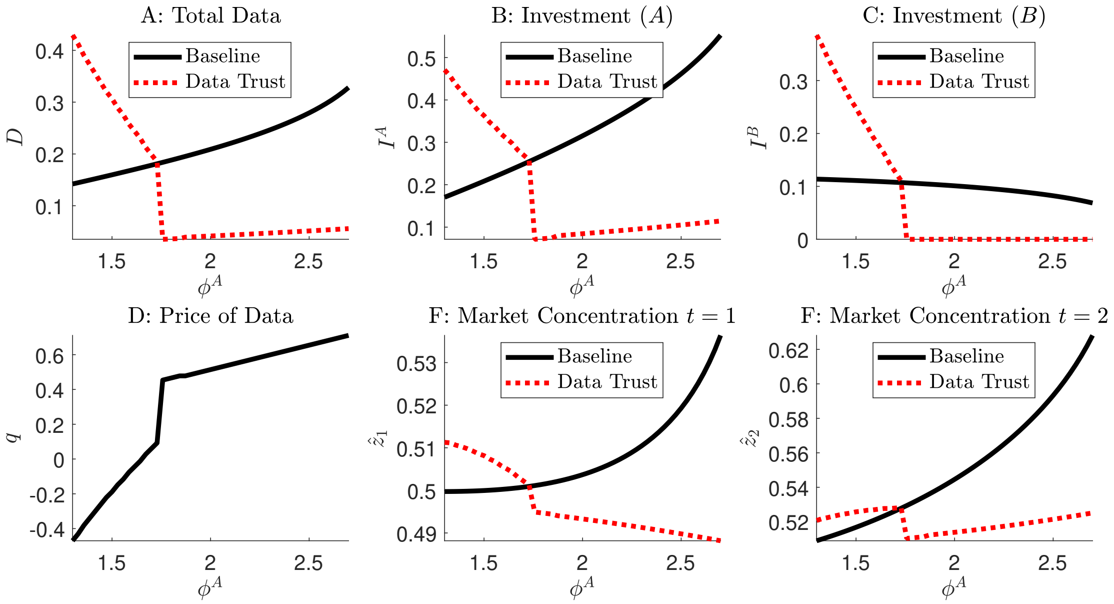

Figure 7: The effects of a data trust. Comparative statics with respect to ϕ^A. The parameters follow Figure 5.

equilibrium, or the equilibria with markets for data. The price for data satisfies $q = \frac{\phi^x}{3} - 2\lambda$.

Interestingly, the data trust from Proposition 9 induces the same investment $I^x = 1$ and user welfare u_1 as the user union from Proposition 8 in the case of symmetric platforms.[47] Finally, Figure 7 plots equilibrium quantities under the data trust against ϕ^A. The findings qualitatively resemble the ones obtained under the user union in Figure 5. When ϕ^A is close to ϕ^B (=1), the data trust stimulates and subsidizes platform data collection by stipulating a low price for data q which, notably, can even become negative. In contrast, when ϕ^A is large, the data trust stipulates a high price for data, thereby curbing data collection in favor of competition. Thus, the price for data q increases in platforms' differences in data collection ability ϕ^x.

Our suggested implementations of user union suffice for illustrating its power. Improving the design or deriving the optimal design constitutes interesting future work. For instance, one straightforward improvement would be to stipulate platform-contingent rewards f^x (so that not necessarily $f^A = f^B$) and $c^x = c - f^x$ or, in the data trust implementation, to allow for price discrimination (i.e., the trust can charge different prices q^x to the two platforms). Moreover, we could also allow the user union to choose θ^x on users' behalf or that user union engages in collective (Nash) bargaining with the two platforms, e.g., to determine the price of data. Having more than one instrument, user union could then address both inefficient under-collection or over-collection of data by the individual platforms simultaneously (e.g., user union could curb data collection by A,

[47] Generally, there is no equivalence between the two implementations.

whilst stimulating it by B).

4.5 User union in practice

Over the past few years, legal entities in the form of data trusts start to emerge (Houser and Bagby, 2022). Willis Tower Watson (WTW) data trust is such a pilot. We thus have used data trust and user union interchangeably. However, user union encompasses but is not restricted to data trusts. User union emphasizes users, whereas data trusts could in principle be owned by large platforms. What we introduce can be viewed as a user-owned data trust.

Before the introduction of data trusts, practitioners have explored the concepts of data pool or cooperative, where users contribute and share data. Examples include Driver's Seat and MIDATA.coop. However, data pools are typically governed by a small subset of users and it is unclear if its goal is to maximize user welfare and distribute revenues from data in an equitable way to users. Corporate and contractual mechanisms as embodied in Nallian, a transportation data platform, focus on interoperability and not privacy protection and data compensation.

One could imagine having the governments or regulators potentially create user unions, but the right incentives and expertise have to be in place. The concept of DataDAO provides an alternative based on distributed ledgers, which can incorporate privacy protection and secure multi-party computation (Sockin and Xiong, 2022; Cao, Cong, and Yang, 2018). Streamr.network is one of the attempts at blockchain-based data union and monetization.

Finally, one additional benefit of a structure like the data trust is outside our model but is relevant in practice. To the extent that individual users often incur attention costs to work with platforms concerning data contribution and sharing arrangements (Holdren and Lander, 2014; Jain, Gyanchandani, and Khare, 2016), user union can reduce the duplication of effort of individual users.

5 Extended Discussions

5.1 Entry and Operating Cost

The previous analysis has assumed that platforms always participate (e.g., because parameters are such that platform payoff is positive). We could formally consider platform entry or platforms' choice to be active in the market. For this sake, one could introduce that at the beginning of period t, any platform x incurs fixed cost ρ_t^x. Then, platform x decides whether or not to cover the fixed cost: If it does, it can operate and sell products in period t; otherwise, x exits the market forever. Under these circumstances, platform x only operates in period t if $\tilde{\mathbb{E}}_t^x[\pi_t^x] \geq \rho_t^x$, where the expectation $\tilde{\mathbb{E}}_t^x[\cdot]$ is over the potential strategic uncertainty on whether the competing platform $-x$ operates in t and π_t^x is the period-t payoff conditional on operating in t.

If one platform, say B, is not active in the market at t, then the other platform A sells products as a monopolist from period t onward. Suppose that platform A covers the entire market, in that $N_t^A = 1$. Then, the price p_t^A is such that consumer $z = 1$ break even, that is:

$$p_t^A = K^A + \gamma^A - \hat{\kappa} + \phi^A D^A \mathbb{I}_{\{t=2\}} + \theta^A I^A (q^A - c^A)\mathbb{I}_{\{t=1\}}. \tag{13}$$

As such, the per-period in $t = 2$ welfare is $\frac{\hat{\kappa}}{2}$. It is immediate that users are better off and per-period user welfare in t is higher under our baseline when both platforms A and B operate in period t.

Consider now the scenario that platform B does not enter, so A covers the entire market as monopolist in both periods $t = 1, 2$ and user welfare is $u_1 = \hat{\kappa}$. Without loss of generality, we consider that $q^A = c^A$, so that platform A compensates users for their privacy cost from sharing data. Given $N_1^A = \theta^A = 1$, platform A then chooses I^A to maximize $(\phi^A - c^A)I^A - \frac{\lambda}{2}(I^A)^2$, so

$$I^A = I^{Mon} = \min\left\{1, \left[\frac{\phi^A - c^A}{\lambda}\right]^+\right\}.$$

It is straightforward to see that when the market is served by one monopolist, then investment in data technology I^A is larger than in the symmetric platform case, i.e., when two platforms split the market equally (see Proposition 2). However, users do not benefit from the surplus that data collection generates and their welfare is strictly lower than when both platforms operate in $t = 1, 2$, since the lack of price competition allows the monopolist to extract this surplus.

Observe that with costly entry cost (for simplicity, only incurred at time $t = 1$, i.e., $\rho_1^x > \rho_2^x = 0$), platform B enters only if its expected profits are sufficiently high, which is when platform A is not too strong. If A is sufficiently strong and has a high market share in period 1 or 2 under duopoly, the payoff of platform B under duopoly (i.e., conditional on entry) is relatively low. Then, with non-trivial entry costs and anticipating the continuation game, platform B might not find it optimal to enter at inception, which then leads to the monopoly outcome to the detriment of users. Put differently, high market power by A stifles entry and competition, thereby harming users as in the baseline (even though the exact channel is different). As such, the trade-offs in a model with entry cost are expected to be similar as in the baseline. Finally, we notice that a user union might be particularly useful in a monopoly market, because it could bargain with the monopolist platform and extract larger fraction of total surplus.

5.2 User Welfare vs. Total Surplus Maximization

Our analysis so far has focused on maximizing user welfare instead of total surplus, broadly in line with the objectives of data-related antitrust and regulatory proposals which often aim for better consumer protection and reducing excessive market power. An illustrating example of why total surplus maximization generally does not benefit users is given in the previous section studying the monopolistic platform case. Then, user welfare becomes κ and the platform can extract all surplus (from data collection) but κ. While the resulting investment and data collection policy is efficient in a sense that it maximizes total surplus, users are generally strictly worse off in the monopoly case than with platform competition featuring lower total surplus. More generally, high market power by one platform may be optimal whilst harming users. While we focus on user welfare, we note that, within our model, one could easily evaluate the effects of antitrust and regulation on different objective functions (e.g., a weighted average of platform payoff and user welfare); we leave this for future research.

5.3 Platform Commitment

Thus far, we have ruled out the possibility for the platforms to commit in period $t = 1$ their actions in period $t = 2$. The lack of commitment is common in practice: Facebook changed data policies over time (Beacon 2007, ToS update 2008, etc.) and settled with the FTC for violating privacy promises in 2011; Amazon engaged in "copycat" practices on two-sided platforms that harmed sellers (Kirpalani and Philippon, 2020). We now evaluate the effects of platform commitments.

Commitment to pricing. Suppose platform x decides on (p_1^x, q^x, I^x, p_2^x) all in period 1 (rather than (p_1^x, q^x, I^x) in period 1 and p_2^x in period 2) to maximize ex-ante payoff from (5) — that is, $\pi_1^x := N_1^x p_1^x - q^x N_1^x I^x \theta^x + N_2^x p_2^x - \frac{\lambda(I^x)^2}{2}$ —whilst taking the choice $(p_1^{-x}, q^{-x}, I^{-x}, p_2^{-x})$ of the other platform as given. Because N_1^x does not depend on p_2^x, one can rewrite x's objective as:

$$\max_{(p_1^x, q^x, I^x, p_2^x)} \pi_1^x = \max_{(p_1^x, q^x, I^x)} \left(N_1^x p_1^x - q^x N_1^x I^x \theta^x - \frac{\lambda(I^x)^2}{2} + \max_{p_2^x} [N_2^x p_2^x] \right),$$

which implies that the choice of p_2^x maximizes $\pi_2^x = N_2^x p_2^x$ just as without commitment to future price p_2^x. That is, whether platform x can commit to p_2^x at time $t = 1$ is irrelevant and does not affect equilibrium and equilibrium quantities, such as N_t^x, p_t^x, or I^x. This outcome is also intuitive: Since users do not incur switching costs, committing to a lower price p_2^x does not help gaining more users (and data) in period $t = 1$.

Commitment to product quantities. Suppose in $t = 1$, the platform can commit to a minimum future quantity of service, N_2^x.[48] The timing of the entire game is as follows: At the very start, the platforms choose to commit (to future quantity, 'C'), or not ('NC'). If 'C' is chosen, the platform decides (q^x, p_1^x, I^x, N_2^x) in period $t = 1$ and decides nothing in the second period. If 'NC' is chosen, the platform decides (q^x, p_1^x, I^x) in the first period and p_2^x in the second period.

The potential benefit of quantity commitment is that under such a commitment, the future pricing power is limited which may be welfare-improving. However, we show in Appendix F.3 that both platforms committing to quantity does not constitute an equilibrium, and we end up with only one of the platforms committing. So instead of effectively limiting the pricing power, price competition is reduced in $t = 2$ and both platforms benefit at the expense of user welfare.

Commitment to data sharing. Suppose in period $t = 1$, the platforms can commit to a data sharing plan, η^x. Patforms may consider such commitments to either coordinate between themselves to improve the quality of services in the future or to discourage the rival platform's investment decision, which benefits the focal platform in the long run. However, we show in Appendix F.3 that neither motive is sufficient to generate commitments in equilibrium. The intuition is that both platforms want to free-ride on the other platform's data sharing commitment.

Overall, we conclude that platform commitment as well as regulation or antitrust proposals advocating such commitment fail to address inefficiencies in data-driven competition.[49] One main driver behind the results is that users are not "sticky" (i.e., they can switch platforms frictionlessly) and cannot coordinate to make the platform commitment matter; similar results would arise if users were short-sighted or myopic. In practice, when interoperability is lacking and switching across platforms are costly, the type of platform commitments we explore could be useful, but user union works with or without user stickiness.

[48] The commitment is about the minimum value of the quantity, but not the definite value of quantity. Otherwise, after the commitment is made in the first period, the other platform has an incentive to aggressively raise price.

[49] Commitments can come in other forms too, especially given recent technological innovations. For example, the distributed ledger technology offers convenient tools for commitment due to the immutability and consensus of blockchains and the algorithmic execution of smart contracts. It has been demonstrated that blockchains, smart contracts, and crypto-tokens can, in principle, facilitate commitment to competition (Goldstein, Gupta, and Sverchkov, 2019), monetary policies of token supply (Cong et al., 2022), production quantity and service pricing (Lyandres, 2019), delegation of control (Sockin and Xiong, 2022), and privacy-preserving computation and payment (Cao et al., 2018; Garratt and Lee, 2021). As mentioned earlier, Streamr.network and other DataDAO platforms are also actively exploring in practice commitments brought forth by blockchains. How blockchain-enabled commitment substitute or complement user unions as a solution to antitrust, privacy protection, and data sharing issues constitutes interesting future research.

6 Conclusion

Firms' production function in the era of digital platforms and big data entails their customers' participation and data contribution, generating data feedback and network effects. We model platform competition with endogenous data collection and sharing, to provide a framework for understanding the key inefficiencies and evaluating data-related antitrust and regulatory policies. We show that data feedback—similar to and interacting with network effects—may concentrate market power while improving service quality. Because users are atomistic, they do not internalize the impact of their data contribution and sharing on (i) future service or product quality which affects all users, (ii) concentration of market power, and (iii) platforms' incentives to innovate and invest in data infrastructure. Neither do they respond to common commitments by the platforms as they can switch platforms easily. We show that data sharing proposals (e.g., open banking and data vendor) and user privacy protections (e.g., GDPR and CCPA) fail to address the resulting inefficiencies regarding data collection and sharing. Finally, we introduce and model user union as a potential solution for improving consumer protection and welfare: A representative governing body can coordinate users' contribution to the platforms and maximize user surplus.

To focus on the economic channels of how users contribute to firm or platform production, we have left out issues related to informational asymmetry, which could be important in, e.g., lending and banking. One may want to encourage upstream firms to share data with downstream firms (e.g., Fang and Kim, 2022). Firms may directly influence the cost of privacy through advertising (Liu et al., 2020). The type of data and structure of information would matter too (e.g., Ichihashi, 2021b), not to mention other potential policy objectives outside our model, such as financial inclusion. In addition, we recognize user union is not a perfect solution, neither is it one-size-fits-all. Successful implementations in the long run may require automated, algorithmic, and contracts that are being developed by computer scientists and proposed in practice (Scholz, 2017; Garside, Wilkinson, Blycha, and Staples, 2021; Casey and Niblett, 2021). There can also be a plurality of data trusts tailored to different types of data or data subjects' preferences. As such, our theory aims to provide an economic foundation and initial benchmark to build upon, not a foregone conclusion.

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2019). Too much data: Prices and inefficiencies in data markets. Technical report, National Bureau of Economic Research.

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

Agarwal, S., P. Ghosh, T. Ruan, and Y. Zhang (2020). Privacy versus convenience: Customer response to data breaches of their information. *Available at SSRN 3729730*.

Akerlof, R., R. Holden, and L. Rayo (2021). Network externalities and market dominance.

Armstrong, M. (2006). Competition in two-sided markets. *The RAND journal of economics 37*(3), 668–691.

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.

Athey, S., C. Catalini, and C. Tucker (2017). The digital privacy paradox: Small money, small costs, small talk. Technical report, National Bureau of Economic Research.

Babina, T., G. Buchak, and W. Gornall (2022). Customer data access and fintech entry: Early evidence from open banking. *Available at SSRN*.

Bajari, P., V. Chernozhukov, A. Hortaçsu, and J. Suzuki (2019). The impact of big data on firm performance: An empirical investigation. In *AEA Papers and Proceedings*, Volume 109, pp. 33–37.

Becker, G. S. (1991). A note on restaurant pricing and other examples of social influences on price. *Journal of political economy 99*(5), 1109–1116.

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

Bergemann, D., A. Bonatti, and A. Smolin (2018). The design and price of information. *American economic review 108*(1), 1–48.

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

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

Bouckaert, J. and H. Degryse (2013). Default options and social welfare: Opt in versus opt out. *Journal of Institutional and Theoretical Economics: JITE*, 468–489.

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

Brynjolfsson, E. and K. McElheran (2016). The rapid adoption of data-driven decision-making. *American Economic Review 106*(5), 133–39.

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.

Cao, S., L. W. Cong, and B. Yang (2018). Auditing and blockchains: Pricing, misstatements, and regulation. *Misstatements, and Regulation (Oct 9, 2018)*.

Carrascal, J. P., C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira (2013). Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 189–200.

Casey, A. J. and A. Niblett (2021). The present and near future of self-driving contracts. *Available at SSRN*.

Chiou, L. and C. Tucker (2017). Search engines and data retention: Implications for privacy and antitrust. Technical report, National Bureau of Economic Research.

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.

Commission, E. et al. (2017). Communication" building a european data economy. *COM (2017) 9.*

Cong, L. W., Y. Li, and N. Wang (2022). Token-based platform finance. *Journal of Financial Economics 144*(3), 972–991.

Cong, L. W., K. Tang, D. Xie, and Q. Miao (2021). Asymmetric cross-side network effects on financial platforms: Theory and evidence from marketplace lending. *Available at SSRN 3461893*.

Cong, L. W., W. Wei, D. Xie, and L. Zhang (2022). Endogenous growth under multiple uses of data. *Journal of Economic Dynamics and Control*, 104395.

Cong, L. W., D. Xie, and L. Zhang (2021). Knowledge accumulation, privacy, and growth in a data economy. *Management Science 67*(10), 6480–6492.

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.

Dewald, W. G., J. G. Thursby, and R. G. Anderson (1986). Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review*, 587–603.

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

Easley, D., S. Huang, L. Yang, and Z. Zhong (2018). The economics of data.

Easley, R. F., H. Guo, and J. Krämer (2017). From network neutrality to data neutrality: A techno-economic framework and research agenda. *Information Systems Research, Forthcoming*.

Economist, T. (2018). Data workers of the world, unite. *July 7*.

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

Elliott, M. and A. Galeotti (2019). The role of networks in antitrust investigations. *Oxford Review of Economic Policy 35*(4), 614–637.

Fainmesser, I. P. and A. Galeotti (2016). Pricing network effects. *The Review of Economic Studies 83*(1), 165–198.

Fainmesser, I. P. and A. Galeotti (2020). Pricing network effects: Competition. *American Economic Journal: Microeconomics 12*(3), 1–32.

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

Fang, H. and S. J. Kim (2022). Data neutrality and market competition. *Working Paper*.

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.

Farrell, J. and P. Klemperer (2007). Coordination and lock-in: Competition with switching costs and network effects. *Handbook of industrial organization 3*, 1967–2072.

Fisher, A. and T. Streinz (2021). Confronting data inequality. *World Development Report*, 21–22.

Gal-Or, E. (1985). Information sharing in oligopoly. *Econometrica: Journal of the Econometric Society*, 329–343.

Garratt, R. and M. Lee (2021). Monetizing privacy with central bank digital currencies. Technical report.

Garside, A., S. Wilkinson, N. Blycha, and M. Staples (2021). Digital infrastructure integrity protocol for smart and legal contracts diip 2021. *Available at SSRN 3814811*.

Glenn, B. C. and M. Ellis Lee (2012). Drug development: raise standards for preclinical cancer research. *Nature 483*(7391), 531–33.

Goldfarb, A. and C. Tucker (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science 30*(3), 389–404.

Goldstein, I., D. Gupta, and R. Sverchkov (2019). Utility tokens as a commitment to competition. *Available at SSRN 3484627*.

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.

Halaburda, H. and Y. Yehezkel (2019). Focality advantage in platform competition. *Journal of Economics & Management Strategy 28*(1), 49–59.

Hauk, E. and S. Hurkens (2001). Secret information acquisition in cournot markets. *Economic Theory 18*(3), 661–681.

He, Z., J. Huang, and J. Zhou (2022). Open banking: credit market competition when borrowers own the data. Technical report, National Bureau of Economic Research.

Holdren, J. and E. Lander (2014). Big data and privacy: a technological perspective. *President's Council of Advisors on Science and Technology*.

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

Ichihashi, S. (2021b). The economics of data externalities. *Journal of Economic Theory 196*, 105316.

Ichihashi, S. (2022). Natural monopoly for data intermediaries. Technical report, Bank of Canada.

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

Jain, P., M. Gyanchandani, and N. Khare (2016). Big data privacy: a technological perspective and review. *Journal of Big Data 3*(1), 1–25.

Jetzek, T., M. Avital, and N. Bjorn-Andersen (2012). The value of open government data: A strategic analysis framework. In *SIG eGovernment pre-ICIS Workshop, Orlando*.

Jin, G. Z. et al. (2018). Artificial intelligence and consumer privacy. *The Economics of Artificial Intelligence: An Agenda; Agrawal, A., Gans, J., Goldfarb, A., Eds*, 439–462.

Jin, Y. and S. Vasserman (2021). Buying data from consumers: The impact of monitoring programs in us auto insurance. Technical report, National Bureau of Economic Research.

Johnson, J. and D. D. Sokol (2020). Understanding ai collusion and compliance. *Cambridge Handbook of Compliance,(D. Daniel Sokol & Benjamin van Rooij, editors),(Forthcoming)*.

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

Katz, M. L. and C. Shapiro (1985). Network externalities, competition, and compatibility. *The American economic review 75*(3), 424–440.

Kirpalani, R. and T. Philippon (2020). Data sharing and market power with two-sided platforms. Technical report, National Bureau of Economic Research.

Klemperer, P. (1987). Markets with consumer switching costs. *The quarterly journal of economics 102*(2), 375–394.

Klemperer, P. (1995). Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. *The review of economic studies 62*(4), 515–539.

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

Lyandres, E. (2019). Product market competition with crypto tokens and smart contracts. *Available at SSRN*.

Martens, B., A. De Streel, I. Graef, T. Tombal, and N. Duch-Brown (2020). Business-to-business data sharing: An economic and legal analysis. *EU Science Hub*.

Merges, R. P. (2008). Ip rights and technological platforms.

Milles, S. (2019). The future of data is political. *Towards Data Science*.

Montes, R., W. Sand-Zantman, and T. Valletti (2019). The value of personal information in online markets with endogenous privacy. *Management Science 65*(3), 1342–1362.

Newman, J. M. (2019). Antitrust in digital markets. *Vand. L. Rev. 72*, 1497.

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

Pentland, A., A. Lipton, and T. Hardjono (2021). *Building the New Economy: Data as Capital.*

Piwowar, H. A., R. S. Day, and D. B. Fridsma (2007). Sharing detailed research data is associated with increased citation rate. *PloS one 2*(3), e308.

Piwowar, H. A. and T. J. Vision (2013). Data reuse and the open data citation advantage. *PeerJ 1*, e175.

Posner, E. A. and E. G. Weyl (2018). Radical markets. In *Radical Markets*. Princeton University Press.

Posner, R. A. (1981). The economics of privacy. *The American economic review 71*(2), 405–409.

Posner, R. A. (2017). Antitrust in the new economy. In *Dominance and Monopolization*, pp. 493–512. Routledge.

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

Richter, H. and P. R. Slowinski (2019). The data sharing economy: on the emergence of new intermediaries. *IIC-International Review of Intellectual Property and Competition Law 50*(1), 4–29.

Rochet, J.-C. and J. Tirole (2003). Platform competition in two-sided markets. *Journal of the european economic association 1*(4), 990–1029.

Rochet, J.-C. and J. Tirole (2006). Two-sided markets: a progress report. *The RAND journal of economics 37*(3), 645–667.

Schaefer, M., G. Sapi, and S. Lorincz (2018). The effect of big data on recommendation quality: The example of internet search.

Scholz, L. H. (2017). Algorithmic contracts. *Stan. Tech. L. Rev. 20*, 128.

Sockin, M. and W. Xiong (2022). Decentralization through tokenization. Technical report, National Bureau of Economic Research.

Stigler, G. J. (1980). An introduction to privacy in economics and politics. *The Journal of Legal Studies 9*(4), 623–644.

Tang, H. (2019). The value of privacy: Evidence from online borrowers. *Available at SSRN 3880119*.

Taylor, C. R. (2004). Consumer privacy and the market for customer information. *RAND Journal of Economics*, 631–650.

Vaitilingam, R. (2020). How leading economists view antitrust in the digital economy. *LSE Business Review*.

Varian, H. R. (2010). Computer mediated transactions. *American Economic Review 100*(2), 1–10.

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

Vives, X. (1988). Aggregation of information in large cournot markets. *Econometrica: Journal of the Econometric Society*, 851–876.

Von Weizsäcker, C. C. (1984). The costs of substitution. *Econometrica: Journal of the Econometric Society*, 1085–1116.

Weyl, E. G. (2010). A price theory of multi-sided platforms. *American Economic Review 100*(4), 1642–72.

Zyskind, G., O. Nathan, and A. Pentland (2015). Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops*, pp. 180–184. IEEE.

Online Appendix

A General Solution

A.1 Derivation of (6)

The marginal user satisfies $u_t^A(\hat{z}_t) = u_t^B(\hat{z}_t)$. Using (3), the marginal user therefore satisfies

$$Y_t^A - p_t^A - \kappa^A(\hat{z}_t) + I^A\theta^A(q^A - c^A)\mathbb{I}_{\{t=1\}} = Y_t^B - p_t^B - \kappa^B(\hat{z}_t) + I^B\theta^B(q^B - c^B)\mathbb{I}_{\{t=1\}}.$$

Next, we use Y_t^x from (2), and $\kappa^x(z)$ as well as $N_t^A = \hat{z}_t$ and $N_t^B = 1 - \hat{z}_t$ from (1), to obtain

$$\begin{aligned} &K^A + \phi^A D^A \quad \mathbb{I}_{\{t=2\}} + \gamma^A \hat{z}_t - p_t^A + I^A\theta^A(q^A - c^A)\mathbb{I}_{\{t=1\}} - \hat{\kappa}\hat{z}_t \\ =&K^B + \phi^B D^B \quad \mathbb{I}_{\{t=2\}} + \gamma^B(1 - \hat{z}_t) - p_t^B + I^B\theta^B(q^B - c^B)\mathbb{I}_{\{t=1\}} - \hat{\kappa}(1 - \hat{z}_t). \end{aligned}$$

Now, we can solve above equation for

$$\hat{z}_t = \frac{1}{2} + \frac{\Delta_K - (p_t^A - p_t^B) + \left[\phi^A D^A - \phi^B D^B\right]\mathbb{I}_{\{t=2\}} + \left[I^A\theta^A(q^A - c^A) - I^B\theta^B(q^B - c^B)\right]\mathbb{I}_{\{t=1\}}}{2\kappa}$$

with

$$\kappa := \hat{\kappa} - \frac{\gamma^A + \gamma^B}{2} \quad \text{and} \quad \Delta_K := K^A - K^B + \frac{\gamma^A - \gamma^B}{2},$$

which was to show.

A.2 Solution in Period 2 — Proof of Lemma 1

Take the marginal user $\hat{z}_2$ from period $t = 2$ in (6), that is,

$$\hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B - (p_2^A - p_2^B)}{2\kappa}.$$

Thus, $N_2^A = \hat{z}_2$ and $N_2^B = 1 - \hat{z}_2$. Platform x chooses price p_2^x to maximize $N_2^x p_2^x$. Hence, the objective of platform A becomes

$$\pi_2^A = p_2^A\left(\frac{1}{2} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B - (p_2^A - p_2^B)}{2\kappa}\right).$$

The first-order condition with respect to price p_2^A reads

$$\frac{\partial \pi_2^A}{\partial p_2^A} = \frac{1}{2} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B - (p_2^A - p_2^B)}{2\kappa} - \frac{p_2^A}{2\kappa} = 0,$$

which we solve for

$$p_2^A = \frac{\Delta_K + \kappa + \phi^A D^A - \phi^B D^B + p_2^B}{2}. \tag{A.1}$$

Next, the objective of platform B becomes

$$\pi_2^B = p_2^B\left(\frac{1}{2} + \frac{\phi^B D^B - \phi^A D^A - \Delta_K - (p_2^B - p_2^A)}{2\kappa}\right).$$

The first-order condition with respect to price p_2^B reads

$$\frac{\partial \pi_2^B}{\partial p_2^B} = \frac{1}{2} + \frac{\phi^B D^B - \phi^A D^A - \Delta_K - (p_2^B - p_2^A)}{2\kappa} - \frac{p_2^B}{2\kappa} = 0,$$

which we solve for

$$p_2^B = \frac{\kappa - \Delta_K + \phi^B D^B - \phi^A D^A + p_2^A}{2}. \tag{A.2}$$

It is immediate to see that the second order conditions are satisfied, i.e., $\frac{\partial^2 \pi_2^x}{\partial (p_2^x)^2} < 0$.

Inserting (A.2) into (A.1), we solve

$$p_2^A = \kappa + \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{3} = \frac{3\kappa + \Delta_K + \phi^A D^A - \phi^B D^B}{3}. \tag{A.3}$$

We now plug (A.3) into (A.2) to calculate

$$p_2^B = \kappa + \frac{-\Delta_K + \phi^B D^B - \phi^A D^A}{3} = \frac{3\kappa - \Delta_K + \phi^B D^B - \phi^A D^A}{3}. \tag{A.4}$$

Thus, $p_2^A - p_2^B = \frac{2}{3}(\Delta_K + D^A\phi^A - D^B\phi^B)$ and therefore

$$\hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{6\kappa} = \frac{3\kappa + \Delta_K + \phi^A D^A - \phi^B D^B}{6\kappa}.$$

Thus, we obtain $p_2^A = 2\kappa \hat{z}_2 = 2\kappa N_1^A$ and $p_2^B = 2\kappa(1 - \hat{z}_2) = 2\kappa N_2^B$.

Next, calculate the profit of platform A, i.e.,

$$\pi_2^A = p_2^A \hat{z}_2 = \frac{(3\kappa + \phi^A D^A - \phi^B D^B + \Delta_K)^2}{18\kappa},$$

and the profit of platform B, i.e.,

$$\pi_2^B = p_2^B(1 - \hat{z}_2) = \frac{(3\kappa + \phi^B D^B - \phi^A D^A - \Delta_K)^2}{18\kappa}.$$

Or, we can write more compactly

$$\pi_2^x = \frac{(\Delta_K - 2\Delta_K \mathbb{I}_{\{x=B\}} + 3\kappa + \phi^x D^x - \phi^{-x} D^{-x})^2}{18\kappa}, \tag{A.5}$$

for $x, -x \in \{A, B\}$ (with $x = A$ implying $-x = B$ and vice versa).

A.3 Solution in Period 1 — Proof of Proposition 1

The first order conditions (8) and (9) follow from the directly differentiating π_1^x from (5) with respect to p_1^x and I^x respectively. The second order condition with respect to price reads

$$\frac{\partial^2 \pi_1^x}{\partial (p_1^x)^2} = 2\left(\frac{\partial N_1^x}{p_1^x}\right) + \frac{\partial}{\partial p_1^x} \sum_{x'=A,B} \left(\frac{\partial \pi_2^x}{\partial D^{x'}} \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial p_1^x}\right) < 0.$$

The second order condition with respect to investment can be written as

$$\frac{\partial^2 \pi_1^x}{\partial (I_1^x)^2} = \mathcal{K} - \lambda,$$

where $\mathcal{K}$ is a finite term with generally unknown sign. It follows that $\frac{\partial^2 \pi_1^x}{\partial (I_1^x)^2} = \mathcal{K} - \lambda < 0$ as long as λ is sufficiently large. Thus, the second order condtions are met and the first order conditions are sufficient as long as λ is sufficiently large.

We next prove that inducing $\theta^x = 1$ is optimal for platforms, and strictly so when $I^x > 0$. The case $\underline{\theta} = 1$ is trivial; hence, consider $\underline{\theta} < 1$. When $I^x = 0$, then the exact value of θ^x is not relevant, so one can induce without loss of generality $\theta^x = 1$. Suppose now to the contrary that in optimum, $\theta^x < 1$ and $I^x > 0$ hold. If $\theta^x = 0$, positive investment, $I^x > 0$, is clearly inefficient; thus, it suffices to consider $\theta^x \in (0,1)$. Then, every user shares $\chi^x := \theta^x I^x > 0$ units of data with platform x. For users to be willing to share $\theta^x \in (0,1)$ data with platform x, it must be that users are indifferent between sharing and not sharing data with platform x, which implies — by means of (4) — that $q^x = c^x$. Notice that the stipulation of $q^x = c^x$ can also induce $\theta^x = 1$; intuitively, platform x could raise q^x marginally to break the indifference and to induce $\theta^x = 1$.

As users are indifferent between sharing and not sharing data (i.e., $q^x = c^x$), it follows that I^x and χ^x do not directly affect $\hat{z}_1$ and adoption N_1^x (see (6)). Moreover, the platform pays every user $q^x \chi^x$ dollars for their total data contribution, which depends on I^x and θ^x only via the product $\chi^x = I^x \theta^x$. The platform x can now set $\theta^x = 1$ and reduce investment I^x to $\hat{I}^x < I^x$, whilst keeping $\chi^x = \theta^x I^x = \hat{I}^x$ and $\chi^x q^x$ as well as $\hat{D}^x = N_1^x \chi^x$ and π_2^x (which depends on I^x and θ^x only via χ^x and $\hat{D}^x$) unchanged. This reduces the cost of investment and increases platform payoff by discrete amount $\frac{\lambda((I^x)^2 - (\hat{I}^x)^2)}{2} > 0$, contradicting the optimality of $\theta^x < 1$. As such, $\theta^x = 1$ is optimal.

Next, we show that (conditional on $\theta^x = 1$) the choice of q^x is payoff relevant, in that N_t^x, π_t^x, I^x as well as user welfare do not depend on $q^{x'}$ for $x, x' \in \{A, B\}$. For this sake, we fix the choice of $q^{x'}$, and conjecture (and later verify) that $\frac{\partial p_1^x}{\partial q^x} = I^x$ as well as $\frac{\partial p_1^x}{\partial q^{-x}} = 0$. Also, conjecture (and later verify) that investment I^x does not depend on $q^{x'}$, i.e., $\frac{dI^x}{dq^{x'}} = 0$, for $x, x' \in \{A, B\}$. Under these conjectures, the expression (6) with $\theta^x = 1$ implies that

$$N_1^A = \hat{z}_t = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B) + \left[I^A(q^A - c^A) - I^B(q^B - c^B)\right]}{2\kappa}$$

is independent of $q^{x'}$ in a sense that $\frac{dN_1^A}{dq^A} = \frac{\partial N_1^A}{\partial p_1^A}\frac{\partial p_1^A}{\partial q^A} + \frac{\partial N_1^A}{\partial q^A} = 0$ and $\frac{dN_1^A}{dq^B} = \frac{\partial N_1^A}{\partial p_1^B}\frac{\partial p_1^B}{\partial q^B} + \frac{\partial N_1^A}{\partial q^B} = 0$. Analogously, $\frac{dN_1^B}{dq^{x'}} = 0$ for $x' \in \{A, B\}$. As a result and due to $\frac{dI^x}{dq^{x'}} = 0$, $\hat{D}^A = N_1^A I^A$ and $\hat{D}^B = N_1^B I^B$ do not depend on $q^{x'}$ for $x' \in \{A, B\}$, i.e., $\frac{d\hat{D}^x}{dq^{x'}} = 0$. Therefore, D^x, which is a function of $N_1^A I^A$ and $N_1^B I^B$, does not depend on $q^{x'}$, i.e., $\frac{dD^x}{dq^{x'}} = 0$. This, in turn, implies that period-2 payoff $\pi_2^x = N_2^x p_2^x$ does not depend on $q^{x'}$, in that $\frac{d\pi_2^x}{dq^{x'}} = 0$ as well as $\frac{dp_2^x}{dq^{x'}} = \frac{dN_2^x}{dq^{x'}} = 0$.

Then, we can differentiate the payoff in period $t = 1$ to obtain

$$\frac{d}{dq^x}\pi_1^x = \frac{d}{dq^x}\left(N_1^x p_1^x - q^x N_1^x I^x + N_2^x p_2^x - \frac{\lambda(I^x)^2}{2}\right) = N_1^x I^x - N_1^x I^x = 0$$

$$\frac{d}{dq^{-x}}\pi_1^x = \frac{d}{dq^{-x}}\left(N_1^x p_1^x - q^x N_1^x I^x + N_2^x p_2^x - \frac{\lambda(I^x)^2}{2}\right) = 0,$$

where it was used (in the first line) that $N_2^x p_2^x - \frac{\lambda(I^x)^2}{2}$ does not depend on q^x, that N_1^x does not depend on q^x, and that $\frac{\partial p_1^x}{\partial q^x} = I^x$. Thus, platform x's payoff does not depend on $q^{x'}$. As a result, we obtain

$$\frac{d}{dq^{x'}}\frac{\partial \pi_1^x}{\partial I^x} = \frac{\partial}{\partial I^x}\frac{d}{dq^{x'}}\pi_1^x = 0.$$

Since optimal investment I^x solves the first-order condition in (F.40), that is, $\frac{\partial \pi_1^x}{\partial I^x} = 0$, it readily follows that I^x does not depend on $q^{x'}$, which verifies our conjecture that optimal investment I^x does not depend on $q^{x'}$.

Lastly, we solve the first-order condition (regarding period-1 price) (8) for the price

$$p_1^x = I^x q^x - \left(\frac{\partial N_1^x}{\partial p_1^x}\right)^{-1} \left[\sum_{x'=A,B} \left(\frac{\partial \pi_2^x}{\partial D^{x'}} \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial p_1^x}\right) + N_1^x\right].$$

Our previous arguments imply that

$$\left(\frac{\partial N_1^x}{\partial p_1^x}\right)^{-1} \left[\sum_{x'=A,B} \left(\frac{\partial \pi_2^x}{\partial D^{x'}} \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial p_1^x}\right) + N_1^x\right]$$

does not depend on $q^{x'}$. As a consequence, it readily follows that $\frac{dp_1^x}{dq^x} = \frac{\partial p_1^x}{\partial q^x} = I^x$ and $\frac{dp_1^x}{dq^{-x}} = \frac{\partial p_1^x}{\partial q^{-x}} = 0$, which verifies our initial conjecture.

Taken together, we have shown that $\frac{\partial p_1^x}{\partial q^x} = I^x$, $\frac{\partial p_1^x}{\partial q^{-x}} = 0$ as well as $\frac{dI^x}{dq^{x'}} = 0$ (i.e., investment does not depend on $q^{x'}$) for $x, x' \in \{A, B\}$. That is, period-1 prices can be written in the form

$$p_1^x = \bar{p}_1^x + I^x q^x,$$

where $\bar{p}_1^x$ does not depend on $q^{x'}$, i.e., $\frac{\partial \bar{p_1}^x}{\partial q^{x'}} = 0$. It follows (from (6)) that N_1^x does not depend on $q^{x'}$. Because I^x does not depend on q^x or q^{-x}, we have that $\hat{D}^x$ does not depend on q^x or q^{-x}. Thus, the levels of $q^{x'}$ do not affect any period-2 equilibrium quantities, such as N_2^x, π_2^x, p_2^x, or u_2. Finally, it remains to show that q^x does not affect period-1 user welfare u_1. However, this is immediate from the facts that $\frac{\partial p_1^x}{\partial q^x} = I^x$, $\frac{\partial p_1^x}{\partial q^{-x}} = 0$, and the fact that no other equilibrium quantities depend on q^x.

Another corollary is that it is without loss of generality to set $q^x = c^x$ which, in turn, incentivizes $\theta^x = 1$ regardless of the value of $\underline{\theta}$. It follows that the value of $\underline{\theta}$ does not affect investment I^x, platform payoff π^x, market shares N_t^x, or user welfare.

B Proof and Derivations for Results of Section 2.2

We present proofs and derivations for the results presented in Section 2.2. The proofs for results which assume symmetric platforms are deferred to Appendix D where we present the model solution and solve for the (symmetric) equilibrium in the symmetric platform case in closed form.

B.1 Proof of Proposition 2

Follows from the more general result in Proposition 4 upon setting $\eta = 0$. The proof of Proposition 4 is presented in Appendix D.

B.2 Proof of Proposition 3

Recall that the solution and equilibrium in period $t = 2$ is characterized in Lemma 1. Notice that because there is neither data sharing nor a market for data, we have $D^x = \hat{D}^x = N_1^x I^x$, where we used that in optimum $\theta^x = 1$ (see Proposition 1 which applies in this context). Also, according to Proposition 1, we consider without loss of generality that $q^x = c^x$ for the following arguments.

As a result, realize that the expression for $\hat{z}_t$ from (6) implies for $t = 1$:

$$\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}. \tag{B.6}$$

Noting that $N_1^A = \hat{z}_1$ and $N_1^B = 1 - \hat{z}_1$, we calculate

$$\frac{\partial N_1^x}{\partial p_1^x} = -\frac{1}{2\kappa} \quad \text{and} \quad \frac{\partial N_1^x}{\partial p_1^{-x}} = \frac{1}{2\kappa}.$$

Next, we differentiate the platforms' period-2 payoff (under equilibrium pricing), characterized in (A.5) or Proposition 1, with respect to $D^{x'}$ for $x, x' = A, B$ to obtain

$$\frac{\partial \pi_2^A}{\partial D^A} = \frac{\phi^A \left(3\,\kappa + D^A\phi^A - D^B\phi^B + \Delta_K\right)}{9\kappa} \quad \text{and} \quad \frac{\partial \pi_2^B}{\partial D^B} = \frac{\phi^B \left(3\,\kappa - D^A\phi^A + D^B\phi^B - \Delta_K\right)}{9\kappa}, \tag{B.7}$$

as well as

$$\frac{\partial \pi_2^A}{\partial D^B} = -\frac{\phi^B \left(3\,\kappa + D^A\phi^A - D^B\phi^B + \Delta_K\right)}{9\kappa} \quad \text{and} \quad \frac{\partial \pi_2^B}{\partial D^A} = -\frac{\phi^A \left(3\,\kappa - D^A\phi^A + D^B\phi^B - \Delta_K\right)}{9\kappa}. \tag{B.8}$$

Next, note that because of $D^x = I^x N_1^x$, it follows that

$$\frac{\partial \pi_2^x}{\partial p_1^x} = -\frac{\partial \pi_2^x}{\partial D^x}\frac{I^x}{2\kappa} + \frac{\partial \pi_2^x}{\partial D^{-x}}\frac{I^{-x}}{2\kappa}. \tag{B.9}$$

As a result, the first-order conditions (8) for period-1 prices p_1^x become

$$\begin{aligned} \frac{\partial \pi_1^A}{\partial p_1^A} &= \left(\frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^A - I^A c^A}{2\kappa}\right) - \frac{1}{2\kappa}\left(\frac{\partial \pi_2^A}{\partial D^A} I^A - \frac{\partial \pi_2^A}{\partial D^B} I^B\right) = 0 \\ \frac{\partial \pi_1^B}{\partial p_1^B} &= \left(\frac{1}{2} - \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^B - I^B c^B}{2\kappa}\right) - \frac{1}{2\kappa}\left(\frac{\partial \pi_2^B}{\partial D^B} I^B - \frac{\partial \pi_2^B}{\partial D^A} I^B\right) = 0. \end{aligned} \tag{B.10}$$

Next, we can insert (B.7) and (B.8) as well as $D^x = N_1^x I^x$ into (B.10) (with N_1^x from (B.6)), and subsequently solve the two first-order conditions in (B.10) — which are two linear equations — for prices p_1^x. These price expressions then imply the expression for market share $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - p_1^A - p_1^B}{2\kappa}$ in $t = 1$. Using the expression for period-2 market share (under equilibrium pricing) from Lemma 1 and $D^A\phi^A = I^A\phi^A\hat{z}_1$ as well as $D^B\phi^B = I^B\phi^B(1 - \hat{z}_1)$, we obtain for $\hat{\phi}^x := \phi^x I^x$:

$$\hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + 2\hat{\phi}^A\hat{z}_1 - \hat{\phi}^B}{2\kappa}.$$

Inserting $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - p_1^A - p_1^B}{2\kappa}$ under period-1 equilibrium prices p_1^x, one obtains, after some algebra omitted here, the following expression for $\hat{z}_2$:

$$\hat{z}_2 = \frac{1}{2} + \frac{3\Delta_K(6\kappa + \hat{\phi}^A + \hat{\phi}^B) + 9\kappa(\hat{\phi}^A - \hat{\phi}^B) - 3c(I^A - I^B)(\hat{\phi}^A + \hat{\phi}^B)}{4\left(27\,\kappa^2 - (\hat{\phi}^A + \hat{\phi}^B)^2\right)}, \tag{B.11}$$

where $\hat{\phi}^x := I^x\phi^x$. Note that $\hat{z}_2$ (partially) increases with $\hat{\phi}^A - \hat{\phi}^B$; it also (partially) increases with $\hat{\phi}^A + \hat{\phi}^B$, when c is sufficiently low.

Next, we show that when c and Δ_K are sufficiently small, and $\phi^A > \phi^B$, then it holds that $I^A > I^B$, $p_1^A < p_1^B$ (when $q^x \leq c^x$), and $p_2^A > p_2^B$, also implying $\hat{\phi}^A > \hat{\phi}^B$. Consider $c = c^x = \Delta_K = 0$ and $q^x = c^x = 0$; the result then follows by continuity once we have establisehd it for $c = \Delta_K = 0$. We conjecture and verify that $N_1^A > N_1^B$ and $D^A\phi^A > D^B\phi^B \iff N_1^A I^A\phi^A > N_1^B I^B\phi^B$. Let us first prove that $p_1^A < p_1^B$. For this sake, we first show that

$$\sum_{x'=A,B}\left(\frac{\partial \pi_2^A}{\partial D^{x'}}\frac{\partial D^{x'}}{\partial N_1^{x'}}\frac{\partial N_1^{x'}}{\partial p_1^x}\right) < \sum_{x'=A,B}\left(\frac{\partial \pi_2^B}{\partial D^{x'}}\frac{\partial D^{x'}}{\partial N_1^{x'}}\frac{\partial N_1^{x'}}{\partial p_1^x}\right). \tag{B.12}$$

Because $D^x = N_1^x I^x$ and $\frac{\partial N_1^x}{\partial p_1^x} = -\frac{1}{2\kappa}$ as well as $\frac{\partial N_1^x}{\partial p_1^{-x}} = \frac{1}{2\kappa}$, we calculate

$$\sum_{x'=A,B}\left(\frac{\partial \pi_2^x}{\partial D^{x'}}\frac{\partial D^{x'}}{\partial N_1^{x'}}\frac{\partial N_1^{x'}}{\partial p_1^x}\right) = -\frac{1}{2\kappa}\left(\frac{\partial \pi_2^x}{\partial D^x}I^x - \frac{\partial \pi_2^x}{\partial D^{-x}}I^{-x}\right). \tag{B.13}$$

To prove inequality (B.12), we therefore need to show that

$$\frac{\partial \pi_2^A}{\partial D^A}I^A - \frac{\partial \pi_2^A}{\partial D^B}I^B > \frac{\partial \pi_2^B}{\partial D^B}I^B - \frac{\partial \pi_2^B}{\partial D^A}I^A.$$

With $\hat{\phi}^x = \phi^x I^x$, we obtain

$$\begin{aligned}
&\frac{\partial \pi_2^A}{\partial D^A}I^A - \frac{\partial \pi_2^A}{\partial D^B}I^B - \left[\frac{\partial \pi_2^B}{\partial D^B}I^B - \frac{\partial \pi_2^B}{\partial D^A}I^A\right]\\
&= \frac{\hat{\phi}^A\left(3\,\kappa + D^A\phi^A - D^B\phi^B\right)}{9\kappa} + \frac{\hat{\phi}^B\left(3\,\kappa + D^A\phi^A - D^B\phi^B\right)}{9\kappa}\\
&\quad - \left[\frac{\hat{\phi}^B\left(3\,\kappa - D^A\phi^A + D^B\phi^B\right)}{9\kappa} + \frac{\hat{\phi}^A\left(3\,\kappa - D^A\phi^A + D^B\phi^B\right)}{9\kappa}\right]\\
&= \frac{2(\hat{\phi}^A + \hat{\phi}^B)(D^A\phi^A - D^B\phi^B)}{9\kappa}.
\end{aligned}$$

As such, given the conjecture $D^A\phi^A > D^B\phi^B$, the inequality (B.12) holds. As prices p_1^x solve the first-order condition (8), inequality (B.12) implies $p_1^A < p_1^B$. It then follows immediately that $N_1^A > N_1^B$ (as $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - p_1^A - p_1^B}{2\kappa}$).

Next, we show that $I^A > I^B$. Notice that, due to $q^x = c^x$, we have $\frac{\partial N_1^x}{\partial I^x} = 0$ and $\frac{N_1^x}{\partial I^{-x}} = 0$. It becomes apparent from the first-order condition for investment (9) (with $q^x = c^x = 0$) that $I^A > I^B$ when

$$\sum_{x'=A,B}\left(\frac{\partial \pi_2^A}{\partial D^{x'}}\left[\frac{\partial D^{x'}}{\partial I^x} + \frac{\partial D^{x'}}{\partial N_1^{x'}}\frac{\partial N_1^{x'}}{\partial I^x}\right]\right) > \sum_{x'=A,B}\left(\frac{\partial \pi_2^B}{\partial D^{x'}}\left[\frac{\partial D^{x'}}{\partial I^x} + \frac{\partial D^{x'}}{\partial N_1^{x'}}\frac{\partial N_1^{x'}}{\partial I^x}\right]\right). \tag{B.14}$$

Because $D^x = N_1^x I^x$, and N_1^x does not (directly) depend on investment (i.e., $\frac{\partial N_1^{x'}}{\partial I^x} = 0$), we have $\frac{\partial D^x}{\partial I^x} = N_1^x$ as well as $\frac{\partial D_2^x}{\partial I^{-x}} = 0$. As such, inequality (B.14) can be rewritten as

$$\frac{\partial \pi_2^A}{\partial D^A}N_1^A > \frac{\partial \pi_2^B}{\partial D^B}N_1^B.$$

Using (B.7) and (B.8), we can calculate

$$\begin{aligned}
&\frac{\partial \pi_2^A}{\partial D^A}N_1^A - \frac{\partial \pi_2^B}{\partial D^B}N_1^B\\
&= \frac{\phi^A\left(3\,\kappa + D^A\phi^A - D^B\phi^B\right)}{9\kappa}N_1^A - \frac{\phi^B\left(3\,\kappa - D^A\phi^A + D^B\phi^B\right)}{9\kappa}N_1^B > 0
\end{aligned}$$

where the last inequality used that $N_1^A > N_1^B$ and $D^A\phi^A > D^B\phi^B$. As a result, inequality (B.14) holds, so that $I^A > I^B$. Because, in addition, $p_1^A < p_1^B$, it follows that $N_1^A > N_1^B$ and $D^A = N_1^A I^B > N_1^B I^B = D^B$. Thus, we have verified the conjecture that $N_1^A > N_1^B$ and $D^A\phi^A > D^B\phi^B \iff N_1^A I^A \phi^A > N_1^B I^B \phi^B$. According to Lemma 1, the fact that $\phi^A D^A > \phi^B D^B$ implies $N_2^A > N_2^B$ (i.e., $\hat{z}_2 > 1/2$) as well as $p_2^A > p_2^B$, which was to show.

B.3 Proof of Lemma 2

Note that $p_2^A = 2\kappa\hat{z}_2$ as well as $p_2^B = 2\kappa(1-\hat{z}_2)$. Thus, we can write user welfare in $t = 2$ as

$$u_2 = \hat{z}_2(Y_2^A - 2\kappa\hat{z}_2) + (1-\hat{z}_2)(Y_2^B - 2\kappa(1-\hat{z}_2)) - \bar{\kappa}_2.$$

Next, note that

$$\begin{aligned}
\hat{z}_2 Y_2^A + (1-\hat{z}_2)Y_2^B &= K^B + \phi^B D^B + \hat{z}_2(K^A - K^B + \phi^A D^A - \phi^B D^B + \gamma^A \hat{z}_2) + \gamma^B(1-\hat{z}_2)^2 \\
&= K^B + \phi^B D^B + \hat{z}_2\left(\Delta_K - \frac{\gamma^A - \gamma^B}{2} + \phi^A D^A - \phi^B D^B + \gamma^A \hat{z}_2\right) + \gamma^B(1-\hat{z}_2)^2 \\
&= K^B + \phi^B D^B + \hat{z}_2\left(6\kappa\hat{z}_2 - 3\kappa - \frac{\gamma^A - \gamma^B}{2} + \gamma^A\hat{z}_2\right) + \gamma^B(1-\hat{z}_2)^2,
\end{aligned}$$

where we used

$$\kappa := \hat{\kappa} - \frac{\gamma^A + \gamma^B}{2} \quad \text{and} \quad \Delta_K := K^A - K^B + \frac{\gamma^A - \gamma^B}{2}.$$

Thus,

$$\begin{aligned}
u_2 = K^B + \phi^B D^B + \hat{z}_2\left(6\kappa\hat{z}_2 - 3\kappa - \frac{\gamma^A - \gamma^B}{2} + \gamma^A\hat{z}_2\right) + \gamma^B(1-\hat{z}_2)^2 \\
- 2\kappa\hat{z}_2^2 - 2\kappa(1-\hat{z}_2)^2 - \frac{\hat{\kappa}(\hat{z}_2)^2 + \hat{\kappa}(1-\hat{z}_2)^2}{2} =: \tilde{u}_2 - 2\kappa\hat{z}_2^2 - 2\kappa(1-\hat{z}_2)^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\partial u_2}{\partial \hat{z}_2} &= 12\kappa\hat{z}_2 - 3\kappa - \frac{\gamma^A - \gamma^B}{2} + 2\gamma^A\hat{z}_2 - 2\gamma^B(1-\hat{z}_2) - \hat{\kappa}\hat{z}_2 + \hat{\kappa}(1-\hat{z}_2) - 4\kappa\hat{z}_2 + 4\kappa(1-\hat{z}_2) \\
&= 4\kappa\hat{z}_2 + \kappa - \frac{\gamma^A - \gamma^B}{2} + 2(\gamma^A + \gamma^B)\hat{z}_2 - 2\gamma^B - \hat{\kappa}(2\hat{z}_2 - 1) \\
&= 4\kappa\hat{z}_2 + 2\kappa + 2(\gamma^A + \gamma^B)\hat{z}_2 - \gamma^B - 2\hat{\kappa}\hat{z}_2 \\
&= 4\kappa\hat{z}_2 + 2\kappa + 2(\gamma^A + \gamma^B)\hat{z}_2 - \gamma^B - 2\left(\kappa + \frac{\gamma^A + \gamma^B}{2}\right)\hat{z}_2 \\
&= 2\kappa\hat{z}_2 + 2\kappa + (\gamma^A + \gamma^B)\hat{z}_2 - \gamma^B.
\end{aligned}$$

As such,

$$\frac{\partial u_2}{\partial \hat{z}_2} < 0 \iff \gamma^B > \frac{2\kappa(1+\hat{z}_2) + \gamma^A\hat{z}_2}{1-\hat{z}_2}.$$

That is, when $\hat{z}_2 \in (0,1)$ and γ^B is sufficiently large relative to γ^A (i.e., when $\gamma^B - \gamma^A$ is sufficiently large), then $\frac{\partial u_2}{\partial \hat{z}_2} < 0$ and user welfare in $t = 2$, i.e., u_2, decreases with platform A's market share $N_2^A = \hat{z}_2$, which concludes the argument.

C Proofs and Derivations for Results of Section 3

We present proofs and derivations for the results presented in Section 3. The proofs for results which assume symmetric platforms are deferred to Appendix D where we present the model solution and solve for the symmetric equilibrium in the symmetric platform case.

C.1 Proof of Proposition 4

The proof of Proposition 4 is presented in Appendix D where we present the solution for the symmetric platform case.

C.2 Proof of Proposition 5

We already use that $\theta^x = 1$ and set without loss of generality $q^x = c^x$ (see Proposition 1 which applies in this context). As $\eta = 1$, we have $D = D_2^x = N_1^A I^A + N_1^B I^B$, so $\frac{\partial D}{\partial I^x} \geq 0$. Using Lemma 1, we obtain for period-2 platform payoffs (under equilibrium pricing):

$$\pi_2^A = \frac{\left(3\,\kappa + \Delta_K + D(\phi^A - \phi^B)\right)^2}{18\,\kappa} \quad \text{and} \quad \pi_2^B = \frac{\left(3\,\kappa - \Delta_K - D(\phi^A - \phi^B)\right)^2}{18\,\kappa}. \tag{C.15}$$

One can calculate (with $\phi^A \geq \phi^B$)

$$\begin{aligned} \frac{\partial \pi_2^A}{\partial I^A} &= (\phi^A - \phi^B)\left(\frac{3\kappa + \Delta K + D(\phi^A - \phi^B)}{9\kappa}\right)\frac{\partial D}{\partial I^A} \geq 0 \\ \frac{\partial \pi_2^B}{\partial I^B} &= -(\phi^A - \phi^B)\left(\frac{3\kappa - \Delta K - D(\phi^A - \phi^B)}{9\kappa}\right)\frac{\partial D}{\partial I^B} \leq 0, \end{aligned}$$

where it was used that — by assumption/parameter condition (7) —- $3\kappa > \Delta_K + \phi^A - \phi^B$ and $D \leq 1$. When $\phi^A > \phi^B$ and $N_1^A > 0$ ($N_1^B > 0$), then $\frac{\partial \pi_2^A}{\partial I^A} > 0 > \frac{\partial \pi_2^B}{\partial I^B}$. As such, the first-order condition with respect to investment (9) readily implies that there exists no interior optimal $I^B \in (0,1)$. In fact, $\frac{\partial \pi_1^B}{\partial I^B} \leq 0$, and $I^B = 0$ is optimal for any $c > 0$ and $\lambda > 0$ (as well as in the limit $c \to 0$ and $\lambda \to 0$).

To derive the expression for platform A's period-2 market share $N_2^A = \hat{z}_2$, first notice that $\frac{\partial N_1^x}{\partial p_1^x} = \frac{-1}{2\kappa}$ and $\frac{\partial N_1^{-x}}{\partial p_1^x} = \frac{1}{2\kappa}$. Then, calculate $\frac{\partial D^x}{\partial p_1^x} = -\frac{I^x - I^{-x}}{2\kappa} = \frac{I^{-x} - I^x}{2\kappa}$ and $\frac{\partial D^{-x}}{\partial p_1^x} = \frac{I^{-x} - I^x}{2\kappa}$, and, due to $D^x = D$, we have $\frac{\partial D}{\partial p_1^x} = \frac{I^{-x} - I^x}{2\kappa}$. Thus,

$$\frac{\partial \pi_2^x}{\partial p_1^x} = \frac{\partial \pi_2^x}{\partial D^x}\left(\frac{I^{-x} - I^x}{2\kappa}\right) = \frac{\partial \pi_2^x}{\partial D^{-x}}\left(\frac{I^{-x} - I^x}{2\kappa}\right) = \frac{\partial \pi_2^x}{\partial D}\left(\frac{I^{-x} - I^x}{2\kappa}\right),$$

where we used $D = D^x = N_1^A I^A + N_1^B I^B = N_1^A I^A$ (due to $I^B = 0$) and (C.15).

Then, the first order conditions with respect to price in period $t = 1$, i.e., $\frac{\partial \pi_1^x}{\partial p_1^x} = 0$ for $x = A, B$, become

$$\begin{aligned} \frac{\partial \pi_1^A}{\partial p_1^A} &= \left(\frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^A - I^A c^A}{2\kappa}\right) - \left(\frac{I^A}{2\kappa}\right)\frac{\partial \pi_2^A}{\partial D} = 0 \\ \frac{\partial \pi_1^B}{\partial p_1^B} &= \left(\frac{1}{2} - \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^B}{2\kappa}\right) + \left(\frac{I^A}{2\kappa}\right)\frac{\partial \pi_2^B}{\partial D} = 0, \end{aligned} \tag{C.16}$$

where we used that $I^B = 0$. Using and differentiating the expression for period-2 payoff in (C.15), we can calculate:

$$\frac{\partial \pi_1^A}{\partial p_1^A} = \left(\frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^A - I^A c^A}{2\kappa}\right) - \left(\frac{I^A(\phi^A - \phi^B)}{2\kappa}\right)\left(\frac{3\,\kappa + \Delta_K + D(\phi^A - \phi^B)}{9\,\kappa}\right)$$
$$\frac{\partial \pi_1^B}{\partial p_1^B} = \left(\frac{1}{2} - \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^B}{2\kappa}\right) - \left(\frac{I^A(\phi^A - \phi^B)}{2\kappa}\right)\left(\frac{3\,\kappa - \Delta_K - D(\phi^A - \phi^B)}{9\,\kappa}\right).$$

Also, recall that $c^A = c^B = (1+\eta)c = 2c$. After some algebra omitted, one obtains equilibrium prices p_1^A and p_1^B by solving (C.16) — which is a system of two linear equations — for p_1^A and p_1^B.

Next, notice that $D = D^x = N_1^A I^A$ and $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K + p_1^A - p_1^B}{2\kappa}$. It follows from Lemma 1 that (under equilibrium pricing) $\hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + N_1^A I^A(\phi^A - \phi^B)}{6\kappa}$. Using period-1 equilibrium prices p_1^x, one can then calculate (after some algebra omitted here) the period-2 market share of A:

$$\hat{z}_2 = \frac{1}{2} + \frac{3\Delta_K(6\kappa + I^A\phi^A - I^A\phi^B) + 9\kappa I^A(\phi^A - \phi^B) - 6c(I^A)^2(\phi^A - \phi^B)}{4\,(27\,\kappa^2 - (I^A\phi^A - I^A\phi^B)^2)}, \tag{C.17}$$

which simplifies to $\hat{z}_2 = 1/2$ when $\phi^A = \phi^B$ and $\Delta_K = 0$.

Next, consider that $\phi^A > \phi^B$, and that $\lambda > 0$ and $c > 0$ are sufficiently small. Then, it is clear that, under the baseline without data sharing, we have that $I^A = I^B = 1$. We can use the expression for period-2 presented in the proof of Proposition 3 — that is, (B.11) — to calculate the period-2 market share of platform A (which we denote by $\hat{z}_2'$) by inserting $I^x = 1$:

$$\hat{z}_2' := \frac{1}{2} + \frac{3\Delta_K(6\kappa + \phi^A + \phi^B) + 9\kappa(\phi^A - \phi^B)}{4\,(27\,\kappa^2 - (\phi^A + \phi^B)^2)}.$$

Under data-sharing ($\eta = 1$), we have $I^A = 1 > 0 = I^B$, when c and λ are sufficiently small. We insert $I^A = 1 > I^B = 0$ into (C.17) to get:

$$\hat{z}_2 = \frac{1}{2} + \frac{3\Delta_K(6\kappa + \phi^A - \phi^B) + 9\kappa(\phi^A - \phi^B) - 6c(\phi^A - \phi^B)}{4\,(27\,\kappa^2 - (\phi^A - \phi^B)^2)}$$

Note that (7) implies $27\,\kappa^2 > (\phi^A + \phi^B)^2$. It is then evident that $\hat{z}_2 < \hat{z}_2'$ when $\phi^A > \phi^B$. As such, by continuity, when λ and c are sufficiently small, then data sharing ($\eta = 1$) reduces platform A's period-2 market share relative to the baseline with $\eta = 0$, which was to show.

C.3 Proof of Proposition 6

The proof of Proposition 6 is presented in Appendix D.2 where we present the solution for the symmetric platform case.

C.4 Proof of Proposition 7

Recall that the timing within period-2 is as follows: First, platforms decide how much data to buy from users and, second, they set prices p_2^x, leading to data-dependent continuation payoff π_2^x characterized in Lemma 1. It is clear that platforms pay (per unit) price c for buying data from individual users, because c is the minimal price at which users are willing to sell data to platforms; offering a higher price would only hurt platform payoffs.

To begin with, take the total stock of data D as given and consider platform x's decision how much data to buy (at per unit price c) at the beginning of $t = 2$ before prices p_2^x. Given $D^{x'}$ for $x' = A, B$, platforms' period-2 (continuation) payoff equals π_2^x from Lemma 1. Platform x maximizes

$$\max_{D^x \in [0,D]} \left(\pi_2^x - cD^x\right),$$

taking the choice of the other platform D^{-x} as given. Because the period-2 platform payoff under equilibrium pricing from Lemma 1, is strictly convex in D^x, i.e., $\frac{\partial^2 \pi_2^x}{\partial (D^x)^2} > 0$, it follows that (in equilibrium/optimum) $D^x \in \{0, D\}$, i.e., platforms either buy no data or the full amount of data and there is no interior optimum.

Now, let us analyze jointly platforms' (equilibrium) choice of investment I^x and their data acquisition in $t = 2$, D^x. If $D^A = D^B = 0$ in period 2 in equilibrium, then clearly $I^A = I^B = 0$. Next, if $D^B = 0$ in equilibrium, then $I^B = 0$ is clearly optimal.

Now, consider $D^B = D$ as well as $D^A = D$ (otherwise, we could relabel platforms; i.e., relabel A to B and then we are back to the previous case of $I^B = 0$). Notice $D = N_1^A I^A + N_1^B I^B$, so $\frac{\partial D}{\partial I^x} \geq 0$. Using the period-2 platform payoff under equilibrium pricing from Lemma 1 for $D^A = D^B = D$, we obtain

$$\pi_2^A = \frac{\left(3\,\kappa + \Delta_K + D(\phi^A - \phi^B)\right)^2}{18\,\kappa} \quad \text{and} \quad \pi_2^B = \frac{\left(3\,\kappa - \Delta_K - D(\phi^A - \phi^B)\right)^2}{18\,\kappa}. \tag{C.18}$$

One can calculate (with $\phi^A \geq \phi^B$)

$$\begin{aligned} \frac{\partial \pi_2^A}{\partial I^A} &= (\phi^A - \phi^B)\left(\frac{3\kappa + \Delta\kappa + D(\phi^A - \phi^B)}{9\kappa}\right)\frac{\partial D}{\partial I^A} \geq 0 \\ \frac{\partial \pi_2^B}{\partial I^B} &= -(\phi^A - \phi^B)\left(\frac{3\kappa - \Delta\kappa - D(\phi^A - \phi^B)}{9\kappa}\right)\frac{\partial D}{\partial I^B} \leq 0, \end{aligned} \tag{C.19}$$

where it was used that — by assumption (7) —- $3\kappa > \Delta_K + \phi^A - \phi^B$ and $D \leq 1$. Then, the fact that optimal investment — if interior — must solve the first-order condition

$$\frac{\partial \pi_1^x}{\partial I^x} = -\lambda I^x + \frac{\partial \pi_2^x}{\partial I^x} = 0$$

with π_1^x from (10) and $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}$ readily implies that there exists no interior optimal $I^B \in (0, 1)$.[50] In fact, $\frac{\partial \pi_1^B}{\partial I^B} \leq 0$, and $I^B = 0$ is optimal for any $c > 0$ and $\lambda > 0$ (as well as in the limit $c \to 0$ and $\lambda \to 0$). In other words, only one platform — say platform A — undertakes investment, in that $I^A \geq 0 = I^B$.

Take $\phi^A > \phi^B$. Consider that c and λ are sufficiently low (e.g., the limit case $c \to 0$ and $\lambda \to 0$), so that $D^x = D$ is optimal and $I^A = 1$ as well as $I^B = 0$, due to $\frac{\partial \pi_1^x}{\partial I^x} = -\lambda I^x + \frac{\partial \pi_2^x}{\partial I^x}$ and (C.19). Then, $D = D^A = D^B = N_1^A I^A + N_1^B I^B = N_1^A I^A$ (as $I^B = 0$), and the model solution becomes similar to data sharing with $\eta = 1$ from Proposition 5, with different privacy cost $c^x = c$ (instead of $c^x = 2c$). The reason is that under data sharing with $\eta = 1$, platform A incurs effectively privacy cost $2c$, as it must also compensate users for their data being used on B in addition to being used on A. With the market for data, platform A compensates users only for their data being used on A, while B compensates them only for their data being used on B.

Consider that $\phi^A > \phi^B$, and that $\lambda > 0$ and $c > 0$ are sufficiently small. In the baseline, we have $I^A = I^B = 1$. We can use the expression for period-2 presented in the proof of Proposition 3 — that is, (B.11) — to calculate then the period-2 market share of platform A when $I^A = I^B = 1$ (which we denote by $\hat{z}_2'$) by inserting $I^x = 1$:

$$\hat{z}_2' = \frac{1}{2} + \frac{3\Delta_K(6\kappa + \phi^A + \phi^B) + 9\kappa(\phi^A - \phi^B)}{4\,(27\,\kappa^2 - (\phi^A + \phi^B)^2)}.$$

With the market for data, we have $I^A = 1 > 0 = I^B$ as well as $D^x = D = N_1^A I^A$. Next, notice that platform x anticipates, when choosing I^x and p_1^x, that it buys any unit data that is generated

[50] Recall, $N_1^x = \hat{z}_1$ if $x = A$ and $N_1^x = 1 - \hat{z}_1$ if $x = B$.

at per unit price c at time $t=2$. That is, with $D^x = D = N_1^A$, first-period payoff from (10) can be written as

$$\pi_1^A = N_1^A p_1^A + \frac{\left(3\,\kappa + \Delta_K + N_1^A(\phi^A - \phi^B)\right)^2}{18\,\kappa} - cN_1^A$$
$$\pi_1^B = N_1^B p_1^B + \frac{\left(3\,\kappa - \Delta_K - N_1^A(\phi^A - \phi^B)\right)^2}{18\,\kappa} - cN_1^A,$$

with $N_1^A = \frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}$ and $N_1^B = \frac{1}{2} - \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}$, where we used (C.18) for expressions for π_2^x.

Now, notice that $\frac{\partial D}{\partial p_1^x} = \frac{I^{-x} - I^x}{2\kappa}$, as $\frac{\partial N_1^x}{\partial p_1^x} = -\frac{1}{2\kappa}$ and $\frac{\partial N_1^x}{\partial p_1^{-x}} = \frac{1}{2\kappa}$ the first order conditions with respect to price in period $t=1$ become (with $D = N_1^A$ and $I^A = 1 > I^B = 0$)

$$\frac{\partial \pi_1^A}{\partial p_1^A} = \left(\frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^A - c}{2\kappa}\right) - \frac{1}{2\kappa}\left(\frac{\partial \pi_2^A}{\partial D}\right) = 0 \tag{C.20}$$
$$\frac{\partial \pi_1^B}{\partial p_1^B} = \left(\frac{1}{2} - \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^B + c}{2\kappa}\right) + \frac{1}{2\kappa}\left(\frac{\partial \pi_2^B}{\partial D}\right) = 0,$$

which — using the expressions for π_2^x with $D^x = D$ — simplifies to

$$\frac{\partial \pi_1^A}{\partial p_1^A} = \left(\frac{1}{2} + \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^A - c}{2\kappa}\right) - \frac{\phi^A - \phi^B}{2\kappa}\left(\frac{3\,\kappa + \Delta_K + D(\phi^A - \phi^B)}{9\,\kappa}\right) = 0$$
$$\frac{\partial \pi_1^B}{\partial p_1^B} = \left(\frac{1}{2} - \frac{\Delta_K - (p_1^A - p_1^B)}{2\kappa}\right) - \left(\frac{p_1^B + c}{2\kappa}\right) - \frac{\phi^A - \phi^B}{2\kappa}\left(\frac{3\,\kappa - \Delta_K - D(\phi^A - \phi^B)}{9\,\kappa}\right) = 0.$$

To obtain equilibrium prices p_1^A and p_1^B, one then solves (C.20) — which is a system of two linear equations — for p_1^A and p_1^B. Solving for period-1 equilibrium prices (after some algebra omitted here), using $I^A = 1$, $\hat{z}_1 = \frac{1}{2} + \frac{\Delta_K + p_1^B - p_1^A}{2\kappa}$, and $D^x = D = N_1^A I^A = I^A \hat{z}_1$ as well as the expression for $\hat{z}_2$ from Lemma 1, one can calculate (after some algebra omitted here) the market share of A in $t=2$ (in the limit when $c \to 0$):[51]

$$\hat{z}_2 = \frac{1}{2} + \frac{3\Delta_K(6\kappa + \phi^A + \phi^B) + 9\kappa(\phi^A - \phi^B)}{4\,(27\,\kappa^2 - (\phi^A - \phi^B)^2)}.$$

It is evident that $\hat{z}_2 < \hat{z}_2'$, i.e., the market for data (in which users sell their data) reduces market concentration relative to the baseline.

C.5 Proofs of Lemma 3 and Details for Section 3.4

We now prove the results of Lemma 3, and provide details for the solution of the model variant in which there is a market for data and the platforms own data.

C.5.1 Proof of Lemma 3

We solve for optimal data sharing at the onset of period $t=2$ to maximize total continuation surplus, i.e., $\pi_2^A + \pi_2^B$, whereby π_2^x is characterized in (A.5). As a preparation, we differentiate

[51] The general expression would be $\hat{z}_2 = \frac{1}{2} + \frac{3\Delta_K(6\kappa+\phi^A+\phi^B)+9\kappa(\phi^A-\phi^B)}{4\left(27\,\kappa^2-(\phi^A-\phi^B)^2\right)} + o(c)$, where the remainder term $o(c)$ tends to zero as $c \to 0$.

(A.5) with respect to $D^{x'}$ for $x, x' = A, B$ to obtain

$$\frac{\partial \pi_2^A}{\partial D^A} = \frac{\phi^A\left(3\,\kappa + D^A\phi^A - D^B\phi^B + \Delta_K\right)}{9\kappa} \quad \text{and} \quad \frac{\partial \pi_2^B}{\partial D^B} = \frac{\phi^B\left(3\,\kappa - D^A\phi^A + D^B\phi^B - \Delta_K\right)}{9\kappa},$$

as well as

$$\frac{\partial \pi_2^A}{\partial D^B} = -\frac{\phi^B\left(3\,\kappa + D^A\phi^A - D^B\phi^B + \Delta_K\right)}{9\kappa} \quad \text{and} \quad \frac{\partial \pi_2^B}{\partial D^A} = -\frac{\phi^A\left(3\,\kappa - D^A\phi^A + D^B\phi^B - \Delta_K\right)}{9\kappa}.$$

Above expressions are identical to (B.7) and (B.8).

Next, notice that $D^x \geq \hat{D}^x$, whereby $\hat{D}^x$ is platform x's stock of data at the beginning of period $t = 2$ *before* data sharing. Using (B.7) and (B.8), we obtain

$$\left(\frac{\partial(\pi_2^A + \pi_2^B)}{\partial D^A}\right) = \phi^A\left(\frac{\left(3\,\kappa + D^A\phi^A - D^B\phi^B + \Delta_K\right) - \left(3\,\kappa - D^A\phi^A + D^B\phi^B - \Delta_K\right)}{9\kappa}\right)$$
$$= \frac{2\phi^A(D^A\phi^A - D^B\phi^B + \Delta_K)}{9\kappa}.$$

Consider $\hat{D}^A\phi^A + \Delta_K \geq \hat{D}^B\phi^B$. Thus, for any $D^A \geq \hat{D}^A$ and $D^B = \hat{D}^B$, we have $\left(\frac{\partial(\pi_2^A+\pi_2^B)}{\partial D^A}\right) \geq 0$, where the inequality is strict if $D^A > \hat{D}^A$.

Symmetrically, we can combine (B.7) and (B.8) to calculate

$$\left(\frac{\partial(\pi_2^A + \pi_2^B)}{\partial D^B}\right) = \frac{2\phi^B(D^B\phi^B - D^A\phi^A - \Delta_K)}{9\kappa}.$$

Thus, for $D^A \geq \hat{D}^A$ and $D^B = \hat{D}^B$, we have $\left(\frac{\partial(\pi_2^A+\pi_2^B)}{\partial D^B}\right) \leq 0$, where the inequality is strict if $D^A > \hat{D}^A$.

It follows that, given $\hat{D}^A\phi^A + \Delta_K \geq \hat{D}^B\phi^B$ and for $\phi^A \geq \phi^B$, total surplus is maximized upon implementing $D^A = \hat{D}^A + \hat{D}^B$ and $D^B = \hat{D}^B$, which was to show.

C.5.2 Solution of Model Variant with Market for Data and Details for Section 3.4

At the beginning of time period $t = 2$, before choosing prices p_2^x, platforms A and B trade data with each other, with endogenous price p_2^D. We assume throughout that (in equilibrium) $\hat{D}^A\phi^A + \Delta_K \geq \hat{D}^B\phi^B$ as well as $\phi^A \geq \phi^B$. Without loss of generality, it suffices to consider that platforms A and B choose the optimal allocation of data through Nash Bargaining at the beginning of period $t = 2$, with equal bargaining weights. The price for data p_2^D is then chosen to implement the split of resulting surplus. According to Lemma 3, total surplus is maximized when one platform x (labeled A) shares data with the other platform $-x$ (labeled B), but not the other way around.

We now derive the Nash bargaining solution. Now, B is the platform that sells data, and A is the platform that buys data. Then, data trade at the beginning of period $t = 2$ implies $D^A = \hat{D}^A + \hat{D}^B$ while $D^B = \hat{D}^B$, whereby $\hat{D}^x = N_1^x I^x \theta^x$ and $\theta^x = 1$ in optimum. As a result, using the expressions from Lemma 1 for period-2 platform payoffs π_2^x under equilibrium pricing, platforms derive the following payoff (just after Nash bargaining and trade and under period-2 equilibrium prices):

$$\pi_2^A = \frac{(3\kappa + \Delta_K + \phi^A\hat{D}^A)^2}{18\kappa} \quad \text{and} \quad \pi_2^B = \frac{(3\kappa - \Delta_K - \phi^A\hat{D}^A)^2}{18\kappa}.$$

On the other hand, absent data trade, we would have $D^x = \hat{D}^x$ and platform (equilibrium) payoffs in period $t = 2$ would be

$$\hat{\pi}_2^A = \frac{(3\kappa + \Delta_K + \phi^A \hat{D}^A - \phi^B \hat{D}^B)^2}{18\kappa} \quad \text{and} \quad \hat{\pi}_2^B = \frac{(3\kappa - \Delta_K + \phi^B \hat{D}^B - \phi^A \hat{D}^A)^2}{18\kappa}.$$

As such, the total surplus created for the platforms from data trade equals

$$S := \pi_2^A + \pi_2^B - (\hat{\pi}_2^A + \hat{\pi}_2^B).$$

We denote by $\tilde{\pi}_2^x$ platform x's payoff at the beginning of period $t = 2$ just before Nash bargaining takes place. At this moment, platform x owns a stock of data $\hat{D}^x = N_1^x I^x$. Just after the data trade, we have $D^A = \hat{D}^A + \hat{D}^B$ and $D^B = \hat{D}^B$ and platform's continuation payoff becomes π_2^x.

It is well-known that, as per Nash bargaining protocol with equal bargaining weights 1/2, platform x's payoff (just before data trade and Nash bargaining) then reads

$$\tilde{\pi}_2^x = \hat{\pi}_2^x + \frac{1}{2} S = \frac{\hat{\pi}_2^x - \hat{\pi}_2^{-x} + \pi_2^A + \pi_2^B}{2}. \tag{C.21}$$

That is, platforms' payoff in $t = 2$ before data trade $\tilde{\pi}_2^x$ is the sum of the "reservation value" $\hat{\pi}_2^x$, which would obtain absent data trade, and half of the surplus generated from data trade S (because the bargaining weights of platforms are 1/2 each). In the context of Nash bargaining and the implementation of the optimal data allocation and payoffs, platform B receives a lump-sum transfer from A equal to $\tilde{\pi}_2^B - \pi_2^B$ as $\tilde{\pi}_2^B$ is the payoff just before data trade/Nash bargaining and π_2^B the payoff just after data trade/Nash bargaining. This transfer would imply a (per unit) price for $p_D = (\tilde{\pi}_2^B - \pi_2^B)/D^B$; this price for data does not play a role in what follows.

Formally, anticipating the continuation equilibrium in period $t = 2$ with Nash bargaining, platforms $x = A, B$ now maximize at time $t = 1$:

$$\max_{p_1^x, I^x, q^x} \left(N_1^x p_1^x - \frac{\lambda (I^x)^2}{2} - q^x \theta^x N_1^x I^x + \tilde{\pi}_2^x \right), \tag{C.22}$$

taking the choice of the other competing platform as given and subject to $N_1^A = \hat{z}_1$ and $N_1^B = 1 - \hat{z}_1$ with $\hat{z}_1$ characterized in (6). As in the baseline, setting $\theta^x = 1$ is optimal, and the choice of q^x is not payoff-relevant (so one can set $q^x = c^x$). Also note that $c^A = c < 2c = c^B$, because users anticipate that their data is used by both platforms when they contribute it to platform B.

Next, we present first order conditions with respect to period-1 prices and investment. Differentiating the objective in (C.22) with respect to p_1, we obtain (for $q^x = c^x$)

$$\frac{\partial \pi_1^x}{\partial p_1^x} = N_1^x + \frac{\partial N_1^x}{\partial p_1^x}(p_1^x - I^x c^x) + \sum_{x' = A, B} \left(\frac{\partial \tilde{\pi}_2^x}{\partial D^{x'}} \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial p_1^x} \right) = 0.$$

The first order condition with respect to investment becomes (due to $\frac{\partial N_1^x}{\partial I^x} = 0$ when $q^x = c^x$):

$$\frac{\partial \pi_1^x}{\partial I^x} = \sum_{x' = A, B} \left(\frac{\partial \tilde{\pi}_2^x}{\partial D^{x'}} \left[\frac{\partial D^{x'}}{\partial I^x} + \frac{\partial D^{x'}}{\partial N_1^{x'}} \frac{\partial N_1^{x'}}{\partial I^x} \right] \right) - \lambda I^x - N_1^x c^x = 0,$$

which holds when investment I^x satisfies $I^x \in (0, 1)$. Provided its existence, we focus on a subgame perfect equilibrium in pure strategies in which $\hat{D}^A \phi^A + \Delta_K \geq \hat{D}^B \phi^B$.

By Lemma 1, platform A's market share in period $t = 2$ (under equilibrium pricing) becomes

$$\hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + D\phi^A - D^B \phi^B}{6\kappa}.$$

When $I^A > 0$, then $\hat{z}_2 > 1/2$ under all parameters and, in particular, even when $\Delta_K = 0$ and $\phi^A = \phi^B$. As a result, when $\phi^A - \phi^B$ and Δ_K are sufficiently small and $I^x > 0$, then market concentration $\hat{z}_2 > 1/2$ is necessarily higher than under the baseline (where market concentration is $1/2$ under platform symmetry).

D Solution for Symmetric platforms — Proofs of Propositions 2, 4, 6, 8, and 9

We now solve the model in the symmetric platform case with data sharing $\eta^x = \eta$, i.e., platform x must share fraction η of its data with its competitor $-x$ and vice versa. We solve for a symmetric equilibrium with $p_t^x = p_t^{-x}$, $I^x = I^{-x}$, $N_1^x = N_1^{-x} = 1/2$, $\theta^x = \theta^{-x}$, and $q^x = q^{-x}$. In this Section, we prove all Lemmata and Propositions which assume symmetric platforms, that is, Propositions 2, 4, 6, 8, and 9.

D.1 Proofs of Propositions 2 and 4

The arguments below therefore prove Proposition 4. The statements from Proposition 2 follow upon setting $\eta = 0$, in that Proposition 2 is a special case of Proposition 4. When platform x must share fraction $\eta \in [0, 1]$ of its data with the competitor platform $-x$, then we have $c^x = c(1 + \eta)$ where c is a constant. We solve for a symmetric equilibrium with $p_t^x = p_t^{-x}$, $I^x = I^{-x}$, $N_1^x = N_1^{-x} = 1/2$, $\theta^x = \theta^{-x}$, and $q^x = q^{-x}$.

D.1.1 Period $t = 2$

In the symmetric platform case, we have $\Delta_K = 0$, $\phi^A = \phi^B$ and $D^A = D^B$, so that — by Lemma 1 — $p_2^x = \kappa$ and $N_2^x = \frac{1}{2}$. As such, $\pi_2^x = \frac{\kappa}{2}$. And, average user utility is

$$u_2 = N_2^A(Y_2^A - p_2^A) + N_2^B(Y_2^B - p_2^B) - \bar{\kappa}_2 = K^x + \phi^x D^x + \frac{\gamma^x}{2} - \kappa - \frac{\hat{\kappa}}{4}.$$

As can be seen, in equilibrium, $\pi_2^x = \kappa/2$ does not depend on D^x and ϕ^x, whereas u_2 increases with D^x and ϕ^x.

In addition, we can use the expressions for π_2^x, i.e., the period-2 platform payoff under equilibrium pricing from Lemma 1, (or alternatively (B.7) and (B.8) for $\Delta_K = 0$ and $D^A = D^B$ as well as $\phi^A = \phi^B$) to derive

$$\frac{\partial \pi_2^x}{\partial D^x} = -\frac{\partial \pi_2^x}{\partial D^{-x}} = \frac{\phi^x}{3}. \tag{D.23}$$

D.1.2 Period $t = 1$

It follows that $\theta^A = \theta^B = 1$ (see Proposition 1). We note that $D^A = N_1^A I^A + \eta N_1^B I^B$ and $D^B = \eta N_1^A I^A + N_1^B I^B$ for $\eta \in [0, 1]$ as well as $c^x = (1 + \eta)c$. Also observe that

$$\pi_1^x = N_1^x \left[p_1^x - q^x I^x\right] + N_2^x p_2^x - \frac{\lambda (I^x)^2}{2}.$$

The two platforms $x = A, B$ solve

$$\max_{p_1^x, I^x, q^x} \pi_1^x,$$

taking the choice of the other platform $(p_1^{-x}, I^{-x}, q^{-x})$ as given.

Equilibrium prices in $t = 1$

We now calculate equilibrium prices in period $t = 1$, i.e., p_1^x for $x = A, B$, taking investments I^x as given. Recall $N_1^A = \hat{z}_1$ and $N_1^B = 1 - \hat{z}_1$ with $\hat{z}_1$ characterized in (6). That is,

$$N_1^A = \frac{1}{2} + \frac{-(p_1^A - p_1^B) + \left[I^A(q^A - c^A) - I^B(q^B - c^B)\right]}{2\kappa} \qquad \text{(D.24)}$$
$$N_1^B = \frac{1}{2} + \frac{(p_1^A - p_1^B) - \left[I^A(q^A - c^A) - I^B(q^B - c^B)\right]}{2\kappa}.$$

Next, we calculate

$$\frac{\partial N_1^x}{\partial p_1^x} = -\frac{1}{2\kappa} \quad \text{and} \quad \frac{\partial N_1^{-x}}{\partial p_1^x} = \frac{1}{2\kappa}$$

as well as

$$\frac{\partial D^{-x}}{\partial p_1^x} = \frac{I^{-x} - \eta I^x}{2\kappa} \quad \text{and} \quad \frac{\partial D^x}{\partial p_1^x} = \frac{\eta I^{-x} - I^x}{2\kappa}.$$

Thus, the first-order condition with respect to price p_1^x reads:

$$\frac{\partial \pi_1^x}{\partial p_1^x} = N_1^x - \frac{p_1^x}{2\kappa} + \frac{I^x q^x}{2\kappa} + \frac{\partial \pi_2^x}{\partial D^x}\left(\frac{\eta I^{-x} - I^x}{2\kappa}\right) + \frac{\partial \pi_2^x}{\partial D^{-x}}\left(\frac{I^{-x} - \eta I^x}{2\kappa}\right) = 0. \qquad \text{(D.25)}$$

Using $N_1^x = \frac{1}{2}$, $I^x = I^{-x}$, and $\frac{\partial \pi_2^x}{\partial D^x} = \frac{\phi^x}{3} = -\frac{\partial \pi_2^x}{\partial D^{-x}}$ (see (D.23)), we can solve

$$p_1^x = \kappa + I^x\left(q^x - \frac{2(1-\eta)\phi^x}{3}\right). \qquad \text{(D.26)}$$

The equilibrium price expression from Proposition 2 follows upon setting $\eta = 0$.

Equilibrium investments

We now calculate equilibrium investments $I^x = I^{-x}$, given the optimal period-1 pricing from (D.26). To start with, recall (D.24) and calculate the partial derivative of N_1^x with respect to investments/investments I^x, I^{-x} (holding p_1^x and p_1^{-x} fixed):

$$\frac{\partial N_1^x}{\partial I^x} = \frac{q^x - c^x}{2\kappa} \quad \text{and} \quad \frac{\partial N_1^{-x}}{\partial I^x} = -\frac{q^x - c^x}{2\kappa}. \qquad \text{(D.27)}$$

Thus,

$$\frac{\partial D^x}{\partial I^x} = N_1^x + \frac{\partial N_1^x}{\partial I^x} I^x + \eta\left(\frac{\partial N_1^{-x}}{\partial I^x}\right) I^{-x}$$
$$\frac{\partial D^{-x}}{\partial I^x} = \eta\left(N_1^x + \frac{\partial N_1^x}{\partial I^x} I^x\right) + \left(\frac{\partial N_1^{-x}}{\partial I^x}\right) I^{-x}.$$

Hence, the partial derivative of period-1 payoff π_1^x with respect to investment I^x becomes

$$\frac{\partial \pi_1^x}{\partial I^x} = \left(\frac{\partial N_1^x}{\partial I^x}\right)(p_1^x - I^x q^x) + \frac{\partial \pi_2^x}{\partial D^x}\left(N_1^x + \frac{\partial N_1^x}{\partial I^x} I^x + \eta\left(\frac{\partial N_1^{-x}}{\partial I^x}\right) I^{-x}\right)$$
$$+ \frac{\partial \pi_2^x}{\partial D^{-x}}\left(\eta\left(N_1^x + \frac{\partial N_1^x}{\partial I^x} I^x\right) + \left(\frac{\partial N_1^{-x}}{\partial I^x}\right) I^{-x}\right) - N_1^x q^x - \lambda I^x$$

If interior, i.e., $I^x \in (0, 1)$, optimal investment/effort solves the first-order condition $\frac{\partial \pi_1^x}{\partial I^x} = 0$.

Recall the optimal price p_1^x from (D.26) so that

$$p_1^x - I^x q^x = \kappa - \frac{2(1-\eta)\phi^x I^x}{3}.$$

Next, note that (in a symmetric equilibrium), $N_1^x = N_1^{-x} = 1/2$, $q^x = q^{-x}$, $I^x = I^{-x}$ as well as $\frac{\partial \pi_2^x}{\partial D^x} = \frac{\phi^x}{3} = -\frac{\partial \pi_2^x}{\partial D^{-x}}$ (see (D.23)), and $\frac{\partial N_1^x}{\partial I^x} = -\frac{\partial N_1^{-x}}{\partial I^x}$ (see (D.27). Using these relations, we obtain after simplifications:

$$\begin{aligned}
\frac{\partial \pi_1^x}{\partial I^x} &= \frac{\partial N_1^x}{\partial I^x}\left(\kappa - \frac{2(1-\eta)\phi^x I^x}{3}\right) - \frac{q^x}{2} \\
&\quad + \frac{\phi^x}{3}\left(\frac{1}{2} + \frac{\partial N_1^x}{\partial I^x} I^x(1-\eta)\right) - \frac{\phi^x}{3}\left(\frac{\eta}{2} - \frac{\partial N_1^x}{\partial I^x} I^x(1-\eta)\right) - \lambda I^x \\
&= \frac{\partial N_1^x}{\partial I^x}\left(\kappa - \frac{2(1-\eta)\phi^x I^x}{3}\right) - \frac{q^x}{2} + \frac{\phi^x}{3}\left(\frac{(1-\eta)}{2} + \left(\frac{\partial N_1^x}{\partial I^x}\right)(2I^x(1-\eta))\right) - \lambda I^x \\
&= \kappa\left(\frac{\partial N_1^x}{\partial I^x}\right) - \frac{q^x}{2} + \frac{\phi^x(1-\eta)}{6} - \lambda I^x.
\end{aligned}$$

Inserting $\frac{\partial N_1^x}{\partial I^x} = \frac{(q^x - c^x)}{2\kappa}$ into above expression for $\frac{\partial \pi_1^x}{\partial I^x}$, we obtain

$$\frac{\partial \pi_1^x}{\partial I^x} = \frac{1}{2}\left(\frac{\phi^x(1-\eta)}{3} - c^x - 2\lambda I^x\right).$$

As a result, equilibrium investment/effort satisfies — if it is interior and solves $\frac{\partial \pi_1^x}{\partial I^x} = 0$ —

$$I^x = \frac{\phi^x(1-\eta) - 3c^x}{6\lambda}.$$

Overall, we therefore obtain

$$I^x = \min\left\{1, \left[\frac{\phi^x(1-\eta) - 3c^x}{6\lambda}\right]^+\right\}. \tag{D.28}$$

The equilibrium investment expression from Proposition 2 follows upon setting $\eta = 0$. The equilibrium investment expression from Proposition 4 follows upon inserting $c^x = c(1+\eta)$.

D.1.3 Payoffs

We now calculate the payoff of platform x, using the derived expressions for prices and investment. Thus,

$$\pi_1^x = N_1^x\left[p_1^x - q^x I^x\right] + N_2^x p_2^x - \frac{\lambda(I^x)^2}{2} = \kappa - \frac{I^x\phi^x(1-\eta)}{3} - \frac{\lambda(I^x)^2}{2},$$

where we used that $N_1^x = \frac{1}{2}$ and the price expression $p_1^x = \kappa + I^x\left(q^x - \frac{2(1-\eta)I^x\phi^x}{3}\right)$ as well as $\pi_2^x = \kappa/2$.

Next, we calculate user welfare. We know from earlier results that users' total payoff in period $t = 2$ reads

$$\begin{aligned}
u_2 &= K^x + \phi^x D^x + \frac{\gamma^x}{2} - \kappa - \frac{\hat{\kappa}}{4} \\
&= K^x + \frac{\phi^x(1+\eta)I^x}{2} + \frac{\gamma^x}{2} - \kappa - \frac{\hat{\kappa}}{4},
\end{aligned}$$

where we used $D^x = (1+\eta)\hat{D}^x = (1+\eta)N_1^x I^x = 0.5(1+\eta)I^x$. Next, total welfare reads

$$\begin{aligned} u_1 &= K^x + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4} - p_1^x + u_2 + I^x(q^x - c^x) \\ &= K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4} + \frac{2I^x(1-\eta)\phi^x}{3} - I^x q^x + I^x(q^x - c^x) + u_2 \\ &= \underbrace{2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right)}_{\equiv const} + I^x\left(\frac{\phi^x(7-\eta)}{6} - c^x\right), \end{aligned} \tag{D.29}$$

whereby investment I^x was previously derived and is characterized in (D.28) and $c^{=}c(1+\eta)$. The first term is simply a model constant, which we denote by "*const*." As a result, we have that $I^x\left(\frac{\phi^x(7-\eta)}{6} - c^x\right) \geq 0$, with the inequality being strict if $I^x > 0$. Since I^x decreases with η, it also follows that total user welfare decreases with η, i.e., $\frac{\partial u_1}{\partial \eta} \leq 0$, and does so strictly when investment $I^x > 0$ is positive. We therefore conclude that required data sharing ($\eta > 0$) reduces user welfare relative to the baseline with $\eta = 0$, and does so strictly when $\phi^x > 3c$.

Finally, also note that any platform's available data in period $t = 2$ (after data sharing) reads $D^x = \frac{(1+\eta)I^x}{2}$, so that

$$\frac{\partial D^x}{\partial \eta} = \frac{I^x}{2} + \frac{1+\eta}{2}\frac{\partial I^x}{\partial \eta}$$

When $I^x \in (0,1)$, then

$$2\left(\frac{\partial D^x}{\partial \eta}\right) = \frac{\phi^x(1-\eta) - 3c(1+\eta)}{6\lambda} - \frac{(1+\eta)(\phi^x+3)}{6\lambda} < 0.$$

As such, when investment is interior (i.e., $I^x \in (0,1)$), then data sharing decreases the amount of data that platforms use in period $t = 2$.

D.2 Market for Data when Users own Data — Proof of Proposition 6

We start by solving the model at the beginning of state $t = 2$, given a total stock of data $D = N_1^A I^A + N_1^B I^B$. We solve for a symmetric equilibrium: Symmetry in equilibrium implies $N_1^x = 1/2$, $D^A = D^B$, and $I^A = I^B = I^x$, so $D = I^x$. At the beginning of period $t = 2$ (before prices are chosen), the two platforms simultaneously choose D^x to maximize

$$\max_{D^x \in [0,D]} \pi_2^x - cD^x, \tag{D.30}$$

taking the choice of the other platform, i.e. D^{-x}, as given. Here, the payoff π_2^x is characterized in Proposition 1. Recall (B.7) and (B.8), and observe that $\frac{\partial^2 \pi_2^x}{\partial (D^x)^2} > 0$. Thus, there exists no interior maximum to the optimization (D.30), and we therefore conjecture $D^x = D^{-x} = D$.

Next, we characterize optimal choice of D^x, given $D = \hat{D}^A + \hat{D}^B$. For this purpose, we use (B.7) and $D^x = D^{-x}$ — which holds in symmetric equilibrium — to take the derivative with respect to D^x in (D.30), yielding

$$\frac{\partial}{\partial D^x}(\pi_2^x - cD^x) = \frac{\phi^x}{3} - c$$

under $D^x = D^{-x}$. We now consider two distinct cases.

First, suppose that $c > \frac{\phi^x}{3}$. Then, for any level of $D \geq 0$, platforms optimally choose $D^x = 0$. Anticipating $D^x = 0$ for any levels of $D = \hat{D}^A + \hat{D}^B$, it is clear that platforms optimally do not exert any investment I^x to collect data, so $I^x = D = 0$. The platform payoff in period $t = 2$ then reads $\pi_2^x = \frac{\kappa}{2}$.

Second, suppose that $\phi^x \geq 3c$. In this case, $D_2^x = D$ and both platforms acquire all available data D. The platform payoff in period $t = 2$ then reads $\pi_2^x = \frac{\kappa}{2} - cD$. Moreover, using $D_2^x = D = I^x/2 + I^{-x}/2$ as well as the expression for π_2^x from Lemma 1, we have

$$\pi_2^x = \frac{\left[3\kappa + \phi^x(I^x + I^{-x})/2 - \phi^{-x}(I^x + I^{-x})/2\right]^2}{18\kappa}$$

so that $\frac{\partial \pi_2^x}{\partial I^x} = 0$ (since $\phi^x = \phi^{-x}$). It is immediate from (5) that $\frac{\partial \pi_2^x}{\partial I^x} = 0$ implies $\frac{\partial \pi_1^x}{\partial I^x} \leq 0$ with the inequality being strict for $I^x > 0$. As such, optimal data collection investment satisfies $I^A = I^B = 0$, so $D = D^x = 0$. In either case, we have established $I^x = D = D^x = 0$ in symmetric equilibrium.

Given that $I^x = 0$ in a symmetric equilibrium, the platform x solves in each period $t = 1, 2$:

$$\max_{p_t^x} N_t^x p_t^x,$$

taking the choice of the other platform, i.e., p_t^{-x}, as given. We have that

$$\frac{\partial N_t^x}{\partial p_t^x} = -\frac{1}{2\kappa} \quad \text{and} \quad \frac{\partial N_t^{-x}}{\partial p_t^x} = \frac{1}{2\kappa}.$$

Thus, price p_1^x solves the first-order condition $N_t^x + \frac{\partial N_t^x}{\partial p_t^x} p_t^x = 0$, which we can solve — using $N_t^x = 1/2$ — for $p_t^x = \kappa$ with $t = 1, 2$.

Platform x's total payoff at $t = 1$ reads $\pi_1^x = \kappa$, while $\pi_2^x = \kappa/2$. Next, we calculate user welfare. We know that users total payoff in period $t = 2$ reads

$$u_2 = K^x - \kappa - \frac{\hat{\kappa}}{4} + \phi^x D_2 = K^x - \kappa - \frac{\hat{\kappa}}{4}.$$

As such, total welfare reads

$$u_1 = 2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right),$$

and it is clear that u_1 is lower than under the baseline, which is characterized in (D.29) for $\eta = 0$, that is,

$$2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right) + I^x\left(\frac{7\phi^x}{6} - c^x\right) \geq u_1.$$

The inequality above is strict if and only if it holds in the baseline that $I^x > 0 \iff \phi^x > 3c$.

E Proofs and Derivations for Section 4

E.1 Proof of Proposition 8

According to the objective (11), the user union chooses at the beginning of time $t = 1$ (before investments and $t = 1$ prices are chosen) f to maximize $u_1 - fI^x$, whereby $c^x = c - f$ and $\eta = 0$. Next, recall the expression for user welfare from (D.29), that is,

$$u_1 = const + I^x\left(\frac{\phi^x(7 - \eta)}{6} - c^x\right),$$

where we define for convenience

$$const := 2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right).$$

We consider $\eta = 0$ (no data sharing), and we rewrite the objective (11) as

$$\hat{u}_1 := u_1 - fI^x = const + I^x\left(\frac{7\phi^x}{6} - c + f\right) - fI^x = const + I^x\left(\frac{7\phi^x}{6} - c\right), \tag{E.31}$$

where I^x is from Proposition 2, that is,

$$I^x = \min\left\{1, \left[\frac{\phi^x - 3(c-f)}{6\lambda}\right]^+\right\}. \tag{E.32}$$

If $6c \geq 7\phi^x$, then (E.31) immediately implies that the user union optimally implements $I^x = 0$, which is achieved by setting $f = 0$.

On the other hand, if $6c < 7\phi^x$, the user union optimally implements $I^x > 0$ and, in fact, the objective $\hat{u}_1 = u_1 - fI^x$ strictly increases with I^x, i.e., $\frac{\partial(u_1 - fI^x)}{\partial I^x} > 0$. As such, the relation (E.31) reveals that the user union optimally chooses f to maximize investment subject to $I^x \leq 1$ (exogenous upper bound of investment); thus, $I^x = 1$. Solving $I^x = 1$ (using (E.32)) for f yields

$$f = f^* = \frac{6\lambda + 3c - \phi^x}{3}.$$

which concludes the proof. The user welfare then reads

$$\hat{u}_1 = 2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right) + \frac{7\phi^x}{6} - c,$$

which is strictly larger than under the baseline (with $\eta = 0$) or data sharing (with $\eta > 0$) yielding user welfare:

$$2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right) + I^x\left(\frac{\phi^x(7-\eta)}{6} - c(1+\eta)\right).$$

Likewise, since we have shown in Proposition 6, that user welfare under the baseline is higher than when users own their data and sell their data in a market for data, it readily follows that welfare under user union is higher than under that scenario too.

E.2 Proof of Proposition 9

To begin with, note that $D^x = N_1^x I^x$ as well as $\theta^x = 1$, where by symmetry $N_1^x = 1/2$ and $I^x = I^{-x}$ so $D^x = D^{-x}$.

E.2.1 Prices

It holds that $N_1^x = N_1^{-x} = \hat{z}_1$, with

$$N_1^x = \frac{1}{2} - \frac{(p_1^x - p_1^{-x})}{2\kappa}.$$

As such, we can calculate

$$\frac{\partial N_1^x}{\partial p_1^x} = -\frac{1}{2\kappa} \quad \text{and} \quad \frac{\partial N_1^{-x}}{\partial p_1^x} = \frac{1}{2\kappa}$$

and

$$\frac{\partial D^{-x}}{\partial p_1^x} = \frac{I^{-x}}{2\kappa} \quad \text{and} \quad \frac{\partial D^x}{\partial p_1^x} = \frac{-I^x}{2\kappa}$$

Thus, the first-order condition with respect to price p_1^x reads:

$$\frac{\partial \pi_1^x}{\partial p_1^x} = N_1^x - \frac{p_1^x}{2\kappa} + \frac{I^x q^x}{2\kappa} - \frac{\partial \pi_2^x}{\partial D^x}\left(\frac{I^x}{2\kappa}\right) + \frac{\partial \pi_2^x}{\partial D^{-x}}\left(\frac{I^{-x}}{2\kappa}\right) = 0. \tag{E.33}$$

Using $N_1^x = \frac{1}{2}$, $I^x = I^{-x}$, and $\frac{\partial \pi_2^x}{\partial D^x} = \frac{\phi^x}{3} = -\frac{\partial \pi_2^x}{\partial D^{-x}}$, we can solve

$$p_1^x = \kappa + I^x\left(q - \frac{2\phi^x}{3}\right).$$

E.2.2 Investment

To start with, note that because of

$$N_1^x = \frac{1}{2} - \frac{(p_1^x - p_1^{-x})}{2\kappa},$$

we have $\frac{\partial N_1^x}{\partial I^{x'}} = 0$ for all $x, x' \in \{A, B\}$. Thus,

$$\frac{\partial D^x}{\partial I^x} = N_1^x = \frac{1}{2} \quad \text{and} \quad \frac{\partial D^{-x}}{\partial I^x} = 0.$$

Hence, the derivative with respect to investment I^x becomes

$$\frac{\partial \pi_1^x}{\partial I^x} = \frac{\partial \pi_2^x}{\partial D^x}\left(\frac{\partial D^x}{\partial I^x}\right) - N_1^x q - \lambda I^x$$

If interior, i.e., $I^x \in (0,1)$, optimal investment solves the first-order condition $\frac{\partial \pi_1^x}{\partial I^x} = 0$. Using $N_1^x = 1/2$ and $\frac{\partial \pi_2^x}{\partial D^x} = \frac{\phi^x}{3}$, we obtain $I^x = \frac{\phi^x - 3q}{6\lambda}$, so optimal investment satisfies

$$I^x = \max\left\{1, \left[\frac{\phi^x - 3q}{6\lambda}\right]^+\right\}.$$

E.2.3 User Welfare

Using the expression for u_1 from (D.29) with $\eta = 0$ and $c^x = c$, total user welfare becomes

$$\hat{u}_1 = u_1 - qI^x = 2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right) + I^x\left(\frac{7\phi^x}{6} - c\right),$$

and depends on q^x only via investment I^x. When $6c \geq 7\phi^x$, then it is optimal to stipulate $I^x = 0$ which is achieved by setting $q = 0$.

When, on the other hand, $6c < 7\phi^x$, $\hat{u}_1$ strictly increases with I^x. As such, optimal investment maximizing user welfare $\hat{u}_1$ is either at the boundary 1 (exogenous upper boundary for investment) or such that the constraint $\pi_1^x \geq 0$ binds (platform participation constraint) and platforms just break even. Using arguments analogous to the ones used in the proof of Proposition 8, this leads to optimal investment $I^x = 1$. The price for data is the same for the two platforms and satisfies

$$q = \frac{\phi^x}{3} - 2\lambda I^x = \frac{\phi^x}{3} - 2\lambda.$$

Total user welfare then reads

$$\hat{u}_1 = u_1 - qI^x = 2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right) + \frac{7\phi^x}{6} - c,$$

which is strictly larger than under the baseline (with $\eta = 0$) or data sharing (with $\eta > 0$) yielding user welfare: $2\left(K^x - \kappa + \frac{\gamma^x}{2} - \frac{\hat{\kappa}}{4}\right) + I^x\left(\frac{\phi^x(7-\eta)}{6} - c(1+\eta)\right)$. Likewise, since we have shown in Proposition 6, that user welfare under the baseline is higher than when users own their data and sell their data in a market for data, it readily follows that welfare under user union is higher than under that scenario too.

E.3 Details for Section 4.3

Take the reward level of the user union f as given, and note $c^x = c - f$. As argued in the main text, given the reward f, the equilibrium from $t = 1$ onward is then characterized in Proposition 1 and Lemma 1. To induce $\theta^x = 1$, platform x sets $q^x \geq c^x$. Assume that given f and c^x, the continuation equilibrium exists and is unique (up to q^x), i.e., $\hat{z}_t$, I^x, p_2^x and $\bar{p}_1^x = p_1^x - I^x q^x$ are unique and do not depend on the exact level of q^x (as long as $\theta^x = 1$ is induced).

We now analyze under what conditions user union participation is privately optimal (i.e., incentive compatible) for any individual user z. To do so, suppose that individual user z deviates by not joining the user union at the beginning of time $t = 1$ while all other users $[0,1] - \{z\}$ join the union; without loss of generality, consider that $z \leq \hat{z}_1$. Thus, user z saves the membership fee $m(z)$, but does not receive the reward for contributing data. Also, user z is quoted the same service prices p_t^x as other users (i.e., there is no service price discrimination), but faces potentially different data prices $\hat{q}^x$ than other users.

More formally, sser z's utility from the deviation is

$$u^{Dev}(z) = \max_{x\in\{A,B\},\hat{\theta}^x\in[0,1],\rho\in\{0,1\}} \rho\big(Y_1^x - p_1^x + \hat{\theta}^x I^x(\hat{q}^x - c) - \kappa^x(z)\big) + \max_{x\in\{A,B\}} \big(Y_2^x - p_2^x - \kappa^x(z)\big),$$

where ρ denotes z's decision to consume at any platform x and $\hat{\theta}^x \in [0,1]$ is z's choice of contributing data at the platform x she joins. For simplicity, we already imposed that z participates in $t = 2$, as — per assumption — the entire market is covered ($N_t^A + N_t^B = 1$) and the deviation of z (and union membership) does not affect period-2 payoff $\max_{x\in\{A,B\}}\big(Y_2^x - p_2^x - \kappa^x(z)\big)$.

In contrast, user z's payoff from joining the union is

$$u^{Union}(z) = -m(z) + \max_{x\in\{A,B\},\hat{\theta}^x\in[0,1]} \big(Y_1^x - p_1^x + \hat{\theta}^x I^x(q^x - c + f) - \kappa^x(z)\big) + \max_{x\in\{A,B\}} \big(Y_2^x - p_2^x - \kappa^x(z)\big),$$

We already impose the assumption that in equilibrium the entire market is covered; thus, if z does not deviate, it must be that she adopts at least one platform. Notice that, being part of the union, z optimally shares data with the platform x she joins, so it is optimal to set $\hat{\theta}^x = 1$ in above maximization.

Without loss of generality, suppose that $z \leq \hat{z}_1$, so $m(z) = I^A f$ and

$$A = \arg\max_{x\in\{A,B\}} \big(Y_1^x - p_1^x + I^x(q^x - c + f) - \kappa^x(z)\big)$$

Then, the gain from deviating equals

$$\begin{aligned} \Delta^z &:= u^{Dev}(z) - u^{Union}(z) \qquad \text{(E.34)} \\ &= -\big[Y_1^A - p_1^A + I^A(q^A - c) - \kappa^A(z)\big] + \max_{x\in\{A,B\},\rho\in\{0,1\},\hat{\theta}^x\in\{0,1\}} \rho\big(Y_1^x - p_1^x - \kappa^x(z) + \hat{\theta}^x I^x(\hat{q}^x - c)\big) \end{aligned}$$

Next, notice that by Proposition 1, we have $\frac{\partial p_1^x}{\partial q^x} = I^x$. That is, period-1 prices can be written in the form

$$p_1^x = \bar{p}_1^x + I^x q^x,$$

where $\bar{p}_1^x$ is unique (by assumption that the continuation equilibrium from Proposition 1 is unique) and does not depend on $I^{x'}$ or $q^{x'}$, i.e., $\frac{\partial \bar{p_1}^x}{\partial I^{x'}} = \frac{\partial \bar{p_1}^x}{\partial q^{x'}} = 0$. Thus, inserting $p_1^x = \bar{p}_1^x + I^x q^x$, we obtain

$$\Delta^z = -\big[Y_1^A - \bar{p}_1^A - I^A c - \kappa^A(z)\big] + \max_{x \in \{A,B\}, \rho \in \{0,1\}, \hat{\theta}^x \in [0,1]} \rho\big(Y_1^x - p_1^x - \kappa^x(z) + \hat{\theta}^x I^x(\hat{q}^x - c)\big). \quad \text{(E.35)}$$

We now need to show $\Delta^z \leq 0$.

If user z is not member of the union, platform x can offer z a potentially different data price $\hat{q}^x$ (e.g., $\hat{q}^x < q^x$) than the data price q^x that it offers to union members. That is, we assume that the platform x can make the price for data that it pays user z contingent on union membership. We assume that the timing underlying the choice of $\hat{q}^x$ is as follows. First, after observing q^x, I^x, p_1^x, the user z joins a platform x; then, x observes whether z is user union member; if yes, z is quoted data price q^x and, if not, x can alter the data price that it quotes to z to any level $\hat{q}^x$ (i.e., x quotes a new price $\hat{q}^x$ to the deviant). When user z joins platform x and x observes that z is not member, it is optimal for x to minimize payments $\hat{q}^x$ whilst inducing z to set $\hat{\theta}^x = 1$. As such, it is optimal for x to choose $\hat{q}^x \leq c$, i.e., not to pay the deviant z more than her required cost of sharing data c. Thus, if x induces z to share data, then $\hat{q}^x = c$. Overall, it follows that $\max_{\hat{\theta}^x \in [0,1]} \hat{\theta}^x I^x(\hat{q}^x - c) = 0 = I^x[\hat{q}^x - c]^+ = 0$ (recall that $[\cdot]^+ = \max\{\cdot, 0\}$). In other words, it is optimal for the platform not to leave rents to the deviant by stipulating $\hat{q}^x = c$.

We now show under what circumstances $\Delta^z \leq 0$ and user union is incentive compatible. We do so separately when $I^A, I^B > 0$ in equilibrium (see Part I) and when $I^A \cdot I^B = 0$ in equilibrium (see Part II)

Part I

We now establish under what circumstances $\Delta^z \leq 0$ when $I^A, I^B > 0$. First, when $\rho = 0$ (i.e., the deviant stays out of the market in $t = 1$), then $\Delta^z = -\big[Y_1^A - \bar{p}_1^A - I^A c - \kappa^A(z)\big]$, which must be negative by the assumption that, conditional on union membership, user z derives positive payoff and participates.[52]

When $\rho = 1$ and

$$A \in \arg \max_{x \in \{A,B\}} \max_{\rho \in \{0,1\}, \hat{\theta}^x \in [0,1]} \rho\big(Y_1^x - p_1^x - \kappa^x(z) + \hat{\theta}^x I^x(\hat{q}^x - c)\big)$$

then (according to (E.35))

$$\Delta^z = -I^A(q^A - c) + I^A[\hat{q}^A - c]^+ = -I^A(q^A - c),$$

which is negative for $q^A \geq c$.

When on the other hand

$$B \in \arg \max_{x \in \{A,B\}} \max_{\rho \in \{0,1\}, \hat{\theta}^x \in [0,1]} \rho\big(Y_1^x - p_1^x - \kappa^x(z) + \hat{\theta}^x I^x(\hat{q}^x - c)\big)$$

and $\rho = 1$, then (according to (E.35))

$$\Delta_z = -\big[Y_1^A - \bar{p}_1^A - I^A c - \kappa^A(z)\big] + (Y_1^B - \bar{p}_1^B - q^B I^B - \kappa^B(z)),$$

which is negative for

$$q^B \geq \frac{(Y_1^B - Y_1^A) - (\bar{p}_1^B - \bar{p}_1^A) - \big(\kappa^B(z) - \kappa^A(z)\big) + I^A c}{I^B}.$$

[52] This must be the case because we assume that the entire market is covered in equilibrium.

Thus, we have $\Delta^z \leq 0$ for $z \leq \hat{z}_1$ if

$$q^A \geq c \quad \text{and} \quad q^B \geq \frac{(Y_1^B - Y_1^A) - (\bar{p}_1^B - \bar{p}_1^A) - \big(\kappa^B(z) - \kappa^A(z)\big) + I^A c}{I^B}.$$

Analogously, we have $\Delta^z \leq 0$ for $z > \hat{z}_1$ if

$$q^B \geq c \quad \text{and} \quad q^A \geq \frac{(Y_1^A - Y_1^B) - (\bar{p}_1^A - \bar{p}_1^B) - \big(\kappa^A(z) - \kappa^B(z)\big) + I^B c}{I^A}.$$

At the same time, to incentivize $\theta^x = 1$, we must have $q^x \geq c - f$.

Altogether, $\Delta^z \leq 0$ and $q^x \geq c - f$ hold for all $z \in [0,1]$ if

$$\begin{aligned} q^A &\geq c + \max_{z \in [0,1]} \left[\max \left\{ -f, \frac{(Y_1^A - Y_1^B) - (\bar{p}_1^A - \bar{p}_1^B) - \big(\kappa^A(z) - \kappa^B(z)\big) + I^B c}{I^A} \right\} \right] \\ q^B &\geq c + \max_{z \in [0,1]} \left[\max \left\{ -f, \frac{(Y_1^B - Y_1^A) - (\bar{p}_1^B - \bar{p}_1^A) - \big(\kappa^B(z) - \kappa^A(z)\big) + I^A c}{I^B} \right\} \right]. \end{aligned} \tag{E.36}$$

As such, if (E.36) holds in equilibrium, then union membership is incentive compatible. Note that (E.36) can be seen as incentive compatibility condition for user union membership.

Recall that the exact values of q^x are not payoff-relevant in a sense made precise in Proposition 1, i.e., the value of q^x does not affect $\hat{z}_t$, p_2^x, $\bar{p}_1^x$, I^x, or u_1. Notice that the right-hand-side of both inequalities in (E.36) does not depend on q^A or q^B and, in particular, involves equilibrium quantities only depending on model parameters. As such, we can always find ("large enough") equilibrium levels of q^A and q^B that satisfy (E.36). That is, provided $I^A, I^B > 0$, there exists an equilibrium (unique up to q^x) in which all users are members of the union and user union participation is incentive compatible in equilibrium (i.e., (E.36) holds). Because, given a level of f and $c^x = c - f$, the continuation equilibrium (outlined in Proposition 1 and Lemma 1) exists with unique $\hat{z}_t$, $\bar{p}_1^x$, p_2^x, and I^x, the user union equilibrium is unique too up to the level of q^x.

E.3.1 Part II

We now establish under what circumstances $\Delta^z \leq 0$ when $I^A \cdot I^B = 0$, i.e., when $I^A = 0$ or $I^B = 0$ or both.

First, when $\rho = 0$ (i.e., the deviant stays out of the market in $t = 1$), then $\Delta^z = -\big[Y_1^A - \bar{p}_1^A - I^A c - \kappa^A(z)\big]$, which must be negative by the assumption that, conditional on union membership, user z derives positive payoff and participates.[53]

Next, when $I^A = 0$, then $p_1^A = \bar{p}_1^A$ as well as

$$A = \arg\max_{x \in \{A,B\}} \big(Y_1^x - p_1^x + I^x(q^x - c + f) - \kappa^x(z)\big),$$

which — owing to $I^B(q^B - c + f) \geq 0$ — implies (for any $z \leq \hat{z}_1$)

$$A = \arg\max_{x \in \{A,B\}} \max_{\hat{\theta}^x \in [0,1]} \big(Y_1^x - p_1^x - \kappa^x(z) + \underbrace{\hat{\theta}^x I^x(\hat{q}^x - c)}_{=0}\big).$$

Inserting this relation into (E.35), we obtain $\Delta^z \leq 0$.

Next, consider $I^B = 0$. Then,

$$A = \arg\max_{x \in \{A,B\}} \big(Y_1^x - p_1^x + I^x(q^x - c + f) - \kappa^x(z)\big),$$

[53] This must be the case because we assume that the entire market is covered in equilibrium.

implies

$$A = \arg \max_{x\in\{A,B\}} \max_{\hat{\theta}^x\in[0,1]} \left(Y_1^x - p_1^x - \kappa^x(z) + \underbrace{\hat{\theta}^x I^x(\hat{q}^x - c)}_{=0}\right)$$

for any $z \leq \hat{z}_1$ if $q^A - c + f = 0$. When $I^B = 0$ and $q^A = c - f$, then $\Delta^z = I^A f \leq 0$ if $f \leq 0$ (see (E.34)).

Analogously, when $I^A = 0 < I^B$, then $\Delta^z \leq 0$ for any $z > \hat{z}_1$ holds if $f \leq 0$ and $q^B = c - f$. As such, when $I^A = 0$ or $I^B = 0$, user union is incentive compatible if $q^x = c - f$ as well as $f \leq 0$. Under these circumstances, the continuation equilibrium (outlined in Proposition 1 and Lemma 1) exists with unique $\hat{z}_t$, $\bar{p}_1^x$, p_2^x, and I^x.

F Extended Discussions

F.1 Market for Data When Platforms own Data — Model Variant with "Symmetric" Equilibrium

Assume ex-ante symmetry, i.e., $\Delta_K = 0$ as well as $\phi^A = \phi^B$. We now introduce a model variant with a market for data, in which platforms own and trade user-generated data. Notably, we make additional assumptions such that the model variant features a symmetric equilibrium in the subgame in period $t = 1$.

F.1.1 Stage 2

At the beginning of time period $t = 2$, platform A and B trade data with each other, with endogenous price p_2^D. Without loss of generality, it suffices to consider that platform A and B choose the optimal allocation of data through Nash Bargaining at the beginning of period $t = 2$, with equal bargaining weights 1/2. The price for data p_2^D is then chosen to implement the split of resulting surplus. According to Lemma 3, total surplus is maximized when one platform x shares data with platform $-x$, but not the other way around.

We now derive the Nash bargaining solution. For this sake, call B the platform that shares data with the other one. That is, B shares data with platform A, but not the other way around, with $\hat{D}^A\phi^A \geq \hat{D}^B\phi^B$. Thus, $D^A = \hat{D}^A + \hat{D}^B$, whereas $D^B = \hat{D}^B$. Using the period-2 platform payoff under equilibrium pricing from Lemma 1 with $D^A = \hat{D}^A + \hat{D}^B$ and $D^B = \hat{D}^B$, we obtain the following platform payoffs just after Nash bargaining but before choosing price p_2^A and p_2^B:

$$\pi_2^A = \frac{(3\kappa + \phi^A\hat{D}^A)^2}{18\kappa} \quad \text{and} \quad \pi_2^B = \frac{(3\kappa - \phi^A\hat{D}^A)^2}{18\kappa}.$$

Then, just after the data trade, total surplus of both platforms (excluding user surplus) reads

$$S := \pi_2^A + \pi_2^B = \frac{(3\kappa + \phi^A\hat{D}^A)^2 + (3\kappa - \phi^A\hat{D}^A)^2}{18\kappa}$$

If there were no data sharing, then $D^x = \hat{D}^x$ and platforms' period-2 payoff under equilibrium pricing would be according to Lemma 1:

$$\hat{\pi}_2^x = \frac{(3\kappa + \phi^x\hat{D}^x - \phi^{-x}\hat{D}^{-x})^2}{18\kappa}.$$

Under these circumstances, total surplus would be

$$\hat{\pi}_2^x + \hat{\pi}_2^{-x} = \frac{(3\kappa + \phi^x\hat{D}^x - \phi^{-x}\hat{D}^{-x})^2 + (3\kappa + \phi^{-x}\hat{D}^{-x} - \phi^x\hat{D}^x)^2}{18\kappa}.$$

Thus, surplus generated through the efficient (ex-post) allocation of data is $S - (\hat{\pi}_2^A + \hat{\pi}_2^B)$. As per

Nash bargaining protocol (among the two parties $x = A$ and $x = B$) with equal bargaining weights $1/2$, the payoff of platform x becomes

$$\tilde{\pi}_2^x := \frac{1}{2}(S - \hat{\pi}_2^A - \hat{\pi}_2^B) + \hat{\pi}_2^x,$$

which is the payoff at the beginning of period $t = 2$ before data trade happens.

As such, the payoff of platform A at inception of period $t = 2$ reads

$$\begin{aligned}\tilde{\pi}_2^A &:= \left(\frac{(3\kappa + \phi^A \hat{D}^A)^2 + (3\kappa - \phi^A \hat{D}^A)^2 + (3\kappa + \phi^A \hat{D}^A - \phi^B \hat{D}^B)^2 - (3\kappa + \phi^B \hat{D}^B - \phi^A \hat{D}^A)^2}{36\kappa}\right)\\ &= \frac{(\phi^A \hat{D}^A)^2 + 6\kappa(\phi^A \hat{D}^A - \phi^B \hat{D}^B)}{18\kappa}.\end{aligned}$$

Likewise, the payoff of platform B at inception of period $t = 2$ becomes

$$\begin{aligned}\tilde{\pi}_2^B &:= \left(\frac{(3\kappa + \phi^A \hat{D}^A)^2 + (3\kappa - \phi^A \hat{D}^A)^2 + (3\kappa + \phi^B \hat{D}^B - \phi^A \hat{D}^A)^2 - (3\kappa + \phi^A \hat{D}^A - \phi^B \hat{D}^B)^2}{36\kappa}\right)\\ &= \frac{(\phi^A \hat{D}^A)^2 + 6\kappa(\phi^B \hat{D}^B - \phi^A \hat{D}^A)}{18\kappa}.\end{aligned}$$

There does not exist a fully symmetric equilibrium. We therefore look for an equilibrium in which both platforms enter period $t = 2$ symmetrically, i.e., $N_1^x = 1/2$ and $\hat{D}^x = I^x/2$, and, with equal probability of $1/2$, A and B are the data "buyers" and "sellers" respectively in the trade of data. Formally, at the beginning of period $t = 2$, nature determines which platform is buyer and seller of data, where each platform is selected by nature with equal probability. After that draw, we possibly relabel platforms such that A is data buyer and B is data seller. This assumption greatly simplifies the analysis. Notice that in the general case, platform A as the "stronger" platform will act as buyer of data with probability one, while B is seller with probability one. However, in the case of this section, not the strength but nature determines who buys/sells data; in the equilibrium we consider, both platforms enter stage 2 symmetrically, so the choice of nature is consistent with the result of Lemma 3 and its implications for the optimal data trade.

From the perspective of platform x, at the very beginning of period $t = 1$, there is a draw (by nature) resulting into $x = A$ (i.e., $-x = B$) and $x = B$ (i.e., $-x = A$) with equal probability $1/2$. Then, above expressions imply that platform x's expected payoff just before this draw becomes

$$\begin{aligned}\bar{\pi}_2^x &= \frac{1}{2}\left(\frac{(\phi^x \hat{D}^x)^2 + 6\kappa(\phi^x \hat{D}^x - \phi^{-x} \hat{D}^x)}{18\kappa} + \frac{(\phi^{-x} \hat{D}^{-x})^2 + 6\kappa(\phi^x \hat{D}^x - \phi^{-x} \hat{D}^{-x})}{18\kappa}\right)\\ &= \frac{(\phi^x \hat{D}^x)^2}{36\kappa} + \frac{\phi^x \hat{D}^x - \phi^{-x} \hat{D}^x}{3} + \frac{(\phi^{-x} \hat{D}^{-x})^2}{36\kappa}.\end{aligned} \tag{F.37}$$

We focus on a subgame perfect equilibrium in which both platforms choose in period $t = 1$ the triple (q^x, p_1^x, I^x) simultaneously to maximize $\bar{\pi}_2^x$. In period $t = 2$, each platform is selected with probability $1/2$ to buy data from the other one, data trade occurs, and after that platforms simultaneously set prices p_2^x. In equilibrium in the subgame in period $t = 1$, the allocation is indeed symmetric and, as it will turn out, platforms x will exert symmetric level of investment and set symmetric prices in stage $t = 1$. The equilibrium in subgame in $t = 2$ — as previously discussed — is no more symmetric.

F.2 Stage $t = 1$

We now study platforms' optimal choice of prices and investments, given the continuation payoff in period $t = 2$, which is characterized in (F.37) and is denoted $\bar{\pi}_2^x$. For this sake, platform x

maximizes

$$\pi_1^x := N_1^x p_1^x - q^x N_1^x I^x + \bar{\pi}_2^x - \frac{\lambda (I^x)^2}{2}, \tag{F.38}$$

where $\bar{\pi}_2^x$ is from (F.37).

F.2.1 Prices

We set $\theta^x = 1$. We start by analyzing optimal pricing in $t = 1$. First, note that $\hat{D}^x = I^x N_1^x = I^x/2$. One calculates using $\bar{\pi}_2^x$ from (F.37):

$$\frac{\partial \bar{\pi}_2^x}{\partial \hat{D}^x} = \frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3} \quad \text{and} \quad \frac{\partial \bar{\pi}_2^x}{\partial \hat{D}^{-x}} = -\left(\frac{(\phi^{-x})^2}{36\kappa} + \frac{\phi^{-x}}{3}\right). \tag{F.39}$$

Next, we calculate (using (D.24)):

$$\frac{\partial N_1^x}{\partial p_1^x} = -\frac{1}{2\kappa} \quad \text{and} \quad \frac{\partial N_1^{-x}}{\partial p_1^x} = \frac{1}{2\kappa}.$$

Thus, the first-order condition of π_1^x from (F.38) with respect to price p_1^x reads:

$$\frac{\partial \pi_1^x}{\partial p_1^x} = N_1^x - \frac{p_1^x}{2\kappa} + \frac{I^x q^x}{2\kappa} + \frac{\partial \pi_2^x}{\partial \hat{D}^x}\frac{\partial \hat{D}^x}{\partial N_1^x}\frac{\partial N_1^x}{\partial p_1^x} + \frac{\partial \pi_2^x}{\partial \hat{D}^{-x}}\frac{\partial \hat{D}^{-x}}{\partial N_1^{-x}}\frac{\partial N_1^{-x}}{\partial p_1^x} = 0. \tag{F.40}$$

From the above relations, we know that — with $I^x = I^{-x}$, $\phi^x = \phi^{-x}$, $\frac{\partial \bar{\pi}_2^x}{\partial \hat{D}^x} = -\frac{\partial \bar{\pi}_2^x}{\partial \hat{D}^{-x}}$ and $\frac{\partial N_1^x}{\partial p_1^x} = -\frac{\partial N_1^{-x}}{\partial p_1^x}$ as well as $\frac{\partial \hat{D}^x}{\partial N_1^x} = I^x$.

Thus, using $N_1^x = \frac{1}{2}$, $I^x = I^{-x}$, the first-order condition (F.40) becomes

$$\frac{\partial \pi_1^x}{\partial p_1^x} = \frac{1}{2} - \frac{p_1^x}{2\kappa} + \frac{I^x q^x}{2\kappa} - \frac{I^x}{\kappa}\left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right) = 0.$$

We can solve above equation for period-1 price

$$p_1^x = \kappa + I^x q^x - 2I^x\left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right). \tag{F.41}$$

F.2.2 Investment

Next, we turn to solving for investment I^x. For this sake, we calculate

$$\frac{\partial N_1^x}{\partial I^x} = \frac{(q^x - 1.5c)}{2\kappa} \quad \text{and} \quad \frac{\partial N_1^{-x}}{\partial I^x} = -\frac{(q^x - 1.5c)}{2\kappa},$$

noting that $c^x = 1.5c$. Observe that when z shares data with platform x, then, with probability 1/2, platform x does not sell data at $t = 2$ and the user's privacy cost is c and otherwise, with probability 1/2, x sells all the data to $-x$ and the realized privacy cost is $2c$. Thus, on average, the user's privacy cost for sharing one unit of data is $c^x = 1.5c$.

We recall that $\hat{D}^x = I^x N_1^x$, so that

$$\frac{\partial \hat{D}^x}{\partial I^x} = N_1^x + \frac{\partial N_1^x}{\partial I^x} I^x \quad \text{and} \quad \frac{\partial \hat{D}^{-x}}{\partial I^x} = \left(\frac{\partial N_1^{-x}}{\partial I^x}\right) I^{-x}.$$

Hence, the derivative of payoff in period $t = 1$ with respect to investment I^x becomes

$$\begin{aligned}\frac{\partial \pi_1^x}{\partial I^x} &= \left(\frac{\partial N_1^x}{\partial I^x}\right)(p_1^x - I^x q^x) + \frac{\partial \bar{\pi}_2^x}{\partial \hat{D}^x}\left(N_1^x + \frac{\partial N_1^x}{\partial I^x} I^x\right) \\ &\quad + \frac{\partial \bar{\pi}_2^x}{\partial D^{-x}}\left(\left(\frac{\partial N_1^{-x}}{\partial I^x}\right) I^{-x}\right) - N_1^x q^x - \lambda I^x\end{aligned}$$

If interior, i.e., $I^x \in (0,1)$, optimal investment solves the first-order condition $\frac{\partial \pi_1^x}{\partial I^x} = 0$.

Recall the price p_1^x from (F.41) so that

$$p_1^x - I^x q^x = \kappa - 2I^x\left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right) = \kappa - 2I^x\left(\frac{\partial \bar{\pi}_2^x}{\partial \hat{D}^x}\right),$$

where the last equality uses (F.39). Next, note that $N_1^x = N_1^{-x} = 1/2$, $I^x = I^{-x}$ as well as $\frac{\partial \pi_2^x}{\partial D^x} = \frac{\phi^x}{3} = -\frac{\partial \pi_2^x}{\partial D^{-x}}$, and $\frac{\partial N_1^x}{\partial I^x} = -\frac{\partial N_1^{-x}}{\partial I^x}$. Using these relations, we obtain after simplifications:

$$\begin{aligned}\frac{\partial \pi_1^x}{\partial I^x} &= \frac{\partial N_1^x}{\partial I^x}\left(\kappa - 2I^x\left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right)\right) - \frac{q^x}{2} \\ &\quad + \left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right)\left(\frac{1}{2} + \frac{\partial N_1^x}{\partial I^x} I^x\right) + \left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right)\left(\frac{\partial N_1^x}{\partial I^x} I^x\right) - \lambda I^x \\ &= \kappa\left(\frac{\partial N_1^x}{\partial I^x}\right) - \frac{q^x}{2} + \frac{1}{2}\left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right) - \lambda I^x\end{aligned}$$

Inserting

$$\frac{\partial N_1^x}{\partial I^x} = \frac{(q^x - 1.5c)}{2\kappa}$$

into above expression for π_1^x, we obtain:

$$\frac{\partial \pi_1^x}{\partial I^x} = \frac{1}{2}\left(\left(\frac{(\phi^x)^2}{36\kappa} + \frac{\phi^x}{3}\right) - 1.5 - 2\lambda I^x\right).$$

As a result, equilibrium investment/effort satisfies — if it is interior and solves the first-order condition $\frac{\partial \pi_1^x}{\partial I^x} = 0$ —

$$I^x = \frac{\phi^x\left(1 + \frac{\phi^x}{12\kappa}\right) - 4.5c}{6\lambda}.$$

That is, optimal investment reads

$$I^x = \min\left\{1, \left[\frac{\phi^x\left(1 + \frac{\phi^x}{12\kappa}\right) - 4.5c}{6\lambda}\right]^+\right\}.$$

Propositions 2 and 4 readily imply that, with a market for data, the investment I^x above is larger than under the baseline and under data sharing (for $\eta > 0$) when c is sufficiently small or ϕ^x is large. Under these circumstances, the service price p_1^x from (F.41) is strictly lower than in the baseline (see Proposition 2) or with data sharing (see Proposition 4), holding q^x fixed in the comparison.

F.3 Commitment Solutions

Commitment to data sharing. Suppose the platform can commit to a future data sharing policy, $\eta^x \in [0,1]$ in period $t = 1$. Formally, in $t = 2$, platforms maximize (as before) $\max_{p_2^x} \pi_2^x$

whereby $D^x = N_1^x \theta^x I^x + \eta^{-x} N_1^{-x} \theta^{-x} I^{-x}$. In $t = 1$, platforms choose simultaneously η^x, p_1^x, I^x, q^x to maximize π_1^x, i.e., they solve $\max_{\eta^x, p_1^x, I^x, q^x} \pi_1^x$. We now show that, when $c^x = c(1 + \eta^x) \geq 0$, then $\eta^x = 0$ holds in equilibrium. For this sake, we consider without loss of generality $q^x = c^x$ so that $\theta^x = 1$, and calculate

$$\frac{\partial \pi_1^x}{\partial \eta^x} = -N_1^x I^x + \frac{\partial \pi_2^x}{\partial D_2^{-x}} N_1^x I^x,$$

which is (strictly) negative (when $I^x > 0$), because the expression for period-2 payoff π_2^x (under equilibrium pricing) in Lemma 1 readily imply $\frac{\partial \pi_2^x}{\partial D_2^{-x}} < 0$. Thus, platforms optimally choose $\eta^x = 0$.

Commitment to product/service quantities. Suppose that the platforms can commit to a minimum future quantity at the beginning. Then compared to the baseline, (i) both the platforms generates higher payoff; (ii) the platform that makes commitment take a larger market share; (iii) the platform that does not commit determines a higher price. We prove these next:

Proposition 10. *In equilibrium, one and only one platform commits to quantity, i.e., the solution involves $\{C, NC\}$ or $\{NC, C\}$. Without loss of generality, suppose platform A chooses 'C' and B chooses 'NC', then*

$$\begin{aligned} N_2^A &= \hat{z}_2 = \frac{5}{8} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{8\kappa}, \\ p_2^A &= \frac{5\kappa}{4} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{4}, \qquad p_2^B = \frac{3\kappa}{2} - \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{2}, \\ \pi_2^A &= \frac{(5\kappa + \Delta_K + \phi^A D^A - \phi^B D^B)^2}{32\kappa}, \qquad \pi_2^B = \frac{(3\kappa - \Delta_K - \phi^A D^A + \phi^B D^B)^2}{16\kappa}. \end{aligned} \tag{F.42}$$

Proof. We first prove that when platform $-x$ chooses 'NC', i.e. not to make a commitment, then the best response for platform x is to commit.

Without loss of generality, suppose platform B chooses 'NC'. If platform A also chooses 'NC', then the case is the same as the baseline, i.e. $p_2^A = \kappa + \frac{\Delta_K}{3} + \frac{D^A\phi^A - D^B\phi^B}{3}$, $N_2^A = \hat{z}_2 = \frac{1}{2} + \frac{\Delta_K + D^A\phi^A - D^B\phi^B}{6\kappa}$. Then consider the case that platform A chooses 'C'. Given andy possible p_2^B, platform A need to decide N_2^A in period $t = 1$ to maximize π_2^A. To fullfill the commitment in period $t = 2$, platform A's product pricing, p_2^A, is restricted to satisfy

$$\begin{aligned} \hat{z}_2 &= \frac{1}{2} + \frac{\Delta_K - p_2^A + p_2^B + D^A\phi^A - D^B\phi^B}{2\kappa}, \\ \Rightarrow p_2^A &= \Delta_K + p_2^B + D^A\phi^A - D^B\phi^B + \kappa(1 - 2N_2^A). \end{aligned} \tag{F.43}$$

Then platform A's optimization problem is

$$\max_{N_2^A} \pi_2^A = p_2^A N_2^A = [\Delta_K + p_2^B + D^A\phi^A - D^B\phi^B + \kappa(1 - 2N_2^A)]N_2^A.$$

Thus we obtain that the optimal response is

$$N_2^{A^*} = N_2^{A^*}(p_2^B) = \frac{1}{4} + \frac{\Delta_K + p_2^B + D^A\phi^A - D^B\phi^B}{4\kappa}. \tag{F.44}$$

Consider platform B. In period $t = 2$, given platform A's commitment $N_2^{A^*}$, then $N_2^B = 1 - N_2^{A^*}$ is also given. Platform B foresees that platform A will make quality commitments based on p_2^B.[54]

[54] Note that the commitment only limits the minimum level of service on platform A. Thus platform B does not have unlimited access to higher prices.

Therefore, platform B faces the following problem:

$$\max_{p_2^B} \pi_2^B = p_2^B(1 - N_2^{A*}(p_2^B)) = p_2^B\left(\frac{3}{4} - \frac{\Delta_K + p_2^B + D^A\phi^A - D^B\phi^B}{4\kappa}\right).$$

We then obtain that the optimal pricing for platform B is

$$p_2^B = \frac{3\kappa}{2} - \frac{\Delta_K}{2} - \frac{\phi^A D^A - \phi^B D^B}{2}. \tag{F.45}$$

Plug (F.45) into (F.43) and (F.44), we have

$$\begin{aligned}
N_2^A &= \hat{z}_2 = \frac{5}{8} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{8\kappa},\\
p_2^A &= \frac{5\kappa}{4} + \frac{\Delta_K + \phi^A D^A - \phi^B D^B}{4},\\
\pi_2^A &= \frac{(5\kappa + \Delta_K + \phi^A D^A - \phi^B D^B)^2}{32\kappa} > \frac{(3\kappa + \Delta_K + \phi^A D^A - \phi^B D^B)^2}{18\kappa} = \pi_{2\ baseline}^A,\\
\pi_2^B &= \frac{(3\kappa - \Delta_K - \phi^A D^A + \phi^B D^B)^2}{16\kappa} > \frac{(3\kappa - \Delta_K - \phi^A D^A + \phi^B D^B)^2}{18\kappa} = \pi_{2\ baseline}^B.
\end{aligned} \tag{F.46}$$

Therefore, the best response for platform A is to make a quantity commitment, i.e. the best response to 'NC' is 'C'. Interestingly, both platforms increase payoff, implying that the best response to 'C' is 'NC'. Moreover, under the sufficient condition that $3\kappa > \max\{\Delta_K + \phi^A - \phi^B, -\Delta_K - \phi^A + \phi^B\}$, N_2^A is larger than the baseline equilibrium N_2^A, and p_2^B is larger than the baseline equilibrium. □

F.4 Institutional Background on Privacy Protection Policy and Open Data Initiatives

Privacy protection policies such as GDPR strengthen individual ownership rights over personal data by granting rights to access, correct, and delete personal data held by firms. While GDPR is a sweeping initiative implemented by the European Union, the U.S. system is piecemeal and multi-layered, with regional initiatives such as CCPA. Generally speaking, firms must minimize personal data processing and can only process personal data under limited and specific circumstances. One such circumstance is an individual's explicit opt-in consent.[55]

While policies such as GDPR have focused on data ownership rights, open data initiatives, e.g., in the form of data sharing initiatives, have emphasized data access. Private data ownership rights should not be confused with access. Privately owned data can allow open access while non-proprietary assets can be de facto closed for access (Merges, 2008). The question of whether or not data should be openly accessible (purely open access to data) is a debated issue in academia (e.g.,

[55] Federal laws on privacy protection tend to be industry and region specific in the United States, with the Department of Health Human Resources (DHHS) enforcing the Health Insurance Portability and Accountability Act of 1996 (HIPPA) in healthcare, the Federal Communication Commission (FCC) regulating telecommunication services, the federal reserve systems monitoring the financial sector through Gramm-Leach-Bliley Act (GLBA), the Security and Exchange Commission (SEC) focusing on public firms and financial exchanges, and the Department of Homeland Security (DHS) dealing with terrorism and cybercrimes related to national security. The Federal Trade Commission (FTC) can address privacy violations and inadequate data security as deceptive and unfair practice, following the 1914 FTC Act whereas the U.S. Constitution, in particular the First and Fourth Amendments, together with the Electronic Communications Privacy Act of 1986 (ECPA), the Stored Communications Act (1986), the Pen Register Act (1986), and the 2001 USA Patriot Act – stipulate when and how the government can collect and process electronic information of individuals. But in practice, it is still case by case. The debate on whether the United States should follow European-style regulation is still ongoing. See "Ad world flocks to Congress urging federal data privacy legislation", The Drum, 26 February 2019, and "Should Congress override state privacy rules? Not so fast," The Washington Post, February 26, 2019.

Dewald, Thursby, and Anderson, 1986) and policy (Commission et al., 2017). Advocates of open access to data argue that it facilitates subsequent research, including replication of existing works, and increases the diffusion of knowledge thereby enhancing the efficiency of the research system (Piwowar, Day, and Fridsma, 2007; Glenn and Ellis Lee, 2012; Piwowar and Vision, 2013). Empirical studies also document benefits of open data (Jetzek, Avital, and Bjorn-Andersen, 2012; Martens et al., 2020).[56] Open data initiatives take various forms. Centralized data commons suffer from privacy issues (Milles, 2019). Real world example includes UPI in India. The Indian authorities have the concept of a data fiduciary, which would be a body that would basically manage the data of individuals and help give them effective control over consent management. South Korean MyData and Brazil's open banking systems use similar setups. The New York Times conducted full-scale investigation in 2018 concerning Facebook (now Meta) forming ongoing partnerships with other firms, including Netflix, Apple, and Microsoft, and granting these companies access to different aspects of consumer data. See full news coverage at https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html. Blockchains and secure-MPC through ZKP etc., constitute an interesting route to explore. Overall, the challenge for data sharing is not only about the technology that enables privacy protection, but also about economic incentives.[57]

Recently, there have been many attempts to promote open data access, including Open Banking and Open Finance initiatives (He et al., 2022; Goldstein et al., 2022). For example, the International Data Spaces Association constitutes a private investment for secure data sharing (Richter and Slowinski, 2019). The EU proposed the Digital Market Act (DMA) in 2020, explicitly emphasizing data sharing for a fair competition.[58] China and South Korea have built open platforms for data sharing to aggregate scattered, isolated, and varied data to help integrate technology and business data to lower information barriers.[59]

Data companies such as Acxiom and Datalogix gather and sell personal information. Policy discussions often suggest that requiring large digital platforms to share data with smaller ones breaks their dominance and leads to more competition among platforms, which is beneficial for users. Such rationale, for instance, underlies the open banking regulation where banks have to share data with FinTech companies (e.g., lenders).

[56] Martens et al. (2020) empirically examine the effectiveness of mandatory data sharing. They find that user welfare is not maximized due to increases in product price, which corroborates our model prediction.

[57] Federated learning and privacy-preserving data sharing infrastrutures build on blockchains may enable decentralized sharing of data (e.g., Sockin and Xiong, 2022). Ocean Protocol, a nonprofit platform developed by a Singapore-based foundation, is a salient example of data marketplaces in which companies consumers, and other parties share or trade data.

[58] See the Digital Markets Act by the European Commission.

[59] See here.