“For the public benefit”: who should control our data?*

Sarit Markovich[†] and Yaron Yehezkel[‡]

January 2023

Abstract

We consider the public-good aspect of platform’s data-collection on users. Data has commercial-benefit to the platform, personal-benefit to the user, and public-benefit to other users. We ask who should decide which data the platform commercializes. We find that the answer depends on the type of heterogeneity in the disutility from data commercialization. When heterogeneity is across users (data-items) and the public-benefit of data is high (low), it is welfare-enhancing to let the platform (users) control the data. Furthermore, dynamic data accumulation strengthens our results.

JEL Classification: L1

Keywords: data regulation, network externalities, platform competition, public good

1 Introduction

Many platforms base their business model on the commercialization of consumers’ data. For example, search engines such as Google can collect data on users’ locations and keyword search. Navigation apps such as Waze can collect data on users’ preferred routes and other driving habits. Media streaming platforms such as Spotify, Pandora, and Deezer can collect data on users’ music preferences and listening habits. Wearables such as Fitbit, Garmin, and Samsung Watch can collect data on users’ sport activities and performances. These

[*] For helpful comments and discussions we thank Ido Eisdorfer, Soo Jin Kim, Michele Polo, Patrick Rey, Michael Riorden, Alex Smolin, Zhijun Chen, and seminar participants at the 2021 IIOC, the Israeli Competition Authority, the 2022 MaCCI conference in Mannheim, the CEPR 2022 meeting in Berlin, the CRESSE 2022 conference in Crete, the Platform Strategy Research Symposium 2022, the EARIE 2022 conference in Vienna and the NUS Economics of Platforms Workshop 2022 in Singapore. For financial support, we thank the Coller Foundation, the NET Institute–www.NETinst.org, and the Henry Crown Institute.

[†] Kellogg School of Management, Northwestern University (s-markovich@kellogg.northwestern.edu)

[‡] Coller School of Management, Tel Aviv University (yehezkel@tauex.tau.ac.il)

platforms can then use the data to improve their services, but at the same time, the data can also be used for commercial purposes such as selling it to advertisers or to other third-party providers. This raises the question of who should own the property rights over users' data? On the one hand, the platform is the party that collects and analyzes the data, and users give their consent to data collection when joining the platform. On the other hand, users are the party that generates the data, and in many cases, bear a disutility from having their data shared. Furthermore, users typically do not have the choice to join the platform without agreeing to give away the rights over their own data.

To study this question, we develop a model with the following features. First, data has three potential benefits: *(i)* Data provides personal benefits. For example, when a driver uses a navigation app and agrees to let the app track their route, the data collected can help direct the driver to un-congested routes. *(ii)* The same data provides the platform with commercial benefit. The navigation app, in our example, can sell the driver's data to advertisers. *(iii)* Data provides a public benefit. For example, data collected from a driver can benefit other drivers that consider taking the same route. Other relevant examples are users that provide their location data on a contact-tracing app benefit others who now know they were in proximity of someone who tested positive for COVID-19;[1] or Fitbit's use of its heart rate data to identify episodes of irregular heart rhythm suggestive of atrial fibrillation (AFib), the most common form of heart rhythm irregularity. Fitbit intends to use this information to alert users about an irregular heart rhythm so that notified individual would connect with a doctor. This third public benefit of data is the most important one for innovation and product improvement, as it implies that data creates positive externalities where users can benefit from other users' data, regardless of whether they share data themselves.

The second main feature of our model is that the platform collects multiple data items. For example, Waze collects data on location, time, and route that users take; Fitbit collects data on steps and heart rate; and Facebook collects data on text and photos users upload as well as posts they read, the people and groups they follow, etc.

The third feature is that users have disutility from having their data shared for commercial benefits. This disutility may differ across users. For example, some users are more sensitive to their privacy than others. Moreover, this disutility may differ across data items. For example, users may not care about Waze sharing information about the route they take but suffer disutility from Waze sharing their exact location at a specific point in time. Similarly, users' disutility from Fitbit sharing one's number of daily steps may be lower than that of sharing their heart rate.

[1] Contact tracing apps use one's phone, or other mobile device, to track and alert individual if they'd crossed paths with someone who within a certain window of time tested positive to COVID-19.

To study who should have the right over users' data, we study two extreme data regimes. In the first regime, the platform has the right to decide which data items to collect and commercialize. Users can only decide whether to join the platform (and agree to its data policy), or stay out. The second regime does not allow the platform to contingent users' participation in the platform on their consent to collect their data.

We find that the different benefits of data create market inefficiencies. The platform only cares about the commercial benefit, and will thus collect data as to maximize this benefit, subject to the constraint that users agree to join it. Users only care about their own private benefit. If given the opportunity to decide which data to provide the platform, users would only provide data that offers them private benefit, as they enjoy the public benefit regardless of their data contribution. Most ill-considered, however, is the the public benefit of data. Although it provides benefits to all on the platform, the public benefit is, at least partially, ignored by both the platform and the users. That is, both parties ignore that while data collected on an individual user may create a disutility for this user, it may benefit the platform's entire user-base. This market inefficiency raises the question of which regime achieves the best balance between the benefits of data (public, personal, and commercial) and disutility to users, as well as whether competition can mitigate these market inefficiencies. We find that giving users full control over their data is not always welfare enhancing, as it may result in too little data collected for the public benefit.

In general, the platform's optimal strategy can take one of three possible outcomes: all data is commercialized but not all users join (i.e., full data coverage but partial user coverage); all users join but not all data is commercialized (full user coverage and partial data coverage); or partial user and data coverage. As it turns out, our results and intuition crucially depend on whether the market is mostly characterized by data coverage or user coverage, which further depend on whether the market is mostly characterized by users with different disutility from the commercialization of their data (hereafter, *heterogeneous users*), or by data items that differ in the disutility that commercializing them inflicts on users (hereafter, *heterogeneous data*).

Consider first the case of heterogeneous users. In this case, we find that the first two regimes are identical when data does not have any public benefit. However, when the public benefit of data is high, it is welfare enhancing to let the platform control the data as otherwise users will share too little data for the public benefit. In contrast, when the public benefit of data is low, giving the platform the control over users' data results in under-participation in the platform and in less data collected for public benefit. In this case, it is welfare enhancing to give the users control on their data. These results highlight the important role the public benefit of data plays when evaluating data regulation.

We find that the opposite conclusion emerges in the case of heterogeneous data items. Then, it is welfare enhancing to let the platform control the data when the public benefit of data is low, while giving the users control on their data is welfare enhancing only when the public benefit of data is high. In this latter case, the platform may "bundle" data items, forcing users to agree that the entire "bundle" of data items is commercialized, or they stay out of the platform.

We find that dynamic accumulation of data intensifies the results of the static game. When the market is characterized by heterogeneous users, it is welfare enhancing to give the platform control over data for a wider range of values of the public benefit of data. In this case when the platform controls the data, the platform serves more and more users over time, all of whom share their data for public benefit which the platform further accumulates. Likewise, when the market is characterized by heterogeneous data, it is welfare enhancing to give users control over data for a wider range of values of the public benefit. In this case, when the platform controls the data, the platform commercializes more and more data items over time, resulting in over-commercialization of data which hurts users and decreases welfare.

Understanding the effects of platforms' data policies on profits and social welfare has important implications for the ongoing debate on the need for data regulation. As Economides and Lianos (2020) point out, existing US laws give the property right over data to the entity that collects it. Platforms can collect and own users' data on the basis of users' consent to join the platform.[2] Yet, when platforms have strong market power, users' voluntary consent to the platform's data policy is controversial. For example, in 2020, the US Department of Justice filed a suit against Google, claiming (among other things) that "American consumers are forced to accept Google's privacy practices, and use of personal data...".[3] Another case in point is Facebook's questionable announcement in 2021, that its users must agree to let Facebook and its subsidiaries collect their personal data on WhatsApp, including phone numbers and locations.[4] If users don't accept the new terms and conditions, they will be forced out of the app.[5] This is especially interesting given that WhatsApp has always positioned itself as a privacy focused service – encrypting all users' messages. Indeed, WhatsApp potentially

[2] See Economides and Lianos (2020), p 4-5.

[3] See The Verge, Oct 20, 2020. Available at: https://www.theverge.com/2020/10/20/21454192/google-monopoly-antitrust-case-lawsuit-filed-us-doj-department-of-justice

[4] In an extension to competing platform, preliminary results show that platforms may choose different data policies. The platform that benefits from a leading position in the market chooses to control the data while the new platform enables users that join it to control their data.

[5] See, for example, The Verge, Feb 22, 2021. Available at: https://www.theverge.com/2021/2/22/22294919/whatsapp-privacy-policy-may-15th-messaging-calls-limited-functionality

has access to many different data items – phone number, contact lists, messages content. Its intention to keep encrypting messages and not sharing this data while sharing other data items, like phone number and location, suggests that WhatsApp believes that users' disutility from sharing phone number information with Facebook is lower than their disutility from sharing messages content.[6]

In contrast to the US, the EU General Data Protection Regulation (GDPR) is designed to provide users with the choice to give data; a choice that does not discriminate those that choose not to provide data. In our model, the GDPR aims to move platforms from a regime that provides the platform with full control over users' data, to a regime that enables users to join a platform without giving their consent to share specific data.

Our results suggest that whether the EU's firmer approach to data regulation as compared to the US enhances welfare, depends on the magnitude of the public benefit of data and the type of heterogeneity in the market. More generally, our paper provides specific conclusions on how to regulate dominant data-driven platforms. When data have significant public benefits and the market is characterized by heterogeneous users, such that users that are relatively sensitive to privacy prefer to stay out, the regulator should not intervene in the platform's data policy. In this case, regulation will result in fewer users giving data for public benefit and may eventually reduce consumer surplus as well as social welfare. When the market is characterized by homogeneous users and is almost fully covered, regulation that requires the dominant platform to give users control over data can enhance social welfare. [7]

We should emphasize that the question of who should control our data is also – perhaps foremost – an ethical question of social morality. Is it ethical to allow a platform to share our personal data items as it wishes? The moral aspects of this question are important but are beyond the scope of our theoretical model. The goal of our paper is to contribute to the debate on data regulation by highlighting some economic forces, specifically, with regards to the public benefit of data. Our results and potential policy implications cannot be placed in isolation from a discussion on the moral aspects of privacy and data protection.[8]

[6] See The Verge, Oct 20, 2020. Available at: https://www.theverge.com/2020/10/20/21454192/google-monopoly-antitrust-case-lawsuit-filed-us-doj-department-of-justice

[7] We focus on platforms that do not have high fixed entry costs into a new market. Naturally, a new platform that needs to cover its fixed entry costs requires sufficient initial profits. Hence, regulating the data policy of such new platforms may deter entry. Another argument against regulating a new platform is that as we show below, an entrant platform may independently choose to give users control over data in order to gain a foothold in the market, if the incumbent does not do so.

[8] In a somewhat related moral debate in Israel, the question is whether to allow public authorities share information concerning the identity of civilians that did not receive the COVID vaccine. Such data may have valuable public benefit in fighting COVID, yet may violate civilians' privacy rights.

Literature Review

This paper combines the literature on privacy and data collection with the literature on platforms. Starting with the literature on privacy, Acquisti et al. (2016) surveys the economic literature on privacy, focusing on the economic value and consequences of protecting and disclosing personal information, and on consumers' understanding and decisions regarding the trade-offs associated with the privacy and the sharing of personal data. O'Brien and Smith (2014) study a model where sellers can commit to privacy policies and consumers have heterogeneous – negative or positive – preferences over privacy. They find that under perfect competition, firms make the socially optimal decision. Furthermore, a positive and sufficiently large correlation between consumers' valuations for the product and privacy is a necessary condition for the under-supply of privacy by firms. Choi et al. (2019) study a model of privacy with negative information externalities where data shared by one user may allow the platform to know more about users that do not share data. They find that the market exhibits excessive data collection. Dosis and Sand-Zantman (2020) consider the effects of property rights of data collected by a monopolistic platform when users have private information about their utility from the platform's service. The platform offers a menu of contracts to screen between users with different valuations. The paper studies how asymmetric information affects the optimal policy of whether to give the platform or users the right over data. Focusing on the improved match between advertisers and consumers data can facilitate, Loertscher and Marx (2020) show that consumer harm arises only by the combination of improved match values due to privacy reduction and more aggressive pricing by the monopoly. For a fixed price, the consumer always benefits from the improved matches that come with a reduction in privacy. Based on this, the authors conclude that competition policy should aim at protecting consumers' information rents rather than their privacy. Jullien, Lefouili, and Riordan (2020) assume a two-stage game where a website monetizes information it collects on its users. Users are unsure about whether the commercialization of their data will increase/decrease/have no effect on their experience. User retention motivates the website to be cautious about its privacy policy—the probability that a user's information is sold in the first period. The authors find that a policy that requires a website to commit ex-post to disclosure leads to less precaution by website. Fainmesser et al. (2020) study how firms' revenue model affect their data policy. Looking at whether a firm's revenues are mostly data-driven or usage-driven—i.e., their main source of revenue stems from selling information to third-parties or from charging users subscription fees—they find that purely usage-driven firms select the socially optimal data policy. All other firms, over-collect users information. The authors then show that this inefficiency in data collection can be corrected with taxes or fines imposed on the firms. Similar to our analysis, Economides and Lianos (2020) emphasize market failure effects of various data policies. As in our regimes 1 and 2 below, the authors examine several different data regimes and find that the requirement to share data in exchange to access to the platform benefits the platform yet decreases consumer surplus. They further find that under a regime that is similar to our regime 2 but where the platform can pay users for data, the price of data would be positive and users would be better off. Ichihashi and Smolin (2022) consider a seller that can request data from a buyer, in the form of an imperfect signal to the buyer's valuation for the seller's product. They find that when the seller has imperfect private information about the product's value, the seller either does not ask the buyer for data, asks for full data collection, or asks for an imperfect signal. Bergemann, Bonatti and Gan (forthcoming) consider a data intermediary that collects data from consumers who are partially informed about their preferences. A consumer's data can predict the preferences of other consumers but can be resold to a price discriminating producer. Hence, it is socially optimal to collect all data and share it with consumers but not with the producer. The paper finds that when the intermediary can preserve the consumers' anonymity, anonymized data is more profitable for the intermediator than complete data if and only if anonymization increases welfare. In a closely related paper, Chen (2022) considers a data-driven platform and users that are heterogeneous in their disutility from having their data collected. The platform can invest in data analytics that improves the users' private benefit from data. The paper finds that when the platform controls the collection of data, the platforms collects too much data which creates a market failure. Giving users control over data enhances consumer surplus but hurts the platform and result in a reduction in its investment in data analytics.

Our paper makes three main contributions to the above literature. First, we introduce and study the role of the public benefit of data, where users benefit from data collected from other users. The paper finds that the comparison between data regimes heavily depend on the degree of the public benefit of data. In particular, to the best of our knowledge we are the first to show that under heterogeneous users (data), it is welfare enhancing to give the platform (users) control over data when the public benefit of data is high, while the opposite case occurs when the public benefit of data is low.

The second main contribution of our paper is in distinguishing between data collection and data commercialization. We assume that users bear a disutility only when their data is commercialized and not when it is collected for the private and public benefit. This distinction enables us to study the case when the platform chooses not to commercialize all the data that it collects, as assumed in previous literature.

The third main contribution of our paper is the consideration of a set of distinct data items. When the platform has the right to collect and commercialize all data items from users that join it, the platform in our model can "bundle" different data items. That is, users agree to commercialize data items with a disutility that exceeds their private benefits, because users have to give their consent to the platform's data policy as a whole, and cannot agree to commercialize some data items but not others. As our model reveals, this feature plays an important role in the comparison between the different data regimes.

Our paper also contributes to the literature on platforms with network externalities and coordination, when users would like to join the same platform other users join. Katz and Shapiro (1986), Caillaud and Jullien (2001; 2003), Jullien (2011), Hałaburda and Yehezkel (2013; 2016; 2019) and Markovich and Yehezkel (2022) consider platform competition and coordination in the context of a static game. Hagiu (2006) considers sequential competition on two sides of a market. Hałaburda et al. (2020) and Biglaiser and Crémer (forthcoming) considers dynamic competition. These papers do not consider data policy. While the public benefit of data exhibits some externalities that are similar to network effects, the two are not identical. In the case of network effects, users benefit from the presence of other users in the same platform, regardless of the platform's data policy. In contrast, when externalities are data-driven, as in our model, the benefit users derive from other users depends on behavior driven by data regulation (either because the data regulation enables the platform to collect the data, or because users are willing to share it with the platform). To evaluate the effect of data regulation on welfare, our model distinguishes between the effect of data regulation on users' participation in the platform, and the effect on the amount of data the platform collects from these users.

2 The Model

Consider a market with a monopolistic/competitive platform/s and a set of potentially heterogeneous users. We describe the model by first defining a general framework of the users' preferences, which allows users' disutility from having their data commercialized to vary across users and across different types of data. Then, we describe the platform's potential strategies. Finally, we define three special cases that our paper focuses on: heterogeneous users, heterogeneous data, and both heterogeneities, and impose some simplifying assumptions.

Users' preferences

Consider a continuum of small users, potentially heterogeneous, with a total mass of one. There is a continuum of *data items* that the platform can potentially collect from each user.

If collected, a data item may provide users with a certain benefit, and if commercialized, can provide value for the platform at a disutility to the user. For example, in the context of a fitness tracker such as Fitbit, data items can be the user's location, number of steps, heart rate, and so on.

Data provide benefits to users. First, a user enjoys a *private benefit*, that we denote by p, when sharing data items with the platform. For example, if users share data on their number of steps and heart rate with a fitness tracker, the platform can help these users to monitor their training and provide them with recommendations concerning healthier training. Second, the data may also benefit all other users that join the platform, regardless of whether they share their data. For example, the data collected by a fitness tracker from an individual user can help the tracker to provide better training recommendations to all other users. We refer to this as the *public benefit* of data collected on an individual user and benefiting all other users and denote it by γ.[9]

The platform can commercialize data. Any data item θ collected from user ε and commercialized results in a disutility to the user of $k_{\varepsilon\theta}$. Users may feel discomfort when their personal data, such as their heart rate, are shared with other commercial firms. Moreover, advertisers may overload users with advertisements and pop-ups. Selling data to advertisers may provide users some positive benefits, for example, users may prefer targeted advertisements over generic ads. Still, we assume that the users' discomfort from the lack of privacy and excessive advertising outweighs any potential benefits, such that users obtain a net disutility from having their data commercialized. Users' disutility from commercializing their data vary across users and across data items. Some users may be more sensitive to their privacy than others. Moreover, users may bear different disutility from the commercialization of different data items.[10] For example, a user of a fitness tracker may incur a higher disutility when their heart rate is commercialized than if their number of steps is shared. To incorporate both types and variations in users' disutility, suppose that $k_{\varepsilon\theta} = \theta + \varepsilon$, where θ represents the common disutility from sharing data item θ among all users and ε is a user idiosyncratic disutility and captures the heterogeneity across users.[11] Suppose that $\varepsilon \sim [0, 1]$ according to a distribution function $f(\varepsilon)$ and $\theta \sim [0, 1]$ according to $g(\theta)$, with cumulative distribution functions $F(\varepsilon)$ and $G(\theta)$, respectively. The platform may commercialize all or part of the data users share. We assume that users bear the disutility $k_{\varepsilon\theta} = \theta + \varepsilon$ only for

[9] In the Online Appendix, we consider the case of heterogeneous private and public benefits, under the assumption that the market is fully covered.

[10] For example, Goldfarb and Tucker (2012) show that privacy costs vary with age. Acemoglu et. al (2022) consider users that vary in their value for privacy for pecuniary and nonpecuniary motives, including political and social reasons for privacy.

[11] The results follow to the case where $k_{\varepsilon\theta} = \theta\varepsilon$.

their data that is commercialized.

Assume a case where the platform controls the data, where all users give the same amount of data, which is announced and set by the platform. In this case, taking these benefits and costs together, we have that the utility of a user with an idiosyncratic disutility ε when other users with $\varepsilon \in [0, \widetilde{\varepsilon}]$ join the platform (hence, a mass of $F(\widetilde{\varepsilon})$), each of which gives data items with $\theta \in [0, \overline{\theta}]$ for public and private benefit (a mass of $G(\overline{\theta})$ per-user), and $\theta \in [0, \widetilde{\theta}]$ data items are commercialized (where $\widetilde{\theta} \leq \overline{\theta}$), is:

$$U(\varepsilon|\widetilde{\varepsilon}, \overline{\theta}, \widetilde{\theta}) = \gamma F(\widetilde{\varepsilon})G(\overline{\theta}) + pG(\overline{\theta}) - \int_0^{\widetilde{\theta}} (\varepsilon + \theta)g(\theta)d\theta, \tag{1}$$

where the first term is the aggregated public benefit, the second term is the private benefit, and the last term is the total disutility from data commercialization. The parameters γ and p measure the magnitude of the public and private benefits, respectively. We can modify the utility function to account for the possibility that not all users give the same amount of data, as we describe later on in the case where users control the data. [12]
As it turned out, our results and intuition crucially depend on whether the market is mostly characterized by heterogeneous users or by heterogeneous data. To disentangle the two types of heterogeneities and deliver the intuition in a clear and tractable manner, it is useful to study the first two cases in isolation and then combine them together. We therefore focus on three special cases:

Case A: Heterogeneous users. In this case, the driving force is partial coverage of users ,while all data items are collected. Users' idiosyncratic disutility from commercializing their data, ε, is distributed between $[0, 1]$ according to $f(\varepsilon)$. There is no heterogeneity in data items and for simplicity suppose that there is one indivisible data item with $\theta = 0$. A user of type ε utility from joining the platform, when users with $\varepsilon \in [0, \widetilde{\varepsilon}]$ join the platform and share data, which is commercialized, is reduced to $U(\varepsilon|\widetilde{\varepsilon}) = \gamma F(\widetilde{\varepsilon}) + p - \varepsilon$.

Case B: Heterogeneous data. In this case, the driving force is partial coverage of data items, while all users join the platform. Here, the disutility from commercializing a data item, θ, is distributed between $[0, 1]$ according to $g(\theta)$. There is no heterogeneity in users' idiosyncratic disutility, and for simplicity we normalize ε to 0. Because all users are identical, they either all join or all stay out, as if there is one representative user. When all users join and share data items with $\theta \in [0, \overline{\theta}]$ for public and private benefits, and data with $\theta \in [0, \widetilde{\theta}]$ for commercial

[12] It is possible to assume that the public and private benefits are general and increasing functions $\Gamma(F(\widetilde{\varepsilon}) \times G(\widetilde{\theta}))$ and $P(G(\widetilde{\theta}))$. Yet, as we consider general cumulative distribution functions $F(\varepsilon)$ and $G(\theta)$, we can assume for simplicity that $\Gamma(F(\widetilde{\varepsilon}) \times G(\widetilde{\theta})) = \gamma F(\widetilde{\varepsilon})G(\widetilde{\theta})$ and $P(G(\widetilde{\theta})) = pG(\widetilde{\theta})$.

benefit, each user's utility is reduced from (1) to:

$$U(\overline{\theta}, \widetilde{\theta}) = \gamma G(\overline{\theta}) + pG(\overline{\theta}) - \int_0^{\widetilde{\theta}} \theta g(\theta) d\theta. \tag{2}$$

Case C: Heterogeneous users and data. Here, the driving force, is partial coverage of both users and data items. The market is equally characterized by both heterogeneous users and data. In this case we assume that both θ and ε are uniformly distributed between $[0, 1]$, respectively. For brevity, we solve case C in the appendix. We show that case C yields the combination of the results obtained in cases A and B.

Platform's strategy

Consider a monopolistic platform. Data has commercial benefit to the platform, when the platform sells some of the data to advertisers, third-party application developers, or to other platforms. When users with $\varepsilon \in [0, \widetilde{\varepsilon}]$ join the platform (a mass of $F(\widetilde{\varepsilon})$) and $\theta \in [0, \widetilde{\theta}]$ data from each user is commercialized (a mass of $G(\widetilde{\theta})$), the platform's commercial benefit, or profits, is $\pi(\widetilde{\varepsilon}, \widetilde{\theta}) = \alpha F(\widetilde{\varepsilon}) G(\widetilde{\theta})$. The parameter α measures the commercial value from a data item of a specific user and the platform's profits are a function of the number of users that join the platform and the number of data items commercialized.[13]

The timing of the game is as follows. In the first stage, the platform (or platforms) sets its data policy: which data items the platform collects and which it commercializes; subject to the data regulation regime described below. Also, the platform sets prices, if applicable. Then, users decide whether they accept the platform's data policy and join the platform or stay out, in which case they get a reservation utility that we normalize to 0. Then, if the regime enables joining users to choose which data to share, users do so. Finally, the platform commercializes the relevant data.

In regulating the platform's control over data, we study three data regulation regimes imposed by the regulator:[14]

Regime 1: the platform controls the data. The platform can contingent platform participation with data collection and commercialization. For all users that choose to join the platform, the platform decides which data items to collect and commercialize. The platform may choose

[13] It is possible to assume that the commercial benefit of data is a general and increasing functions $A(F(\widetilde{\varepsilon})G(\widetilde{\theta}))$. Yet, as we consider general cumulative distribution functions $F(\varepsilon)$ and $G(\theta)$, we can assume for simplicity that $A(F(\widetilde{\varepsilon})G(\widetilde{\theta})) = \alpha F(\widetilde{\varepsilon})G(\widetilde{\theta})$.

[14] Since digital platforms, typically, provide their service for free and do not compensate users for their data, we only consider regimes with zero or negative price.

not to commercialize all data items, in order to attract more users to join. For any data item, the platform informs users whether it plans to collect this item, and if so, whether it plans to commercialize it. The platform's data policy is publicly observable and the platform is committed to it. Upon joining the platform, users give their consent to this data policy as a whole. Users can reject the data policy, stay out and earn the utility of 0.

Regime 2: users control their data. The platform *cannot* contingent platform participation with data collection and commercialization. The platform needs the users' consent to collect and commercialize each data item. Users choose which data items they wish to share with the platform. For each data item, the platform informs users whether it would commercialize it, given that users agree to share it. Users that join the platform give individual consent for the collection of each data item, recognizing that by agreeing to share a data item, it might be commercialized (unless the platform states otherwise).

Regime 3: users control, and can be compensated for, the commercialization of their data. Just like in regime 2, this regime prohibits the platform from tying users' participation to the consent to collect data. Here, however, it is the users' decision whether a data item they agreed to be collected can also be commercialized. That is, a user can give the platform the consent to collect a specific data item for private and public benefit, while denying it the right to commercialize it. Note that this regime provides users with even stronger control over their data relative to regime 2. A user can agree to share their data in order to receive the private benefit, yet refuse to have this data commercialized to save on the costs of privacy. Obviously, in this regime no user agree to commercialize data unless compensated. Platforms can incentivize users to agree to commercialize their data by offering users compensation for the right to commercialize their data.

To disentangle the different effects of heterogeneous users and data, the next two chapters study regimes 1 and 2. For each regime we start with the heterogeneous users case and then analyze heterogeneous data. We then compare the two regimes and show how the comparison depends on the type of heterogeneity. Section 3 extends the analysis to both heterogeneous users and data. We considers regime 3 in the appendix.

To make the problem meaningful, we restrict the parameters as follows. First, suppose that $0 < \alpha < 1$. This assumption implies that under heterogeneous users or data, commercializing data items with either $\varepsilon \in [0, \alpha]$ (heterogeneous users) or $\theta \in [0, \alpha]$ (heterogeneous data) enhances welfare, while commercializing data items with $\varepsilon \in [\alpha, 1]$ or $\theta \in [\alpha, 1]$ is welfare reducing. Intuitively, we allow the disutility of some users or some data items to exceed or be under the commercial benefit. Second, suppose that $0 < p < 1$. This implies that when users control their data, some users (or for some data items) will not give data for commercial benefit, even though depriving data prevents the platform from providing the users with the associated private benefit. Finally, we assume that $0 < \gamma < 1 - p$. This assumption ensures that under heterogeneous users, the market is not fully covered in both users and data.

Given the above assumptions, in all three cases (heterogeneous users, data, or both), total social welfare is maximized when all users join the platform and share all data for private and public benefits. The platform commercializes data item θ of users ε if and only if $k_{\varepsilon\theta} < \alpha$.

3 Regime 1: The platform controls the data it collects

Recall that under regime 1, regulation permits the platform to contingent participation in the platform with users' consent for the collection and commercialization of their data.

Case *A*: Heterogenous users

Suppose that users' idiosyncratic disutility from commercializing their data, ε, is distributed between $[0, 1]$ according to $f(\varepsilon)$. There is no heterogeneity in data items, and there is one indivisible data item with $\theta = 0$. If users with $\varepsilon \in [0, \widetilde{\varepsilon}]$ join the platform and share commercialized data, user ε's utility is $U(\varepsilon|\widetilde{\varepsilon}) = \gamma F(\widetilde{\varepsilon}) + p - \varepsilon$.

Under regime 1 the platform commercializes the data item from each user that joins it. Users are aware of this policy and can choose whether to join the platform or stay out and earn 0. Because users' utility deceases in ε, there is a threshold, $\widetilde{\varepsilon}$, such that users with $\varepsilon \in [0, \widetilde{\varepsilon}]$ join and give data, while users with $\varepsilon \in [\widetilde{\varepsilon}, 1]$ stay out, where $\widetilde{\varepsilon}$ solves:

$$U(\widetilde{\varepsilon}|\widetilde{\varepsilon}) = 0 \quad \Longleftrightarrow \quad \gamma F(\widetilde{\varepsilon}) = \widetilde{\varepsilon} - p. \tag{3}$$

Equation 3 provides initial results with respect to the effect the presence of a public benefit of data has on users' behavior. It is easy to see that when $\gamma = 0$, $\widetilde{\varepsilon} = p$; that is, users join the platform as long as the private benefit from sharing data, p, is larger than their disutility from sharing it, ε. Once the public benefit of data becomes positive, even users with $p < \varepsilon$ join the platform as users want to enjoy the public benefit, $\gamma F(\widetilde{\varepsilon})$, from data collected on other users on the platform. The following proposition characterizes how the number of users (hence, the amount of data collected), is affected by the public benefit (all proofs are in the Appendix).

Proposition 1. *(Regime 1 with heterogeneous users: The effect of the public benefit) A solution to equation (3) exists and is unique when $F(\varepsilon)$ does not exhibit an extreme unimodal distribution. Moreover:*

(i) *when data has no public benefit, $\gamma = 0$, $\widetilde{\varepsilon} = p$ and users with $\varepsilon \in [0, p]$ join the platform and share data;*

(ii) *the number of users that join the platform and share data increases in the public benefit of data: when $\gamma > 0$, $\widetilde{\varepsilon} > p$ and is increasing with γ;*

(iii) *when $\gamma = 1 - p$, all users join and share data: $\widetilde{\varepsilon} = 1$.*

In what follows, we assume that $F(\varepsilon)$ is not "too" unimodal, such that there is a unique solution to (3). We note that our results hold even when there are multiple solutions to (3), because all of these solutions have the qualitative features that we discuss below. We comment on this assumption in remark 1 in the proof of Proposition 1.

Proposition 1 shows that even though each user takes the equilibrium public benefit of data as given, the presence of the public benefit motivates users to join the platform and share data, even if their personal discomfort from doing so exceeds their personal benefit from data. Notice that in the case of heterogeneous users, data collection in regime 1 plays the same role as network effects. This is because each user that joins the platform shares data with the remaining users. Below we show that this will no longer be the case in the other scenarios that we investigate, in which there is no direct mapping between the number of users that join the platform and the amount of data collected.

Consumer surplus, $CS_{1,users}$, and profits, $\pi_{1,users}$, under Regime 1 when there are heterogenous users are:

$$CS_{1,users} = \gamma \cdot F(\widetilde{\varepsilon}) \cdot F(\widetilde{\varepsilon}) + \int_0^{\widetilde{\varepsilon}} (p - \varepsilon) f(\varepsilon) d\varepsilon, \qquad \pi_{1,users} = \alpha F(\widetilde{\varepsilon}), \tag{4}$$

and total welfare is given by $W_{1,users} = CS_{1,users} + \pi_{1,users}$.

Case B: Heterogenous data

Suppose now that the disutility from commercializing a data item, θ, is distributed between $[0, 1]$ according to $g(\theta)$ and that $\varepsilon = 0$. Because all users are identical, they either all join or all stay out. When all users join and give data items with $\theta \in [0, \overline{\theta}]$ for public and private benefits, and data with $\theta \in [0, \widetilde{\theta}]$ for commercial benefit, each user's utility is given by 2.

In Regime 1, the platform decides which data items to collect and commercialize. Because now there is a set of heterogeneous data items, the platform can choose to commercialize only a subset of the data it collects. Users, can only decide whether to join the platform and accept its data policy, or stay out. Given that users are identical, they make the same decision.

Since the platform bears no cost for data collection, yet data collected provides users with $p > 0$ and $\gamma > 0$, the platform collects all data items: $\overline{\theta} = 1$. Suppose that the platform chooses to commercialize a subset of data items with $\theta \in [0, \widetilde{\theta}]$. The platform would like to commercialize as many data items as possible, subject to the users' participation constraint. Let $\widetilde{\theta} = min\{\widetilde{\theta}', 1\}$, where $\widetilde{\theta}'$ is the solution to:

$$U(1, \widetilde{\theta}') = 0 \quad \Longleftrightarrow \quad \gamma + p = \int_0^{\widetilde{\theta}'} \theta g(\theta) d\theta. \tag{5}$$

That is, as in the case of heterogenous users, equation (5) shows that the presence of a public benefit of data has an important effect on market efficiency. With heterogeneous data items and identical users, all users join and give data for public use. The platform, then, takes advantage of its ability to contingent participation with data collection and commercialization and commercializes more data items than optimal for users; i.e., data items with $\theta > p$. Moreover, since users get private benefit for all data items, it is easy to show that $\theta > p$ even for $\gamma = 0$. This result already points to the first difference between the case with heterogeneous users and the case with heterogeneous data items. When users are heterogeneous, not all users join the platform and thus not all users contribute to the public benefit. In this case, welfare may be harmed by too little users' participation. In the case of heterogenous data items, all users contribute to the public benefit and the negative effect on welfare is driven by the platform's exploitation of its ability to contingent participation with data commercialization to commercialize too many data items.

The following proposition characterizes how the number of data items is affected by the public benefit.

Proposition 2. *(Regime 1 with heterogeneous data: The effect of the public benefit)*

(*i*) *The platform collects all data items, and commercializes data with $\theta \in [0, \widetilde{\theta}]$, where $\widetilde{\theta} > p$ for all $\gamma \geq 0$;*

(*ii*) *the number of data items that the platform commercializes increases with the public benefit: $\widetilde{\theta}$ is increasing with γ;*

(*iii*) *there is a threshold γ_{data}, $0 < \gamma_{data} < 1 - p$, such that the platform commercializes only a subset of the data items if $\gamma < \gamma_{data}$ and all data items otherwise. That is, $\widetilde{\theta} < 1$ if $\gamma < \gamma_{data}$ and $\widetilde{\theta} = 1$ otherwise.*

The ability to contingent participation with the provision of data for commercialization allows the platform to "bundle" the provision of less "costly" data – data items with $\theta < p$ – with the provision of more "costly" data – data items with $\theta < \widetilde{\theta}$, where $\widetilde{\theta} > p$. This bundling allows the platform to demand that users either agree to commercialize all data items with $\theta < \widetilde{\theta}$, or stay out. As γ increases, the platform can add more costly data items with $\theta > p$ to the bundle, and maintain the users' consent to commercialize them. Recalling that it is welfare enhancing to commercialize data items with $\theta < \alpha$ and that $\alpha \leq 1$, we have that when γ is high enough, regime 1 renders users to give more data for commercial use than the efficient level. The higher the public benefit, the more the platform can extract from users.

Consumer surplus with heterogeneous data items, $CS_{1,data}$, and profits, $\pi_{1,data}$, are:

$$CS_{1,data} = \gamma + p - \int_0^{\widetilde{\theta}} \theta g(\theta) d\theta, \qquad \pi_{1,data} = \alpha G(\widetilde{\theta}).$$

Total welfare is $W_{1,data} = CS_{1,data} + \pi_{1,data}$.

4 Regime 2: Users control their data

In this regime, regulation does not permit the platform to contingent participation on data sharing. Users can choose whether to join the platform and if they join, whether to share their data with the platform, knowing that shared data might be commercialized. For example, a navigation app can inform users that it plans to commercialize their location, if shared. Under regime 2, users can decide to refuse sharing their location (opt-out). In this case, the platform is still obligated to give users access to the public benefit (e.g., maps, current traffic) but will not be able to monitor the user's actual location. In the context of our model, this implies that should the user decline to share a certain data item, the user will not receive the private benefit for this particular data item, p, but will receive the total public benefit that the platform provides. We further assume that the platform cannot distinguish, ex-ante, between users that plan to share data, and block users that do not. As with regime 1, below we first solve the model with heterogeneous users and then move to heterogeneous data items.

Heterogenous users

The platform commercializes the data of any user that gives it the right to collect it. Users that join the platform yet choose not to share their data, only receive the public benefit of data collected on users who shared their data with the platform–i.e., γn_2, where n_2 is the number of users that share their data. Users that share their data with the platform enjoy, in addition to the public benefit, the private benefit from sharing data, p, yet bear the disutility ε. That is, under regime 2, a user can save on its disutility from having a data item commercialized by refusing to share the data, but by doing so also gives up on the data items' private benefit.

Given that joining the platform without sharing data bears no cost yet delivers benefits, under Regime 2, all users join the platform, however only users with $\varepsilon < p$ share data. The number of users that share data is therefore $n_2 = F(p)$. Total welfare under Regime 2 is then $W_{2,users} = CS_{2,users} + \pi_{2,users}$, where:

$$CS_{2,users} = \gamma F(p) + \int_0^p (p-\varepsilon) f(\varepsilon) d\varepsilon, \qquad \pi_{2,users} = \alpha F(p).$$

Heterogenous data

As in the case with heterogeneous users, in regime 2 with heterogeneous data items, all users join the platform yet agree to the commercialization of only data items with $\theta < p$. The platform collects all data items, as it is costless for it to do, but commits not to commercialize data items with $\theta > p$, because if the platform commercializes a data item with $\theta > p$, users will not agree to share it. Users enjoy the public and private benefits from all data items, yet bear the disutility of data items with $\theta < p$, which the platform commercializes. Total welfare is $W_{2,data} = CS_{2,data} + \pi_{2,data}$, where:

$$CS_{2,data} = \gamma + p - \int_0^p \theta g(\theta) d\theta, \qquad \pi_{2,data} = \alpha G(p).$$

5 Comparison between regimes 1 and 2

This section compares between the two data collection and commercialization regimes. We show that the comparison depends on the interaction between the magnitude of the public benefit of data, γ, and the type of heterogeneity in users' disutility. In particular, with heterogeneous users, it is welfare enhancing to let the platform (users) control the data when the public benefit of data is high (low). The opposite holds with heterogeneous data. We start with comparing the two regimes under heterogeneous users and then analyze the heterogeneous data case.

Heterogeneous users

Comparing the number of users, total data collected, and total data commercialized, the following corollary follows directly from the two sections above:

Corollary 1. *(Heterogeneous users: regime 1 collects more data than regime 2) In regime 1, the platform serves fewer users than in regime 2. Moreover, when $\gamma > 0$ ($\gamma = 0$), the platform collects more (same level of) data for public and commercial benefits in regime 1 than in regime 2.*

Intuitively, in regime 1 users have to share data knowing that it will be commercialized, so not all users agree to join the platform. Yet, when data has public benefit, i.e., $\gamma > 0$, in regime 1 the platform can exploit the public benefit to attract users to join the platform and share their data even though their disutility from data commercialization is higher than their private benefit. These users join the platform in regime 2, but in this regime they do not share their data.

Next, we turn to comparing total welfare, consumer surplus and the platform's profits in the two regimes.

Proposition 3. *(Heterogeneous users: the effect of the public and commercial benefits on the comparison between regime 1 and 2) When $\gamma = 0$, regimes 1 and 2 are identical in terms of consumer surplus, platform's profits, and total welfare. When $\gamma > 0$:*

(i) the platform's profits in regime 1 are higher than in regime 2;

(ii) consumer surplus in regime 1 is higher (lower) than in regime 2 when γ is large (small);

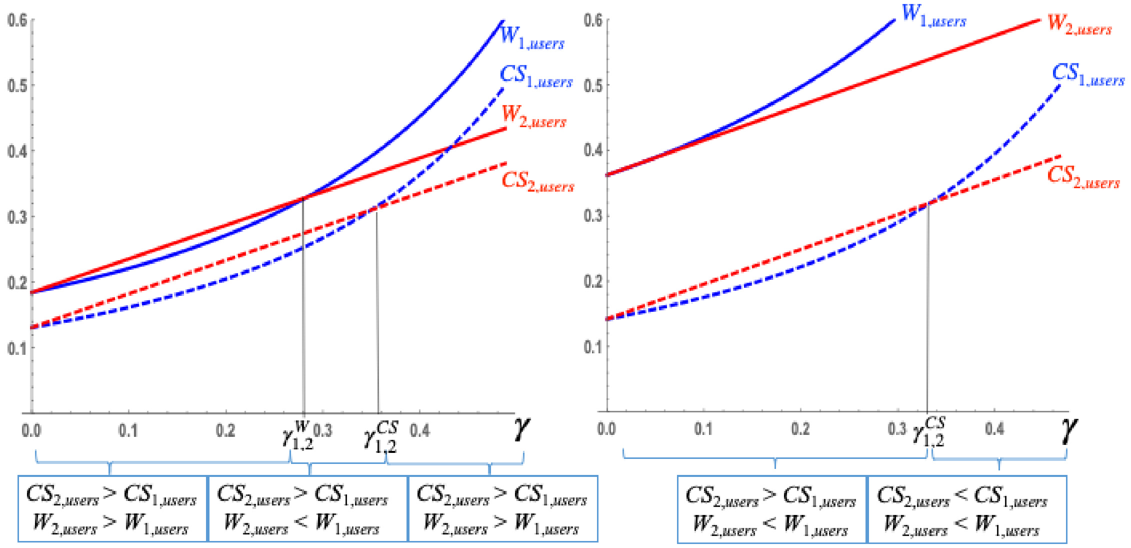
(iii) if $CS_{1,users}$ is convex and has no inflection points then there is a unique threshold, $0 < \gamma_{1,2}^{CS} < 1$, such that consumer surplus in regime 1 is higher than in regime 2 if $\gamma > \gamma_{1,2}^{CS}$ and lower otherwise;

(iv) welfare in regime 1 is higher than in regime 2 when γ is large. If $W_{1,users}$ is convex and has no inflection points then there exists a unique threshold, $0 \leq \gamma_{1,2}^{W} < 1$, such that total welfare in regime 1 is higher than in regime 2 if $\gamma > \gamma_{1,2}^{W}$ and lower otherwise;

(v) when data has no commercial benefit, i.e., $\alpha = 0$, $0 < \gamma_{1,2}^{W} = \gamma_{1,2}^{CS}$. As α increases, $\gamma_{1,2}^{CS}$ remains constant while $\gamma_{1,2}^{W}$ decreases. Moreover, $\gamma_{1,2}^{W} = 0$ if α is high enough.

Figure 1 illustrates the results of Proposition 3 for a uniform distribution $F(\varepsilon)$. The figure shows consumer surplus and welfare as a function of the public benefit of data. Notice that with a uniform $F(\varepsilon)$, both $CS_{1,users}$ and $W_{1,users}$ are convex and have no inflection points, resulting in unique thresholds of $\gamma_{1,2}^{CS}$ and $\gamma_{1,2}^{W}$.

The intuition for Proposition 3 is that regime 1 has an advantage and disadvantage, in comparison with regime 2. The disadvantage is that not all users join under regime 1, as data-sensitive users prefer to stay out of the platform, while all users join under regime 2.



Panel (a): data has small commercial benefit (α = 0.1)

Panel (b): data has high commercial benefit (α = 0.5)

Figure 1: Consumer surplus and welfare as a function of γ for a uniform $F(\varepsilon)$ ($p = 0.5$)

The advantage is that all users that do join under regime 1, share their data, among other things, for the public benefit. As γ increases, the disadvantage of regime 1 becomes weaker because more users join the platform, yet the advantage of regime 1 becomes stronger because the public benefit of the data of users that do join under regime 1 becomes more valuable.

Notice first that at $\gamma = 0$, the figure shows that both regimes are identical, because in both regimes the platform collects data only from users for whom the disutility from the commercialization of their data is lower or equal to their private benefit from providing data. This result highlights the role the public benefit of data plays in users' behavior under these two regimes. This result also highlights the distinction between the public benefit of data and network effects. Recall that more users join the platform in regime 2 than in regime 1, regardless of the level of γ. In contrast to our model, in the presence of network effects that are based on participation in the platform, these users would make regime 2 superior to regime 1.

Next consider the comparison in consumer surplus. As γ becomes positive (but small enough), consumer surplus is higher under regime 2. While regime 1 provides more data for the public benefit than regime 2, in regime 2 more users participate and can benefit from it, making regime 2 superior. In contrast, there is a threshold in γ, $\gamma_{1,2}^{CS}$, such that if γ is high, consumers actually benefit when the platform control their data because then more consumers that join the platform share their data and thus there is more data for public benefit.

As for the platform's profit, the platform always prefer regime 1 over regime 2, as it commercializes more data under regime 1. Consequently, there is a second threshold, $\gamma^W_{1,2}$, such that welfare is higher under regime 1 if $\gamma > \gamma^W_{1,2}$ (as shown in panel (a)). If the commercial benefit of data is high, then because the platform earns higher profits under regime 1, this regime provides higher welfare for all values of γ (as shown in panel (b)).

Heterogeneous data

Next we move to compare between the two regimes when the model exhibits heterogeneous data items. The proposition below shows that heterogeneous data items case yields the opposite conclusion than the heterogeneous users case. In particular, now regime 1 enhances welfare when the public benefit of data is low, while regime 2 offers higher welfare otherwise.

Proposition 4. *(Heterogeneous data: the effect of the public and commercial benefits of data on the comparison between regime 1 and 2) The platform prefers regime 1 while users prefer regime 2 for all values of γ and α. Moreover, there are two thresholds $\underline{\alpha}^W_{1,2}$ and $\overline{\alpha}^W_{1,2}$, where $0 < \underline{\alpha}^W_{1,2} < \overline{\alpha}^W_{1,2} < 1$, such that:*

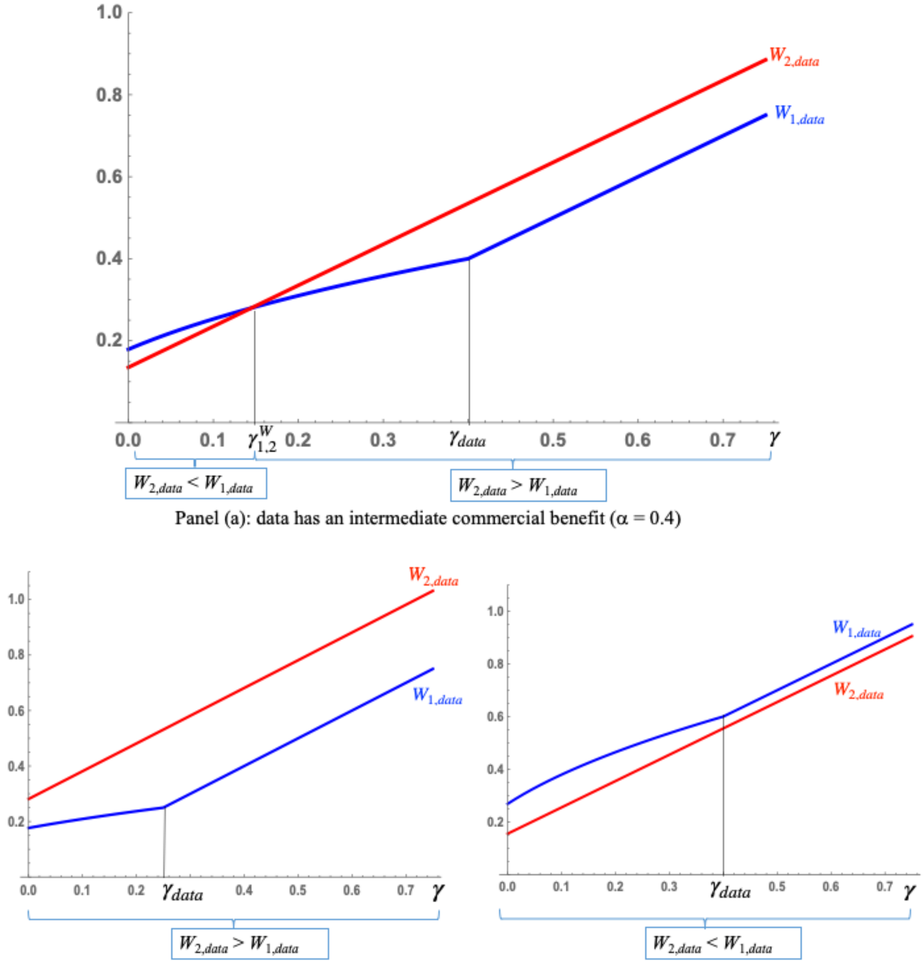
(*i*) *for intermediate values of the commercial benefit of data, regime 1 is welfare enhancing (reducing) when the public benefit is low (high). That is, when $\alpha \in [\underline{\alpha}^W_{1,2}, \overline{\alpha}^W_{1,2}]$, there is a threshold, $\gamma^W_{1,2}$, such that $W_{1,data} > W_{2,data}$ iff $\gamma < \gamma^W_{1,2}$;*

(*ii*) *for low values of the commercial benefit of data, regime 2 is welfare enhancing. That is, when $\alpha \in [0, \underline{\alpha}^W_{1,2}]$, $W_{2,data} > W_{1,data}$* for all γ;

(*iii*) *for high values of the commercial benefit of data, regime 1 is welfare enhancing. That is, when $\alpha \in [\overline{\alpha}^W_{1,2}, 1]$, $W_{1,data} > W_{2,data}$* for all γ.

Figure 2 illustrates the results of Proposition 4 for a uniform $G(\theta)$. Panel (a) shows part (i) of the proposition, when data has an intermediate commercial benefit. In this case, in contrast to the case of heterogeneous users, regime 1 is welfare enhancing when data has low public benefit, while the opposite holds for high values of public benefit. Panel (b) illustrates part (ii), where data has small commercial benefit and regime 2 is always welfare enhancing. Likewise, panel (c) illustrates part (iii): when data has high commercial benefit, regime 1 is always welfare enhancing.

The intuition behind these results is the following. Recall that when users are identical and data items are heterogeneous, regime 1 enables the platform to require that users consent to the commercialization of a "bundle" of data items with $\theta < \widetilde{\theta}$, where $\widetilde{\theta} > p$. Users agree to commercialize "costly" data items with $\theta > p$ because they gain a positive net private benefit $p - \theta$ on other data items and because they gain the public benefit γ. As the public benefit



Panel (a): data has an intermediate commercial benefit ($\alpha = 0.4$)

Panel (b): data has small commercial benefit ($\alpha = 0.1$)

Panel (c): data has high commercial benefit ($\alpha = 0.6$)

Figure 2: Welfare as a function of γ for a uniform $G(\theta)$ ($p = 0.1$)

increases, the platform's ability to commercialize more data items with $\theta > p$ increases, while maintaining users' consent to commercialize them. This aggressive "bundling" results in too many data items being commercialized (relative to the first best) when γ is high. That is, the potential inefficiency under regime 1 is that the platform can take advantage of the public benefit to commercialize a too large set of data items. While a central planner would commercialize data items with $\theta < \alpha$, the platform commercializes data items with $\theta < \widetilde{\theta}$ regardless of whether $\widetilde{\theta}$ is larger or smaller than α. Given that in regime 2 the users choose how many data items the platform can commercialize, the platform cannot employ the same type of aggressive bundling. Indeed, as panel (a) shows, when the public benefit of data is high, in regime 1 the platform takes advantage of γ to commercialize too many data items as a bundle, making regime 2 welfare enhancing. In contrast, when γ is small, regime 2 under-performs regime 1 because in regime 2 users agree to commercialize too little data.

Note that the slope of $W_{2,data}$ is constant at 1. This is not the case in regime 1, for γ values that are smaller than γ_{data} – i.e., γ values where under regime 1, the platform commercializes only a subset of the data items. In this case, as γ increases the platform faces a tradeoff between extracting value by commercializing more data items and the negative effect this may have on user participation. The platform does not face the same tradeoff under regime 2, as the number of data items it can commercialize is set by the users.

Panel (a) holds for intermediate values of α. When the commercial benefit of data is small (panel (b)), the platform's ability to bundle data in regime 1 reduces welfare in comparison to regime 2 for all values of γ, because the platform commercializes too many data items that have small commercial benefit. Alternatively, when the commercial benefit of data is high (panel (c)), the platform's ability to bundle data in regime 1 enhances welfare in comparison to regime 2 for all γ, because in regime 2 users agree to commercialize too little data items that have high commercial benefit.

Comparison between types of heterogeneity

The distinction between heterogeneous users and heterogeneous data yields different conclusions with respect to the effect of the public benefit of data on data regulation. We summarize these differences in Table 1.

Table 1: Comparison between types of heterogeneity

Type of Heterogeneity	Inefficiency	Effect of an increase in γ	Result
Heterogeneous users	The platform attracts too little users and thus collects too little data for public benefit	Regime 1 mitigates this problem as an increase in γ increases its ability to attract more users and collect their data	For high γ, welfare in regime 1 is higher than in regime 2
Heterogeneous data items	The platform collects too much data for commercial benefit	Regime 1 exacerbates this problem as an increase in γ increases its ability to demand that more data be commercialized	For high γ, welfare in regime 2 is higher than in regime 1

As noted in the table, with heterogeneous users, in regime 1 not all users join the platform, yet those who provide data also give data for public benefit. Hence, the potential inefficiency is that too little data is collected for public benefit. For example, given the choice to join a contact tracing app, many users choose the outside option of not joining the platform than bearing the cost of their data being shared; despite knowing that they also won't be able to enjoy the benefits of knowing whether they were in proximity of an infected individual. Regime 1 better mitigates this problem when the public benefit of data is high. With heterogeneous data, in regime 1 all users give data for public benefit, but the platform may commercialize too many data items. Hence, the inefficiency concerns too much data being commercialized. Regime 1 exacerbates this problem when the public benefit of data is high.

6 Dynamic data accumulation

An important aspect of data is that it may accumulate over time such that platforms can use past data to offer higher public benefit for current users. For example, data on a driver's location in the past can help a navigation app to predict future traffic. Yet, the value of data may depreciate between periods. For example, data collected by a navigation app on traffic conditions becomes partially obsolete after some time. One may wonder whether our results change once we allow for data to accumulate over time. Below we show that accounting for dynamics through data accumulation, in fact, strengthens the results of our base model. Specifically, we show that as the degree of data accumulation increases, regime 1 (2) becomes attractive for a wider set of parameters when heterogeneity is mostly driven by users (data items). Moreover, we find that under heterogeneous users, if data does not depreciate much over time, the platform initially chooses not to commercialize data at all. In this case, the platform first serves all users and accumulates their data. This allows the platform to then exploit the accumulated data in future period to offer high public benefit, thereby attracting many users even though it commits to commercialize all their data.

Let Δ_t denote the amount of data for public benefit accumulated at the beginning of period t. The platform starts in period $t = 1$ with no data accumulated: $\Delta_1 = 0$. For each period $t > 1$, if the platform starts the period with Δ_t data for public benefit, and collected, in time t, $F(\widetilde{\varepsilon}_t)G(\widetilde{\theta}_t)$ data for public benefit, then in the next period the platform starts the period with $\Delta_{t+1} = \delta\left(\Delta_t + F(\widetilde{\varepsilon}_t)G(\widetilde{\theta}_t)\right)$, where δ $(0 < \delta < 1)$ is the degree to which data accumulates between periods. When $\delta = 0$, all previous data becomes obsolete and the game is equivalent to a static game. As δ increases, more data is transferred across periods and dynamics become more important.

We look for a steady state where $\Delta_{t+1} = \Delta_t$. That is, while in the short run, the platform accumulates more and more data along time, the value of data depreciates over time at a rate of $(1 - \delta)$ such that, in the long run, the platform may reach a steady state where it starts all periods with the same amount of data. This is in contrast to the case where the platform keeps growing and $\Delta_{t+1} > \Delta_t$ for all t. We identify how data policy affects the convergence or agglomeration of data accumulation. For simplicity, in this section we assume that ε and θ are uniformly distributed and as before, we study the two cases of heterogeneous users and data separately.

6.1 Case A: Heterogeneous Users

Suppose that users' idiosyncratic disutility from commercializing their data, ε, is uniformly distributed between $[0, 1]$. There is no heterogeneity in data items, and there is one indivisible data item with $\theta = 0$. If, in period t, the platform accumulated at the beginning of the period data of size Δ_t and users with $\varepsilon \in [0, \widetilde{\varepsilon}_t]$ join the platform and share data which is then commercialized, user ε's utility is $U(\varepsilon|\widetilde{\varepsilon}_t) = \gamma(\Delta_t + \widetilde{\varepsilon}_t) + p - \varepsilon$.

Regime 1: The platform controls the data it collects.

Under regime 1, the platform commercializes data from all users that joins it. As in our base model, users are aware of this policy and can choose whether to join the platform or stay out and earn 0. Equation 3 then becomes:

$$U(\widetilde{\varepsilon}_t|\widetilde{\varepsilon}_t) = 0 \quad \Longleftrightarrow \quad \gamma(\Delta_t + \widetilde{\varepsilon}_t) = \widetilde{\varepsilon}_t - p. \tag{6}$$

Unless in a steady state, $\widetilde{\varepsilon}_t$ increases over time because the platform can utilize the data accumulated from previous periods to attract more users. That is, the platform grows over time in terms of data and consequently in terms of users. We look for a steady state where at some point in time the platform stops growing because the amount of data that depreciates balances out the amount of new data collected in each period.

Lemma 1. ***(The steady state under regime 1 and heterogenous users).*** *The market achieves a steady state if $\delta < 1 - \gamma$, where in each period, the platform starts with $\Delta_t = \frac{\delta p}{1-\delta-\gamma}$ data for public benefit, serves $\widetilde{\varepsilon}_t = \frac{(1-\delta)p}{1-\delta-\gamma}$ users and provides total public benefit of $\Delta_t + \widetilde{\varepsilon}_t = \frac{p}{1-\delta-\gamma}$. In the beginning of the next period, the platform starts with $\Delta_{t+1} = \delta\Delta_t = \frac{\delta p}{1-\delta-\gamma}$, and so on.*

Notice that, if there is no public benefit ($\gamma = 0$), then $\widetilde{\varepsilon}_t = p$ and dynamics has no effect. Furthermore, as δ increases, in the steady state, both $\widetilde{\varepsilon}_t$ and Δ_t increase. Intuitively, the more data for the public benefit accumulates along time, the more users the platform can attract, which in turn provide even more data for public benefit. If δ is sufficiently high ($\delta > 1 - \gamma$), there is no steady state and the platform keeps growing, until it serves all users.

Regime 2: Users control their data The platform commercializes the data of any user that gives it the right to collect it. Users that join the platform yet choose not to share their data, only receive the public benefit of data collected on users who shared their data with the platform. Users that share their data with the platform enjoy, in addition to the public benefit, the private benefit from sharing data, p, yet bear the disutility ε. We have the following result:

Lemma 2. ***(The steady state under regime 2 and heterogenous users).*** *The market achieves a steady state for all $0 < \delta < 1$. In each period, the platform starts with $\Delta_t = \frac{\delta p}{1-\delta}$ data for public benefit, serves p users and provides total public benefit of $\Delta_t + p = \frac{p}{1-\delta}$. In the beginning of the next period, the platform starts with $\Delta_{t+1} = \delta \Delta_t = \frac{\delta p}{1-\delta}$, and so on.*

Under regime 2, the amount of data collected in each period is constant and equals to p, hence there is always a steady state. Data accumulation in the steady state is increasing with δ but unlike regime 1, is independent of γ.

Comparison of the two data regimes Notice first that, if there is no public benefit, i.e., $\gamma = 0$, the two regimes are identical, as in the static case.

We start with comparing the market's tendency to reach a steady state.

Corollary 2. *The market is more likely to converge to a steady state under regime 2 than under regime 1. In particular, for $\delta > 1 - \gamma$, under regime 1 the platform keeps growing over time, while under regime 2, there is convergence.*

Intuitively, under regime 1 the platform can utilize the public benefit to attract more users and thus to collect more data. Hence, the platform has more of a potential to grow over time. Under regime 2, the platform always attracts the same number of users and collects the same amount of data; thereby staying stagnant after reaching a certain amount of data.

Next we turn to evaluate the social welfare under the two regimes:

Proposition 5. ***(Heterogeneous users: how data accumulation affects the comparison between regime 1 and 2)*** *Data accumulation makes regime 1 more welfare enhancing in comparison with regime 2. There exists a unique threshold, $0 \leq \gamma^W_{1,2}(\delta) < 1$, such that total welfare in regime 1 is higher than in regime 2 if $\gamma > \gamma^W_{1,2}(\delta)$ and lower otherwise; where $\gamma^W_{1,2}(\delta) = 0$ if α is high enough. Moreover, $\gamma^W_{1,2}(\delta)$ is decreasing in δ.*

Proposition 5 shows that dynamics make regime 1 more attractive. That is, as the degree of data accumulation, δ, increases, $\gamma^W_{1,2}(\delta)$ decreases and regime 1 becomes attractive for a wider set of parameters. Recall that regime 1 has the disadvantage over regime 2 that not all users join, and the advantage that all users that do join share their data for the public benefit. As data accumulates along time, the disadvantage becomes weaker and the advantage becomes stronger because the increase in data for public benefit attracts more users to join the platform. As a result, regime 1 becomes the superior regime for a wider range of γ. This result indicates that a data policy that gives the platform control over data is more desirable in markets where users are heterogeneous, the public benefit of data is high, and data accumulates along time.

6.2 Case B: Heterogeneous data

Suppose now that data items differ in users' disutility from commercializing them, while all users are identical. For simplicity, suppose that θ is uniformly distributed between $[0,1]$, while $\varepsilon = 0$ for all users. Given that in a certain period the platform collects data up to $\overline{\theta}_t$ for public and private private benefit and commercializes data up to $\widetilde{\theta}_t$, a user's utility is:

$$U(\overline{\theta}, \widetilde{\theta}) = \gamma \left(\Delta_t + \overline{\theta}_t\right) + p\overline{\theta}_t - \int_0^{\widetilde{\theta}_t} \theta d\theta.$$

Regime 1: The platform controls the data it collects.

Under regime 1, the platform collects all data items for public and private benefit. Hence, data for public benefit accumulates at a fixed amount of 1 in each period. Yet, the amount of data commercialized in each period increases along time.As in the base model, the platform commercializes data in each period up to the users' participation constraint. This enables the platform to bundle data items such that $\theta < p$ with data items such that $\theta > p$. The more data accumulates, the more data items with $\theta > p$ that the platform can bundle. Consequently, in regime 1 the platform takes advantage of data accumulation to gradually increase the amount of data commercialized along time, until reaching the steady state with the following features:

Lemma 3. ***(The steady state under regime 1 and heterogenous data).*** *The market achieves a steady state for all $0 < \delta < 1$. In each period, the platform starts with $\Delta_t = \frac{\delta}{1-\delta}$ data for public benefit, serves all users and provides total public benefit of $\Delta_t + 1 = \frac{1}{1-\delta}$. In the beginning of the next period, the platform starts with $\Delta_{t+1} = \delta\Delta_t = \frac{\delta}{1-\delta}$, and so on. The amount of data commercialized, $\widetilde{\theta}_t$ increases along time and in the steady state is: $\widetilde{\theta}_t = \sqrt{\frac{2\gamma}{1-\delta} + 2p}$.*

Regime 2: The platform controls the data it collects.

Under regime 2, the platform collects all data items for private and public benefit. As in under regime 1, data accumulates in each period at a fixed rate of 1. Now, however, the platform can only commercialize data items with $p > \theta$ because users will not agree to share data items with $p < \theta$ if these data items are commercialized. Hence, the amount of data commercialization is also fixed in each period and equals to p.

Comparison of the two data regimes In both regimes, the platform accumulates the same amount of data. The two regimes differ in the amount of data commercialized. In regime 2, the platform can only commercialize data up to p. Regime 1 has the welfare reducing effect that the platform "bundles" data items, including data items with $\theta > p$, and takes advantage of data accumulation to increase the amount of data commercialization along time. Therefore, as the following proposition shows, under heterogeneous data, data accumulation increases the range of parameters under which regime 2 is the superior regime:

Proposition 6. ***(Heterogeneous data: how data accumulation affects the comparison between regime 1 and 2)*** *Data accumulation makes regime 2 more welfare enhancing in comparison with regime 1. For intermediate values of α, there exists a unique threshold, $0 \leq \gamma^W_{1,2}(\delta) < 1$, such that total welfare in regime 2 is higher than in regime 1 if $\gamma > \gamma^W_{1,2}(\delta)$ and lower otherwise. Moreover, $\gamma^W_{1,2}(\delta)$ is decreasing in δ.*

Proposition 6 shows that the results of the static game are reinforced when the platform accumulates data along time. When the public benefit of data is high, under heterogeneous data it is optimal to give users control over data, and the range of γ under which regime 2 is welfare enhancing is increasing in the data accumulation parameter, δ. This result indicates that a data policy that gives users control over data is more desirable in markets where data items are heterogeneous, the public benefit of data is high, and data accumulates along time.

Two-stage Game

Our analysis above compares focuses on the long-run steady state under the two regimes. In this case, the platform always prefers regime 1 over regime 2. Yet, in the short-run, a platform may prefer not to commercialize data in its early days in order to attract more users, accumulate more data, and then use this data to commercialize data as it grows. We illustrate this intuition by folding the dynamic game into a two-period game. We show that under heterogeneous users, if data substantially accumulates along time (i.e., high δ), then the platform finds it optimal not to commercialize data in the first period, serve all users and accumulate their data, and then exploit this data accumulation in the second period in order to commercialize data. For brevity, we relegate this analysis to the appendix.

7 Conclusion

The paper studies the "public good" aspect of data–i.e., the data digital platforms collect on a specific user provide benefit to other users of the platform, regardless of whether they also share their data. The commercialization of this data, however, inflicts a cost to the users whose data were commercialized. This raises the question of whether policy makers should regulate the platforms' ability to collect and commercialize data. We consider the interaction between a platform and users, when the platform can collect and commercialize data. We develop a model where, in addition to personal benefit, data also provide public benefits to other users. The platform collects a set of data items and can "bundle" data items by requiring users to either accept to share all of them or not join the platform. We allow users' disutility from the commercialization of their data to vary across users (the case of heterogeneous users) and across data items (heterogeneous data). We use this model to examine three extremes of data regulation regimes that vary in terms of who controls the data (users or the platform) and whether users can be compensated for having their data commercialized.

We find that the preferable regime for social welfare depends on the magnitude of the data's public benefit and on the type of heterogeneity in users' disutility from the commercialization of their data. With heterogeneous users, giving the *platform control* over data enhances welfare when the public benefit is high. In contrast, with heterogeneous data, it is welfare enhancing to give *users control* over their data when the pubic benefit is high. The difference in results is driven by the type of market inefficiency the two types of heterogeneity exhibit. With heterogeneous users, the main market inefficiency is that the platform attracts too few users and thus collects too little data for public benefit. Giving the platform the control over data enables it to exploit the public benefit to attract more users. With heterogeneous data, the main market inefficiency is that the platform collects too much data for commercial benefit, and giving users the control over data enables them to limit the level of data that the platform can commercialize. Whether compensating users for their data enhances or harms welfare depends on the type of heterogeneity as well as on the magnitude of the commercial and public benefits of data.

Interestingly, dynamic accumulation of data strengthens our results. Specifically, under heterogeneous users (data), it is welfare enhancing to give the platform (users) control over data for a wider range of values of the public benefit of data.

References

[1] Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. 2022. "Too Much Data: Prices and Inefficiencies in Data Markets." *American Economic Journal: Microeconomics*, 14 (4): 218-56.

[2] Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. 2016. “The Economics of Privacy.” *Journal of Economic Literature* 54(2): 442-492.

[3] Bergrmann, Dirk, Alessandro Bonatti and Tan Gan. “The Economics of Social Data.” *The Rand Journal of Economics* (forthcoming).

[4] Biglaiser, Gary, and Jacques Crémer. “The value of incumbency in heterogenous networks.” *American Economic Journal: Micro* (forthcoming).

[5] Caillaud, Bernard, and Bruno Jullien. 2001. “Competing cybermediaries.” *European Economic Review* 45 (4-6): 797–808.

[6] Caillaud, Bernard, and Bruno Jullien. 2003. “Chicken & egg: Competition among intermediation service providers.” *The RAND Journal of Economics* 34 (2): 309–328.

[7] Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim. 2019. “Privacy and personal data collection with information externalities.” *Journal of Public Economics* 173:113-124.

[8] Dosis Anastasios and Wilfried Sand-Zantman. 2021. “The Ownership of Data.” Working paper.

[9] Economides, Nicholas, and Ioannis Lianos. “Restrictions on Privacy and Exploitation in the Digital Economy: A Market Failure Perspective.” *Journal of Competition Law and Economics* 17 (4): 765–847.

[10] Fainmesser, Itay, Andrea Galeotti, and Ruslan Momot. 2022. “Digital Privacy.” *Management Science, forthcoming.*

[11] Goldfarb, Avi, and Catherine Tucker. 2012. "Shifts in Privacy Concerns." *American Economic Review*, 102 (3): 349-53.

[12] Hagiu, Andrei. 2006. “Pricing and commitment by two-sided platforms.” *The RAND Journal of Economics* 37 (3): 720–737.

[13] Hałaburda, Hanna, and Yaron Yehezkel. 2013. “Platform competition under asymmetric information.” *American Economic Journal: Microeconomics* 5 (3): 22–68.

[14] Hałaburda, Hanna, and Yaron Yehezkel. 2016. "The role of coordination bias in platform competition." *Journal of Economics and Management Strategy* 25 (2): 274–312.

[15] Hałaburda, Hanna, and Yaron Yehezkel. 2019. "How beliefs affect platform competition." *Journal of Economics and Management Strategy* 28 (1), 49-49.

[16] Hałaburda, Hanna, Bruno Jullien, and Yaron Yehezkel. 2020. "Dynamic platform competition: how history matters?" *The RAND Journal of Economics* 51 (1): 3-31.

[17] Ichihashi, Shota and Alex Smolin. 2022. "Data Collection by an Informed Seller." Working paper.

[18] Jullien, Bruno. 2011. "Competition in multi-sided markets: Divide and conquer." *American Economic Journal: Microeconomics* 3 (4): 186–220.

[19] Jullien, Bruno, Yassine Lefouili, and Michael Riordan. 2020, "Privacy protection, security, and consumer retention." Working paper.

[20] Katz, Michael L., and Carl Shapiro. 1986. "Technology adoption in the presence of network externalities." *Journal of Political Economy* 94 (4): 822-841.

[21] Loertscher, Simon, and Leslie Marx. 2020. "Digital monopolies: Privacy protection or price regulation?" *Industrial Journal of Industrial Organization* 71.

[22] Markovich, Sarit, and Yaron Yehezkel. 2022. "Group Hug: Platform Competition with User-groups." *American Economic Journal: Micro,* 14 (2): 139-175.

[23] O'Brien, Daniel and Doug Smith. 2014. "Privacy in online markets: A welfare analysis of demand rotations." Working Paper No. 323

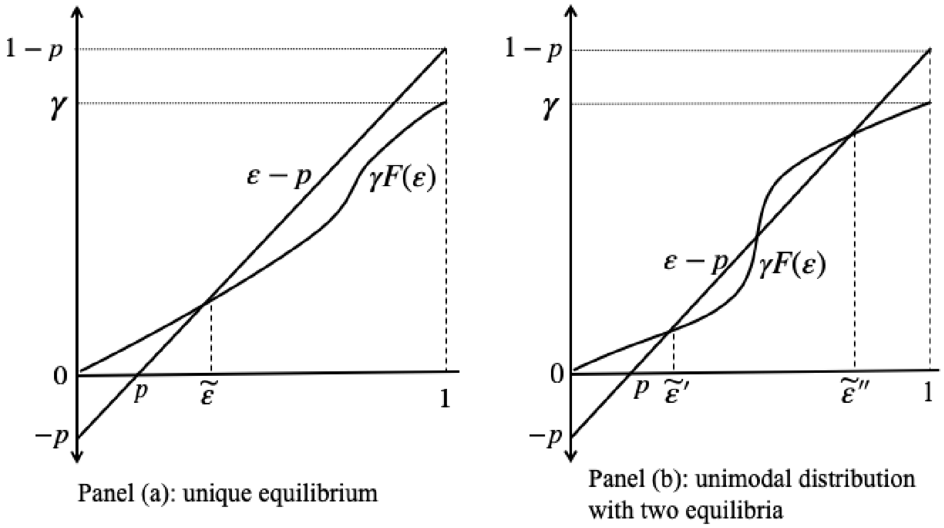
Appendix

Below are the proofs for all lemmas and propositions in the text.

Proof of Proposition 1:

We first show that there is at least one solution to (3). Evaluated at $\varepsilon = 0$, the left-hand side (hereafter LHS) of (3) is $\gamma F(0) = 0$ while the right hand side (hereafter RHS) is $0 - p < 0$, hence $\gamma F(\varepsilon) > \varepsilon - p$ if ε is sufficiently close to 0. Evaluated at $\varepsilon = 1$, the LHS of (3) is $\gamma F(1) = \gamma$ while the RHS is $1 - p \geq \gamma$ (recall that we assume that $\gamma \leq 1 - p$), hence $\gamma F(\varepsilon) < \varepsilon - p$ if ε is sufficiently close to 1, and at the highest possible γ, $\gamma = 1 - p$, the solution to (3) is at $\widetilde{\varepsilon} = 1$. This implies that there is at least one intersection point between $\gamma F(\varepsilon)$ and $\varepsilon - p$.

Next, we show the conditions under which this intersection point is unique. Figure 3 (panel (a)) shows the solution to $\widetilde{\varepsilon}$ when $F(\varepsilon)$ is not too unimodal (we can derive a qualitatively similar figure for a $F(\varepsilon)$ that is not unimodal). In this case, there is a unique solution to $\widetilde{\varepsilon}$, hence a unique equilibrium. Panel (b) shows the case of a strong unimodal $F(\varepsilon)$, in which case there are three solutions to (3). The middle one is not stable while in the two stable solutions, $\widetilde{\varepsilon}'$ and $\widetilde{\varepsilon}''$, $\gamma F(\varepsilon)$ intersects $\varepsilon - p$ from below, hence the comparative statics of $\widetilde{\varepsilon}'$ and $\widetilde{\varepsilon}''$ are qualitatively the same. That is, both solutions are higher than p, and both solutions are increasing with γ. Notice that with unimodal distribution, there can be at most three solutions to (3). When $F(\varepsilon)$ is not unimodal, there can be more than three solutions, yet all solutions in which $\varepsilon - p$ intersects $\gamma F(\varepsilon)$ from "above" are stable so have the same features as in the unimodal case.



$1-p$
γ
$\varepsilon - p$
$\gamma F(\varepsilon)$
0
p
$\widetilde{\varepsilon}$
1
$-p$
Panel (a): unique equilibrium
$1-p$
γ
$\varepsilon - p$
$\gamma F(\varepsilon)$
0
p
$\widetilde{\varepsilon}'$
$\widetilde{\varepsilon}''$
1
$-p$
Panel (b): unimodal distribution with two equilibria

Figure 3: The solution to $\widetilde{\varepsilon}$

Finally, the comparative statics of $\widetilde{\varepsilon}$ with respect to γ follow directly from the feature that $\gamma F(\varepsilon)$ intersects $\varepsilon - p$ from below and because $\gamma F(\varepsilon)$ is increasing in γ.■

Proof of Proposition 2:

part *i*

The platform collects all data items because the platform bears no cost for data collection, yet the data collected provides users with $p > 0$ and $\gamma > 0$. Next, we show that there is a unique solution to $\widetilde{\theta}$.

Evaluating (5) at $\theta = 0$, the RHS of eq. (5) is 0. The LHS is $1 > \gamma + p > 0$. Given that the LHS is independent of θ, while the RHS is increasing in θ, there are two possibilities. First, if the RHS evaluated at $\theta = 1$ is higher than $\gamma + p$, then there is a unique intersection point between the two sides and $\widetilde{\theta}$ is the solution to (5). Second, if the RHS evaluated at $\theta = 1$ is lower than $\gamma + p$, then $\widetilde{\theta} = 1$.

Next, we show that $\widetilde{\theta} > p$. Because:

$$\gamma + p - \int_0^{\widetilde{\theta}'} \theta g(\theta) d\theta = \gamma + p \int_0^{\widetilde{\theta}'} g(\theta) + p \int_{\widetilde{\theta}'}^1 g(\theta) - \int_0^{\widetilde{\theta}'} \theta g(\theta) d\theta$$

$$> \gamma + \int_0^{\widetilde{\theta}'} (p - \theta) g(\theta) d\theta,$$

and $\int_0^{\widetilde{\theta}'} (p - \theta) g(\theta) d\theta$ is positive at $\widetilde{\theta}' = p$, it follows that $\widetilde{\theta}' > p$ for all $\gamma \geq 0$.

part *ii*

To show that $\widetilde{\theta}'$ is increasing in γ. Using the implicit function theorem and defining $\Gamma(\gamma, \theta) \equiv \gamma + p - \int_0^{\theta} \theta g(\theta) d\theta = 0$,

$$\frac{d\theta}{d\gamma} = -\frac{\frac{\partial \Gamma(\gamma,\theta)}{\partial \gamma}}{\frac{\partial \Gamma(\gamma,\theta)}{\partial \theta}} = \frac{1}{\theta g(\theta)} > 0.$$

part *iii*

To prove this part, it is enough to show that for $\gamma = 1 - p$, the solution to eq. (5) is at $\widetilde{\theta}' = 1$. Evaluating eq. (5) at $\gamma = 1 - p$:

$$(1 - p) + p - \int_0^{\widetilde{\theta}'} \theta g(\theta) d\theta = 1 - \int_0^{\widetilde{\theta}'} \theta g(\theta) d\theta > 0,$$

where the inequality follows as θ is distributed between $[0, 1]$, implying that $\int_0^{\widetilde{\theta}'} \theta g(\theta) d\theta <$

1.■

Proof of Proposition 3:

The result that evaluated at $\gamma = 0$, $CS_{1,users} = CS_{2,users}$ and $\pi_{1,users} = \pi_{2,users}$ follows directly from the result that evaluated at $\gamma = 0$, $\widetilde{\varepsilon} = p$. When $\gamma > 0$:

***part** i*

The first part is a direct result of Corollary (1). If the platform collects more data for commercial benefit under regime 1 than regime 2 then $F(\widetilde{\varepsilon}) > F(p)$ and $\pi_{1,users} > \pi_{2,users}$.

***part** ii*

We first show that for γ values close to 0, $CS_{2,users} > CS_{1,users}$. When $\gamma > 0$, yet still very small, the derivative of consumer surplus with respect to γ:

$$\left.\frac{dCS_{1,users}}{d\gamma}\right|_{\gamma=0} = \left[F(\widetilde{\varepsilon})^2 + 2\gamma F(\widetilde{\varepsilon})f(\widetilde{\varepsilon})\frac{\partial\widetilde{\varepsilon}}{\partial\gamma} + (p-\widetilde{\varepsilon})f(\widetilde{\varepsilon})\frac{\partial\widetilde{\varepsilon}}{\partial\gamma}\right]_{\gamma=0}$$

$$= \left[F(\widetilde{\varepsilon})^2 + \gamma F(\widetilde{\varepsilon})f(\widetilde{\varepsilon})\frac{\partial\widetilde{\varepsilon}}{\partial\gamma}\right]_{\gamma=0} = F(p)^2. \tag{7}$$

where the first equality follows by substituting $\widetilde{\varepsilon} = \gamma F(\widetilde{\varepsilon}) + p$ and the last equality follows because at $\gamma = 0$, $\widetilde{\varepsilon} = p$. Looking at regime 2, $\frac{dCS_{2,users}}{d\gamma} = F(p)$. Since $0 < F(p) < 1$, it follows that when γ is positive yet very small, $\frac{dCS_{1,users}}{d\gamma} < \frac{dCS_{2,users}}{d\gamma}$. Since for $\gamma = 0$, $CS_{1,users} = CS_{2,users}$, it follows that for γ values slightly higher than 0, $CS_{2,users} > CS_{1,users}$.

To prove that for high values of γ, $CS_{1,users} > CS_{2,users}$, we evaluate consumer surplus in both regimes at the other extreme: $\gamma = 1-p$. Under regime 1, when $\gamma = 1-p$, all users join the platform and $\widetilde{\varepsilon} = 1$. Substituting $\gamma = 1-p$ into $CS_{1,users}$ and $CS_{2,users}$, we get that,

$$CS_{1,users}\big|_{\gamma=1-p} = F(1)(1-p) + \int_0^1 (p-\varepsilon)f(\varepsilon)d\varepsilon,$$

$$CS_{2,users}\big|_{\gamma=1-p} = F(p)(1-p) + \int_0^p (p-\varepsilon)f(\varepsilon)d\varepsilon.$$

It follows that when $\gamma = 1-p$:

$$\Delta CS_{users} \equiv CS_{1,users}\big|_{\gamma=1-p} - CS_{2,users}\big|_{\gamma=1-p} = (1-F(p))(1-p) + \int_p^1 (p-\varepsilon)f(\varepsilon)d\varepsilon$$

$$= (1 - F(p))(1 - p) + p(1 - F(p)) - \int_p^1 \varepsilon f(\varepsilon) d\varepsilon$$

$$= (1 - F(p)) - \int_p^1 \varepsilon f(\varepsilon) d\varepsilon.$$

From integrating by parts,

$$\int_p^1 \varepsilon f(\varepsilon) d\varepsilon = \varepsilon F(\varepsilon)\Big|_p^1 - \int_p^1 F(\varepsilon) d\varepsilon = 1 - pF(p) - \int_p^1 F(\varepsilon) d\varepsilon. \tag{8}$$

Substituting (8) into ΔCS_{users},

$$\Delta CS_{users} = \int_p^1 F(\varepsilon) d\varepsilon - F(p)(1 - p).$$

To show that this difference is positive, note that evaluated at $p = 0$, ΔCS_{users} is positive because the second term vanishes. Moreover, ΔCS_{users} is decreasing with p because $\frac{d\Delta CS_{users}}{dp} = -(1 - p)f(p) < 0$. Finally, evaluated at $p = 1$, $\Delta CS_{users} = 0$ because both terms vanishes. We therefore have that when $\gamma = 1 - p$, $\Delta CS_{users} > 0$ for all $0 < p < 1$.

***part** iii*

We first note that both $CS_{1,users}$ and $CS_{2,users}$ are increasing with γ, because (7) indicates that $\frac{dCS_{1,users}}{d\gamma} = F(\widetilde{\varepsilon})^2 + \gamma F(\widetilde{\varepsilon}) f(\widetilde{\varepsilon}) \frac{\partial \widetilde{\varepsilon}}{\partial \gamma} > 0$ and it is straightforward to see that $\frac{dCS_{2,users}}{d\gamma} = F(p) > 0$. Because at $\gamma = 0$, $CS_{1,users} = CS_{2,users}$, and $CS_{2,users}$ is linear in γ, it suffices to show that $\frac{d^2 CS_{1,users}}{d\gamma^2} > 0$, which holds if $CS_{1,users}$ is convex. Note that convexity of $CS_{1,users}$ is a sufficient but not a necessary condition for uniqueness of the threshold to hold.

***parts** iv*

We showed that for high values of γ , $CS_{1,users} > CS_{2,users}$, and that $\pi_{1,users} > \pi_{2,users}$ for all values of γ. It follows that for high values of γ: $W_{1,users} > W_{2,users}$. At the other extreme, for $\gamma = 0$, we know that $CS_{1,users} = CS_{2,users}$ and $\pi_{1,users} = \pi_{2,users}$. It follows that when $\gamma = 0$, $W_{1,user} = W_{2,user}$. When $W_{1,users}$ is convex with no inflection points, $\frac{d^2 W_{1,users}}{d\gamma^2} > 0$. Given that $\frac{d^2 W_{2,users}}{d\gamma^2} = 0$, it follows that there exists a unique threshold $\gamma_{1,2}^W \geq 0$ such that $W_{1,users} > W_{2,users}$, if $\gamma > \gamma_{1,2}^W$. As we show below, when α is small and γ is close to 0, $\frac{dW_{1,users}}{d\gamma} < \frac{dW_{2,users}}{d\gamma}$, while for larger values of α, when γ is close to 0, $\gamma_{1,2}^W = 0$. It follows then that for $\gamma < \gamma_{1,2}^W$, $W_{1,users}$ is smaller or equal to $W_{2,users}$.

***part** v*

When $\alpha = 0$, $W_{1,user} = CS_{1,user}$ and $W_{2,user} = CS_{2,user}$ and thus $\gamma_{1,2}^W = \gamma_{1,2}^{CS}$. Since at $\alpha = 0$ and γ values close to 0, $CS_{2,users}$ is strictly higher than $CS_{1,users}$, it follows that $\gamma_{1,2}^{CS} > 0$.

As α increases, since $\frac{dCS_{1,users}}{d\alpha} = 0$, $\gamma_{1,2}^{CS}$ remains constant. To show that as α increases, $\gamma_{1,2}^{W}$ decreases, we have: $\frac{d(W_{1,users} - W_{2,users})}{d\alpha} = F(\widetilde{\varepsilon}) - F(p) > 0$, where the inequality follows because $\widetilde{\varepsilon} > p$. This implies that if there is a unique $\gamma_{1,2}^{CS}$, there is a unique $\gamma_{1,2}^{W}$, such that $\gamma_{1,2}^{W} = \gamma_{1,2}^{CS}$ for $\alpha = 0$ and $\gamma_{1,2}^{W}$ is decreasing in α while $\gamma_{1,2}^{CS}$ is constant in α. Consequently, $0 \leq \gamma_{1,2}^{W} < \gamma_{1,2}^{CS}$ for $\alpha > 0$.

Finally, to show that for a sufficiently high α, $\gamma_{1,2}^{W} = 0$, we look at $\frac{dW_{1,users}}{d\gamma}$ and $\frac{dW_{2,users}}{d\gamma}$ evaluated at $\gamma = 0$. Since $W_{1,users} > W_{2,users}$, if $\gamma > \gamma_{1,2}^{W}$, it suffices to show that evaluated at $\gamma = 0$, $\frac{dW_{1,users}}{d\gamma} > \frac{dW_{2,users}}{d\gamma}$.

$$\frac{dW_{1,users}}{d\gamma} \mid_{\gamma=0} = F^2(p) + \alpha f(p)\widetilde{\varepsilon}', \qquad \frac{dW_{2,users}}{d\gamma} \mid_{\gamma=0} = F(p). \tag{9}$$

It follows that $\frac{d(W_{1,users} - W_{2,users})}{d\gamma} \mid_{\gamma=0} = F^2(p) + \alpha f(p)\widetilde{\varepsilon}' - F(p)$ and is positive if

$$\alpha > \frac{F(p)(1 - F(p))}{\frac{d\varepsilon}{d\gamma} f(p)}. \tag{10}$$

We can further simplify condition (10) by using the implicit function theorem. Let $\Gamma(\gamma, \varepsilon) \equiv \gamma F(\varepsilon) + p - \varepsilon = 0$. Hence,

$$\frac{d\varepsilon}{d\gamma} = \frac{\frac{d\Gamma}{d\gamma}}{\frac{d\Gamma}{d\varepsilon}} = -\frac{F(\varepsilon)}{\gamma f(\varepsilon) - 1} \mid_{\gamma=0} = F(p).$$

Substituting this into eq. (10), we get that $\frac{d(W_{1,users} - W_{2,users})}{d\gamma} \mid_{\gamma=0} > 0$ if: $\alpha > \frac{1 - F(p)}{f(p)}$, in which case $\gamma_{1,2}^{W} = 0$.$\blacksquare$

Proof of Proposition 4

The platform prefers regime 1 over regime 2 because $\pi_{1,data} - \pi_{2,data} = \alpha(G(\widetilde{\theta}) - G(p)) > 0$, which holds because $\widetilde{\theta} > p$ for all γ and α. Consumers prefer regime 2 because:

$$CS_{2,data} - CS_{1,data} = \int_0^{\widetilde{\theta}} \theta g(\theta) d\theta - \int_0^{p} \theta g(\theta) d\theta > 0,$$

where the inequality follows because $\widetilde{\theta} > p$. It follows then that users prefer regime 2 while the platform prefers regime 1, for all value of γ and α.

Next, consider welfare. Let:

$$\Delta W_{data} \equiv W_{2,data} - W_{1,data} = \int_0^{\widetilde{\theta}} \theta g(\theta) d\theta - \int_0^{p} \theta g(\theta) d\theta + \alpha(G(p) - G(\widetilde{\theta}))$$

$$= \int_p^{\widetilde{\theta}} \theta g(\theta)d\theta - \alpha \left(G(\widetilde{\theta}) - G(p)\right)$$

$$= \int_p^{\widetilde{\theta}} \theta g(\theta)d\theta - \alpha \left(\int_0^{\widetilde{\theta}} g(\theta)d\theta - \int_0^p g(\theta)d\theta \right) = \int_p^{\widetilde{\theta}} \theta g(\theta)d\theta - \alpha \left(\int_p^{\widetilde{\theta}} g(\theta)d\theta \right)$$

$$= \int_p^{\widetilde{\theta}} (\theta - \alpha)\, g(\theta)d\theta. \tag{11}$$

Evaluating at $\alpha \to 0$, the gap ΔW_{data} is positive because $\widetilde{\theta} > p$, implying that when α is small, $W_{2,data} > W_{1,data}$ for all γ. In contrast, when $\alpha \to 1$ the gap ΔW_{data} is negative for all γ because:

$$\Delta W_{data}\big|_{\alpha=1} = \int_p^{\widetilde{\theta}} (\theta - 1)\, g(\theta)d\theta \leq \int_p^{\widetilde{\theta}} \left(\theta - \widetilde{\theta}\right) g(\theta)d\theta < 0,$$

where the first inequality follows because $\widetilde{\theta} < 1$ and the last inequality follows because $p < \widetilde{\theta}$. Moreover, notice that ΔW_{data} is decreasing in α. Therefore, comparing $W_{2,data}$ with $W_{1,data}$ as functions of γ yields that if α is small, $W_{2,data}$ is higher than $W_{1,data}$ for all values of γ, as illustrated in panel (b) of Figure 2. When α is high, $W_{2,data}$ is lower than $W_{1,data}$ for all values of γ, as illustrated in panel (c) of Figure 2. For intermediate values of α, there is a unique intersection point between $W_{2,data}$ and $W_{1,data}$ at some γ, as illustrated in panel (a) of Figure 2. This intersection point exists and is unique because ΔW_{data} is decreasing in α. It is left to verify that in this intersection point $W_{2,data}$ crosses $W_{1,data}$ "from below", as illustrated in the figure. That is, evaluated at this intersection point, $\frac{d\Delta W_{data}}{d\gamma} > 0$. To this end, let $\alpha_{1,2}(\gamma)$ denote the solution to $\Delta W_{data} = 0$. It has to be that $\widetilde{\theta} > \alpha_{1,2}(\gamma) > p$. The effect of γ on ΔW_{data} is:

$$\frac{d\Delta W_{data}}{d\gamma} = (\widetilde{\theta} - \alpha)g(\theta)\frac{d\widetilde{\theta}}{d\gamma}.$$

We then have that $\frac{d\Delta W_{data}}{d\gamma}\Big|_{\alpha=\alpha_{1,2}(\gamma)} > 0$ because $\widetilde{\theta} > \alpha_{1,2}(\gamma)$ and $\frac{d\widetilde{\theta}}{d\gamma} > 0$. This implies that when $\alpha = \alpha_{1,2}(\gamma)$ such that $W_{2,data} = W_{1,data}$, an increase in γ result in $W_{2,data} > W_{1,data}$ while a decrease in γ result in $W_{2,data} < W_{1,data}$ as shown in panel (a) of Figure 2. Because $W_{2,data}$ crosses $W_{1,data}$ "from below" for intermediate values of α, there are two thresholds $\underline{\alpha}_{1,2}^W$ and $\overline{\alpha}_{1,2}^W$, that are the solutions to $\Delta W_{data} = 0$ evaluated at $\gamma = 0$ and $\gamma = \gamma_{data}$, respectively, such that for $\alpha \in [\underline{\alpha}_{1,2}^W, \overline{\alpha}_{1,2}^W]$, there is a threshold, $\gamma_{1,2}^W$, such that $W_{1,data} > W_{2,data}$ iff $\gamma < \gamma_{1,2}^W$ (part (*i*) of Proposition 4). For $\alpha \in [0, \underline{\alpha}_{1,2}^W]$, $W_{2,data} > W_{1,data}$ for all γ (part (*ii*)) and for $\alpha \in [\overline{\alpha}_{1,2}^W, 1]$, $W_{1,data} > W_{2,data}$ for all γ (part (*iii*)).■

Appendix B: Heterogenous users *and* data

In this appendix, we combine the two types of heterogeneities – user and data – in order to study how the interaction of both heterogeneities affects market outcomes under the different data regimes. We find that the intuition from the cases where the two heterogeneities are analyzed in isolation applies. Specifically, when the user market is not fully covered, the platform's behavior resembles the heterogenous users case. Once the user market is fully covered, the platform focuses on commercializing more data items and its behavior resembles the heterogeneous data case.

Consider a continuum of users and data items, each with total mass of 1. Users and data items are heterogeneous, so both ε and θ are distributed between $[0, 1]$ according to $G(\theta)$ and $F(\varepsilon)$, respectively. Given the complexity of the model with two types of heterogeneity, and in order to keep the analysis simple and clear, we further assume that $G(\theta)$ and $F(\varepsilon)$ follow a uniform distribution and that users bear the idiosyncratic cost from the commercialization of their data, ε, regardless of the number of data items the platform chooses to commercialize. For example, the idiosyncratic component of the user's disutility, ε, can represent the user's identity, that once revealed to advertisers, inflicts a costs on the user in addition to the costs of each data item, θ. Hence, if the platform chooses to commercialize $\widetilde{\theta}$ data items, the disutility user of type ε bears from the commercialization of their data is: $\varepsilon + \int_0^{\widetilde{\theta}} \theta d\theta$. We further assume that $\gamma + p < 3/2$, to ensure that the market cannot be fully covered in both users and data. That is, if the platform commercializes all data items, some users will not join it.

Below we analyze regime 1 and regime 2, in turn, and then compare the two regimes.

Regime 1

As before, the platform bears no cost for data collection, yet data collected provides users with positive p and γ, thus, the platform collects all data items but may choose to commercialize data items up to $\widetilde{\theta}_1 \in [0, 1]$. Given the platform's choice of $\widetilde{\theta}_1$, users choose whether to join the platform or stay out and earn 0. Hence, there is a threshold, $\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$, such that users with $\varepsilon < \widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ join the platform, where $\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ is the solution to

$$\gamma\varepsilon + p - \varepsilon - \int_0^{\widetilde{\theta}_1} \theta d\theta = 0 \quad \Longleftrightarrow \quad \widetilde{\varepsilon}_1(\widetilde{\theta}_1) = \min\left\{ \frac{2p - \widetilde{\theta}_1^2}{2(1-\gamma)}, 1 \right\}.$$

The platform faces the tradeoff that the more it collects data from each user (increases $\widetilde{\theta}_1$), the less users join it ($\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ decreases). The platform, thus, sets $\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ as to maximize

$\pi_{1,both} = \alpha \times \widetilde{\theta}_1 \times \widetilde{\varepsilon}_1(\widetilde{\theta}_1)$. Hence:

$$\widetilde{\theta}_1 = \begin{cases} \sqrt{\frac{2}{3}p}, & \text{if } \gamma + p < 1, \\ \max\Big\{\sqrt{\frac{2}{3}p}, \sqrt{2}\sqrt{\gamma + p - 1}\Big\}, & \text{if } \gamma + p \geq 1. \end{cases} \tag{12}$$

$$\widetilde{\varepsilon}_1(\widetilde{\theta}_1) = \begin{cases} \min\Big\{\frac{2}{3}\frac{p}{1-\gamma}, 1\Big\}, & \text{if } \gamma < 1, \\ 1, & \text{if } \gamma \geq 1. \end{cases} \tag{13}$$

Figure 4 illustrates $\widetilde{\theta}_1$ and $\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ as a function of γ. Notice that starting from $\gamma = 0$, the market has partial user and data coverage. As γ increases, the platform takes advantage of the higher public benefit to increase users coverage. Once the market becomes fully covered in users, the platform takes advantage of further increases in γ to increase the data coverage. Hence, the driving force of the platform's optimization when γ is small is the presence of heterogeneous users, while for high values of γ, it is the heterogeneity in data items that determines the platform's strategy.

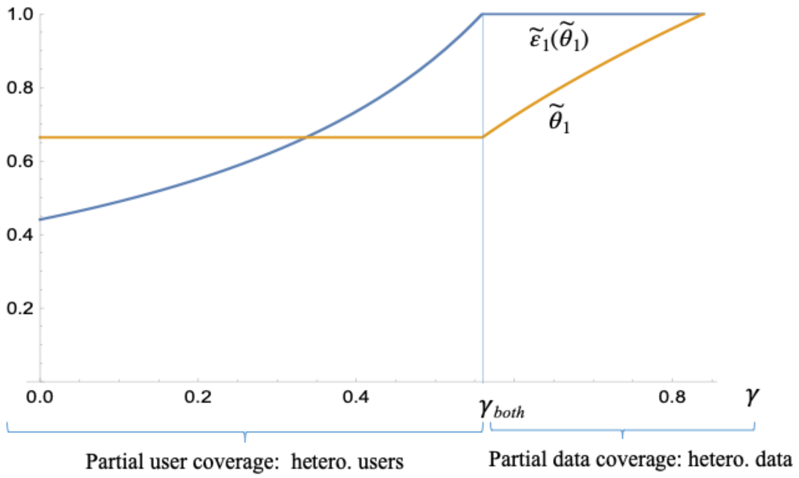


Figure 4: $\widetilde{\theta}_1$ and $\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ as a function of γ ($p = 2/3$)

Social welfare in regime 1 is:

$$W_{1,both} = \int_0^{\widetilde{\varepsilon}_1(\widetilde{\theta}_1)} \left(\gamma\widetilde{\varepsilon}_1(\widetilde{\theta}_1) + p - \varepsilon - \int_0^{\widetilde{\theta}_1} \theta d\theta \right) d\varepsilon + \alpha \times \widetilde{\theta}_1 \times \widetilde{\varepsilon}_1(\widetilde{\theta}_1),$$

where $\widetilde{\theta}_1$ and $\widetilde{\varepsilon}_1(\widetilde{\theta}_1)$ are given by (12) and (13), respectively and the first term in $W_{1,both}$ is consumer surplus while the second term is the platform's profits.

Regime 2

Under regime 2, the platform cannot contingent users' participation with sharing data. Moreover, users can choose to share certain data items and not others. For each data item, the platform declares whether it plans to commercialize it, should the user consent to sharing it. For each commercialized data item the user agrees to share, the user gains p, yet bears the disutility θ. On top of it, the user bears the disutility ε, unless the user declines to share any commercialized data item. If a data item is not commercialized, the user will always agree to share it because doing so provides the user with p at no costs.

The platform commercializes only data items with $\theta \in [0, p]$ because users never agree to share commercialized data with $\theta > p$. In equilibrium, all users join. Users with $\widetilde{\varepsilon}_2$ agree to share the commercialized data items with $\theta \in [0, p]$, where $\widetilde{\varepsilon}_2$ is the solution to:[15]

$$-\varepsilon + \int_0^p (p - \theta) d\theta = 0, \quad \Longleftrightarrow \quad \widetilde{\varepsilon}_2 = \frac{p^2}{2}.$$

Moreover, all users share the non-commercialized data items with $\theta \in [p, 1]$.

Total public data collected is $(1 - p) + \widetilde{\varepsilon}_2 p$, out of which total commercialized data is $\widetilde{\varepsilon}_2 p$. Welfare is:

$$\begin{aligned} W_{2,both} = & \int_0^{\widetilde{\varepsilon}_2} \left(\gamma \left((1 - p) + \widetilde{\varepsilon}_2 p \right) + p - \varepsilon - \int_0^p \theta d\theta \right) d\varepsilon \\ & + \int_{\widetilde{\varepsilon}_2}^1 \left(\gamma \left((1 - p) + \widetilde{\varepsilon}_2 p \right) + p(1 - p) \right) d\varepsilon + \alpha \widetilde{\varepsilon}_2 p, \end{aligned}$$

where the first term is the surplus of users with $\varepsilon < \widetilde{\varepsilon}_2$ who agree to share all data, out of which $\theta \in [0, p]$ is commercialized, the second term is the surplus of users with $\varepsilon > \widetilde{\varepsilon}_2$ who share only the non-commercialized data with $\theta \in [p, 1]$ and the last term is the platform's profit.

Comparison

Figure 5 presents welfare under the two regimes. For intermediate values of α, the comparison is qualitatively identical to a combination between the heterogeneous users and the heterogeneous data cases. Specifically, for low values of public benefit, the graph resembles the heterogeneous users case (see Figure 1). That is, when γ is low, the main driving force is heterogeneity in users and as long as the market is not fully covered, the platform focuses on attracting more users. In this case, there is a threshold in γ such that regime 1 performs

[15] We verified that the platform will not want to commercialize less data items than $\theta \in [0, p]$, even though doing so would increase the number of users that agree to give commercialized data.

better than regime 2 when γ is above this threshold. The intuition is the same as in the heterogeneous users case: the platform takes advantage of the increase in γ in order to attract more users to join and give data for public benefit, which enhances welfare. Once we reach full user coverage, at γ_{both}, the platform responds to further increases in γ by bundling more data items. In this case, the figure resembles the heterogeneous data case (see Figure 2) and there is a threshold in γ such that regime 2 outperforms regime 1 when γ is above this threshold. The intuition is the same as in the heterogeneous data case. When the user market is fully covered, under regime 1, an increase in γ essentially unfolds the bundling effect – allowing the platform to commercialize more data. As in the heterogeneous data case, the platform becomes aggressive and commercializes too many data items.

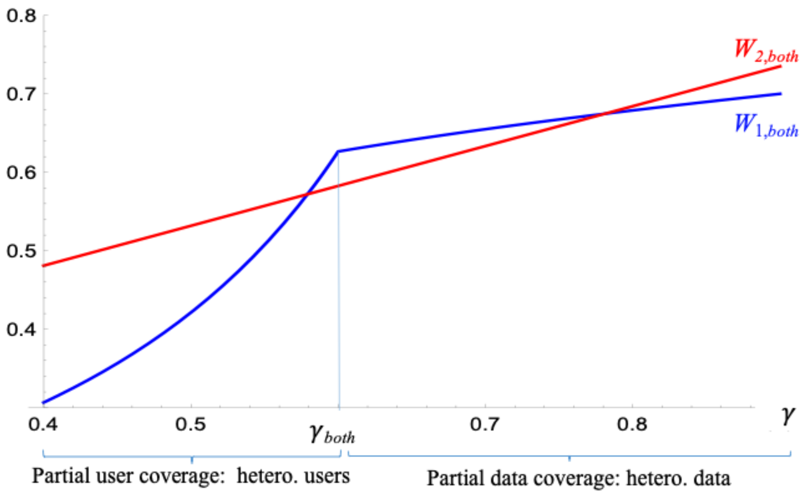
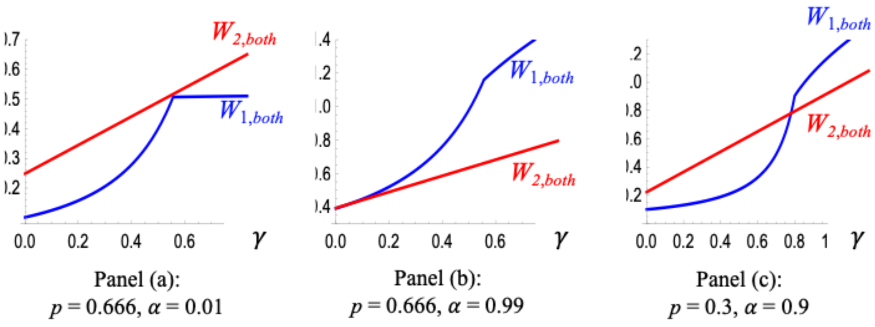


Figure 5: Welfare as a function of γ for a uniform $F(\varepsilon)$ and $G(\theta)$ ($p = 2/3$ and $\alpha = 0.2$)



Panel (a): $p = 0.666$, $\alpha = 0.01$

Panel (b): $p = 0.666$, $\alpha = 0.99$

Panel (c): $p = 0.3$, $\alpha = 0.9$

Figure 6: Welfare as a function of γ for a uniform $F(\varepsilon)$ and $G(\theta)$

Figure 6 further shows that for low values of α (panel (b)), just like in the heterogeneous data case, regime 2 is superior to regime 1 for all values of γ, because regime 1 commercializes more data than regime 2. We find that this effect dominates when both heterogeneities are present. Likewise, for high values of α (panel (b)), as in the heterogeneous data case, regime 1 is superior to regime 2 for all values of γ (regime 2 commercializes too little data) and this effect dominates when both heterogeneities are present. Finally, when p is small, as shown in panel (c), the effect of heterogeneous users dominates for all values of γ and the comparison is qualitatively similar to heterogeneous users case.