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

Sarit Markovich† and Yaron Yehezkel‡

October 2021

Abstract

We consider a platform that collects data from users. Data has commercial benefit to the platform, personal benefit to the user, and public benefit to other users. We ask whether the platform, or users, should have the right to decide which data the platform commercializes. We find that when users differ in their disutility from the commercialization of their data and the public benefit of data is high (low), it is welfare enhancing to let the platform (users) control the data. In contrast, when heterogeneity is in the disutility from the commercialization of different data items, it is welfare enhancing to let users (the platform) control the data when the public benefit of data is high (low). Furthermore, we find that allowing the platform to compensate users for their data is not always welfare enhancing and competition does not necessarily result in the efficient outcome.

JEL Classification: L1

Keywords: data regulation, network externalities, platform competition

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

*For helpful comments and discussions we thank Ido Eisdorfer and participants at the 2021 IIOC. 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)

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 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 platforms. This raises the question of who should own the property rights over users' data? On 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. The first feature is that data has three potential benefits. First, data provides personal benefits. For example, when a driver uses a navigation app and agrees to let the app track the route, the data collected can help direct the driver to un-congested routes. Second, the same data provides the platform with a commercial benefit. The navigation app, in our example, can sell the driver's data to advertisers. Third, 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. Media streaming platforms can utilize the data on listeners' habits in order to improve the recommendation to other listeners. 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 provide data themselves.

The second main feature of our model is that the platform collects multiple data

[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.

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 of our model 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.

We study three 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. The third regime does not allow the platform to contingent users' participation *and* data collection on their consent to the commercialization their data. In order to incentivize users to allow the platform to commercialize their data, we allow the platform to compensate users for selling their data.

We find that the different benefits of data create a market inefficiency. 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. Specifically, the optimal regime depends on the magnitude of the public benefit of data and on whether the

market is characterized by heterogeneous users (i.e., users differ in their disutility from commercializing their data), or heterogeneous data items (i.e., data items differ in the disutility that commercializing them inflicts on homogeneous users). Consider first the case of heterogeneous users. In this case, we find that both regimes are identical when data does not have any public benefit. When the public benefit of data is high, it is welfare enhancing to let the platform control the data, while giving the users control on their data is welfare enhancing when the public benefit of data is low. These results highlights 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.

Interestingly, the third regime which in essence provides users with all the control over their data is not always welfare enhancing. In this regime all users join and share data for public and private benefits, but the platform needs to pay users for agreeing to share data for commercial benefit. We find that doing so achieves the first best under heterogeneous data. Still, once users are heterogeneous, this regime may underperform the first two regimes that we consider, especially when the commercial and pubic benefits of data are high.

Moreover, our results show that, even under competition, the market does not necessarily achieve the efficient outcome. Studying platform competition with an incumbent that enjoys focality advantage and an entrant with a quality advantage, we find that competition does not motivate either platform to implement the welfare-maximizing regime. Indeed, the entrant has stronger incentives than the incumbent to give users control over data. Yet, the entrant may win the market adopting a regime that gives users control over data when it is welfare enhancing to give the platform the control over data.

Understanding the effects of platforms' data policies on profits and social welfare has important implications for the ongoing debate on the necessity of 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.

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

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 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 value of data and the type of heterogeneity in the market.

Our results suggest that when the platform owns the data, even under competition, a platform would not permit users to choose whether to commercialize their data for at least some data items. When users are heterogeneous, giving the platform the control over users' data results in under-participation in the platform and in less data collected for public benefit. When data items vary in users' disutility from commercializing them, we show that the platform may "bundle" data items, forcing users to either agree that

[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

[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

the entire "bundle" of data items is commercialized, or they stay out of the platform. Still, giving users the property right over their data is not always welfare enhancing because users ignore the public benefit of data as well as its commercial value. In particular, it is welfare enhancing to let the platform control the data when either users are heterogeneous and the public benefit of data is high, or when data items are heterogeneous and the public benefit of data is low. Compensating users for their data results in the platform internalizing users' disutility from the commercialization of their data. While in the case of heterogeneous users this leads to the first best, in the case of heterogeneous data, the platform under-collects data for commercial benefit.

We should emphasize that the question of who should control our data is also – perhaps foremost – a question of social morality. Is it moral to allow a platform to collect our personal data items? 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 value 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.[7]

Literature Review

There is a growing literature on data regulation and its effect on welfare, consumer surplus, and firm profitability. 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. 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

[7]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.

optimal policy of whether to give the platform or users the right over 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. 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.

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.

The two papers closest to ours are Fainmesser et al. (2020) and Economides and Lianos (2020). 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.

Our paper makes two main contributions to the above literature. First, we highlight the role of the public benefit of data, where users benefit from data collected from other users. While the public benefit of data exhibit 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.[8]

The second 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 different data regimes.

2 The Model

We start with a general framework for studying data collection and commercialization. Then, to detangle some of the effects, we introduce some simplifying assumptions. Consider a market with a monopolistic platform and a set of potentially heterogeneous users. We start by describing users' preferences and then move to the platform's strategies.

Users' preferences

Consider a set N of small users, potentially heterogeneous, with a total mass of one. There is a finite set of *data items,* $\Theta = \{1, 2, 3, \ldots, \overline{\Theta}\}$, that the platform can potentially collect from each user. If collected, a data item may provide users with a certain benefit,

[8]For the literature on network effects with coordination problems, see Katz and Shapiro (1986), Caillaud and Jullien (2001; 2003), Jullien (2011), Hałaburda and Yehezkel (2013; 2016; 2019), Hałaburda et al. (2020), Biglaiser and Crémer (forthcoming), Markovich and Yehezkel (forthcoming).

and if commercialized, can provide value for the platform. For example, in the context of a fitness tracker such as Fitbit, item 1 can represent the user's location, item 2 can represent the user's number of steps, item 3 can represent the users' heart rate, and so on. In the context of navigation app such as Waze, data item 1 can represent the driver's location, data item 2 can represent the time it takes the driver to navigate from one location to the other, data item 3 can represent the frequency in which the driver uses a certain route.

Suppose that each data item $\theta \in \Theta$ provides the following benefits. First, if collected, user $i \in N$ enjoys a *private value* $p_{i\theta}$ when sharing data item θ 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. When drivers share their current location with a navigation app, the app can notify the driver about unexpected congestion, and divert the driver to an alternative route. When users share their location with Google, Google can provide these users with more helpful search results.

The data the platform collects on users that share their 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. A driver's location may help the navigation app direct other drivers to a less congested route. Finally, Google uses search data to improve its algorithm that even users that use the Incognito mode can enjoy.[9] We refer to this as the *public benefit* from data collected by the platform on an individual user and denote by $\gamma_{i\theta}$ the public benefit that the collection of data item θ on user i provides to each user that joins the platform.[10]

A platform may choose to commercialize the data that it collects. The platform can sell data to advertisers, third-party application developers, or to other platforms. Let $\alpha_{i\theta}$ denote the platform's profit from selling data item θ collected from user i. Users, however, have disutility when their data is commercialized that we denote by $k_{i\theta}$. Users may feel discomfort when their personal data, such as their heart rate, driving habits, or online search inquiries are shared with other commercial firms. Moreover,

[9]Incognito mode or private browsing mode is a privacy feature offered by some browsers where browsing history and local data associated with the session are not saved.

[10]Note that our setting assumes separability across the different data items. A more general formulation would allow dependency in the value the different data items provide. For example, a user that provides Google with their age and location could get better (or potentially worse) search results than the sum of the separate values.

advertisers may overload users with advertising and pop-ups. We assume that users bear this disutility only if their data is shared with other firms. So if the platform commercializes data item θ by user i, the platform gains the profit $\alpha_{i\theta}$ while the user bears a cost of $k_{i\theta}$, in which case the user's total private benefit from data item θ is $p_{i\theta} - k_{i\theta}$, which can be positive or negative.[11]

The above payoffs characterize an N-by-4 data matrix for each user i that specifies, for each data item θ, vectors of $\{\gamma_{i\theta}, p_{i\theta}, k_{i\theta}, \alpha_{i\theta}\}$:

$$\begin{Bmatrix} \gamma_{i1} & p_{i1} & k_{i1} & \alpha_{i1} \\ \gamma_{i\theta} & p_{i\theta} & k_{i\theta} & \alpha_{i\theta} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{i\overline{\Theta}} & p_{i\overline{\Theta}} & k_{i\overline{\Theta}} & \alpha_{i\overline{\Theta}} \end{Bmatrix}.$$

A user's total benefit is the sum over the benefits from each data item collected on the user, and the sum of the public benefits from the data items that the platform collects from other users.

Notice that this framework allows for two types of heterogeneity. First, heterogeneity between users: users may differ in the benefits that their data provide. Second, heterogeneity between data items: for a specific user, data items may vary in their benefits. In Section 8, we solve the model given the above matrix and both types of heterogeneity, when the market is fully covered: all users join a platform that collects all data items. In the case a partially covered market discussed in sections 3 - 7, we distinguish between two extreme cases of the model above, that we refer to as "heterogeneous users" and "heterogeneous data." In both cases we assume, for simplicity, that all data items by all users provide the same social, private, and commercial benefit: $\gamma_{i\theta} = \gamma > 0$, $p_{i\theta} = p > 0$ and $\alpha_{i\theta} = \alpha > 0$ for all i and θ. In order to focus on interior solutions, in what follows, we assume that $\gamma \leq 1 - p$, $p < 1$ and $\alpha < 1$.

Our two cases focus on variations in $k_{i\theta}$. In general, heterogeneity in the disutility from selling data may be of two types. First, for a given data item, users' disutility from the commercialization of that data item may vary. Some users may be more sensitive to their privacy than others. Alternatively, users may be identical in their preferences, however, may bear different disutility from the commercialization of different

[11] It is possible to assume that collecting data inflicts a discomfort on the user even when this data is not commercialized. In such a case, $p_{i\theta}$ represents the personal benefit minus personal disutility. When the latter is higher than the former, we have that $p_{i\theta} < 0$.

data items. 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. Facebook users' disutility from sharing their photos may be lower than their disutility from the commercialization of the amount of time they spend on Facebook. Hence, we assume that $k_{i\theta} = k$, where k is distributed $k \in [0, 1]$ with a probability distribution function $f(k)$ and a cumulative distribution function $F(k)$. In the case of heterogeneous users, suppose that there is only one data item, and the variations in k emerges because users differ in their disutility from commercializing it. In the second case of heterogeneous data items, suppose that all users are identical (i.e., they have the same $k_{i\theta}$), but there is a continuum of data items that differ in the disutility that they inflict on users. While, in general, both types of heterogeneity exist, we show below that in a partially covered market, they differ dramatically in the resulting behavior and thus in terms of profits, consumer surplus, and overall welfare. To highlight these differences and the intuition behind them, we analyze each case in isolation.

Given the above assumptions, the outcome that maximizes social welfare is identical under our two interpretations of variation in k. With either heterogeneous users or heterogeneous data, it is optimal to serve all users and collect all data for public and private benefit, and commercialize data with $\alpha > k$. Hence, social welfare is:

$$W^* = \gamma + p - \int_0^{\alpha} kf(k)dk + \alpha F(\alpha). \tag{1}$$

Platform's strategy

The platform profits from commercializing data and potentially from charging users. 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:

Regime 1: *the platform controls the data.* The platform can collect and commercialize all data items of all users that choose to join the platform. That is, the platform can

contingent platform participation with data collection and commercialization. For any data item $\theta \in \Theta$, 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 needs the users' consent to collect and commercialize their data. Users choose which data items they wish to share with the platform. Users can join the platform and choose not to allow the platform to collect and/or commercialize any of their data yet still enjoy the public benefit of data collected on other users. For each data item, the platform informs users whether, given the user's consent, it would collect this data item and whether it would commercialize 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 with 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. Platforms can incentivize users to agree to commercialize their data by offering users compensation for the right to commercialize their data.

In our analysis below, we make several simplifying assumptions. First, we assume that the data set is common knowledge. That is, users know which data items the platform can extract from them. For future research, it is possible to extend this model the case where the platform has private information concerning the set of data items that can be collected from each user. Second, we assume that the platform can credibly commit to collecting only certain data items, and not collecting others. Moreover, the platform can commit not to commercialize certain data items. Finally, we assume that users' data matrixes are common knowledge, but users are anonymous. That is, the

platform cannot collect different data items from different users, and cannot commit to selling data items of some users but not others.

Next, we study each regime in turn. For each regime we start with heterogeneous users case and then analyze heterogeneous data.

3 Regime 1: The platform controls the data it collects

Recall that under regime 1, regulation permits for the platform to contingent participation in the platform with users' consent for the collection and commercialization of their data. While the platform may give users the option to choose whether to give data, as we show below, in the case of one data item the platform will never profit from doing so.

Heterogenous users

Consider a market with one data item and a mass 1 of users. Users are heterogeneous in terms of the disutility from the commercial use of their data, k. The platform contingents participation with data provision. Since there is only one data item and the platform cannot discriminate across users, the platform collects data from all users that join it, and commercializes it. Users are aware of this policy and can choose whether to join the platform or stay out and earn 0. Suppose that $n_1 \leq 1$ users join. Then, the platform collects the data item from these users, and provides each user with a total public benefit of γn_1. Users take the total public benefit, γn_1, as given, when choosing whether to join the platform. The utility for a user from joining the platform is then $\gamma n_1 + p - k$ and 0 if they choose not to join. Given that the utility deceases in k, there is a threshold, $\widetilde{k}$, such that users with $k \in [0, \widetilde{k}]$ join and give data, while users with $k \in [\widetilde{k}, 1]$ stay out. Therefore, $n_1 = F(\widetilde{k})$. Solving for users' decision whether to join the platform, the marginal user sets:

$$\gamma F(\widetilde{k}) + p - \widetilde{k} = 0 \iff \gamma F(\widetilde{k}) = \widetilde{k} - p. \tag{2}$$

Equation 2 provides interesting results with respect to the effect of the existence of a public benefit from data on users' behavior. It is easy to see that when $\gamma = 0$, $\widetilde{k} = p$; that is, users join the platform as long as the private benefit from sharing data, p, is larger than their disutility from sharing it, k. Once the public benefit of

data becomes positive, even users with $p < k$ join the platform as users want to enjoy the public benefit, $\gamma F(\widetilde{k})$, 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 on data collection) A solution to equation (2) always exists and is unique when $F(k)$ does not exhibit an extreme unimodal distribution. Moreover:*

(i) when data has no public benefit, $\gamma = 0$, $\widetilde{k} = p$ and users with $k \in [0, p]$ 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 $\gamma > 0$, $\widetilde{k} > p$ and is increasing with γ;

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

In what follows, we assume that $F(k)$ is not "too" unimodal, such that there is a unique solution to (2). We comment on this assumption in remark 1 below.[12]

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 one data item 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 one-to-one 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:

$$CS_{1,users} = \gamma \cdot F(\widetilde{k}) \cdot F(\widetilde{k}) + \int_0^{\widetilde{k}} (p - k) f(k) dk, \qquad \pi_{1,users} = \alpha F(\widetilde{k}), \tag{3}$$

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

[12] We note that our results hold even when there are multiple solutions to (2), because all of these solutions have the qualitative features that we discuss below. We comment on the case of an extreme unimodal $F(k)$ in remark 1 below.

Remark on multiple equilibria. Proposition 1 shows that there is a unique equilibrium when $F(k)$ is not too unimodal. In the proof, we show that in the case of a unimodal distribution, it is possible to have two solutions to (2), at $\widetilde{k}'$ and $\widetilde{k}''$, where $p < \widetilde{k}' < \widetilde{k}'' < 1$. In this case, there are two equilibria that depend on users' beliefs. In both equilibria, at least $F(\widetilde{k}_1)$ users join the platform and share data. If users expect that users with $k \in [\widetilde{k}', \widetilde{k}'']$ do not join, then in equilibrium users with $k \in [\widetilde{k}', \widetilde{k}'']$ stay out because they expect the gain from the public benefit not to be large enough to compensate for their personal disutility from sharing data (recall that for these users, $k > p$). In the second equilibrium users are optimistic that other users with $k \in [\widetilde{k}', \widetilde{k}'']$ join and thereby increase the value of the public benefit, making it worthwhile for them to do the same. Intuitively, with a unimodal function, users' beliefs concerning the decisions of users in the "center" (i.e., with $k \in [\widetilde{k}', \widetilde{k}'']$) are more important than the beliefs concerning users at the "tails". Yet, the proof shows that even in the case of such multiplicity of equilibria, both equilibria satisfy the same features described in parts (i) - (iii) of Proposition 1.

Heterogenous data

Suppose now that users have identical preferences in terms of the disutility from the commercialization of each data item $\theta \in \Theta$, however, there is a continuum of data items that are heterogeneous in terms of the disutility their commercialization imposes on users, k. As before, we assume that k is distributed $k \in [0, 1]$. In addition, we assume that all data items are identical in terms of the public and private benefits: γ and p.

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. Suppose that the platform chooses to commercialize a subset of data items with $k \in [0, \widehat{k}]$. The platform would like to commercialize as many data items as possible, subject to the users' participation constraint. Let $\widehat{k} = min\{\widehat{k}', 1\}$, where $\widehat{k}'$ is the solution to:

$$\gamma + p - \int_0^{\widehat{k}'} kf(k)dk = 0 \tag{4}$$

That is, as in the case of heterogenous users, equation (4) shows that the existence of a public benefit from 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 $k > p$. Moreover, since users get private benefit for all data items, it is easy to show that $k > p$ 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 market power 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 data collection in regime 1)*

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

(*ii*) *the number of data items that the platform commercializes increases with the public benefit:* $\widehat{k}$ *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,* $\widehat{k} < 1$ *if* $\gamma < \gamma_{data}$ *and* $\widehat{k} = 1$ *otherwise.*

The last part of Proposition 2 shows that high values of γ allow the platform to take advantage of regime 1 and commercialize all data items. That is, the ability to contingent participation on the provision of data for commercialization allows the platform to "bundle" the provision of less "costly" data – data items with $k < p$ – with the provision of more "costly" data – data items with $k < \widehat{k}$, where $\widehat{k} > p$. This bundling allows the

platform to demand that users either agree to commercialize all data items with $k < \widehat{k}$, or stay out. As γ increases, the platform can add more costly data items with $k > p$ to the bundle, and maintain the users' consent to commercialize them. Recalling that it is welfare enhancing to commercialize data items with $k < \alpha$ 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 in the first best case. 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:

$$CS_{1,data} = \gamma + p - \int_0^{\widehat{k}} k f(k) dk, \qquad \pi_{1,data} = \alpha F(\widehat{k}).$$

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 give the platform the right to collect and commercialize their data. 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

As there is only one data item, 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 k.

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 $k < p$ 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:

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

Heterogenous data

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

$$CS_{2,data} = \gamma + p - \int_0^p kf(k)dk, \qquad \pi_{2,data} = \alpha F(p).$$

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 from commercializing their data. 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, where it is welfare enhancing to let the users (the platform) control the data when the public benefit of data is high (low). 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: $n_1 > n_2$ ($n_1 = n_2$).*

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 from data to attract users that are willing to 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, $\gamma_{1,2}^{CS}$, $0 < \gamma_{1,2}^{CS} < 1$, 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, $\gamma_{1,2}^{W}$, $0 \leq \gamma_{1,2}^{W} < 1$, such that total welfare in regime 1 is higher than in regime 2 if $\gamma > \gamma_{1,2}^{W}$. For $\gamma < \gamma_{1,2}^{W}$ welfare in regime 1 is lower or equal to welfare in regime 2;*

(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 $F(k)$. The figure shows consumer surplus and welfare as a function of the public benefit of data. Notice that with a uniform $F(k)$, 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 $\gamma_{1,2}^{W}$. Starting with $\gamma = 0$, 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 of providing data. This result highlights

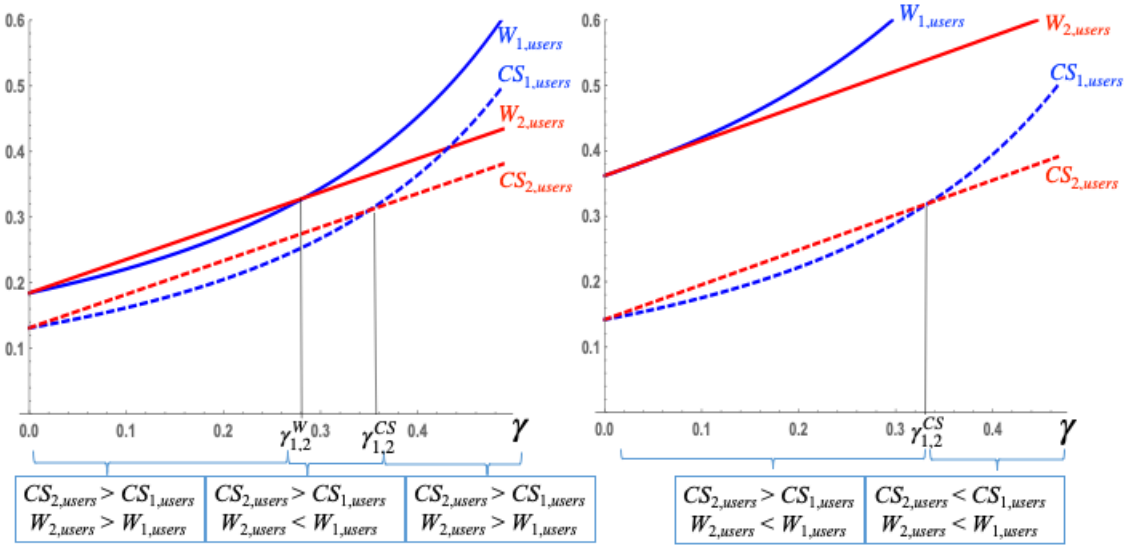

Figure 1: Consumer surplus and welfare as a function of γ for a uniform $F(k)$ ($p = 0.5$)

the role the public benefit of data plays in users' behavior under these two regimes. This result also highlights the distinction between 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 the presence of network effects that are based on participation in the platform, these users would make regime 2 superior to regime 1.

As γ increases above 0 (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. The platform, on the other hand, always prefers regime 1 for its superior level of data commercialized. When the commercial benefit of data is small (panel (a)), the users' benefit from regime 2 outweighs the commercial benefit from regime 1, making regime 2 welfare enhancing. The opposite holds when the commercial benefit of data is high (panel (b)), where the commercial benefit from regime 1 outweighs the users' benefit from regime 2, making regime 1 welfare enhancing. Moreover, welfare under regime 1 is also higher when the commercial benefit of data is small (panel (a)) and the magnitude of the public benefit from data is intermediate. In this case, the gap in consumer surplus between the two regimes is small, while regime 1 improves profits.

When γ is sufficiently high, regime 1 improves both consumer surplus and profits regardless of the level of the commercial benefit from data, making regime 1 superior for welfare. In this case, regime 1 pushes a larger number of users with $p < k$ to give

data. These data then highly benefit all users that join the platform. That is, while individually each user would like to have the choice to share data only if $p > k$, as in regime 2, collectively, when γ is high all users benefit from the platform's ability to force other users to give data. That is, data becomes a "public good", which is better supplied when each user is required to give it, for the benefit of others. In such a case, regime 1 performs better, in terms of welfare, than regime 2.

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 yields the opposite conclusion than in the previous case of heterogeneous users. In particular, now regime 1 enhances welfare when the public benefit of data is low, while regime 2 offers a 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 commercialize 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 commercialize 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 commercialize 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 $F(k)$. 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.

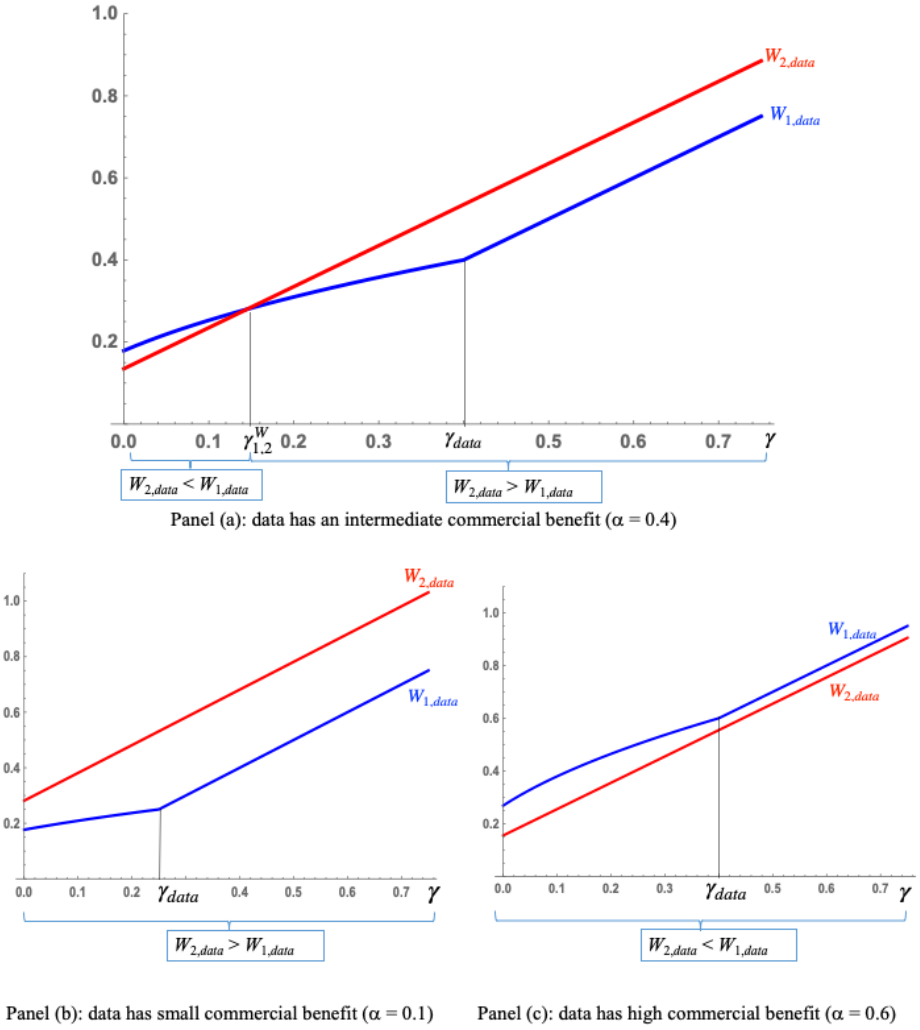

Figure 2: Welfare as a function of γ for a uniform $F(k)$ $(p = 0.1)$

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 $k < \widetilde{k}$, where $\widetilde{k} > p$. Users agree to commercialize "costly" data items with $k > p$ because they gain a positive net private benefit $p - k$ on other data items and because they gain the public benefit γ. As the public benefit increases, the platform's ability to commercialize more data items with $k > p$ 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 $\alpha > k$, the platform commercializes data items with $k < \widetilde{k}$ regardless of whether $\widetilde{k}$ 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.

Type of Heterogeneity	Inefficiency	Effect of an increase in γ	Result
Heterogeneous users	The platform collects too little data for public benefit	Regime 1 reduces 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

Table 1: Comparison between types of heterogeneity

As noted in the table, with heterogeneous users, in regime 1 not all users give data for public benefit, yet those who provide data also give data for commercial 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 is of too much data being commercialized. Regime 1 exacerbates this problem when the public benefit of data is high.

6 Regime 3: Users control their data and can be compensated for it

Suppose now that the platform cannot contingent the collection of data for private and public benefit with users' consent to commercialize it. The platform has to ask the user's permission for two separate decisions. First, whether the user agrees to give data. Second, given a positive answer to the first question, whether the user agrees to commercialize it. In the context of this model, all users agree to share their data for

private and public benefit (the first question). Yet, as $k > 0$, no user agrees to commercialize the data (the second question) without compensation. Hence, we assume that the platform can compensate users for the commercialization of their data. Users that share their data with the platform receive the private benefit. The platform commits not to commercialize users' data without the user's permission; in which case, the user receives a compensation, denoted by ϕ, but bears the disutility k. This compensation can be monetary, or can be of kind such as coupons, additional services, or other benefits.

In the first stage, the platform sets the compensation fee ϕ for each data item. We assume that the platform cannot price discriminate between users but can offer different compensations for different data items. Then, users decide whether to join, give data, and agree to commercialize it.

Heterogenous users

Given the platform's commitment, all users join the platform, share their data, and receive the private and public benefit, $p+\gamma$. Users with $k \leq \phi$ agree to have their data commercialized and receive an additional value of $\phi - k$. The number of users that agree to have their data commercialized is, therefore, $n_3 = F(\phi)$. The platform sets ϕ to maximize:

$$\pi_{3,users}(\phi) = (\alpha - \phi)F(\phi).$$

Solving the first order condition, we get that

$$\phi = \alpha - \frac{F(\phi)}{f(\phi)}. \tag{5}$$

Define the solution to (5) as a function of α by $\phi_{users}(\alpha)$. Suppose that $\frac{F(\phi)}{f(\phi)}$ is increasing in ϕ and that $\alpha < \frac{1}{f(1)}$. As expected, $\phi_{users}(0) = 0$ and increases in α. That is, as the commercial benefit of data increases, the platform is willing to pay users more for their data, in order to convince more users (i.e., users with higher k) to agree to commercial their data. Yet, notice that while it is welfare-maximizing to commercialize the data of all users with $k \in [0, \alpha]$, under regime 3 only users with $k \in [0, \phi_{users}(\alpha)]$ agree to commercialize their data, where (5) implies that $\phi_{users}(\alpha) < \alpha$. That is, since the platform needs to compensate users for selling their data, it sells less data than optimal. This holds for all α, including the extreme case where $\alpha = 1$. Consumer

surplus and the platform's profit under regime 3 are given by:

$$CS_{3,users} = \gamma + p + \int_0^{\phi_{users}(\alpha)} (\phi(\alpha) - k) f(k) dk, \quad \pi_{3,users} = (\alpha - \phi_{users}(\alpha)) F(\phi_{users}(\alpha)).$$

Heterogenous data

As in the heterogenous users case, here too all users join the platform and share all their data items, yet agree to commercialize only data items with $k < \phi$. While in the case of heterogeneous users, the platform cannot price discriminate across users, as it cannot distinguish between different users' costs, in the case of heterogenous data, the platform knows which data items have higher costs and which ones have lower costs.[13] For each data item with $k < \alpha$, the platform would be willing to pay users up to k for the commercialization of that data item. Hence, the platform pays $\phi_{data}(k) = k$ for $k \leq \alpha$ and $\phi_{data}(k) = \alpha$ for $k > \alpha$. The platform, thus, commercializes the welfare-maximizing level of data items with $k < \alpha$ and consumer surplus and profit are:

$$CS_{3,data} = \gamma + p, \qquad \pi_{3,data} = \textstyle\int_0^{\alpha} (\alpha - k) f(k) dk \ .$$

It is easy to see that total welfare under regime 3 with heterogenous data, $W_{3,data} = CS_{3,data} + \pi_{3,data}$, is the same as under the first best case; i.e., $W_{3,data} = W^*$.

Does paying users for data improve efficiency?

We compare regime 3 with the first two regimes. A common feature of regime 3 under both heterogeneous users and heterogeneous data is that the platform serves all users and collects all data for public and private benefits. So the regime achieves the first best in terms of the public and private benefit. The two cases differ in the platform's ability to commercialize the efficient level of data.

Consider first heterogeneous users. As stated above, regime 3 under-supplies the commercial benefit. When this commercial benefit is negligent (i.e., $\alpha \to 0$), this effect is negligent and regime 3 is close to the first-best outcome. Indeed, it is straightforward to see that when $\alpha \to 0$, $W_{3,users} = \gamma + p = W^*$. Yet, when the commercial benefit

[13] While it makes sense to assume that platforms do not know each user's costs, the knowledge about whether a certain data item is more sensitive and thus associated with higher commercialization costs is, typically, of common knowledge.

is very high, and the public benefit is high, we have that regime 1 outperforms regime 3. This is because if $\alpha \to 1$ and $\gamma \to 1-p$, $W_{3,users} < W^* = W_{1,users}$. The following corollary summarizes this result:

Corollary 2. *(Heterogeneous users: the effect of the public and commercial benefits of data on the comparison between regimes 1, 2, and 3). When the commercial benefit α is sufficiently small, regime 3 provides the highest social welfare: $W_{3,users} > \max\{W_{1,users}, W_{2,users}\}$. When α and γ are sufficiently high, regime 1 provides the highest social welfare: $W_{1,users} > \max\{W_{2,users}, W_{3,users}\}$.*

Next, consider the case of heterogeneous data.

Corollary 3. *(Heterogeneous data: Regime 3 always outperforms regimes 1 and 2) Welfare under regime 3 with heterogeneous data achieves the first best.*

The intuition is simple. Since the platform must compensate users for their cost, and can price discriminate across the different data items based on the cost selling them imposes, it internalizes the users' commercialization costs and the market achieves the first best.

To summarize the two results above, this section shows that a regime that provides users with all the control over their data is not always efficient, depending on the type of heterogeneity and the value of the commercial and public benefits of data. With homogeneous users, regime 3 is indeed the optimal policy, even in the face of heterogeneous data. Yet, once users become heterogeneous, regime 3 may underperform relative to regimes 1 and 2, especially when the commercial and pubic benefits of data are high.

7 Platform Competition

As we show above, under monopoly, the platform always prefers regime 1, even when regime 2 provides higher welfare. One might expect, however, competition to induce platforms to achieve the first-best regime. To answer this question, we study platform competition between two platforms: an incumbent, I, and an entrant, E. The entrant has a quality advantage, while the incumbent enjoys focality–meaning, users believe that it would be the dominant platform in the market. We assume that users are heterogeneous in terms of the disutility they bear from their data being commercialized and that each platform can choose which data regime to implement – regime 1 or 2.

We find that even under competition, the focal platform always prefers regime 1. Whether the entrant prefers regime 1 or 2 depends on the quality advantage it enjoys. Specifically, when the quality advantage is very large, the entering platform wins the market while implementing regime 1. When the quality advantage is small, the incumbent platform wins the market (implementing regime 1 as well). If, however, the quality advantage of the entrant is of intermediate value, the entrant chooses regime 2 and wins the market, while the incumbent chooses regime 1 but is unable to attract users.

In what follows, suppose that the platform in our model, (hereafter, the incumbent, I), faces the threat of entry by an entrant platform, E. The benefits from data are the same as in our base model, but each user gains an additional base quality of $q > 0$ from joining the entrant. Users can join one of the platforms, or stay out. Users are heterogeneous in their k and for simplicity, we assume that k is uniformly distributed along $[0, 1]$. In the first stage, the two platforms announce their data polices – regime 1 or 2 – simultaneously and non-cooperatively. In the second stage, users decide which platform to join, simultaneously and non-cooperatively and share data according to the relevant regime. In this second stage, users' decisions to join a platform depend on their beliefs concerning the decisions of other users, through the data that other users provide for public benefit. Hence, as is usually the case in platform competition, the second stage may have multiple outcomes. Following the literature on platform competition, suppose that the incumbent is "focal".[14] That is, when the second stage yields two outcomes: one in which all users join the incumbent and a second in which all users join the entrant, then users join the incumbent, expecting other users to do the same. Intuitively, users' inertia pushes them to join the incumbent platform, whenever this is an equilibrium outcome. The entrant can attract users when there is no outcome in which users join the incumbent, and there is an outcome in which users join the entrant.

To solve for the equilibrium outcome, we start by solving the second stage: the market outcome taking the platforms' data policies as given. Then, we solve for the equilibrium data policy.

Consider the case where in the first stage, both platforms adopt regime 1. Then, in the second stage, there is an equilibrium in which $n_I = F(\widetilde{k})$ users join platform I, where $\widetilde{k}$ is given by Proposition 1. Using our assumption of uniform distribution,

[14]See, for example, Caillaud, and Jullien (2003) and Halaburda and Yehezkel (2019).

$\widetilde{k} = p/(1-\gamma)$. By the focality of platform I, users play this equilibrium even if there is another outcome in which users join E. Users will join E only when the outcome of joining I cannot be an equilibrium. User of type k receives $\gamma n_I + p - k$ from joining platform I, $q + p - k$ from joining E, and 0 otherwise. Solving for the users' decision, the equilibrium in which users join I fails when:

$$\gamma n_I + p - k < q + p - k \quad \iff \quad q > \frac{\gamma p}{1-\gamma}. \tag{6}$$

According to eq (6), if platform E's quality advantage is small, $q < \frac{\gamma p}{1-\gamma}$, platform I wins the market – i.e., $n_I = \frac{p}{1-\gamma}$ users join I and no user joins E. When $q > \frac{\gamma p}{1-\gamma}$, there is no equilibrium in which users join I, as given such a putative equilibrium, by (6), a user will deviate and join E. In this case, there is an equilibrium where users with k such that $\gamma n_E + q + p - k > 0$ join E. Letting $n_E = k$ and solving, $n_E = \frac{q+p}{1-\gamma}$. We summarize these results in the following lemma:

Lemma 1. *Suppose that both platforms adopt regime 1 in the first stage. Then, in the second stage, if $q \leq \frac{\gamma p}{1-\gamma}$, the incumbent wins the market and serves $n_I = \frac{p}{1-\gamma}$ users that join and give data. If $q > \frac{\gamma p}{1-\gamma}$, the entrant wins the market an serves $n_E = \frac{q+p}{1-\gamma}$ users that join and give data.*

As in the monopolistic case, it is easy to see here the important role the public benefit of data plays. Specifically, if $\gamma = 0$, and both platforms choose regime 1, E wins the market for all $q > 0$. Once the public benefit of data becomes positive, the focality advantage becomes important and may allow I to win the market, despite its inferior quality.

Next, consider the case where the incumbent adopts regime 1 and the entrant adopts regime 2. In an equilibrium where only I attracts $\widetilde{n}_I$ data-sharing users, users join I if $\gamma \widetilde{n}_I + p - k \geq q$. Letting $\widetilde{n}_I = k$ and solving, $\widetilde{n}_I = \frac{p-q}{1-\gamma}$ users join I and share data, while the remanning users join E just to benefit from E's base quality without sharing data. This equilibrium exists if:

$$\gamma \widetilde{n}_I + p - k \geq q + \max\{p - k, 0\}, \tag{7}$$

which holds iff $q \leq \gamma p$. Otherwise, there is an equilibrium in which all users join E, and users with $k \leq p$ share data. We summarize these results in the following lemma:

Lemma 2. *Suppose that the incumbent adopts regime 1 and the entrant adopts regime 2. Then, if $q \leq \gamma p$, the incumbent wins the market and serves $\widetilde{n}_I = \frac{p-q}{1-\gamma}$ users that join*

and give data. If $q > \gamma p$, the entrant wins the market, serves all users and collects data of size $\widetilde{n}_E = p$.

Next, consider the case where both platforms adopt regime 2. Under regime 2, only users with $k < p$ agree to share their data, and users can choose to join the platform without sharing their data. There is an equilibrium in which all users join I and users with $k < p$ share data when:

$$\gamma p + \max\{p - k, 0\} \geq q + \max\{p - k, 0\} \iff \quad q \leq \gamma p.$$

If $q > \gamma p$, there is an equilibrium in which all users join E and users with $k < p$ share data. We summarize these results in the following lemma:

Lemma 3. *Suppose that both platforms adopt regime 2. Then, if $q \leq \gamma p$, the incumbent wins the market, serves all users and collects data of size p. If $q > \gamma p$, the entrant wins the market, serves all users and collects data of size $\widetilde{n}_E = p$.*

For simplicity, we do not show the case where the incumbent chooses regime 2 while the entrant chooses regime 1, as this case is irrelevant. Given the results above, the following proposition presents the equilibrium outcome of the first stage, when platforms choose their data regimes:

Proposition 5. *(Platform competition affects the equilibrium data regime) Assume platform competition where each platform can choose whether to implement regime 1 or 2. If one platform, I, enjoys a focality advantage and the second, E, enjoys a quality advantage, q, then:*

(*i*) *if $0 < q \leq \gamma p$, the incumbent focal platform chooses regime 1, wins the market and serves $n_I = \frac{p}{1-\gamma}$ users who all share data. Regardless of the chosen regime, the entrant cannot attract data-sharing users;*

(ii) *if $\gamma p < q \leq \frac{\gamma p}{1-\gamma}$, the entrant chooses regime 2, wins the market and serves all users, while only $\widetilde{n}_E = p$ users share data. The focal incumbent is indifferent between regimes 1 and 2 and either way cannot win the market;*

(iii) *if $\frac{\gamma p}{1-\gamma} < q$, the entrant chooses regime 1, wins the market and serves $n_E = \frac{q+p}{1-\gamma}$ users that also share data. The focal incumbent is indifferent between regimes 1 and 2 and either way cannot win the market.*

Our results have important implications for platforms and policy makers. First, we find that an entrant platform has a stronger incentive, than the incumbent platform, to give users control over data. Intuitively, the entrant may suffer from a non-focal position, hence, needs to provide value for users by giving them control over data. That is, data regime can serve as a tool to become more competitive in the market. Second, we find that competition does not necessarily motivate either platform to choose the welfare-maximizing regime. The equilibrium regimes depend on the quality gap between the two platforms and not exclusively on the public and private benefits of data.

For policy makers, imposing regulation that requires platforms to implement regime 2 would not be helpful in facilitating entry by a more efficient, yet non-focal platform. To see why, suppose that a regulator restricts the two platforms to adopt only regime 2. From Lemma 3, we have that this restriction does not prevent a focal incumbent from winning the market when $0 < q < \gamma p$. This holds despite the entrant's quality advantage. When $\gamma p < q$, the entrant wins the market by adopting regime 2 regardless of the incumbent's regime, and again restricting both platform to regime 2 would not facilitate entry.

8 Fully covered markets

The analysis above focused on partially covered markets and demonstrated the different effects the public and private benefit of data have on the equilibrium outcomes and welfare in such markets. In this section we focus on the case of fully covered markets: in equilibrium, all users join the platform that collects and commercializes in regime 1 all data items. The main result of this section is that the distinctions between the public and the private benefit of data is meaningful even when the market is fully covered. An increase in the public benefit of data increases the relative efficiency of regime 1, in comparison with regime 2. Yet, the private benefit of data has conflicting effects on the comparison between the two regimes. This last result highlights the qualitative difference between the public and the privet benefit of data.

When the market is fully covered, the distinction between the type of heterogeneity: heterogeneous users and heterogeneous data, no longer matters. Instead, the relevant heterogeneity is between any *data element* $i\theta$: a data item $\theta \in \Theta$ collected from user $i \in N$. Therefore, this section studies the general data matrix considered in Section 2 that allows for both heterogeneous users and data. Suppose that there is a discrete set

of data items Θ and users, N (as before, the number of data items is $\overline{\Theta}$ and we normalize the mass of users to 1). The total set of data elements, $i\theta$, is $\Omega = \{i\theta \mid i \in N, \theta \in \Theta\}$: a set of size $1 \times \overline{\Theta}$. To make the problem of data regulation meaningful, suppose that there are some data items and some users for whom the disutility from commercialization is higher than the private benefit. Likewise, some users and data items bear higher disutility from commercialization than their commercial benefits:

Assumption 1: $\exists i\theta \in \Omega$ such that $k_{i\theta} > p_{i\theta}$.

Assumption 2: $\exists i\theta \in \Omega$ such that $k_{i\theta} > \alpha_{i\theta}$.

To ensure that the market is fully covered, suppose that:

Assumption 3: $u_i = \sum_{i\theta \in \Omega} \gamma_{i\theta} + \sum_{\theta \in \Theta} (p_{i\theta} - k_{i\theta}) \geq 0, \quad \forall i \in N.$

The first term in u_i is the total public benefit of data from all users and all data items, which is always positive. The second term is user i's private benefit minus the disutility from commercializing i's data items, which can be positive or negative for some users. Assumption 3 is more likely to hold when the public benefit of data is high, and when users are sufficiently homogeneous in their disutilities from data commercialization. We further assume that for each data item, there is at least one user with private benefit that is higher than the disutility from commercializing it. That is:

Assumption 4: $\forall \theta \in \Theta$, $\exists i \in N$ such that $p_{i\theta} > k_{i\theta}$.

This assumption implies that for each data item, there is at least one user that agrees to commercialize it in order to enjoy the data's private benefit.

In the first-best outcome, it is optimal to serve all users, collect all data items for public and private benefit, and commercialize data item $i\theta$ if and only if $\alpha_{i\theta} > k_{i\theta}$. Define the set $\Omega_{fb} = \{i\theta \mid i \in N, \theta \in \Theta \text{ and } \alpha_{i\theta} \geq k_{i\theta}\}$. Notice that by Assumption 2, Ω_{fb} is a subset of Ω. First-best welfare is:

$$W_{fb} = \sum_{i\theta \in \Omega} (\gamma_{i\theta} + p_{i\theta}) + \sum_{i\theta \in \Omega_{fb}} (\alpha_{i\theta} - k_{i\theta}) . \tag{8}$$

Regime 1: the platform controls the data

Under regime 1, by Assumption 3 the platform's strategy is simple: the platform announces that it collects and commercializes all data items from users that join it. Given

Assumption 3, all users join the platform.[15] Hence, the platform's total data set under regime 1 is Ω. Social welfare is:

$$W_1 = \sum_{i\theta \in \Omega} (\gamma_{i\theta} + p_{i\theta} - k_{i\theta}) + \sum_{i\theta \in \Omega} \alpha_{i\theta}, \tag{9}$$

where the first term is consumer surplus and the second term is the platform's profit. Comparing (9) with the first-best welfare in (8), notice that regime 1 collects the optimal level of data items for public and private benefit, but collects too much data for commercial benefit.

Regime 2: users control the data

In this regime, the platform announces which data items it plans to collect and commercialize. A user can join the platform yet refuse to give a certain data item θ, in which case the platform will not be able to provide the user with the private benefit that data item θ offers the user, $p_{i\theta}$, but at the same time the user will not bear the disutility from data commercialization, $k_{i\theta}$.

Consider first the users' decisions. As users can choose whether to give data or not, all users join the platform. Moreover, users agree to give any data item that the platform commits not to commercialize, as users bear the costs $k_{i\theta}$ only if the data item is commercialized. User i agrees to give data item θ that the platform declares to commercialize, if and only if $p_{i\theta} \geq k_{i\theta}$. Next consider the platform's decision. By Assumption 2, the platform declares that it collects and commercializes all data items. This is because for each data item, there is at least one user that will agree to give it. Let $\Omega_2 = \{i\theta \mid i \in N, \theta \in \Theta \text{ and } p_{i\theta} \geq k_{i\theta}\}$ denote the subset of data that users agree to share with the platform in regime 2. By Assumption 1, we have that the subset of data that users do not agree to share, $\Omega - \Omega_2$, is non-empty. However, we cannot compare between Ω_2 and Ω_{fb}. Total welfare in regime 2 is, therefore:

$$W_2 = \sum_{i\theta \in \Omega_2} (\gamma_{i\theta} + p_{i\theta} - k_{i\theta}) + \sum_{i\theta \in \Omega_2} \alpha_{i\theta}. \tag{10}$$

In comparison with the first-best welfare, regime 2 provides less data for public and private benefit. Yet, regime 2 may provide more or less data elements for commercial

[15] We assume that there is no coordination failure. Each user expects that all other users join the platform if it is mutually beneficial to do so.

benefit than the first-best level. Data elements with $\alpha_{i\theta} > k_{i\theta} > p_{i\theta}$, if exist, are welfare enhancing to commercialize but are not collected. Likewise, data elements with $p_{i\theta} > k_{i\theta} > \alpha_{i\theta}$, if exist, are commercialized under regime 2 even though it is welfare enhancing to collect them only for public and private benefit.

Comparison between regime 1 and regime 2

Since neither regime implements the welfare maximizing outcome, the comparison between the two regimes is nontrivial. Moreover, the results above suggest that the divergence between the welfare in each regime and the first-best welfare depends on variations in $k_{i\theta} - p_{i\theta}$. Below we study how this variation affects the comparison between the two regimes.

We can write the gap in welfare between the two regimes as:

$$W_1 - W_2 = \sum_{i\theta \in \Omega - \Omega_2} (\gamma_{i\theta} + \alpha_{i\theta}) - \sum_{i\theta \in \Omega - \Omega_2} (k_{i\theta} - p_{i\theta}) \,. \tag{11}$$

Intuitively, equation (11) shows that in comparison with regime 2, regime 1 has two conflicting effects on welfare. First, regime 1 enables the platform to collect more data for public and commercial benefits. This effect is represented by the first term in (11), which is strictly positive. Second, regime 1 enables the platform to commercialize data elements in the set $\Omega - \Omega_2$, which users would rather not commercialize because $k_{i\theta} > p_{i\theta}$ for all $i\theta \in \Omega - \Omega_2$. This effect is represented by the second term in (11), which is strictly negative. Hence, regime 1 enhances welfare when the sum of the public and commercial benefits of the additional data collected by regime 1, $\Omega - \Omega_2$, is higher than the costs that this regime inflicts on users.

To see how that gap $W_1 - W_2$ varies with changes in the public and private benefit of data, notice first that a constant increase in the public benefit of each data elements, $\gamma_{i\theta}$, increases the benefit from regime 1, in comparison with regime 2. Intuitively, this is because regime 1 enables the platform to collect all data elements for public benefit, while under regime 2, users do not internalize the positive externaility their data have on other users. The effect of an increase in the private benefit of data, $p_{i\theta}$, is more subtle. Similar to an increase in $\gamma_{i\theta}$, a constant increase in $p_{i\theta}$ increases the gap $W_1 - W_2$ because $k_{i\theta} - p_{i\theta}$ decreases. Yet, unlike the public benefit of data, an increase in the private benefit of data also increases the set Ω_2, as more data elements (more users and more data items) are shared and commercialized under regime 2. This, in turn,

decreases the gap $W_1 - W_2$. Proposition 1 summarizes this result:

Proposition 6. *(the distinction between the public and private benefit of data when the market is fully covered)*

(*i*) *A constant increase (across all users and data items) in the public benefit of data, $\gamma_{i\theta}$, increases the gap $W_1 - W_2$.*

(*ii*) *A constant increase in the private benefit of data, $p_{i\theta}$, has conflicting effects on the gap $W_1 - W_2$. Suppose that the private benefit of data increases to $\widetilde{p}_{i\theta} = p_{i\theta} + \varepsilon$ ($\varepsilon > 0$) for $\forall i\theta \in \Omega$. Let $\widetilde{\Omega}_2 = \{i\theta \mid i \in N, \theta \in \Theta \text{ and } p_{i\theta} + \varepsilon \geq k_{i\theta}\}$ and let $\widetilde{n}$ denote the number of data items in $\Omega - \widetilde{\Omega}_2$. Then, the gap $W_1 - W_2$ increases if and only if:*

$$\widetilde{n}\varepsilon - \sum_{i\theta \in \widetilde{\Omega}_2 - \Omega_2} (\gamma_{i\theta} + \alpha_{i\theta} - k_{i\theta} + p_{i\theta}) > 0. \tag{12}$$

Proposition 6 highlights the distinction between the public benefit of data and the private benefit, for evaluating policy for data regulations. While an increase in the public benefit increases the relative efficiency of regime 1, an increase in the private benefit has two effects, represented by (12). The first term in (12) is the increase in consumer surplus due to higher private benefit from data elements that are collected only in regime 1. The second term is the sum of the benefits from the data elements in $\widetilde{\Omega}_2 - \Omega_2$: data elements that the increase in the private benefit motivate users to give in regime 2. This second term can be positive or negative, as $k_{i\theta} > p_{i\theta}$ for all $i\theta \in \widetilde{\Omega}_2 - \Omega_2$.

This suggests that in markets where the public benefit of data is high, like in the case of contact tracing apps, giving the platform the control over users' data would likely result in higher welfare. Furthermore, from the platform's perspective, in markets like the U.S., where regulators give platforms the right over data they collect, platforms should invest more in increasing the public benefit of their data as opposed to investing a lot in increasing the private benefit.

9 Conclusion

Many digital platforms collect our data to improve the services they provide to users but at the same time to commercialize it by selling it to advertisers or third-party

firms. This raises the question of whether policy makers should regulate the platforms' ability to collect and commercialize data. This paper considers the interaction between a platform and users, when the platform can collect and commercialize data. Our model has three main features. First, in addition to personal benefit to the user, data also provide public benefits to other users. Second, 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. Third, users have disutility from commercializing their data, which can vary across users (the case of heterogeneous users) and across data items (heterogeneous data).

We consider three extremes of data regulation regimes. In the first regime, the platform controls the data. Users can either stay out of the platform, or give their consent to collect and commercialize their data. In the second regime, users control whether their data is collected. In the third regime, users control whether their data is collected and whether it is commercialized and 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 collects too little data for public benefit, and 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.

Competition does not necessarily results in the efficient outcome. Under some market conditions, it incentivizes entrants to choose a data regime that provides users with more control over their data, which act as tool for the entrant to better compete with market incumbents. Yet, the entrant may choose to give users control over data, when

the efficient outcome is for the platform to fully control the data. Moreover, we find that data regulation that restricts competing platforms to give users control over data does not facilitate entry.

References

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

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

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

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

[5] 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.

[6] Dosis Anastasios and Wilfried Sand-Zantman. 2021. "The Ownership of Data." Working paper.

[7] Economides, Nicholas, and Ioannis Lianos. "Restrictions on Privacy and Exploitation in the Digital Economy: A Market Failure Perspective." *Journal of Competition Law and Economics* (forthcoming)

[8] Fainmesser, Itay, Andrea Galeotti, and Ruslan Momot. 2020. "Digital Privacy." Working paper.

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

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

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

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

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

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

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

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

[17] Markovich, Sarit, and Yaron Yehezkel. "Group Hug: Platform Competition with User-groups." *American Economic Journal: Micro,* (forthcoming).

[18] 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 (2). Evaluated at $k = 0$, the left-hand side of (2) is $\gamma F(0) = 0$ while the right hand side is $0 - p < 0$, hence $\gamma F(k) > k - p$ if k is sufficiently close to 0. Evaluated at $k = 1$, the left-hand side of (2) is $\gamma F(1) = \gamma$ while the right hand side is $1 - p \geq \gamma$ (recall that we assume that $\gamma \leq 1 - p$), hence $\gamma F(k) < k - p$ if k is sufficiently close to 1, and at the highest possible γ, $\gamma = 1 - p$, the solution to (2) is at $\widetilde{k} = 1$. This implies that there is at least one intersection point between $\gamma F(k)$ and $k - p$.

Next, we show the conditions under which this intersection point is unique. Figure 3 (panel (a)) shows the solution to $\widetilde{k}$ when $F(k)$ is not too unimodal (we can derive a qualitatively similar figure for a $F(k)$ that is not unimodal). In this case, there is a unique solution to $\widetilde{k}$, hence a unique equilibrium. Panel (b) shows the case of a strong unimodal $F(k)$, in which case there are three solutions to (2), the middle one is not stable while in the two stable solutions, $\widetilde{k}_1$ and $\widetilde{k}_2$, $\gamma F(k)$ intersects $k - p$ from below, hence the comparative statics of $\widetilde{k}_1$ and $\widetilde{k}_2$ are qualitatively the same. That is, both solutions are higher than p, and both solutions are increasing with γ.

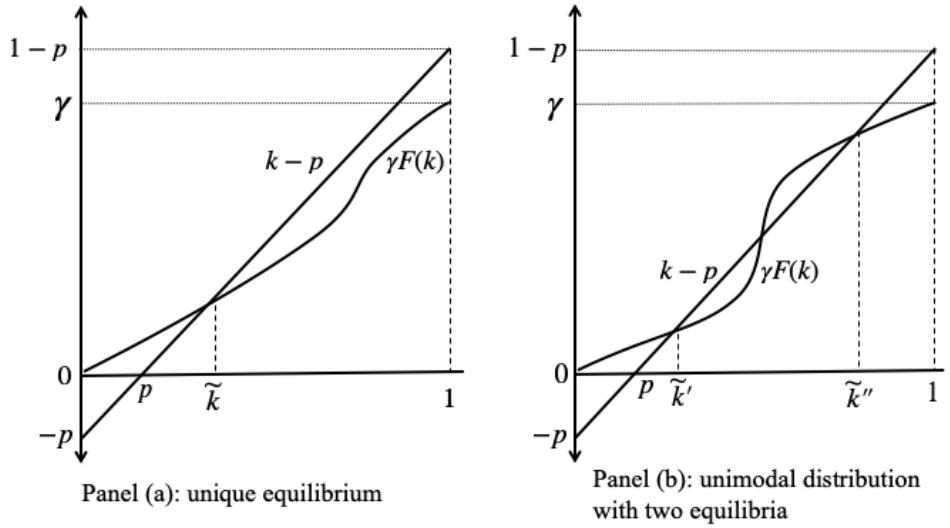

Figure 3: The solution to $\widetilde{k}$

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

Proof of Proposition 2:

part *i*

Since the platform bears no cost for data collection, yet the data collected provides users with $p > 0$ and $\gamma > 0$, the platform collects all data items. We first show that there is at least one solution to (4), or rearranging order to:

$$\gamma + p = \int_0^{\widehat{k}'} kf(k)dk \tag{13}$$

Evaluated at $k = 0$, the right-hand side of eq. (13) is 0. Since $\gamma + p > 0$ then if k is sufficiently close to 0, the left-hand side of eq. (13) is larger than its right-hand side. Evaluated at $k = 1$, the right-hand side of eq. (13) larger than 1. Hence, if k is sufficiently close to 1, and at the highest possible γ, $\gamma = 1-p$, the right-hand side of eq. (13) is larger than its left-hand side. This implies that there is at least one intersection point where the two are equal. Since the right hand side of eq. (13) is increasing in k, this intersection point is unique.

Since

$$\gamma + p - \int_0^{\widehat{k}'} kf(k)dk > \gamma + \int_0^{\widehat{k}'} (p-k)f(k)dk$$

and we know that $\int_0^{\widehat{k}'} (p-k)f(k)dk$ is positive at $\widehat{k}' = p$, it follows that $\widehat{k}' > p$ for all $\gamma \geq 0$.

part *ii*

To show that $\widetilde{k}'$ is increasing in γ. Using the implicit function theorem, $G = \gamma + p - F(k) = 0$

$$\frac{dk}{d\gamma} = -\frac{\frac{dG}{d\gamma}}{\frac{dG}{dk}} = -\frac{1}{-kf(k)} = \frac{1}{kf(k)} > 0$$

part *iii*

To prove this part, it is enough to show that for $\gamma = 1 - p$, the solution to eq. (4) is at $\widehat{k}' = 1$. Evaluating eq. (4) at $\gamma = 1 - p$:

$$(1-p) + p - \int_0^{\widehat{k}'} kf(k)dk = 1 - \int_0^{\widehat{k}'} kf(k)dk$$

Using integration by parts:

$$\int_0^1 kf(k)dk = kF(k)\mid_0^1 - \int_0^1 F(k)dk = 1 - 0 - \int_0^1 F(k)dk$$

Substituting this in, we get:

$$1 - \left(1 - \int_0^1 F(k)dk\right) = \int_0^1 F(k)dk > 0$$

It follows that at values of γ that are close to $1-p$, $\widehat{k}' = 1$.

■

Proof of Proposition 3:

The consumer surplus and the platform's profit in regime 1 are:

$$CS_{1,users} = \gamma \cdot F(\widetilde{k}) \cdot F(\widetilde{k}) + \int_0^{\widetilde{k}} (p-d)f(k)dk, \qquad \pi_{1,users} = \alpha F(\widetilde{k}). \tag{14}$$

Likewise, in regime 2, consumer surplus and the platform's profit are:

$$CS_{2,users} = \gamma F(p) + \int_0^{p} (p-k)f(k)dk, \qquad \pi_{2,users} = \alpha F(p). \tag{15}$$

_Evaluated at $\gamma = 0$, $\widetilde{k} = p$, hence $CS_{1,users} = CS_{2,users}$ and $\pi_{1,users} = \pi_{2,users}$. When $\gamma > 0$:

part i

The first part is a direct result of Corollary (1). If $n_1 > n_2$ then $F(\widetilde{k}) > F(p)$ 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:

$$\frac{dCS_{1,users}}{d\gamma} \mid_{\gamma=0} = 2\gamma F(\widetilde{k})f(\widetilde{k})\widetilde{k}' + F^2(p) + (p-\widetilde{k})f(\widetilde{k})\widetilde{k}' = F^2(p)$$

where the last equality follows from $\gamma = 0, p = \widetilde{k}$. 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 $\widetilde{k} = 1$. Substituting $\gamma = 1 - p$ we get that, $CS_{1,users}(\gamma = 1 - p) = F(1)(1 - p) - \int_0^1 (d - p)f(k)dk$ and $CS_{2,users}(\gamma = 1 - p) = F(p)(1 - p) - \int_0^p (d - p)f(k)dk$. It follows that when $\gamma = 1 - p$:

$$CS_{1,users} - CS_{2,users} = (F(1) - F(p))(1 - p) - \int_p^1 (d - p)f(k)dk$$

To show that this difference is positive, note that the utility for the user with the highest disutility is $1 - p$; that's the least utility users may receive. If all users in the range $[p, 1]$ received this utility then overall surplus for users in this range is $(1 - p)(1 - F(p))$. We know, however, that users in this range have lower disutility than $1 - p$ and thus receive higher utility. It follows that $(F(1) - F(p))(1 - p) > \int_p^1 (d - p)f(k)dk$ and thus that at $\gamma = 1 - p$, $CS_{1,users} > CS_{2,users}$.

part iii

Since at $\gamma = 0$, $CS_{1,users} = CS_{2,users}$, and $\frac{d^2 CS_{2,users}}{d^2\gamma} = 0$, it suffices to show that $\frac{d^2 CS_{1,users}}{d^2\gamma} > 0$, 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, $\frac{d^2 W_{1,users}}{d^2\gamma} > 0$. Given that $\frac{d^2 W_{2,users}}{d^2\gamma} = 0$, 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 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 $\gamma = 0$, $\frac{dW_{1,users}}{d\gamma} > \frac{dW_{2,users}}{d\gamma}$.

$$\frac{dW_{1,users}}{d\gamma} \mid_{\gamma=0} = F^2(p) + \alpha f(p)\widetilde{k}', \qquad \frac{dW_{2,users}}{d\gamma} \mid_{\gamma=0} = F(p). \tag{16}$$

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

$$\alpha > \frac{F(p)(1-F(p))}{\widetilde{k}' f(p)}. \tag{17}$$

Using the implicit function theorem, $G = \gamma F(k) + p - k = 0$

$$\widetilde{k}' = \frac{dk}{d\gamma} = \frac{\frac{dG}{d\gamma}}{\frac{dG}{dk}} = -\frac{F(k)}{\gamma f(k) - 1} \mid_{\gamma=0} = F(k)$$

Substituting this into eq. (17), we get that $\frac{d(W_{1,users}-W_{2,users})}{d\gamma} \mid_{\gamma=0} > 0$ if:

$$\alpha > \frac{F(p)(1-F(p))(1-\gamma f(p))}{F(p)f(p)} = \frac{(1-F(p))(1-\gamma f(p))}{f(p)}$$

evaluated at $\gamma = 0$, we get that $\alpha > \frac{1-F(p)}{f(p)}$. That is, when $\alpha > \frac{1-F(p)}{f(p)}$ and γ is small, $\frac{d(W_{1,users}-W_{2,users})}{d\gamma} > 0$. That is, for large enough α, $\gamma_{1,2}^W = 0$.■

Proof of Proposition 4

We know that $\pi_{1,data} - \pi_{2,data} = \alpha(F(\widetilde{k}) - F(p)) > 0$, as $F(\widetilde{k}) > F(p)$ for all γ and α. $CS_{2,data} - CS_{1,data} = \int_0^{\widetilde{k}} kf(k)dk - \int_0^p kf(k)dk > 0$ as $\widetilde{k} > p$. It follows then that users prefer regime 2 while the platform prefers regime 1, for all value of γ and α.

When $\alpha = 0$, the platform's profit is zero, so $W_{2,data} > W_{1,data}$ if $CS_{2,data} > CS_{1,data}$. Under regime 1, no users join the platform and $CS_{1,data} = 0$. In regime 2, all users join the platform and the platform does not commercialize any of the data items, so $W_{2,data} = \gamma + p$. It follows that when α is close to 0, $W_{2,data} > W_{1,data}$ for all γ. That is, there is a threshold, $\underline{\alpha}_{1,2}^W$, such that $W_{2,data} > W_{1,data}$ for all γ.

To show that for high values of α, $W_{2,data} < W_{1,data}$, we look at

$$W_{2,data} - W_{1,data} = \int_0^{\widetilde{k}} kf(k)dk - \int_0^p kf(k)dk - \alpha(F(\widetilde{k}) - F(p)) = \int_p^{\widetilde{k}} (kf(k) - \alpha)dk$$

Using integration by part:

$$\int_p^{\widetilde{k}} kf(k)dk = kF(k) \mid_p^{\widetilde{k}} - \int_0^k f(k)dk < (\widetilde{k} - p)$$

where the last inequality follows since $k < 1, F(k) < 1$, and $F(\widetilde{k}) - F(p) < 1$, resulting in $kF(k) \mid_p^{\widetilde{k}} < 1$. Consequently, for $\alpha \to 1$, $\int_p^{\widetilde{k}} (kf(k) - \alpha)dk < 0$ as $\int_p^{\widetilde{k}} kf(k)dk < \int_p^{\widetilde{k}} \alpha dk = \alpha(1-p)$. Letting $\overline{\alpha}_{1,2}^W$ be the solution to $\int_p^{\widetilde{k}} (kf(k) - \alpha)dk$, it follows that there is a threshold, $\overline{\alpha}_{1,2}^W$, such that $W_{1,data} < W_{2,data}$ for all γ.

So far we showed that when $\alpha \to 0$, $W_{2,data} > W_{1,data}$, while the opposite holds for $\alpha \to 1$. Next, we prove that the gap between welfare under regime 2 and welfare under regime 1 is decreasing in the commercial value of data. To do that, we look at

$$\frac{d(W_{2,data} - W_{1,data})}{d\alpha} = F(p) - F(\widetilde{k}) < 0$$

where the last inequality is driven by $F(\widetilde{k}) > F(p)$ proven above.

Next we look at the case where the values meet; that is when $\underline{\alpha}_{1,2}^W = \overline{\alpha}_{1,2}^W \equiv \alpha_{1,2}^W$.

Recall that $\overline{\alpha}_{1,2}^W$ is the solution to $\int_p^{\widetilde{k}} (kf(k) - \alpha)dk$. Given that $kf(k)$ increases in k, for $\alpha = \widetilde{k}f(\widetilde{k})$, $\int_p^{\widetilde{k}} (kf(k) - \alpha)dk$ is still negative, and thus $\alpha_{1,2}^W < \widetilde{k}f(\widetilde{k})$. Since

$$\frac{dW_{1,data}}{d\gamma} = \frac{a}{\widetilde{k}(\gamma)f(\widetilde{k}(\gamma))}$$

then $\frac{dW_{1,data}}{d\gamma} \mid_{\alpha=\alpha_{1,2}^W} < 1$. Since $\frac{dW_{2,data}}{d\gamma} = 1$, we get that at $\alpha = \alpha_{1,2}^W$, $\frac{d(W_{1,data} - W_{2,data})}{d\gamma} \mid_{\alpha=\alpha_{1,2}^W} < 0$. It follows then that there exits a $\widetilde{\gamma}$ such that $W_{1,data} > W_{2,data}$ for $\gamma < \widetilde{\gamma}$.

■